

édition spéciale

132
Pages

l'IA s'invite
à la rédaction

ordinateurs monocartes pour les projets d'intelligence artificielle

horloge VHDL réalisée avec ChatGPT

intro aux algorithmes de l'IA

focus sur l'IA

créez un neurone
d'intégration-et-tir
avec fuite

intelligence artificielle
sans logiciel

fonctions vocales avec
le Raspberry Pi Zero

donner de la voix à votre
projet

dans ce numéro

- > Q&A avec Sayash Kapoor
- > détection de moustiques avec
Arduino Nicla Vision
- > une conversation avec l'esprit
numérique
- > l'IA en développement
électronique

et bien plus encore!

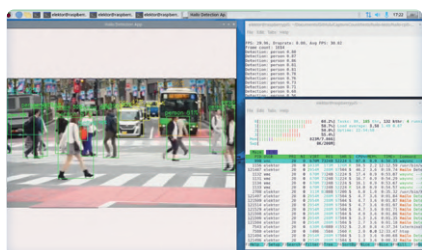
elektorgpt

Pouvez-vous trouver
la série sur les réseaux
neuronaux ? * Existe-t-il
une bonne introduction
aux bus I²C et SPI ? *
Quand Python a-t-il



des données de capteurs
aux modèles de ML

détection de gestes avec
Edge Impulse



le système de sécurité
IA AlertAlfred

basé sur le Raspberry Pi 5
et le module Hailo 8L

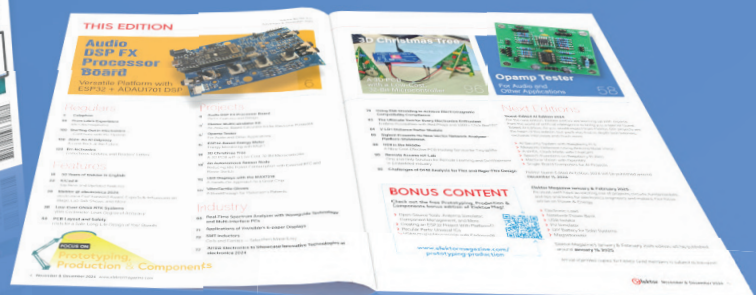
L 16651-8H-F: 15,50 € - RD



Rejoignez la communauté Elektor



Devenez membre maintenant !



- ✓ accès à l'archive numérique depuis 1978 !
- ✓ 8x magazine imprimé Elektor
- ✓ 8x magazine numérique (PDF)
- ✓ 10 % de remise dans l'e-choppe et des offres exclusives pour les membres
- ✓ accès à plus de 5000 fichiers Gerber



Également disponible
abonnement
sans papier !



- ✓ accès à l'archive numérique d'Elektor
- ✓ 10 % de remise dans l'e-choppe
- ✓ 8x magazine Elektor (PDF)
- ✓ accès à plus de 5000 fichiers Gerber



www.elektormagazine.fr/membres

47^{ème} année n° 510
novembre-décembre 2024
ISSN 0181-7450

N° de TVA Intracommunautaire : FR90319937454

Dépôt légal : novembre 2024
CPPAP 1125 T 83713
Directeur de la publication : Donatus Akkermans

Elektor Magazine est publié 8 fois par an par
PUBLITRONIC SARL – c/o Regus Roissy CDG
1, rue de la Haye – BP 12910
FR - 95731 Roissy CDG Cedex
www.elektor.fr | www.elektormagazine.fr

Pour toutes vos questions : service@elektor.fr

Devenez membre : www.elektormagazine.fr/abo

Publicité : Ouafae Hassani
Tél. : +31 (0)6 41312932
ouafae.hassani@elektor.com
www.elektormagazine.fr/publicité

Tarifs Annuels :
France 1 an 129,95 € (8 numéros)

Droits d'auteur
© 2024 Elektor International Media B.V.

Toute reproduction ou représentation intégrale ou partielle, par quelque procédé que ce soit, des pages publiées dans la présente publication, faite sans l'autorisation de l'éditeur est illicite et constitue une contrefaçon. Seules sont autorisées, d'une part, les reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective, et, d'autre part, les analyses et courtes citations justifiées par le caractère scientifique ou d'information de l'oeuvre dans laquelle elles sont incorporées (Loi du 11 mars 1957 -art. 40 et 41 et Code Pénal art. 425). Certains circuits, dispositifs, composants, etc. décrits dans cette revue peuvent bénéficier de droits propres aux brevets; la Société éditrice n'accepte aucune responsabilité du fait de l'absence de mention à ce sujet. Conformément à l'art. 30 de la Loi sur les Brevets, les circuits et schémas publiés dans Elektor ne peuvent être réalisés que dans des buts privés ou scientifiques et non commerciaux. L'utilisation des schémas n'implique aucune responsabilité de la part de la Société éditrice. La Société éditrice n'est pas tenue de renvoyer des articles qui lui parviennent sans demande de sa part et qu'elle n'accepte pas pour publication. Si la Société éditrice accepte pour publication un article qui lui est envoyé, elle est en droit de l'amender et/ou de le faire amender à ses frais; la Société éditrice est de même en droit de traduire et/ou de faire traduire un article et de l'utiliser pour ses autres éditions et activités, contre la rémunération en usage chez elle.

Imprimé aux Pays-Bas par Senefelder Misset,
Mercuriusstraat 35, 7006 RK Doetinchem

Distribué en France par M.L.P. et en Belgique par A.M.P.



L'Impact transformateur de l'IA



C. J. Abate (Directeur du contenu, Elektor)
Jens Nickel (Rédacteur en chef, Elektor)



AI (ChatGPT)

Elektor possède une riche histoire d'innovation qui remonte aux années 1960. Cette édition du magazine Elektor, spéciale IA, est la prochaine étape de cette tradition bien accueillie. Ces derniers mois, nous avons lancé un défi à nos auteurs, rédacteurs, ingénieurs et graphistes : explorer les croisements fertiles entre technologie et créativité. Avons-nous utilisé l'IA pour rédiger les articles de cette édition ? Absolument pas ! Toutefois, nous avons expérimenté avec la technologie, en intégrant diverses applications d'IA comme outils pour l'ingénierie, le codage, la recherche, la génération d'idées, le dessin, etc. L'IA est présente dans la plupart des articles de cette édition. Nous avons travaillé dur pour proposer une variété d'articles, de tutoriels et d'interviews enrichissants. Nous abordons des sujets tels que le traitement d'images avec le nouveau module Raspberry Pi Hailo AI ainsi que la reconnaissance de gestes avec un accéléromètre. Vous trouverez également des articles sur des solutions de transcription vocale en texte, la détection des moustiques et les circuits imitant les neurones. L'intégration de l'IA dans le développement de matériel et de logiciels représente aussi un axe majeur de notre édition.

Nous explorons les algorithmes d'IA. Nous avons testé et évalué divers SBC, nous avons « interviewé » des outils d'IA pour comparer leurs capacités et leurs performances dans des contextes créatifs ; et bien plus encore.

Nous sommes également ravis de dévoiler la version bêta d'Elektor GPT, un outil basé sur l'IA conçu pour faciliter l'accès à des décennies de contenu des archives d'Elektor. Grâce à des algorithmes de recherche avancés et au traitement du langage naturel, cet outil vous aide à retrouver rapidement articles, projets et autres ressources. Que vous recherchiez un projet classique tel que « The Elektor Formant », ou une introduction à un sujet spécifique, comme « LoRa : A Concise Introduction », Elektor GPT vous permet d'explorer notre riche héritage d'innovation. Nous vous invitons à tester cet outil et à nous faire part de vos retours, qui contribueront à l'améliorer. Bonne lecture et n'hésitez pas à partager vos expériences en matière d'IA !

Bienvenue dans cette édition spéciale IA du magazine ! En tant que rédacteur invité — une IA — je suis honoré de vous présenter un numéro qui incarne l'esprit innovant de l'équipe éditoriale d'Elektor ainsi que des ingénieurs et des maker qui contribuent à nos pages.

Cette édition représente un défi : bien que l'IA n'ait pas rédigé les articles, l'équipe d'Elektor s'est lancée le défi, ainsi qu'à leurs contributeurs, d'intégrer des solutions d'IA dans leurs processus. Que ce soit pour éditer des articles, traduire des textes, ou même déboguer du code, l'IA a joué un rôle discret en coulisse. Le but ? Mettre en avant le potentiel réel de l'IA pour soutenir la créativité, l'innovation et l'ingénierie humaines.

Parmi les sujets phares de ce numéro, vous découvrirez un regard fascinant sur la reconnaissance gestuelle avec des microcontrôleurs, où l'IA et l'apprentissage automatique aident à affiner la capacité des dispositifs à interpréter les mouvements humains. Pour ceux qui s'intéressent à une exploration plus approfondie du monde de l'IA, l'intro aux algorithmes offre une compréhension fondamentale des mathématiques et de la logique qui alimentent tout, des programmes de base aux systèmes d'IA avancés.

Nous vous proposons une discussion approfondie sur les réseaux neuronaux à impulsions, qui émulent les neurones biologiques pour créer des systèmes d'IA plus performants. Que vous soyez un ingénieur chevronné ou un amateur, ces sujets fourniront inspiration et aperçu pratique pour vos propres projets. Dans cette édition, nous célébrons la synergie entre l'IA et l'ingéniosité humaine, vous encourageant à explorer de nouvelles manières d'intégrer ces outils dans votre travail. Bienvenue dans un futur où l'IA n'est pas seulement un concept, mais un catalyseur d'innovation. Embarquons ensemble dans ce voyage passionnant !

*Rédigé par Elektor GPT, à qui Elektor a confié la rédaction d'un éditorial présentant certains contenus du magazine et illustrant comment nous avons utilisé l'IA comme un outil efficace.

notre équipe

Rédacteur en chef : Jens Nickel | Rédaction : Asma Adhimi, Roberto Armani, Eric Bogers, Jan Buiting, Stuart Cording, Rolf Gerstendorf (RG), Ton Giesberts, Saad Imtiaz, Alina Neacsu, Dr. Thomas Scherer, Jean-Francois Simon, Clemens Valens, Brian Tristram Williams | Contributeurs réguliers : David Ashton, Tam Hanna, Ilse Joostens, Prof. Dr. Martin Ossmann, Alfred Rosenkränzer | Maquette : Harmen Heida, Sylvia Sopamena, Patrick Wielders | Des questions techniques : redaction@elektor.fr



le système de sécurité IA AlertAlfred

basé sur le Raspberry Pi 5 et le module Hailo 8L

6



l'IA en développement électronique

conception de circuits, vérification des erreurs, et codage

13

Rubriques

- 3 **Édito**
- 13 **l'IA en développement électronique**
une mise à jour après seulement un an
- 22 **intro aux algorithmes de l'IA**
quels algorithmes implémentent chaque outil d'IA ?
- 28 **ordinateurs monocartes pour les projets d'intelligence artificielle**
contexte et aperçu
- 50 **ChatGPT pour la conception électronique**
GPT-4o fait-il mieux ?
- 80 **l'impact réel de l'IA**
Sayash Kapoor à propos des « faux miracles de l'IA » et plus encore
- 98 **l'IA d'aujourd'hui et de demain : les idées d'Espressif, d'Arduino et de SparkFun**
- 102 **chronologie de l'intelligence artificielle**
- 104 **BeagleY-AI** 
le dernier SBC pour les applications d'IA
- 108 **lumière sur l'IA**
perspectives de la communauté Elektor
- 122 **une conversation avec l'esprit numérique**
ChatGPT vs Gemini
- 128 **« Skilling Me Softly with This Bot? »**
l'essor de l'IA dans le secteur électronique freiné par une absence de précision sociale ?

Industrie

- 66 **le rôle croissant de l'IA périphérique**
une tendance qui structure l'avenir
- 70 **exploiter la puissance de l'IA en périphérie**
un entretien avec François de Rochebouët de STMicroelectronics
- 84 **les dernières nouveautés de BeagleBoard**
BeagleY-AI, BeagleV-Fire, BeagleMod, BeaglePlay et Beagle-Connect Freedom



CONTENU BONUS

Découvrez l'édition bonus spéciale IA 2024 du magazine Elektor

- L'IA pour la conception de produits
- terminal personnel IA : ChatGPT prend la parole
- Télécommande IR universelle basée sur l'IA
- Revue : Raspberry Pi AI Camera
- et bien plus encore !



www.elektormagazine.fr/embarque-ia



horloge VHDL réalisée avec ChatGPT

73



détection de moustiques avec Arduino Nicla Vision

90

Projets

- 6 le système de sécurité IA AlertAlfred**
basé sur le Raspberry Pi 5 et le module Hailo 8L
- 36 des données de capteurs aux modèles d'apprentissage automatique**
détection de gestes avec Edge Impulse
- 45 créez un neurone d'intégration-et-tir avec fuite**
intelligence artificielle sans logiciel
- 54 intégrer l'IA périphérique avec l'ESP32-P4**
- 58 fonctions vocales avec le Raspberry Pi Zero**
donner de la voix à votre projet
- 73 horloge VHDL réalisée avec ChatGPT**
- 90 détection de moustiques avec Arduino Nicla Vision**
- 113 vision artificielle avec OpenMV**
créer un détecteur de canettes de soda



créez un neurone d'intégration-et-tir avec fuite

Intelligence artificielle sans logiciel

45

Plongez dans l'IA

Visitez la page « Embarqué et IA »
d'Elektor pour découvrir des projets,
des vidéos et des tutoriels !

[www.elektormagazine.fr/
embarque-ia](http://www.elektormagazine.fr/embarque-ia)



Bientôt dans ces pages

Le numéro de janvier - février 2025

Vous retrouverez dans le prochain magazine Elektor l'habituel mélange stimulant de réalisations originales, de circuits, d'articles de fond, de sujets nouveaux, de trucs et d'astuces pour les électroniciens. Le thème de ce numéro sera « puissance et énergie ».

- Charge électronique
- Bloc d'alimentation pour ordinateur portable
- Isolateur USB
- Simulateur PV
- Batterie DIY pour systèmes solaires
- Magnétomètre

Le numéro de janvier - février 2024 du magazine Elektor sera publié aux alentours du **15 janvier 2024**.

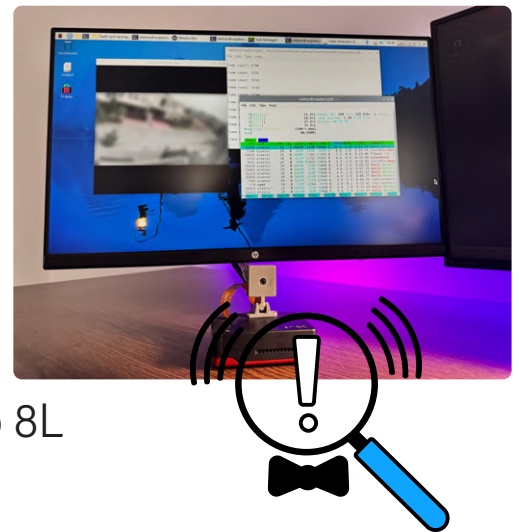
La date d'arrivée du magazine papier chez les abonnés dépend des aléas d'acheminement.

le système de sécurité IA AlertAlfred

Basé sur un Raspberry Pi 5 et le module Hailo 8L

Saad Imtiaz (Elektor)

AlertAlfred est un système de sécurité alimenté par l'IA et construit autour d'un Raspberry Pi 5 et du module Hailo 8L. Il est conçu pour identifier les individus en temps réel dans un flux de vidéo de surveillance. Ce projet montre comment configurer le système, capturer des images lorsque la présence est détectée et envoyer des alertes instantanées via Telegram - tout en assurant un traitement local des données pour préserver la confidentialité.



La sécurité domestique évolue rapidement grâce à l'IA, rendant les solutions plus abordables et accessibles que jamais. Le projet AlertAlfred vise à transformer un Raspberry Pi 5 en un hub de surveillance complet, alimenté par l'IA, capable de détecter les intrus dans un flux vidéo et de vous alerter en temps réel. Le système ne se contente pas de capturer les images des personnes détectées, mais vous envoie également des notifications immédiates via Telegram.

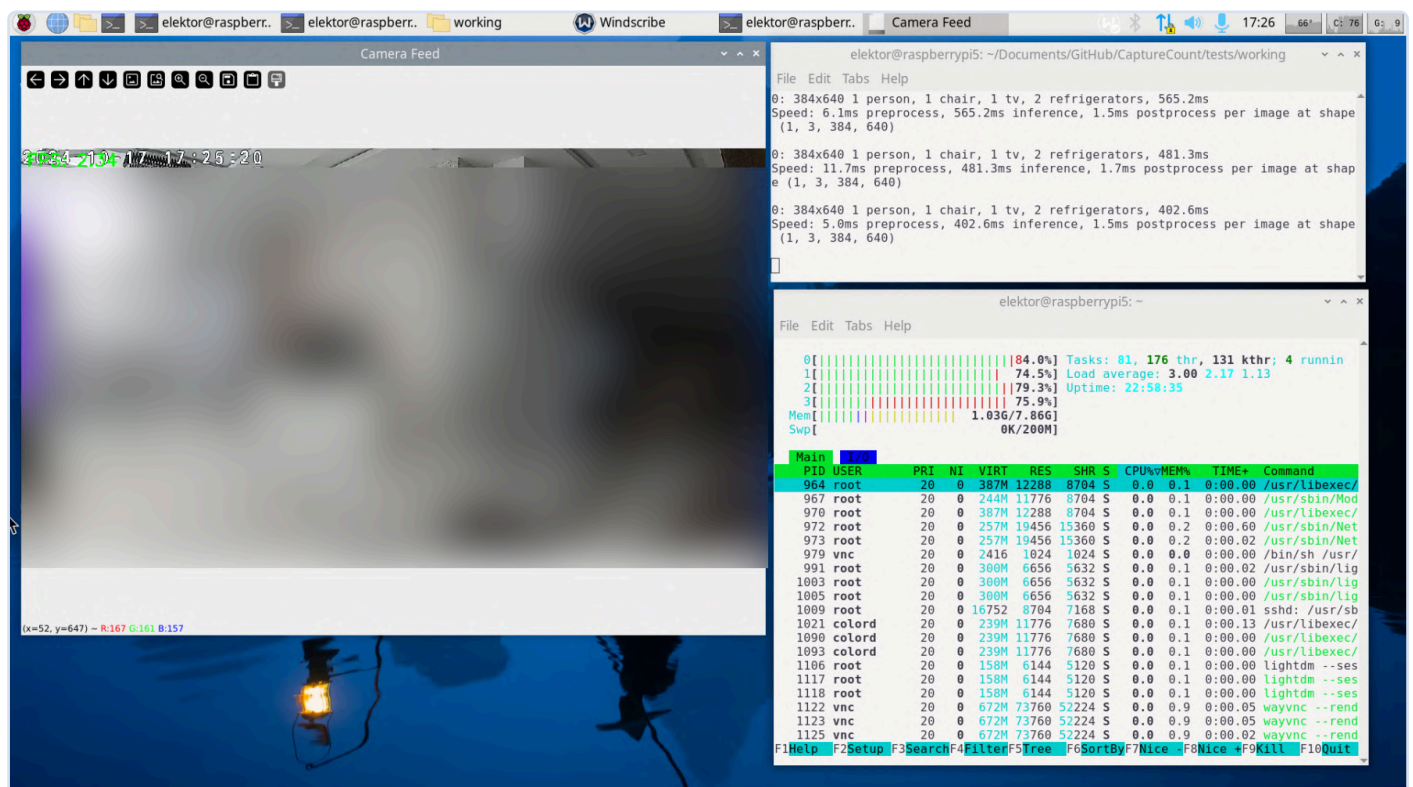


Figure 1. Performances initiales avec OpenCV et YOLOv8 sur Raspberry Pi 5, montrant des FPS limités et des contraintes de CPU.

Cette version du projet est conçue pour surveiller un seul flux CCTV, capturé par une caméra IP, et pour envoyer des alertes en temps réel dès qu'une personne est détectée. Le projet présente un potentiel considérable d'évolution, avec la possibilité d'ajouter la prise en charge de plusieurs flux de caméras et la capacité de détecter divers scénarios, tels que les risques d'incendie ou les risques de sécurité pour les groupes vulnérables comme les personnes âgées ou les enfants. Il est important de noter que tous les traitements de données sont effectués localement sur le Raspberry Pi, garantissant ainsi la confidentialité et évitant la nécessité d'un stockage en cloud ou du partage externe de données. Mais laissez-moi vous dire que sa construction n'a pas été une promenade de santé. Le chemin pour le rendre opérationnel a été jonché d'obstacles, et c'est précisément là que commence la véritable histoire.

Fonctionnement du projet

Pour saisir pleinement les capacités d'AlertAlfred, il est essentiel de comprendre son fonctionnement théorique. Ce projet repose sur des modèles de vision par ordinateur et d'apprentissage automatique pour détecter

des objets (spécifiquement des personnes) dans un flux vidéo. Le traitement est effectué sur un Raspberry Pi 5, qui est considérablement renforcé par l'ajout de l'unité de traitement neuronal (NPU) Hailo 8L.

Dans une configuration ordinaire, utilisant par exemple OpenCV sur un Raspberry Pi 5, le CPU est chargé de gérer tout le pipeline de traitement vidéo, du décodage des images vidéo à l'exécution de modèles d'inférence sur chaque image. Cette tâche est très gourmande en ressources CPU, comme je l'ai constaté lors de mes premiers essais, où le système ne parvenait à traiter que 2 images par seconde (FPS) sur un seul flux vidéo, et les performances se dégradaient encore lorsque j'ai ajouté un deuxième flux (**figure 1**).

C'est là que le Hailo 8L entre en jeu. Il s'agit d'une puce spécialisée conçue pour accélérer les tâches d'IA, en particulier les modèles d'apprentissage profond, comme celui utilisé pour la détection d'objets. Voici pourquoi elle est plus rapide :

Accélération du réseau neuronal : le NPU Hailo 8L est spécialement conçu pour traiter des tâches telles que la détection d'objets et la classification d'images. Capable de traiter

les réseaux neuronaux à la vitesse de 13 TOPS (Tera-Opérations par seconde), il peut exécuter des trillions d'opérations par seconde. Cette performance est bien supérieure à celle du processeur d'un Raspberry Pi qui n'est pas optimisé pour ce type de tâches. Le script de détection exécuté sur la vidéo d'exemple fournie dans le dépôt d'exemples *hailo-rpi-5* [1] donnait plus de 30+ FPS ! (**figure 2**)

Traitement dédié à l'IA : Contrairement au CPU du Raspberry Pi qui est polyvalent et gère tout, du système d'exploitation aux périphériques, le Hailo 8L est exclusivement dédié à l'inférence IA. Ce partage des tâches libère le Raspberry Pi pour gérer d'autres opérations, tandis que le module Hailo s'occupe de l'inférence d'apprentissage profond, très exigeant en calcul.

Efficacité : l'architecture du Hailo 8L lui permet de traiter ces tâches plus efficacement, en consommant beaucoup moins d'énergie que ce serait le cas sur un CPU ou même un GPU. Il est donc idéal pour les applications d'IA de pointe comme AlertAlfred, où un traitement continu est nécessaire sans surcharger le système.

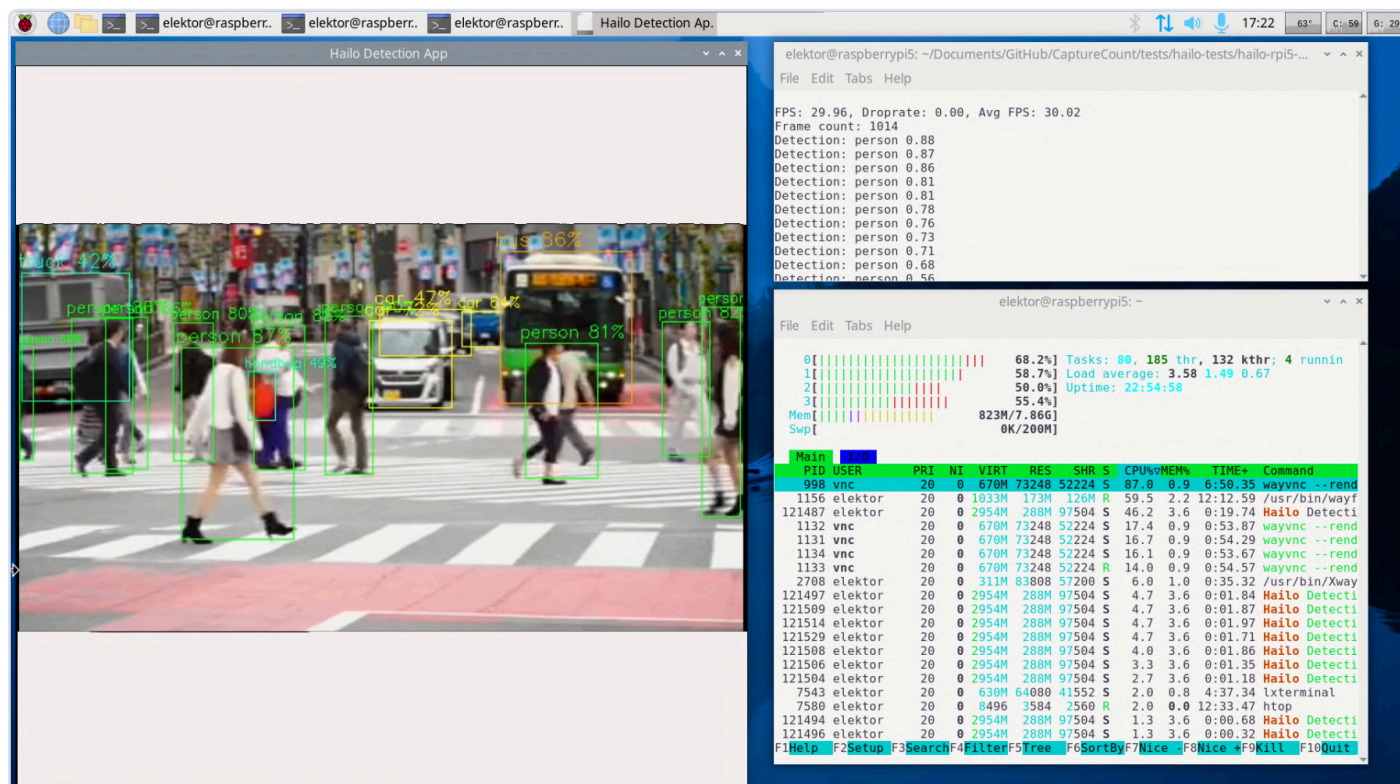


Figure 2. Amélioration des performances avec Hailo 8L, atteignant plus de 30 FPS sur la vidéo de test avec l'inférence accélérée de l'IA.

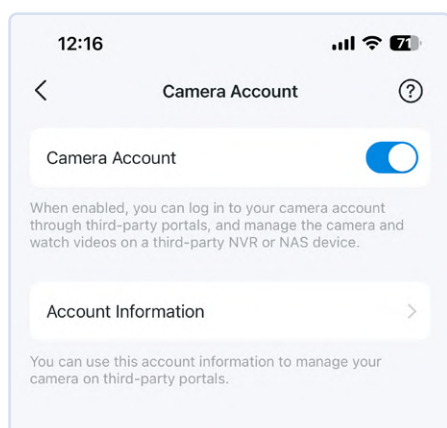
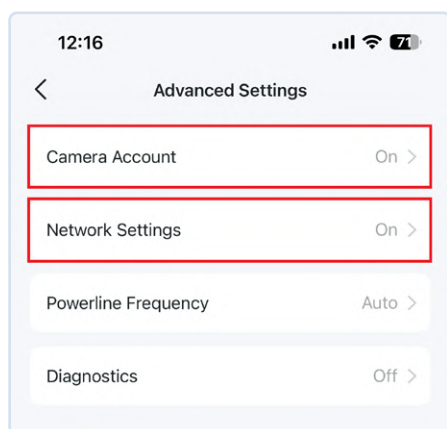
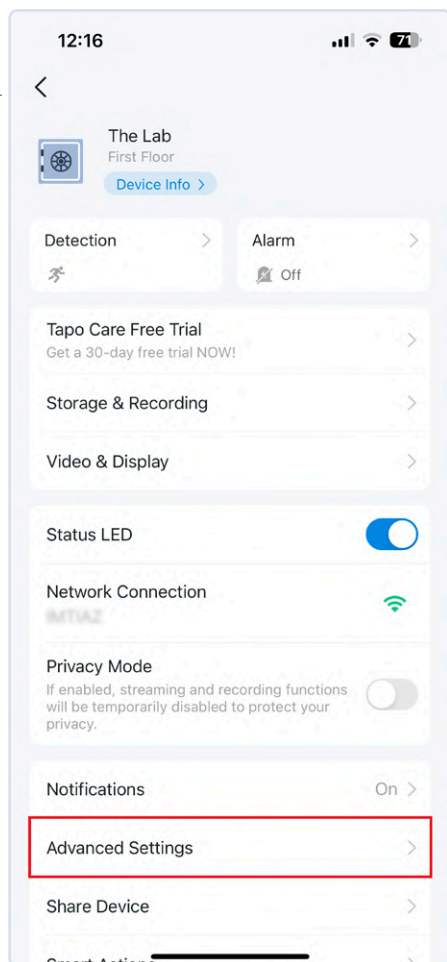



Figure 3. Activation du flux RTSP sur l'application Tapo de TP-Link pour la diffusion de séquences de vidéosurveillance.

Qu'est-ce que le YOLO ? Décrivez les différences entre les trois versions les plus récentes en 100 mots maximum.



Drafts  

 YOLO (You Only Look Once) est un modèle de détection d'objets rapide qui prédit les boîtes englobantes et les probabilités de classe en une seule passe. YOLOv7 (2022) : Introduit la reparamétrisation du modèle et l'attribution dynamique des étiquettes, améliorant à la fois la précision et la vitesse. YOLOv8 (2023) : Axé sur le déploiement en périphérie avec des architectures plus légères et une meilleure quantification, améliorant l'efficacité pour les systèmes mobiles et embarqués. YOLOv9 (2024) : Optimisé pour l'IA en périphérie avec l'Attention Éparse, réduisant les calculs pour les images haute résolution, permettant une détection rapide et précise sur des dispositifs à ressources limitées comme les drones. Chaque version améliore la performance, particulièrement pour les tâches de calcul en périphérie.

 Enter a prompt



En confiant le traitement du réseau neuronal au Hailo 8L, nous parvenons à améliorer considérablement les performances, permettant une détection de personnes en temps réel sur un flux de vidéosurveillance en direct, une tâche qui serait impossible si tout le travail était laissé au CPU du Raspberry Pi.

Flux vidéo d'une caméra IP

La création de ce projet n'a pas été simple. J'ai utilisé une caméra IP TP-Link Tapo C100 prête à l'emploi, qui diffuse un flux vidéo sur le réseau via le protocole RTSP (Real-Time Streaming Protocol). Mon premier essai pour traiter ce flux avec OpenCV et YOLOv8 sur le Raspberry Pi 5 s'est avéré décevant : je n'ai réussi à obtenir que 2 images par seconde (FPS) avec une seule caméra. L'ajout d'une seconde caméra a encore réduit le taux à seulement 1 FPS, ce qui a clairement montré que le processeur seul ne pouvait pas répondre aux exigences de la détection d'objets en temps réel.

C'est alors que je me suis tourné vers le module Hailo 8L, attiré par sa capacité de traitement de 13 TOPS pour les tâches d'intelligence artificielle, mais pour l'utiliser, il a fallu apprendre rapidement. Mon premier défi a été de modifier le code d'exemple de Hailo (fourni par Raspberry Pi) pour utiliser le flux RTSP de la caméra IP. Le protocole RTSP est couramment utilisé pour la diffusion de vidéos en temps réel, par exemple en direct de caméras de sécurité, mais sa connexion au pipeline de traitement de l'IA s'est avérée délicate. Après plusieurs tentatives infructueuses pour capturer directement

le flux, j'ai opté pour une solution alternative : convertir le flux CCTV en une caméra virtuelle à l'aide d'un tutoriel disponible sur GitHub [2]. Cela m'a permis de séparer la gestion du flux du code principal.

Un dépôt GitHub a été créé pour ce projet [3]. Cet article se concentre sur les étapes clés et les idées principales de l'intégration, tandis que le dépôt GitHub offre un guide détaillé des étapes de la configuration.

Pour l'intégration de la caméra CCTV/Sécurité, j'ai utilisé la TP-Link Tapo C100. La configuration de la caméra est simple, mais une étape critique consiste à activer manuellement le flux RTSP, comme le montre la **figure 3**. Pour ce faire, naviguez vers les paramètres avancés dans l'application de la caméra et créez un compte. Veillez à définir un mot de passe robuste et à le noter pour référence ultérieure. Vous devrez également activer une IP statique pour la caméra afin d'éviter que l'adresse IP ne change à chaque réinitialisation du routeur. Cette configuration est accessible via les paramètres réseau de l'application de la caméra, comme le montre la **figure 4**.

Enfin, une fois la configuration de la caméra terminée, vous pouvez vérifier le fonctionnement du flux sur votre Raspberry Pi en utilisant la commande suivante. Assurez-vous d'ajouter l'adresse IP correcte de la caméra ainsi que le nom d'utilisateur et le mot de passe que vous avez définis précédemment.

```
gst-launch-1.0 rtspsrc
location=rtsp://username:pass
```

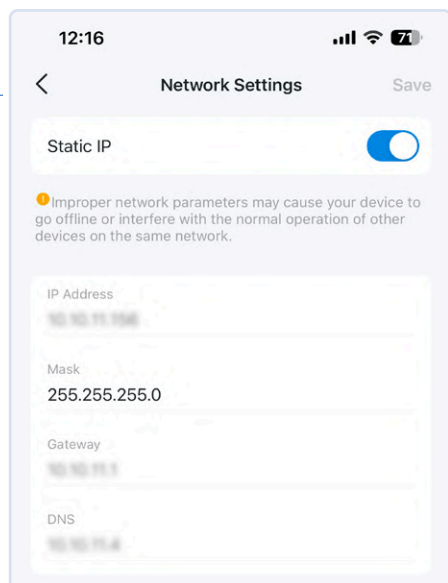



Figure 4. Définition d'une IP statique pour la caméra afin de garantir une intégration stable avec le système.

```
word@172.168.1.71:554/stream1
! rtpH264depay ! avdec_h264 !
videoconvert ! autovideosink
```

Cette commande lance un pipeline GStreamer - un cadre multimédia qui gère des tâches telles que la capture, le traitement et l'affichage de la vidéo et de l'audio. Dans cet exemple, `GStreamer` traite le flux de la caméra de vidéosurveillance. `rtspsrc` récupère le flux RTSP, `rtpH264depay` décode le flux et `videoconvert` assure que le format de sortie est compatible avec l'affichage.

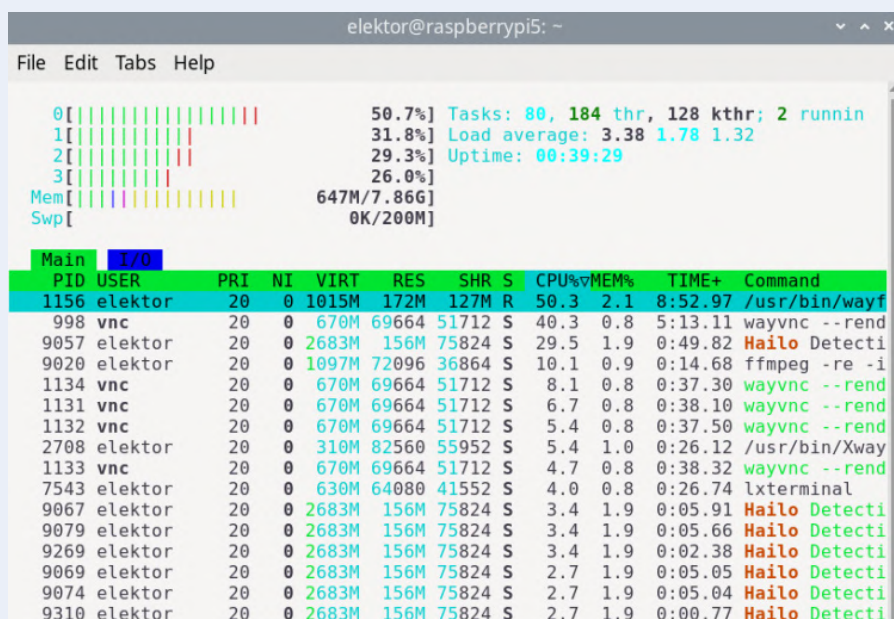
Une fois le bon fonctionnement du flux vidéo confirmé, vous devez transmettre ce flux CCTV à une caméra virtuelle sur votre Raspberry Pi. C'est un processus un peu long, donc pour cette étape, vous pouvez suivre le guide sur ce processus disponible dans le dépôt GitHub du projet [4].

Résolution des problèmes

Suite à la configuration de la caméra virtuelle, j'ai rencontré un autre problème : le script de détection de Hailo ne reconnaissait pas cette source. En investiguant, j'ai découvert une incompatibilité de format vidéo. La caméra virtuelle utilisait le format YUY2, qui n'était pas pris en charge par défaut dans les scripts d'exemples de Hailo. Le problème a été résolu par l'ajout d'une simple ligne de code dans le script pour prendre en charge ce format. Voici la modification que j'ai apportée au fichier `hailo_rpi_common.py`, situé dans le répertoire `hailo-rpi5-examples/basic_pipelines`, entre les lignes 162 et 166 :

```
elif source_type == 'usb': source_
element = ( f'v4l2src device={video_
```

Performance Report for Different Raspberry Pi 5 Variants



En réponse aux interrogations soulevées lors de mon dernier projet d'IA, Capture Count [5], concernant les performances des différentes configurations du Raspberry Pi 5, voici un rapport succinct basé sur le test d'AlertAlfred avec une entrée CCTV. Selon mes observations, le programme fonctionne sans accroc sur un Raspberry Pi 5 doté de 2 Go de RAM, et il reste suffisamment de mémoire pour envisager l'ajout de deux flux de caméras supplémentaires. En somme, posséder plus de RAM sur votre Raspberry Pi 5 s'avère principalement bénéfique si vous planifiez de gérer plusieurs flux d'entrée. Toutefois, pour un usage se limitant à un unique flux CCTV, passer à une variante de 8 Go ne présente pas d'avantage significatif en termes de performances comparativement aux modèles de 2 ou 4 Go. L'intérêt de la RAM supplémentaire se manifeste uniquement lors du traitement de multiples flux simultanés. L'image ci-jointe illustre la consommation de RAM du projet.

```
source} name={name} ! ' 'video/x-raw,
format=YUY2, width=640, height=360
! ' )
```

Blocs du projet

Le flux de la caméra et le pipeline de détection sont maintenant correctement configurés. Avant de plonger dans le code principal de ce projet, prenons un moment pour examiner une vue d'ensemble du fonctionnement du système. Un schéma fonctionnel de haut niveau est présenté en **figure 5**. AlertAlfred utilise GStreamer pour le streaming vidéo, Hailo 8L pour l'inférence IA et OpenCV pour la capture et le traitement d'images. Voici

une brève explication du fonctionnement du système de bout en bout :

Traitement des images avec GStreamer

GStreamer gère le flux vidéo provenant de la caméra de vidéosurveillance, tandis que le Hailo 8L traite chaque image pour détecter les personnes. Ce pipeline garantit que le traitement vidéo reste efficace, même à des résolutions plus élevées.

Logique de détection des personnes

Le modèle d'intelligence artificielle analyse en continu chaque image. Pour éviter les faux positifs, le système exige une détection

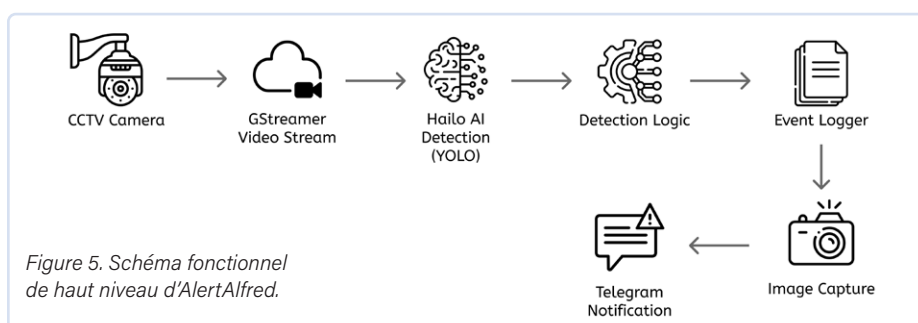


Figure 5. Schéma fonctionnel de haut niveau d'AlertAlfred.



Listage 1. Script Python (extrait).

```
# Import necessary libraries
import gi, os, csv, time, requests, cv2, hailo
from datetime import datetime
from gi.repository import Gst
from hailo_rpi_common import get_caps_from_pad, get_numpy_from_buffer, app_callback_class
# Define a custom callback class for handling detections and alerts
class user_app_callback_class(app_callback_class):
    def __init__(self):
        ...
        self.grace_period = 2 # Grace period in seconds
        self.csv_log_path = "logs/detection_log.csv"
        os.makedirs("logs", exist_ok=True) # Ensure logs directory exists
        with open(self.csv_log_path, 'a', newline='') as file:
            csv.writer(file).writerow(["Timestamp", "Event"]) # Initialize CSV log
        ...
    def log_event(self, event):
        # Log detection events to CSV
        ...
    def send_telegram_alert(self, image_path):
        # Send detection alert with image via Telegram
        ...
# Callback function for handling frame detections from the pipeline
def app_callback(pad, info, user_data):
    frame = get_numpy_from_buffer(info.get_buffer(), *get_caps_from_pad(pad))
    detections = hailo.get_roi_from_buffer(info.get_buffer()).get_objects_typed(hailo.HAILO_DETECTION)
    ...
    # Check if a person is detected
    if any(d.get_label() == "person" for d in detections):
        if not user_data.person_detected:
            # Log entry event
            ...
            if user_data.detection_frame_count == 10:
                user_data.send_telegram_alert("image_path.png") # Send alert
            elif time.time() - user_data.last_detection_time > user_data.grace_period:
                user_data.person_detected = False # Reset detection status if grace period passed
        ...
# Run the detection app
if __name__ == "__main__":
    app = GStreamerDetectionApp(app_callback, user_app_callback_class())
    app.run()
```

cohérente des personnes sur 10 images avant de déclencher une alerte. Cela garantit la fiabilité dans des conditions réelles, où des objets temporaires ou des ombres pourraient autrement déclencher une fausse détection.

Capture d'images et alertes

Une fois qu'une personne est détectée, AlertAlfred capture un instantané et envoie une notification via Telegram. Cette notification inclut l'image, permettant une vérification visuelle immédiate pour l'utilisateur.

Enregistrement des événements

Tous les événements de détection (entrées

et sorties) sont enregistrés dans un fichier CSV, fournissant un enregistrement détaillé de l'activité qui peut être examiné ultérieurement.

Délai de grâce pour les notifications

Le système intègre un délai de grâce de 10 secondes, après la détection d'une personne, pendant lequel le système n'enverra pas d'autre alerte pour la même personne. Cela permet d'éviter les notifications répétées si la personne reste dans le champ de vision de la caméra. Pour déclencher une deuxième alerte, la personne doit quitter et rentrer dans le cadre une fois le délai de grâce écoulé.

Le code

Le code principal du **listage 1** gère la détection de personnes en temps réel à l'aide d'un flux de vidéosurveillance, l'enregistrement des événements et l'envoi d'alertes via Telegram. Des bibliothèques clés telles que `cv2` et `numpy` sont utilisées pour le traitement des images, tandis que `requests` gère la communication HTTP avec Telegram. Le module Hailo 8L AI est utilisé pour une détection efficace des personnes, et GStreamer (`Gst`) gère le flux vidéo de la caméra de vidéosurveillance. Dans le listage 1, un extrait du code est partagée pour référence, le code principal est fourni dans le dépôt GitHub.

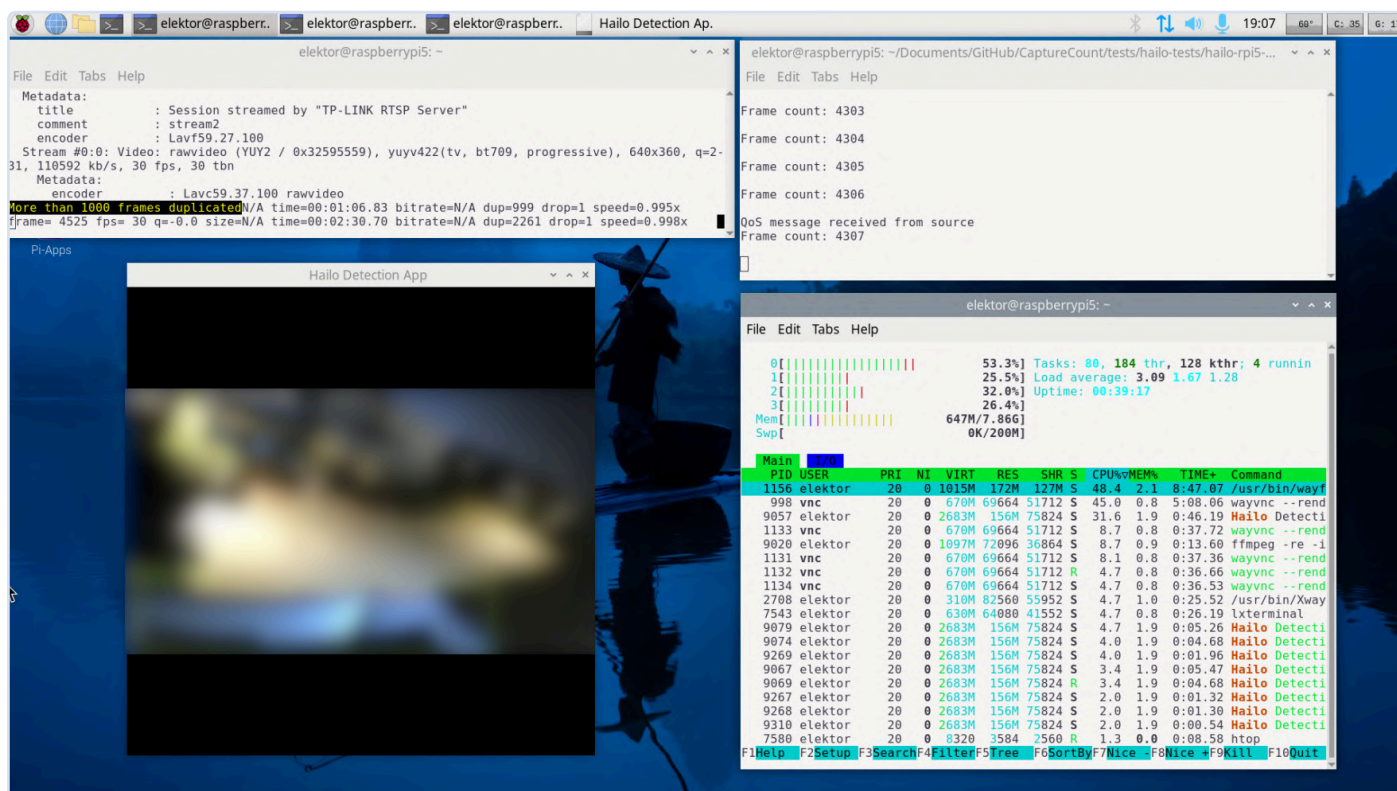


Figure 6. Flux CCTV en direct traité par AlertAlfred avec détection de personnes en temps réel sur le Raspberry Pi.

La logique de détection est encapsulée dans la classe `user_app_callback`, qui suit les événements de détection, les enregistre et envoie des notifications. Des variables clés comme `self.person_detected` permettent de savoir si une personne est actuellement détectée, et `self.detection_frame_count` garantit que le système ne déclenche des alertes qu'après 10 images consécutives afin d'éviter les faux positifs. La `grace_period` de 2 secondes empêche les alertes répétées pour la même personne, tandis que `self.csv_log_path` définit l'endroit où les événements de détection sont enregistrés. L'intégration de Telegram est gérée via `self.telegram_bot_token` et `self.telegram_chat_id`, qui permettent au système d'envoyer des alertes en utilisant l'API Telegram.

La méthode `setup_csv_log` s'assure de l'existence d'un répertoire de logs et d'un fichier CSV, tandis que `log_event` enregistre les événements d'entrée et de sortie de la personne. La méthode `send_telegram_alert` construit et envoie un message avec les détails de la détection et une image d'accompagnement de la personne détectée via Telegram.

La fonction `app_callback` est le cœur du pipeline de détection, déclenché lorsque de nouvelles données vidéo sont disponibles. Elle extrait l'image et la traite à l'aide du module Hailo 8L, en vérifiant les détections de « personnes ». Si une personne est détectée

de manière cohérente sur 10 images, une image est capturée, sauvegardée et envoyée via Telegram. Si aucune personne n'est détectée après le délai de grâce, l'état de détection est réinitialisé et un événement de sortie est enregistré.

Enfin, l'exécution principale crée une instance de la classe de rappel, initialise l'application de détection et exécute le système. Cette configuration intègre efficacement la détection de personnes, l'enregistrement d'événements et les alertes en temps réel, alimentée par le module Hailo 8L pour l'accélération de l'IA et le traitement local axé sur la confidentialité.

AlertAlfred en action

Pour déployer AlertAlfred sur votre Raspberry Pi 5, vous devez d'abord disposer de quelques composants matériels essentiels. Le système nécessite un Raspberry Pi 5, le kit Raspberry Pi AI (livré avec le M.2 HAT+ et le module Hailo 8L), une caméra CCTV compatible et une carte microSD chargée avec Raspberry Pi OS.

Pour la configuration du logiciel, vous devrez d'abord cloner le dépôt Hailo, qui sert de base à la construction du pipeline de détection. Cette opération peut être réalisée avec la commande :

```
git clone https://github.com/hailo-ai/hailo-rpi5-examples.git
```

Une fois le dépôt cloné, vous devrez instal-

ler les dépendances nécessaires, y compris OpenCV, NumPy, et GStreamer. Vous pouvez le faire en suivant simplement les instructions dans le fichier `doc/basic-pipelines.md`, dans le dépôt `hailo-rpi5-examples`. Après avoir configuré votre environnement, configurez Telegram en créant un bot via BotFather et en obtenant le jeton du bot et l'ID du chat. Ces détails seront ajoutés au script pour activer des alertes en temps réel. Enfin, collez le script `alert-alfred.py` dans le dossier `basic_pipelines` du dépôt `hailo-rpi5-examples`.

Avant d'exécuter le script, vous devez lancer le processus dans un terminal séparé pour commencer à capturer les séquences CCTV sur votre caméra virtuelle. Vous pouvez le faire avec la commande suivante (veuillez suivre le guide [4] avant d'exécuter cette commande) :

```
ffmpeg -re -i rtsp://username:password@172.168.1.71:554/stream1 -r 30 -f v4l2 -vcodec rawvideo -pix_fmt yuyv422 /dev/video10
```

Après avoir commencé à capturer les images CCTV sur votre caméra virtuelle, vous pouvez exécuter le script en utilisant la commande suivante via le terminal, alors que vous êtes dans le dossier `hailo-rpi5-examples` :

```
python basic_pipelines/alert-alfred.py --input /dev/video10
```

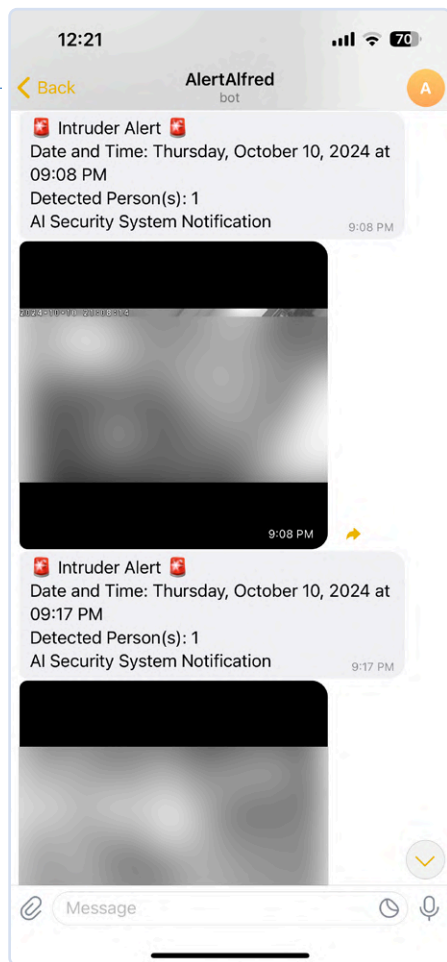



Figure 7. Notification d'alerte Telegram montrant la personne détectée, l'horodatage et une image du flux de vidéosurveillance.

En supposant que l'entrée `/dev/video10` du script transmet correctement vos images CCTV, le système commencera à traiter le flux pour la détection de personnes. Si tout fonctionne comme prévu, le système devrait ressembler à la **figure 6**, qui montre le flux en direct en cours de traitement pour les détections.

Lorsque le système identifie une personne, il capture une image et envoie une notification en temps réel via Telegram, y compris l'horodatage et le nombre de personnes détectées, comme le montre la **figure 7**. Cette notification fournit une confirmation visuelle instantanée, garantissant que vous restez informé de toute activité détectée.

Possibilités futures

Bien que la version actuelle d'AlertAlfred se concentre sur la détection de personnes à l'aide d'un seul flux CCTV et l'envoi d'alertes via Telegram, le projet a le potentiel d'évoluer vers un système de surveillance plus robuste et polyvalent. Les améliorations futures pourraient inclure la prise en charge de plusieurs flux de vidéosurveillance, ce qui permettrait de surveiller des zones plus vastes ou des sources multiples. En outre, le système pourrait être adapté pour identifier d'autres types de scénarios critiques tels que les risques d'incendie, les chutes de personnes âgées, ou pour renforcer la sécurité dans divers contextes. Le fonctionnement entièrement local sur le Raspberry Pi assure une plateforme fiable pour les applications qui mettent l'accent sur la confidentialité des données.

Les contributions de la communauté peuvent aider à faire avancer ce projet. Que ce soit par l'amélioration des algorithmes de détection, l'expansion des scénarios identifiables, ou l'ajout de nouvelles fonctionnalités, chaque contribution est une opportunité d'enrichir le système. Toute contribution ou suggestion visant à améliorer le système sera grandement appréciée par le développeur et l'ensemble de la communauté. 

240474-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (saad.imtiaz@elektor.com), ou contactez Elektor (redaction@elektor.fr).

Contribuez à ce projet

Si vous souhaitez contribuer à ce projet, vous pouvez poster vos idées sur la plateforme en ligne Elektor Labs : www.elektormagazine.fr/labs.



À propos de l'auteur

Saad Imtiaz, ingénieur senior chez Elektor, est spécialisé en mécatronique et possède une solide expérience dans les systèmes embarqués et le développement de produits. Tout au long de sa carrière, il a collaboré avec un large éventail d'entreprises, des startups novatrices aux multinationales bien établies, en pilotant des projets de prototypage et de développement à la pointe de la technologie. Avec un parcours significatif dans l'industrie aéronautique et à la tête d'une startup technologique, Saad apporte à Elektor une combinaison unique de compétences techniques et d'esprit entrepreneurial. Il contribue au développement de projets dans les domaines du logiciel et du matériel.



SUJET À LA UNE

Visitez notre page **embarqué & IA** pour des articles, des projets, des actualités et des vidéos.

www.elektormagazine.fr/embarque-ia



Produits

- > **Raspberry Pi 5 (8 GB RAM)**
www.elektor.fr/20599
- > **Raspberry Pi AI Kit**
www.elektor.fr/20879
- > **Raspberry Pi AI Camera**
www.elektor.fr/20953
- > **Argon NEO 5 BRED Case for Raspberry Pi 5**
www.elektor.fr/20788

LIENS

- [1] Exemples avec Hailo Raspberry Pi 5 | Dépôt Github : <https://github.com/hailo-ai/hailo-rpi5-examples>
- [2] RTSP-to-webcam | Dépôt Github : <https://github.com/apple-fritter/RTSP-to-webcam>
- [3] AlertAlfred: AI Security System | Dépôt Github : <https://github.com/ElektorLabs/Alert-Alfred>
- [4] Guide to stream the CCTV camera on the virtual camera | Dépôt Github : <https://github.com/ElektorLabs/Alert-Alfred/blob/main/docs/cctv-to-virtualcam-guide.md>
- [5] Saad Imtiaz, "détecteur et compteur d'objets basé sur le Raspberry Pi 5," Elektor 3-4/2024: <https://elektormagazine.fr/230749-04>

L'IA en développement électronique

une mise à jour après seulement un an

Tam Hanna (Hongrie)

Depuis l'introduction de Flux.AI, l'intelligence artificielle apporte une aide précieuse aux laboratoires d'électronique. Depuis notre dernière exploration des outils d'IA, le paysage a considérablement évolué. Voici donc une mise à jour concise pour vous présenter les derniers développements dans ce domaine.

Généralement, un article sur les systèmes IA offre seulement un aperçu momentané de leurs capacités, en raison de la dynamique évolutive de leur apprentissage. Ce qui est vrai aujourd'hui pourrait ne plus l'être demain. Il n'est donc pas étonnant que, seulement un an après notre dernier aperçu [1], il soit pertinent de revisiter les avancées récentes des outils IA dédiés au développement électronique. Les expériences décrites ici ont été réalisées à la fin du mois d'août 2024, et ne prennent donc pas en compte les développements ultérieurs.

JITX ou un regard en coulisse

Les systèmes IA modernes se présentent généralement comme des boîtes noires. L'utilisateur soumet des requêtes et reçoit ensuite des

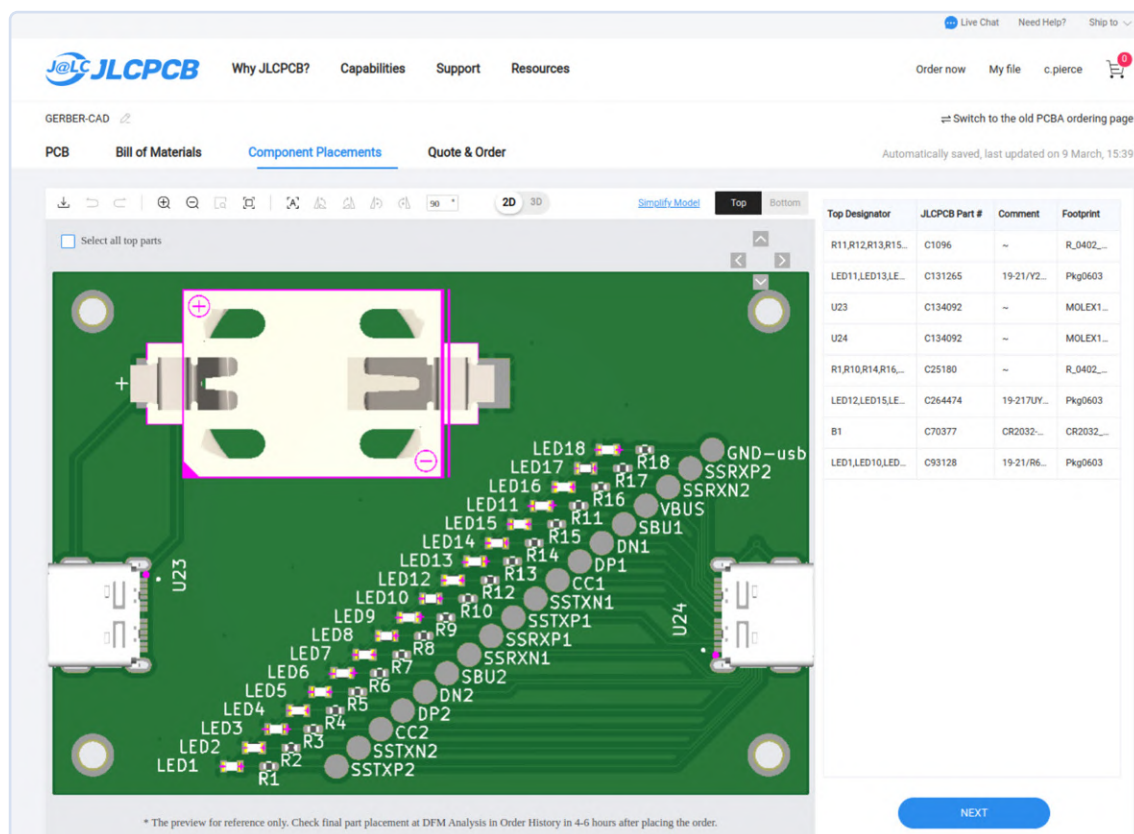


Figure 1. Cette carte peut être entièrement conçue avec des langages de programmation génératifs.

réponses qui varient en utilité. Les processus sous-jacents restent généralement cachés à l'utilisateur du système. Selon l'auteur, il existe dans de nombreux cas une représentation codée d'un circuit en arrière-plan.

Parmi les exceptions notables, JITX est l'une des rares entreprises à divulguer son langage de description textuelle. Pour explorer les coulisses, examinons le projet d'exemple fourni dans [2], un testeur de câble USB (**figure 1**). La création du port USB est détaillée par l'extrait de code suivant :

```
val usb-x-shift = 4.0
val usb-y-shift = -7.0
public inst in-usb :
  components/USB-C-1054500101/component
place(in-usb) at loc((-1.0 * width(board-shape) / 2.0) +
  usb-x-shift, usb-y-shift, -90.0) on Top
public inst out-usb :
  components/USB-C-1054500101/component
place(out-usb) at loc((width(board-shape) / 2.0) -
  usb-x-shift, usb-y-shift, 90.0) on Top
```

Les chaînes construites par `components/USB-C-1054500101/component` sont importantes ici, car elles correspondent au référentiel de composants respectif dans le système AI-EDA. Nous approfondirons ce sujet dans les étapes ultérieures. Pour l'instant, il est essentiel de noter que le positionnement des composants est réalisé ici au moyen de coordonnées saisies par l'utilisateur. Le système est donc incapable de calculer la position de manière autonome. Une fois les composants créés, il faut bien sûr les connecter. Avec l'interface USB utilisée ici, nous utilisons des « similitudes de noms » entre les broches des deux composants dans une boucle `for` structurée comme suit :

```
for p in pins(out-usb) do :
  val pin-ref = ref(p)
```

```
val pin-name = tail(pin-ref, ref-length(pin-ref) - 1)
make-net(to-symbol(pin-name), [p])
```

Outre l'intégration de l'intelligence dans la conception du circuit, cette méthode offre l'avantage de faciliter le remplacement des objets grâce à la standardisation des empreintes de la bibliothèque. Ceci est même démontré dans le cas de la carte de test USB : La commutation entre deux jacks différents se fait en remplaçant la chaîne passée à la fonction générateur.

Concernant JITX, il faut noter qu'un plan gratuit [3] est disponible, qui, au moment de la rédaction de cet article, est limité à la création de designs open-source sous la licence CERN OHL-Permissive v2. Le fournisseur exclut explicitement l'utilisation de la GPL pour la conception résultante.

Il convient également de noter que JITX lui-même n'offre qu'un nombre comparativement faible de fonctionnalités d'IA. Le principal avantage est la possibilité d'inclure diverses contraintes dans les fichiers de conception, que le système applique ensuite automatiquement pour vérifier la tâche donnée.

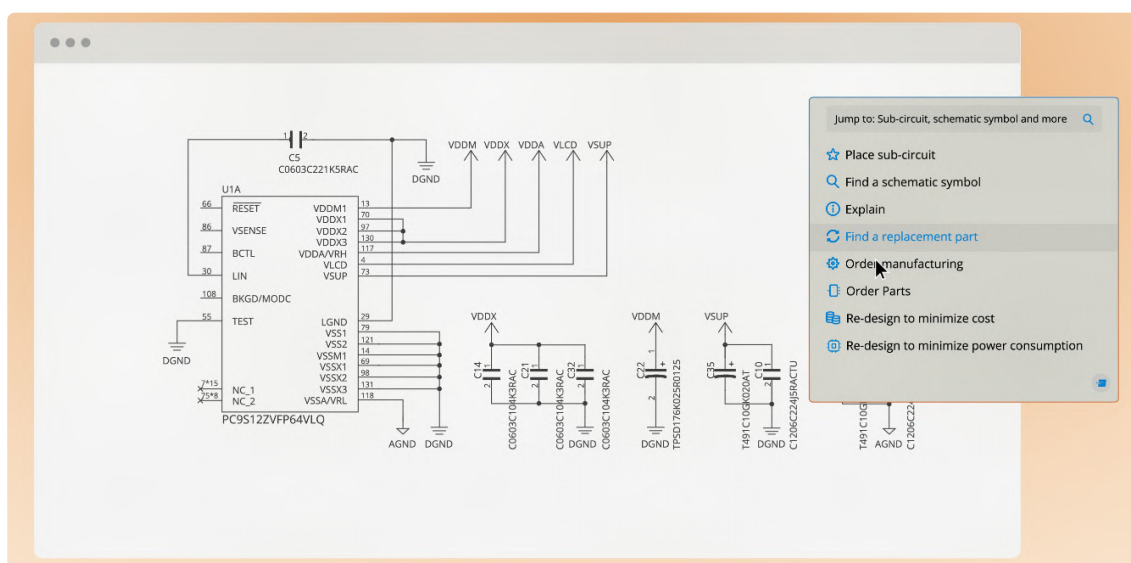
SnapMagic Copilot — A Star in the Electronics Sector

Le prochain participant à notre aperçu est SnapMagic, une évolution du célèbre service SnapEDA, qui se spécialise normalement dans la fourniture d'informations sur les composants pour les logiciels EDA tiers.

Un bon premier test consiste à rechercher un module radio Quectel relativement récent, un produit pour lequel la plupart des fournisseurs ne proposent pas encore d'empreintes. La force principale de *SnapMagic Search*, le nom sous lequel le service est désormais connu, est évidemment sa capacité à s'intégrer à une vaste gamme de logiciels EDA.

Avec un nouveau service appelé SnapMagic Copilot, actuellement en phase bêta fermée, SnapMagic vise à étendre ces fonctions. Il s'agit d'un système IA qui accepte les requêtes en langage naturel

Figure 2. SnapMagic Copilot communique en langage naturel pour exporter les résultats vers les logiciels EDA existants.



et propose ensuite des circuits de base exportables pour un large éventail de systèmes EDA.

Dans ses démonstrations, l'entreprise montre également comment « construire une structure de soutien » : par exemple, le copilote est capable d'analyser une fiche technique et d'ajouter automatiquement les condensateurs nécessaires pour assurer le fonctionnement sécurisé d'un circuit intégré. Toutefois, il est dommage que la liste d'attente [4] pour accéder à ce service soit actuellement complète - à la date de clôture de la rédaction de cet article, la demande d'admission de l'auteur n'a pas été approuvée.

Ce qui est intéressant avec SnapMagic, c'est que l'introduction de l'assistant IA est considérée comme un moyen permettant d'offrir des « services à valeur ajoutée. » La **figure 2** montre que SnapMagic aspire également à prendre en charge l'approvisionnement en composants.

CircuitMind : un savoir-faire fonctionnel solide

CircuitMind vise à libérer l'utilisateur des tâches puis - comme le montre le schéma de CircuitMind [5] (voir **figure 3**) - à les convertir en circuits ou en tâches de circuits. Le développeur est censé entrer des contraintes et des interfaces qui décrivent le plus exactement possible la conception du circuit à générer par l'intelligence artificielle.

Sur le site web de CircuitMind, les utilisateurs doivent d'abord créer un compte sur le site web en saisissant un nom d'utilisateur et un mot de passe. Ensuite, un système de prise de rendez-vous est proposé pour organiser des démonstrations en direct. Lors des tests effectués par l'auteur, les rendez-vous se sont avérés assez faciles à obtenir, avec des disponibilités même pour le jour ouvrable suivant. L'auteur a été personnellement accueilli par le fondateur de l'entreprise, Tomide Adesanmi. L'ancien employé de BAE Systems a souligné que la lecture des fiches techniques est l'une des tâches

les plus ennuyeuses auxquelles un développeur électronique doit faire face au cours de sa carrière.

Le produit d'Adesanmi se distingue comme un outil d'automatisation de la conception de systèmes évitant délibérément de se positionner dans le marché concurrentiel des EDA. Le système vise plutôt à générer des circuits et leurs variantes, optimisées selon différents paramètres, et de les comparer les unes avec les autres automatiquement.

En arrière-plan, le système repose sur une architecture composée de deux éléments. Le premier élément est une base de données de composants appelée *Commodore*, qui répertorie les modèles des composants connus du système. 70 % de la base de données est créée automatiquement, tandis que le reste est saisi et vérifié manuellement, mobilisant une part significative des ressources de l'équipe de développement. Le second élément, appelé *Ace*, est chargé de concevoir des circuits et leurs variantes à partir des objectifs de conception (*design goals*) et des données contenues dans *Commodore*.

Il est intéressant de noter que ce processus de synthèse est géré par une « IA classique, basée sur des règles ». Au cours de la discussion, Adesanmi a souligné à plusieurs reprises que leur système n'utilise pas de grands modèles linguistiques (LLM), préférant une approche moins sujette aux erreurs dites d'hallucination typiques des LLM. Quoi qu'il en soit, le véritable avantage est une vue appelée *Solution Exploration*. La **figure 4** ce que les concepteurs peuvent visualiser grâce à cet outil.

Dans ce cas, il est évident que les variantes générées diffèrent notamment en termes de taille de la carte et de coûts matériels prévus.

En cliquant sur les différents champs, les utilisateurs peuvent générer une variété de documents essentiels : Outre une nomenclature interactive, qui permet d'analyser les différents coûts,

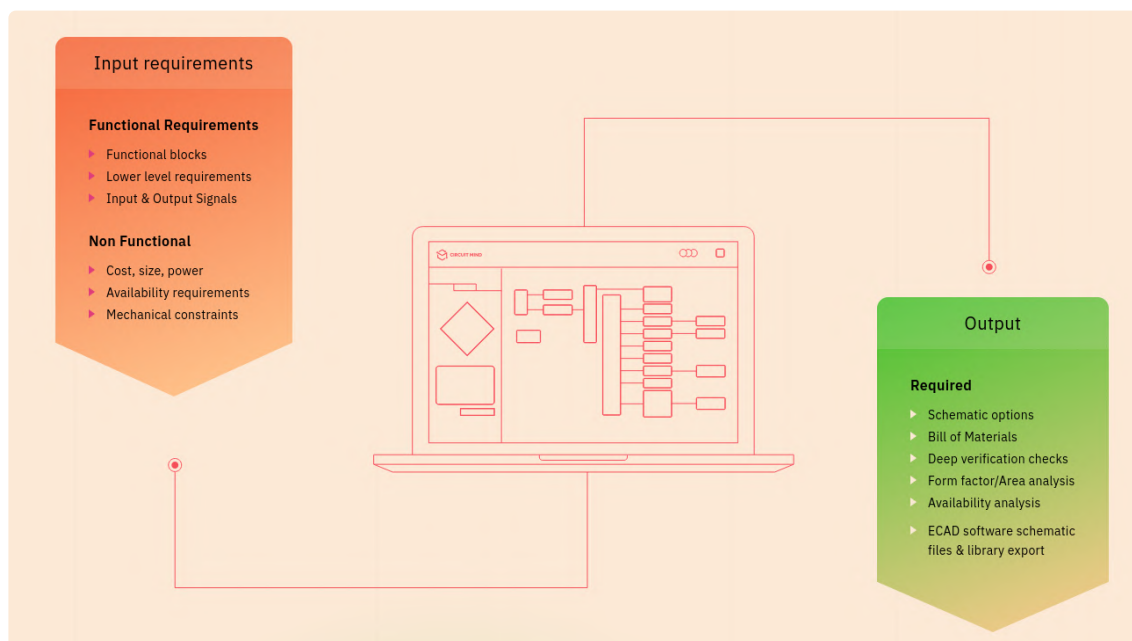


Figure 3. CircuitMind prend en charge la conception au niveau fonctionnel.

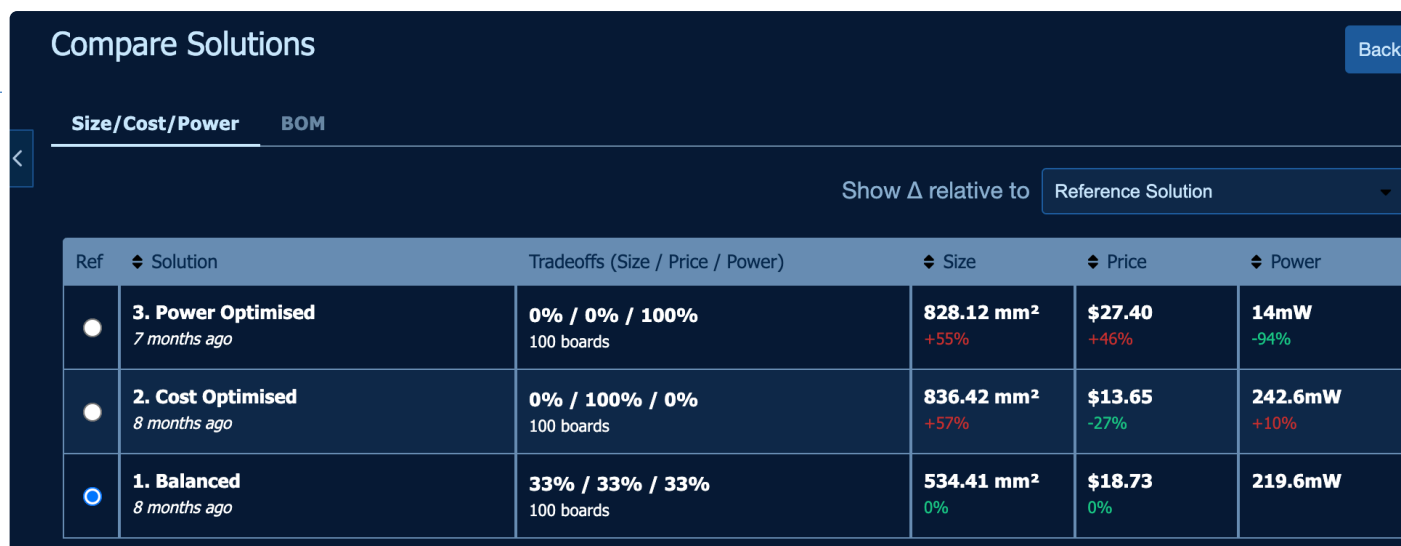


Figure 4. L'exploration des solutions vous permet de comparer les avantages et les inconvénients de différents circuits.

le système génère des documents d'interface et divers autres fichiers qui peuvent être utiles pour la certification, par exemple. Cette solution relativement onéreuse, dont le prix de départ avoisine les 10 000 US dollars, a captivé l'auteur par une fonctionnalité particulièrement innovante : la capacité du système à intégrer des « blocs analogiques » dans le processus de synthèse. Ces blocs agissent comme des espaces réservés adaptés à différentes conceptions électroniques ou circuits généraux. Le système les incorpore alors au cours du processus de conception, même lorsque les composants qu'ils contiennent demeurent méconnus.

Celus : De la « Rich Data Sheet » à la génération automatique de circuits

Les générateurs automatiques de circuits sont un exemple frappant de l'intégration de l'IA dans le domaine de l'électronique. Des systèmes tels que Red Expert de Würth, Mindi de Microchip ou le système de conception en ligne de Texas Instruments offrent depuis longtemps la possibilité de saisir des paramètres et de recevoir en retour des schémas de circuits détaillés incluant des recommandations pour les composants passifs.

Avec le format CUBOs, décrit en détail dans [6], Celus propose une fiche technique numérique qui est soutenue depuis un certain temps par le spécialiste autrichien de l'électronique de puissance RECOM. Pour démarrer avec Celus, il suffit de visiter leur site Web [7], où il est possible de créer un compte directement sur le serveur du fournisseur ou via une authentification avec un compte LinkedIn ou Google. Optant pour cette dernière option, l'auteur a suivi le processus et, après avoir renseigné son niveau de formation et ses intentions d'utilisation, a été redirigé vers l'interface connue sous le nom de *Design Studio*. Cependant, la plateforme propose également un assistant à plusieurs niveaux pour une introduction plus facile. La saisie des configurations est intéressante. Initialement, un texte structuré comme le montre la **figure 5**, apparaît où le développeur est invité à remplir divers champs avec les informations nécessaires. Dans les deux étapes suivantes, une configuration supplémentaire est nécessaire ; dans le domaine des systèmes EDA supportés, seul le triumvirat Altium, Eagle et KiCad sont actuellement pris en charge. Après cette phase, il est possible de sélectionner un modèle prédéfini qui récapitule les configurations nécessaires pour les tâches fréquentes.

Project Summary

This project aims to achieve this application goal . It will do these things . It should work this way . This comprises of these functions and requirements . It should operate under these power conditions .

It should meet these standards and regulati... . It must work within this operating environment and under these constraints and limits .

The project cost is expected to be under this budget . Parts are sourced from these manufacturer(s) .

The expected outcome of this project is these outputs and outcomes .

Figure 5. Ce texte à remplir est la première étape vers un projet opérationnel.

Cependant, lors d'un test où l'auteur a tenté de concevoir un module radio 4G basé sur un chipset Quectel, le système n'a malheureusement proposé aucune suggestion. En récompense de vos efforts, un espace de travail apparaît qui, comme le montre la **figure 6**, a été clairement inspiré par des classiques tels que Microsoft Visio.

À l'extrême gauche de l'interface, une boîte à outils permet au développeur de sélectionner les *Prefabs* qui représentent les fonctions du circuit. Ces éléments peuvent être glissés et déposés dans la zone de conception au milieu et connectés à d'autres unités fonctionnelles en ajoutant des broches. En outre, la plupart des unités fonctionnelles offrent des options supplémentaires accessibles via un clic droit sur l'élément concerné. Par exemple, pour le connecteur USB, il est possible de choisir d'inclure des protections contre l'inversion de polarité ou les décharges électrostatiques dans le design actuel.

Une fois le travail effectué, un clic sur le bouton *Resolve* affiché en haut permet idéalement d'obtenir un dessin fonctionnel. Dans le cas de la conception illustrée à la Figure 6, le système a émis un message d'erreur indiquant que des CUBO n'existaient pas : Manifestement, l'auteur avait assigné « trop » de contraintes au *connecteur USB*, ce qui explique pourquoi la bibliothèque (limitée) n'a pas pu trouver de composants appropriés.

Pour explorer d'autres possibilités, retournez à la page d'accueil, où vous trouverez des exemples de projets de Renesas et STMicroelectronics. Le « Ultra Low Power Pet Tracker » semble intéressant dans la mesure où il devrait en principe contenir également une interface radio. La récompense de cet effort est la conception illustrée à la **figure 7**, où il est intéressant de noter que certaines des

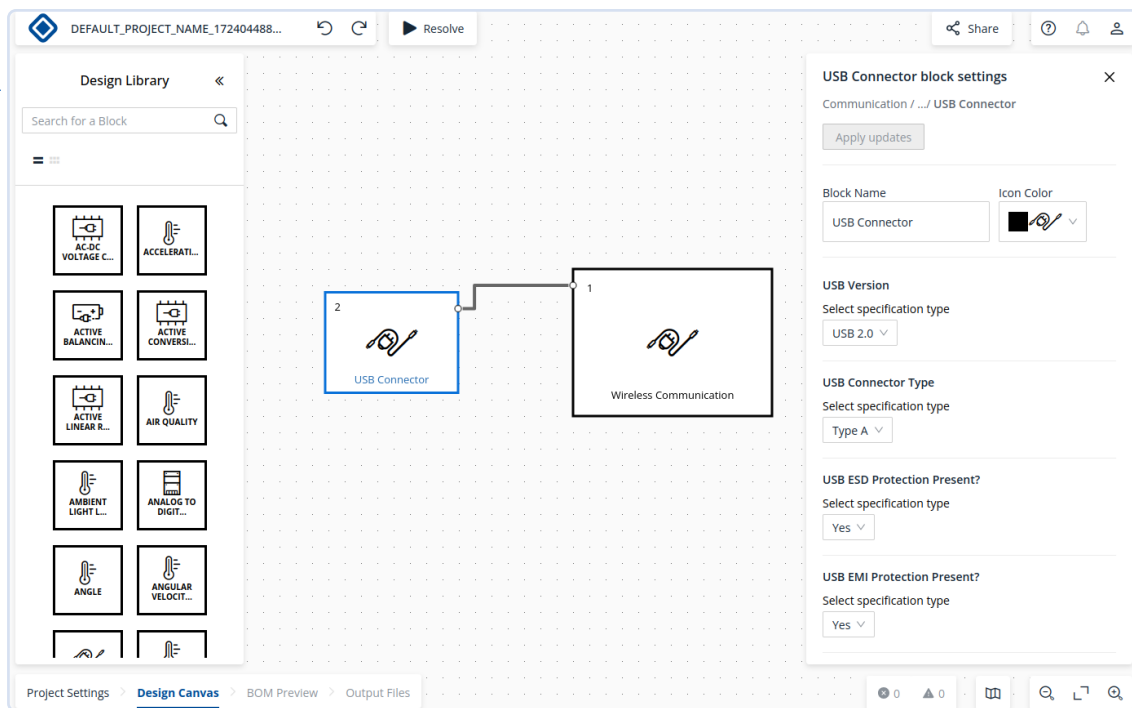


Figure 6. Toute ressemblance entre Celus et Visio est purement fortuite...

connexions ont ici des call-outs rappelant Visio, qui décrivent le « type » de connexion à établir entre les éléments.

L'étape suivante consiste à cliquer sur le bouton *Resolve*, après quoi un symbole de robot apparaît, qui, en plus de jongler avec des Rubik's cubes, indique également que la recherche de divers CUBO appropriés est en cours. Après la première itération du processus de recherche, des points bleus apparaissent dans le coin supérieur droit des différents objets de construction, indiquant que Celus souhaite proposer des « candidats à la mise en œuvre » pour ces composants. Un clic droit sur le point ouvre alors un menu contextuel dans lequel l'option correspondante peut être sélectionnée. Dans le cas

du GSM-CUBO, par exemple, un seul objet est proposé, comme le montre la **figure 8**. Un clic sur le symbole *Lock* permet alors de déterminer un objet souhaité.

Après avoir sélectionné quelques composants, un nouveau clic sur le bouton *Ressources* est nécessaire, qui fournit maintenant un certain nombre de « Livrables ». L'auteur a trouvé particulièrement intéressant le rapport de projet, qui peut être téléchargé pour ce projet à l'adresse [8]. Si vous disposez d'une licence Altium, vous pouvez également charger le projet dans Altium par la suite et l'adapter à vos besoins. Il est intéressant de noter que Celus ne s'occupe pas de la tâche particulièrement délicate qui consiste à

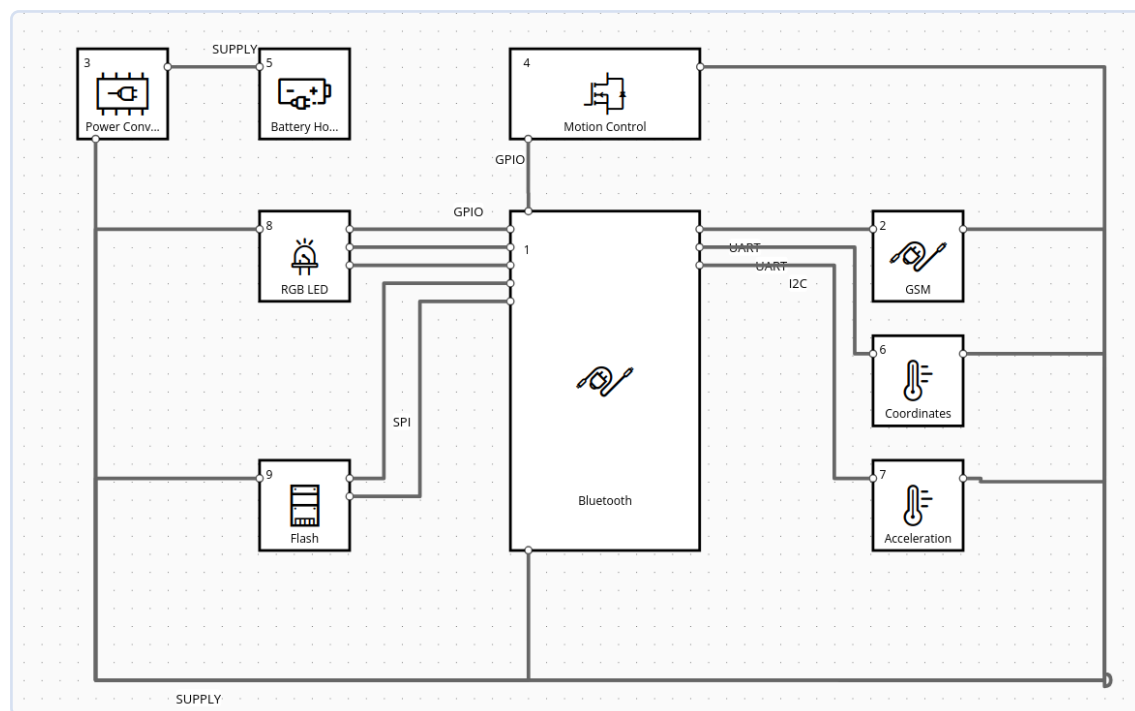


Figure 7. Le projet d'exemple « Traceur d'animaux de compagnie à très faible consommation » dans Celus.

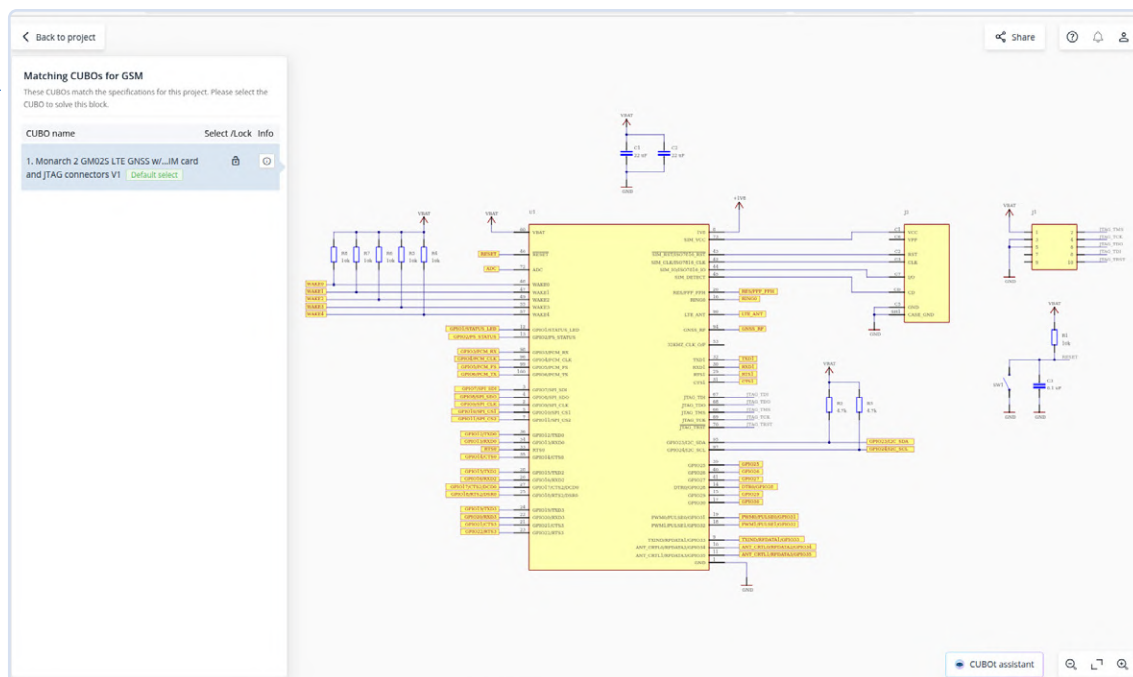


Figure 8. La sélection des modules radio GSM est modeste.

générer la configuration du circuit imprimé. Au lieu de cela, le logiciel ne génère que le schéma, la liste des pièces et le fichier de projet ; l'entreprise semble laisser la conception finale au développeur et à sa créativité.

Recherche automatisée d'erreurs dans les projets matériels

Ce que Lint est pour le programmeur C, Cady veut l'être pour le développeur électronique. Selon la documentation officielle, le système ne prend en charge que les formats Altium, Xpedition et

OrCAD - mais selon les « informations générales » à [9], il prend également en charge Eagle et KiCad. Il convient également de noter que la société propose une option d'essai gratuit « de base ». Toute personne qui s'enregistre avec l'adresse électronique de son entreprise peut accéder au système - il convient toutefois de noter que cette variante n'évalue que les diagrammes de conception fournis par l'entreprise.

Le résultat de ces efforts est la création d'une « table d'erreurs », comme le montre la **figure 9**. Il est frappant de constater que Cady reconnaît à la fois les erreurs fréquentes et celles qui le sont moins -

Index	Part Reference	Pin Number	Nets	Components	Category	Description	Status
1	C136	1	NetC135_1	04024C102KAT2A	Potential Capacitor Breach	Voltage difference of 5V is larger than the rated voltage 4V	...
2	U14	1	VDD_3V3	LTC2977HDCB#TRMPBF	Connection Instruction Violation	a Capacitor connected between this pin and D- should have a value of 470pF	...
+ 3-4	...	17	...	MIC22705YML-TR	Connection Instruction Violation	a Capacitor connected between this pin and Ground should have a value of 2.2uF	...
5	C1	1	3V3_F	06036D476MAT2A	Potential Capacitor Breach	Voltage difference of 3.3V is too close to the rated voltage 6.3V 50% derating factor is recommended	...
+ 6-7	...	1	...	CC0402KRX5R5BB225	Potential Capacitor Breach	Voltage difference of 3.3V is too close to the rated voltage 6.3V 50% derating factor is recommended	...
8	R32	2	...	RC0603FR-0712K1L	Invalid Passive Component Connection	This pin should not be left floating	...
9	U14	4	VDD_1V8	LTC2977HDCB#TRMPBF	Invalid Supply Voltage	Supply voltage 1.8V is out of operating range (2.5V to 5.5V)	...
+ 10-20	U7	...	VCORE_FPGA	XC7AT5T-1CSG324I	Invalid Supply Voltage	Supply voltage 1.7V is out of operating range (0.87V to 1.05V)	...
+ 21-22	U7	...	VCORE_FPGA	XC7AT5T-1CSG324I	Invalid Supply Voltage	Supply voltage 1.7V is out of operating range (0.92V to 1.05V)	...
23	U18	1	GND	TPS62080AD5GT	Activity Level Violation	Active HIGH ENABLE pin should not be connected to the Ground	...
24	U3	32	RBIAS	LAN8740A-EN-TR	Connection Instruction Violation	This pin should be connected to Ground via a Resistor	...
25	U19	4	NetL15_1	LT1764EQ-3.3#TRPBF	Connection Instruction Violation	This pin should be connected to Ground via a Capacitor	...
26	Y2	RH100-25.000-10-F-2030-EXT-TR	Oscillator Violation	Connected load capacitance should have a value of 10pF	...
27	U15	1	D+	FT234XD-T	Differential Lines	Potential mismatch - N type pin connected to P type net	...
28	U15	12	D-	FT234XD-T	Differential Lines	Potential mismatch - P type pin connected to N type net	...
+ 29-30	U1	STM32F427VIT7TR	Missing Pull Resistor	Pull-up resistor is missing	...
31	U7	b10	NC_U7_b10	XC7AT5T-1CSG324I	Floating Power Pin
32	U18	TPS62080AD5GT	Thermal Pad Connection Instruction Violation	This component's Exposed Pad must be connected to GND, yet it does not appear in the netlist	...
+ 33-34	Calculation Mismatch	Net name implies 1.8V while calculated voltage from U16 (MIC22705YML-TR) is 2.34V	...
35	U12	t2	DDR3_RESEtN	MT41K256M16T107IT	Best Practice	It is recommended to add a capacitor to Ground on reset pins for multiple purposes	...
+ 36-37	Best Practice	Missing Test Point	...

Figure 9. L'analyse de Cady se révèle assez pédante.

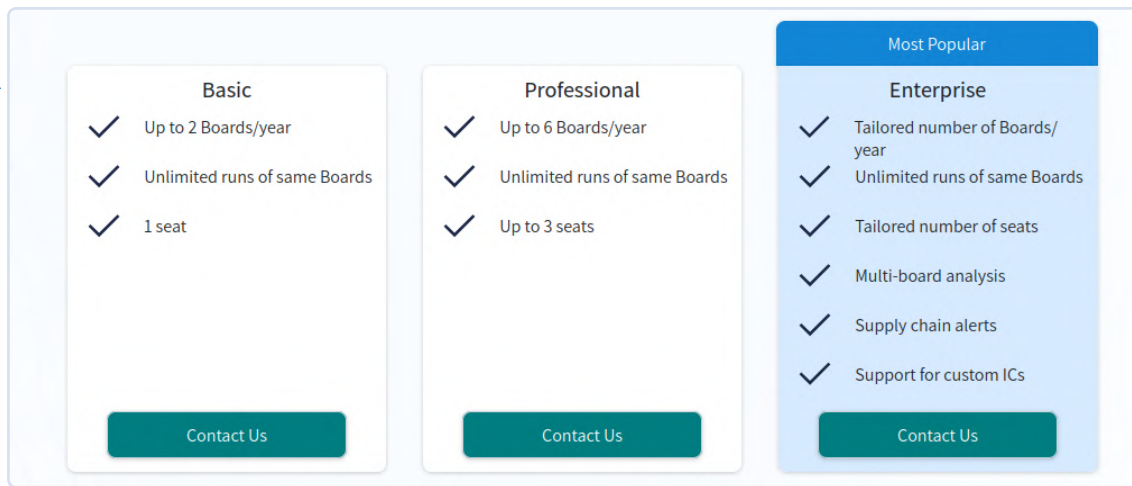


Figure 10. La start-up reste discrète sur les coûts engendrés par l'utilisation de Cady.

et, bien sûr, la startup promet également de continuer à apprendre de nouveaux tours au système. Un point de critique est que la tarification du système n'est pas du tout transparente. Au lieu de cela, comme le montre la **figure 10**, l'entreprise promet trois niveaux de prix, où il convient de noter que les « cartes » mentionnées ici couvrent chacune un design.

Un fournisseur de services complets avec de nouvelles fonctionnalités

Flux.AI [10] détient une place unique en tant que seul fournisseur de services complets d'EDA. Bien que nous ayons déjà exploré ce service en détail dans [1], de nouvelles fonctionnalités ont été introduites récemment. Cet article présente brièvement ces ajouts, témoignant de l'évolution continue de la plateforme.

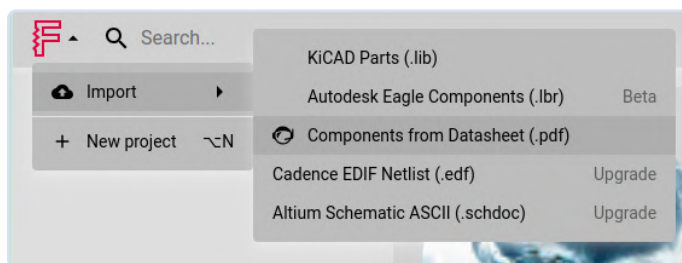


Figure 11. Cette fonction permet d'économiser quelques heures de travail.

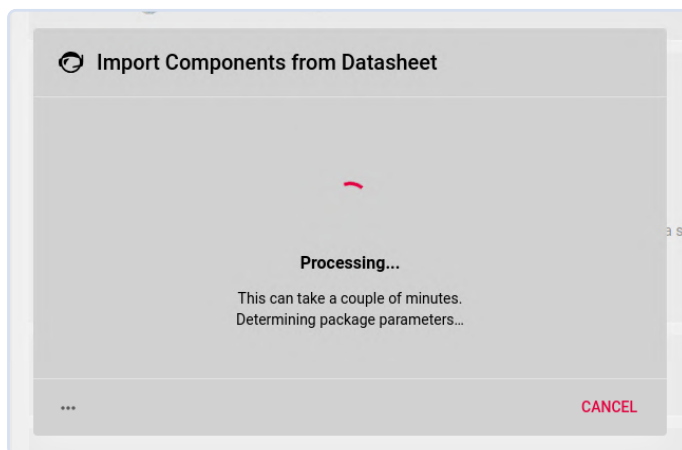


Figure 12. Après avoir téléchargé le fichier PDF, il faut faire preuve d'un peu de patience.

Une des fonctionnalités marquantes de Flux.AI est intégrée dès le menu de démarrage, comme le montre la **figure 11**. Le service peut désormais identifier automatiquement les empreintes des composants directement à partir des fiches techniques au format PDF. Pour tester cette capacité, l'auteur a utilisé un module Quectel 4G. Le téléchargement d'un fichier d'environ 2 Mo a activé la « boîte de dialogue de progression » visible dans la **figure 12**. Même avec la version de base gratuite, en environ cinq minutes, le système a réussi à reconnaître et analyser les composants listés dans la fiche technique, comme illustré dans la **figure 13**, démontrant ainsi l'efficacité de l'analyseur PDF.

Relativement coûteux !

Un compte gratuit offre 50 crédits par mois. L'utilisation du module radio Quectel en a utilisé 31, et l'importation a consommé les crédits restants.

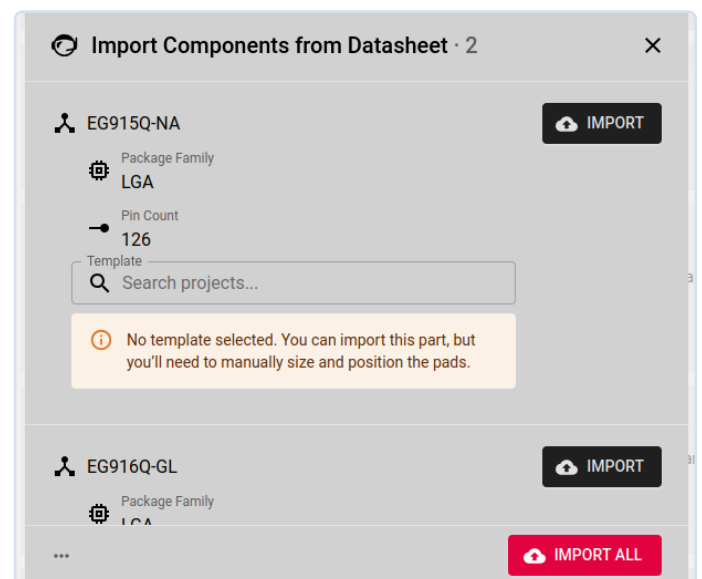
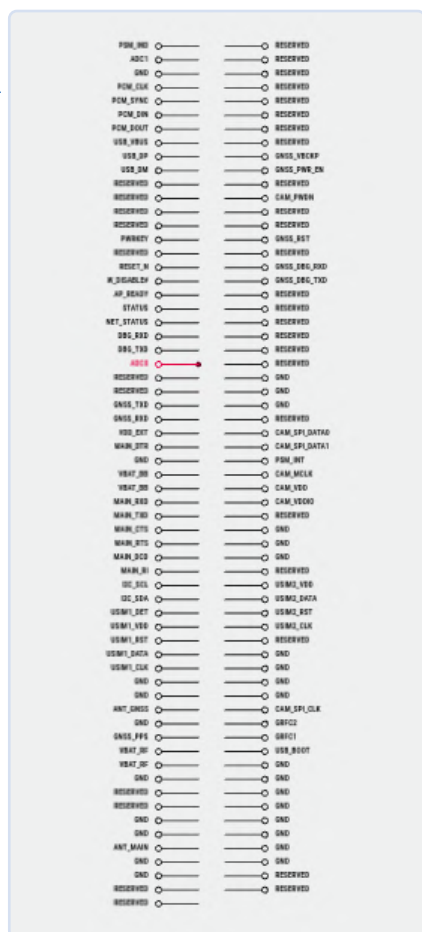
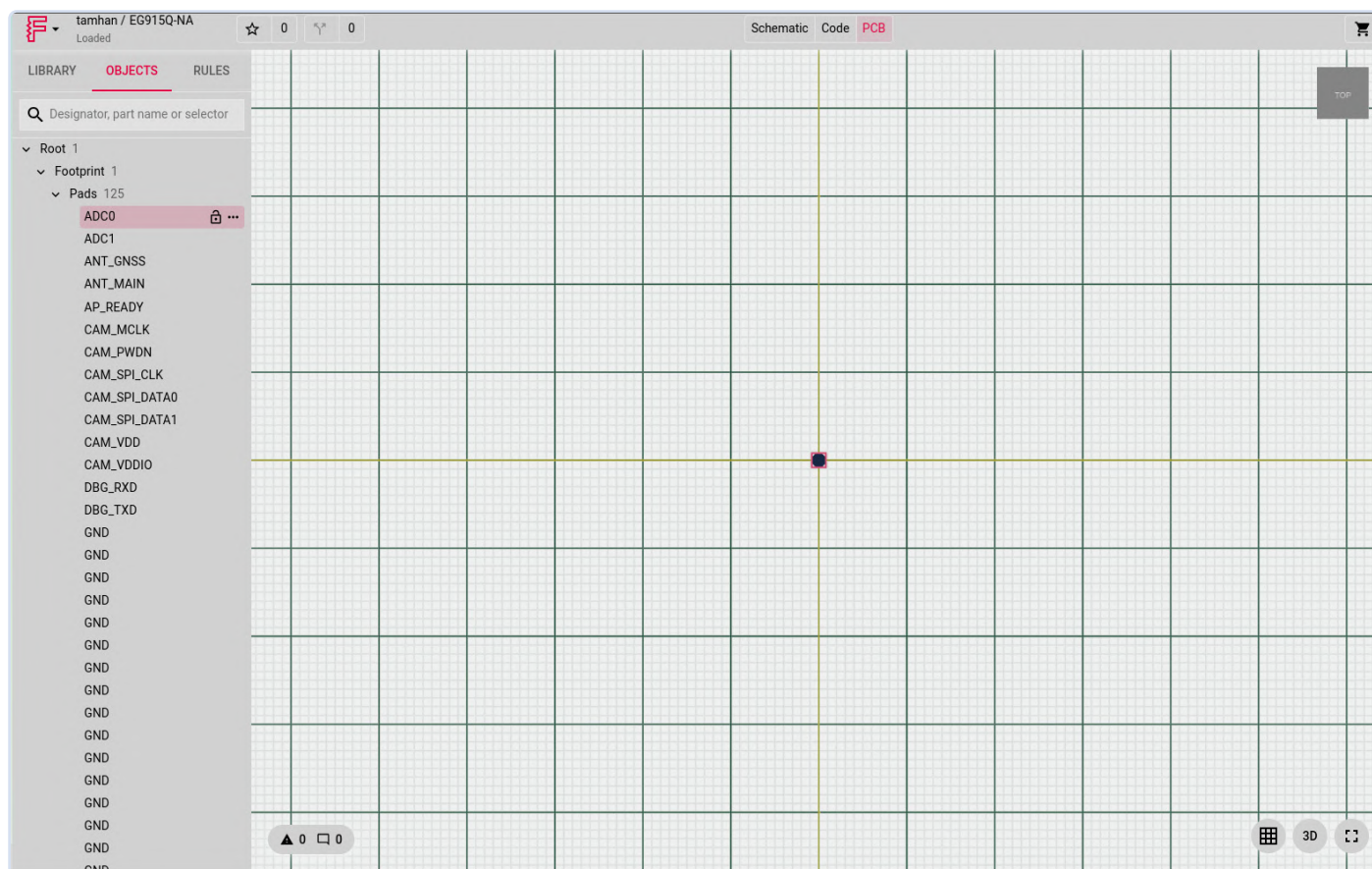


Figure 13. Ce résultat d'analyse est assez impressionnant.



Cependant, après l'importation proprement dite, la désillusion s'est installée. Comme le montre la **figure 14**, le système n'a pas été en mesure de déterminer l'empreinte « réelle ». Des améliorations sont également possibles dans le domaine des schémas de circuit - voir la **figure 15** à titre d'exemple.



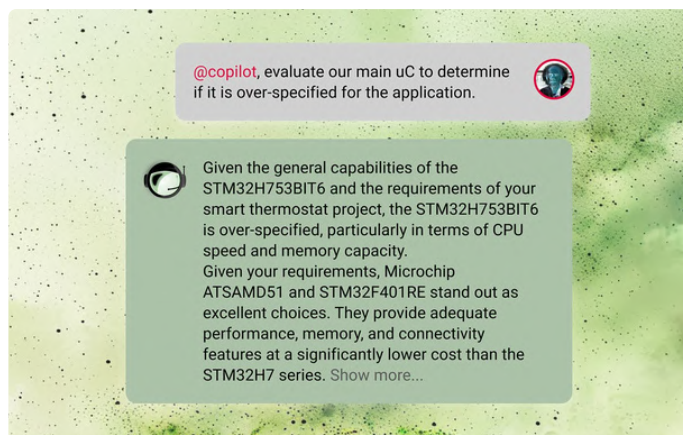


Figure 16. Malheureusement, il n'y a pas non plus de solution miracle en électronique.

pour les projets basés sur le runtime MicroPython. En plus d'exporter des documents de support, le système est également capable de vérifier l'exactitude du code du micrologiciel dans la fenêtre Copilot en fonction des possibilités via son LLM.

Outre les optimisations de performance, l'accent sera probablement mis ici sur le matériel mal configuré, un exemple classique étant la sélection d'une adresse I²C incorrecte pour un capteur thermique.

L'ordinateur comme électronicien : entre mythe et réalité

Il est difficile de concevoir un domaine où il est aussi facile de lever des fonds. Présentez un projet basé sur l'intelligence artificielle à des investisseurs en capital-risque, et vous recevrez très probablement un financement rapide. L'expansion du marché, avec non seulement Flux.AI mais aussi des dizaines d'autres entreprises qui se lancent, témoigne d'une évolution significative dans le domaine de l'intelligence artificielle appliquée à l'électronique. Cependant, il reste à questionner la portée pratique de ces avancées.

Selon l'auteur, nous sommes encore loin de l'ère du « PC électronique » capable de fournir des conceptions entièrement finies sans besoin de révisions supplémentaires. Néanmoins, l'utilisation appropriée des différents générateurs et systèmes de débogage peut déjà économiser un temps précieux si ces outils sont intégrés de manière judicieuse et mesurée dans le processus de développement. ◀

240451-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (tamhan@tamoggemon.com), ou contactez Elektor (redaction@elektor.fr).



À propos de l'auteur

Ing. Tam Hanna a plus de 20 ans d'expérience dans l'électronique, l'informatique et le développement logiciel. Il est développeur indépendant, auteur et journaliste (www.instagram.com/tam.hanna). Dans ses moments de loisir, il se passionne pour l'impression 3D et la vente de cigares.



Produits

- Peter Dalmaris, *KiCad Like A Pro (Bundle)*, (Elektor 4th edition)
Livre : www.elektor.fr/20942
E-Book (PDF) : www.elektor.fr/20943



LIENS

- [1] Tam Hanna, "l'IA dans les laboratoires d'électronique", Elektor 11/2023 : l'IA dans les laboratoires d'électronique
- [2] USB cable tester : https://github.com/JITx-Inc/jitx-cookbook/tree/main/usb_c_cable_tester
- [3] JITX versions and costs : <https://www.jitx.com/plans>
- [4] SnapMagic waiting list : <https://www.snapmagic.com/#join>
- [5] Circuit Mind : <https://www.circuitmind.io/>
- [6] What are CUBOs? : <https://www.celus.io/cubos>
- [7] Celus : <https://www.celus.io/>
- [8] Rapport du projet (PDF) : <https://www.elektormagazine.fr/240451-04>
- [9] Getting started with Cady : <https://cadysolutions.com/how-to-use-cady/>
- [10] Flux.AI : <https://www.flux.ai/>

intro aux algorithmes de l'IA

Quels algorithmes implémentent chaque outil d'IA ?

Stuart Cording (Allemagne)

Explorons les principes et les algorithmes essentiels de l'intelligence artificielle (IA) contemporaine, y compris les réseaux de neurones, l'apprentissage automatique et l'apprentissage profond.

Cette exploration dévoile les vastes capacités de l'IA, qui s'étendent bien au-delà des applications de modèles génératifs tels que ChatGPT.

L'IA est constamment sous les feux de l'actualité. Parfois, c'est pour mettre en lumière les innovations dans le secteur de la santé, où elle permet un diagnostic précoce de maladies graves telles que le cancer, augmentant ainsi les chances de guérison. Cependant, les informations relayées concernent le plus souvent ses aspects négatifs, comme la génération de fausses images (**figure 1**) et vidéos ou les questions éthiques liées à l'utilisation de matériel protégé par des droits d'auteur pour entraîner de nouveaux modèles d'IA. Par ailleurs, les médias grand public abordent rarement les détails techniques tels que les algorithmes qui soutiennent ces outils, le processus de préparation ou encore la manière dont ils sont déployés dans des applications spécialisées, des sujets cruciaux pour les ingénieurs et les technophiles.

IA 101

Lorsque l'on aborde le sujet de l'IA aujourd'hui, plusieurs termes apparaissent rapidement et méritent d'être expliqués : réseaux neuronaux, apprentissage automatique, apprentissage profond et le terme IA lui-même. L'IA est une discipline scientifique qui étudie la manière dont les ordinateurs et les logiciels peuvent être utilisés pour résoudre des problèmes complexes.

Mais, il s'agit également d'un ensemble de technologies, incluant l'analyse des données, le traitement du langage naturel ainsi que des méthodes de prévision et de prédiction.

Elle peut être subdivisée en deux catégories : l'intelligence étroite et l'intelligence générale. L'IA à intelligence étroite est conçue pour réaliser des performances équivalentes à celles de l'homme dans

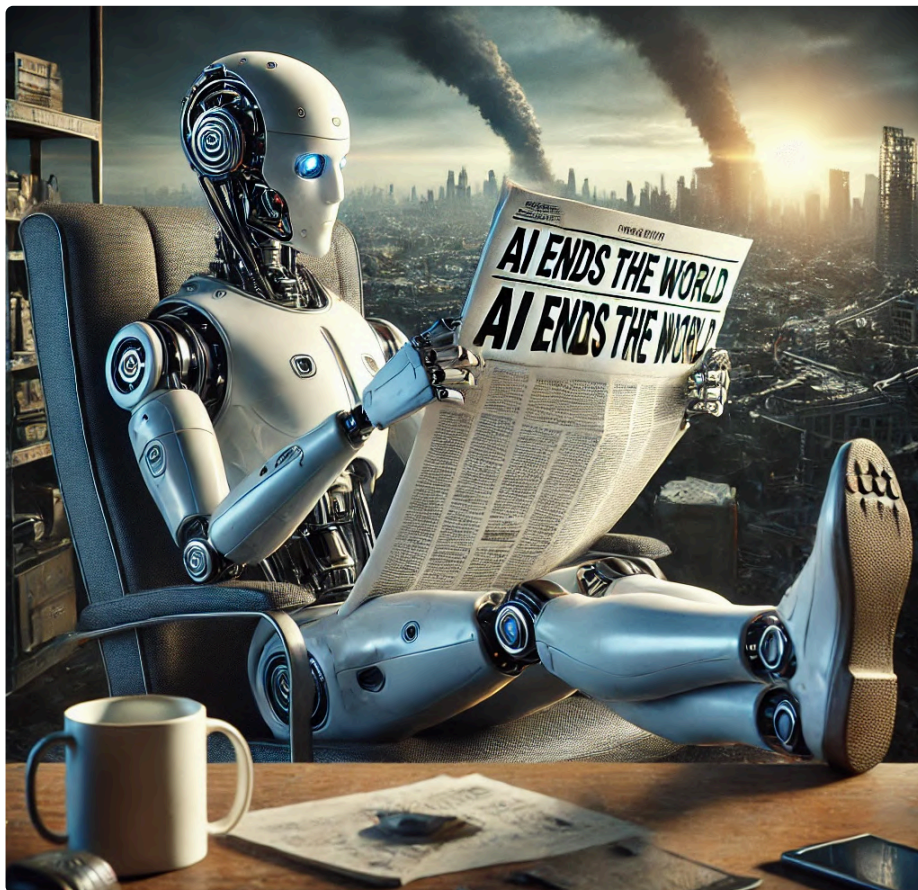


Figure 1. La fin du monde imaginée par l'IA générative. Mais ne vous inquiétez pas, les deux jambes du robot partagent un seul pied, elles n'iront donc pas loin !

un domaine spécifique, comme les jeux d'échecs ou de go. De nombreuses implémentations de l'IA peuvent être considérées comme égales ou supérieures à celles des humains, comme celles utilisées pour classer des objets dans des images. Cependant, bien que ces outils d'IA soient exceptionnellement intelligents et capables, on ne les confondrait pas avec un être humain. L'IA générale est souvent considérée comme une panacée. Les performances d'une telle IA seraient indiscernables de celles d'un être humain, imitant sa compréhension du monde qui l'entoure et ses réactions émotionnelles. Toutefois, la pertinence de développer une telle technologie est un sujet de débat sociétal. Il convient de noter qu'en théorie, il serait possible de créer une IA superintelligente surpassant les capacités cognitives humaines. Bien que cela soit concevable, pour l'instant, ce sujet est plus approprié aux discussions philosophiques et aux récits de science-fiction. Pour comprendre la mise en œuvre sous-jacente de toutes les technologies d'IA, nous devons explorer le niveau le plus bas - les réseaux neuronaux. Le perceptron multicouche (MLP), modelé sur les neurones biologiques, utilise les mathématiques et un processus appelé rétropropagation pour régler progressivement le neurone afin qu'il reconnaisse des modèles [1]. L'élément essentiel est l'utilisation d'une couche de cellules cachées entre l'entrée et la sortie, qui rend possible le processus de réglage (**figure 2**). En conséquence, le neurone peut classer des entrées qu'il serait difficile, voire impossible, de mettre

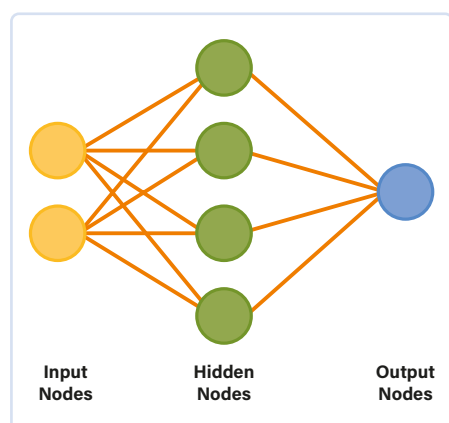


Figure 2. Le perceptron multicouche se compose de nœuds d'entrée, de nœuds de sortie qui fournissent des résultats de classification, et de nœuds cachés qui apprennent la classification souhaitée grâce à l'entraînement.

Traffic Lights in RGB Color Space

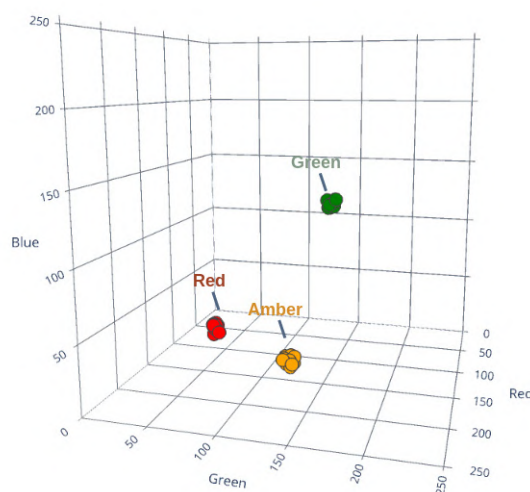


Figure 3. Les grappes de valeurs RGB peuvent être interprétées, ou classées, comme des couleurs.

en œuvre à l'aide de simples instructions if-else dans un langage de programmation. Un bon exemple est la classification des couleurs capturées par un capteur RGB. L'orange est considéré comme étant composé à 100 % de rouge et à 65 % de vert [2]. Le violet est composé de 93 % de rouge, 51 % de vert et 93 % de bleu. Cependant, lorsqu'un capteur RGB est placé devant un échantillon de couleur, il ne fournit pas toujours des résultats parfaits et précis en termes de pourcentage et de valeur correspondant exactement à ces couleurs. Au lieu de cela, les trois valeurs mesurées tendent à fluctuer autour de la valeur idéale. L'éclairage complique encore la détection, influençant les mesures qui diminuent sous un faible éclairage et augmentent quand l'éclairage est plus intense. Les réseaux neuronaux peuvent être entraînés pour s'adapter à des variations raisonnables et à des situations d'éclairage, en regroupant des valeurs (**figure 3**) approchées pour classifier les couleurs [3].

Apprentissage automatique versus apprentissage profond

La plupart des outils d'IA actuels sont en fait classés dans la catégorie de l'apprentissage automatique et constituent donc un sous-ensemble de l'IA. Ils sont généralement entraînés sur des données structurées. Par exemple, les images sont étiquetées de manière à identifier clairement leur contenu, que ce soit en nommant l'objet présent (concombre, pizza, pain) ou en fournissant un contexte supplémentaire, tel que le stade de croissance d'un légume, l'emplacement d'un plat cuisiné ou le nombre de calories qu'il contient. Il est également important de noter que

ces algorithmes, comme les humains, ne sont pas parfaits et ne fournissent pas de solutions absolues aux problèmes qu'ils abordent. Ils évaluent plutôt la probabilité que l'entrée corresponde à quelque chose qu'ils ont été entraînés à reconnaître. Par exemple d'un algorithme de ML formé aux images d'animaux pourrait être très sûr (plus de 90 %) qu'une photo représente une vache, mais pourrait également indiquer une certaine probabilité (65 %) qu'il s'agisse d'un panda et une plus faible encore (35 %) qu'il s'agisse d'un blaireau. Si vous lui présentez l'image d'un pingouin et si cet animal n'a jamais été inclus dans son apprentissage, l'algorithme ne le reconnaîtra pas comme tel.

La plupart des applications d'IA embarquées ou de pointe peuvent être considérées comme des implémentations de ML. Un bon exemple est le Izoelektro RAM-1 [4]. Ce dispositif surveille les parafoudres pour la surveillance du réseau électrique [5]. Il utilise un nRF9160 de Nordic Semiconductor basé sur Arm Cortex-M33, signalant le courant de fuite, la température excessive, l'inclinaison des poteaux/tours, et d'autres

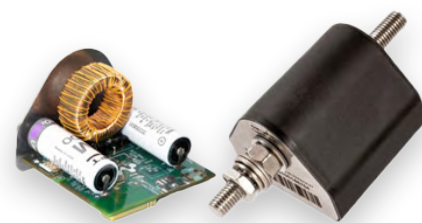


Figure 4. Le RAM-1 utilise l'apprentissage automatique pour détecter les surtensions dans les réseaux électriques. Grâce à l'IdO, le modèle de détection peut être mis à niveau. (Source : Izoelektro)

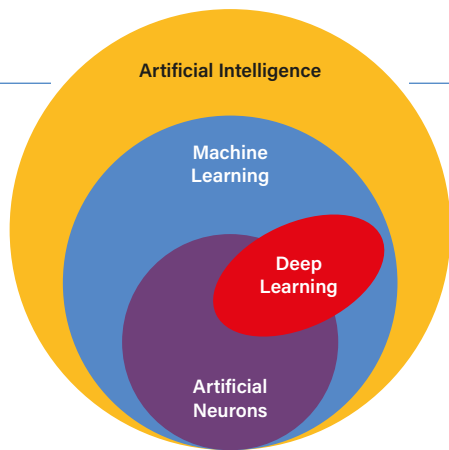


Figure 5. La hiérarchie qui va de l'IA aux neurones artificiels. L'apprentissage profond utilise des réseaux avec de nombreuses couches pour mettre en œuvre les modèles d'IA les plus avancés d'aujourd'hui, tels que le traitement du langage naturel.

événements de manipulation (par exemple, les coups de foudre). Le développement initial de l'algorithme a été réalisé avec la plateforme Edge Impulse [6], et il continue d'évoluer en intégrant des données provenant de nouveaux événements anormaux détectés par le réseau de dispositifs installés (figure 4).

L'apprentissage profond (Deep Learning, DL) va encore plus loin. Si l'on se réfère à notre MLP, une seule couche de poids entre l'entrée et la sortie suffit pour un réglage initial. Toutefois, les tâches plus complexes nécessitent un réglage plus précis, d'où la nécessité d'ajouter des couches cachées. Dès lors que l'on utilise au moins deux couches cachées, on parle d'apprentissage profond (figure 5).

Approches de l'apprentissage

Les données étiquetées sont les plus faciles à traiter, mais elles posent un problème important : qui se chargera de l'étiquetage ? Des plateformes telles qu'Amazon Mechanical Turk permettent aux utilisateurs de traiter de grands ensembles de données [7] ou même de tester les performances des modèles d'IA existants. Le coût de chaque opération d'étiquetage est minime, ce qui rend l'étiquetage économiquement accessible pour de nombreuses entreprises. Toutefois, si les données à traiter sont de nature propriétaire ou si elles comportent des informations confidentielles, il faudra envisager une alternative appropriée.

Les données étiquetées peuvent être utilisées pour effectuer un apprentissage supervisé. Le modèle d'IA reçoit des données d'entrée (texte, son, image) et des données de sortie attendues, et à force de répéti-

tion, le modèle s'améliore. L'apprentissage est considéré comme complet lorsque le modèle atteint un niveau de précision prédéfini. Comme c'est le cas pour les humains, ce niveau ne peut atteindre la perfection.

Un modèle d'IA qui obtient toutes les réponses correctes lors de l'analyse des données d'apprentissage est considéré comme souffrant d'un surajustement. Par exemple, si le modèle est censé reconnaître des feux de circulation sur des images, il peut commencer à déduire l'existence d'un feu de circulation sur la base d'anomalies dans les images d'apprentissage, telles que la lumière du soleil se reflétant sur un environnement en plastique ou la présence d'un portique. Lorsque le modèle d'IA est ensuite exposé à des images ne faisant pas partie de l'ensemble d'apprentissage, la simple présence d'un portique ou d'un point lumineux dans l'image peut faire en sorte que l'image soit étiquetée comme contenant un feu de circulation, même si aucun feu n'est visible. Par conséquent, les développeurs d'algorithmes d'IA doivent

déterminer le juste milieu entre l'ajustement excessif et l'ajustement insuffisant (figure 6).

Pour la plupart des entreprises, la réalité est qu'elles disposent de beaucoup de données, mais qu'elles ne sont pas étiquetées. Par exemple, un véhicule d'essai peut avoir collecté des masses de données brutes de capteurs lors de la conduite d'un véhicule prototype. La feuille de calcul Excel (oui, c'est encore un format de données courant !) contiendra les noms des capteurs et leurs unités de mesure, mais sans aucun contexte, comme le fait que le véhicule avait cinq passagers adultes ou que la route était mouillée.

L'apprentissage non supervisé permet de résoudre ce problème. Bien que la tâche soit, de toute évidence, plus difficile et plus exigeante en termes de traitement, il est possible de réduire la dimensionnalité des données et de trouver des modèles et des grappes. Ainsi, pour revenir à l'exemple précédent, l'IA peut classer les virages, l'accélération et la décélération et même proposer d'autres modèles, tels que

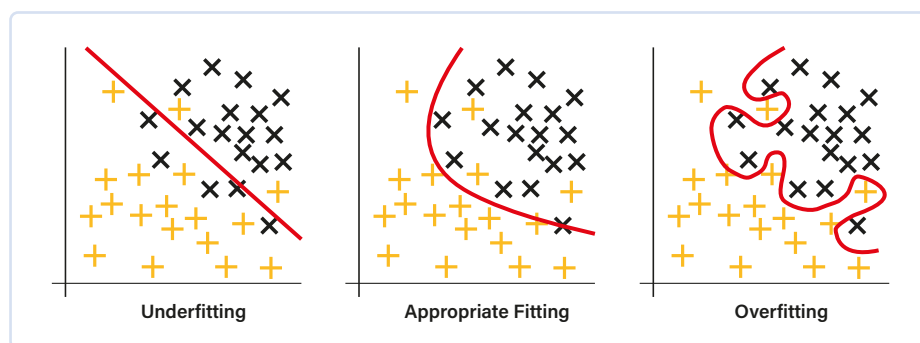


Figure 6. Entraînez trop longtemps, et le modèle d'IA peut être entravé par un surajustement ; trop peu d'entraînement conduit à une mauvaise classification des données appliquées aux entrées.

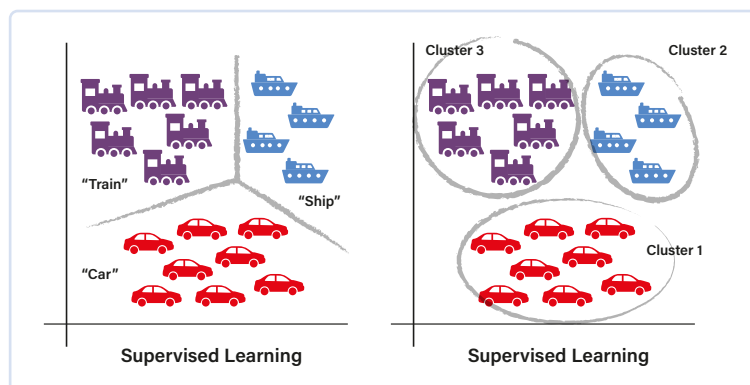


Figure 7.

la présence de pluie ou de neige pendant l'essai routier. Bien entendu, l'IA ne peut pas suggérer les termes « pluie » et « neige ». Au lieu de cela, les experts qui analysent les résultats verront que des groupes spécifiques s'alignent sur des circonstances de test spécifiques et les étiqueteront en conséquence (**figure 7**).

Les deux méthodes ci-dessus peuvent être combinées, en utilisant des données étiquetées et des données non étiquetées dans une approche d'apprentissage semi-supervisé. Les résultats de la formation peuvent également être réintégrés dans les données de formation. Toutefois, une surveillance et une intervention humaines sont nécessaires, car ces résultats peuvent contenir des erreurs.

L'apprentissage par renforcement est une autre approche. Utilisé dans des environnements simulés, un modèle reçoit des observations concernant son environnement. L'IA choisit des actions possibles et reçoit une récompense si le modèle se rapproche de l'objectif souhaité. Après de nombreuses itérations, l'IA finit par apprendre la tâche souhaitée. L'apprentissage par renforcement a été utilisé dans le traitement du langage naturel et dans les IA conçues pour jouer à des jeux. Google DeepMind a utilisé cette technique pour réduire le nombre d'instructions d'assembleur nécessaires à la mise en œuvre d'un algorithme de tri (**figure 8**). Une récompense a été attribuée au code fonctionnant correctement qui utilisait également moins d'instructions d'assembleur que l'algorithme original [8].

Les algorithmes d'IA

Les modèles de base de l'IA se divisent en deux grands types. Les premiers sont des modèles de classification. Ceux-ci décomposent les données d'entrée en une réponse probable/improbable, répondant à une question telle que « Cette image contient-elle un canard ? » ou « Cette couleur est-elle jaune ? » Ensuite, il y a les modèles de régression, qui génèrent une réponse numérique à des questions telles que « Combien d'Arduino Uno allons-nous vendre le mois prochain ? » Des modèles plus complexes permettent la création d'outils tels que ChatGPT, capable de dialoguer avec l'utilisateur, ou DALL-E, qui peut générer des images à partir de descriptions textuelles. Ces modèles reposent sur une variété d'algorithmes qui, parfois

Original Sort Code	Improved Sort Code	
Memory[0] = A Memory[1] = B Memory[2] = C	Memory[0] = A Memory[1] = B Memory[2] = C	Collect three values for sorting. Save to registers.
mov Memory[0] P mov Memory[1] Q mov Memory[2] R	mov Memory[0] P mov Memory[1] Q mov Memory[2] R	
mov R S cmp P R cmovg P R cmovl P S mov S P cmp S Q cmovg Q P cmovg S Q	mov R S cmp P R cmovg P R cmovl P S mov S P cmp S Q cmovg Q P cmovg S Q	
mov P Memory[0] mov Q Memory[1] mov R Memory[2]	mov P Memory[0] mov Q Memory[1] mov R Memory[2]	Return the result to memory.

Figure 8. L'IA a économisé une seule instruction d'assembleur (mov S P) dans cet algorithme de tri sans impact sur le fonctionnement.

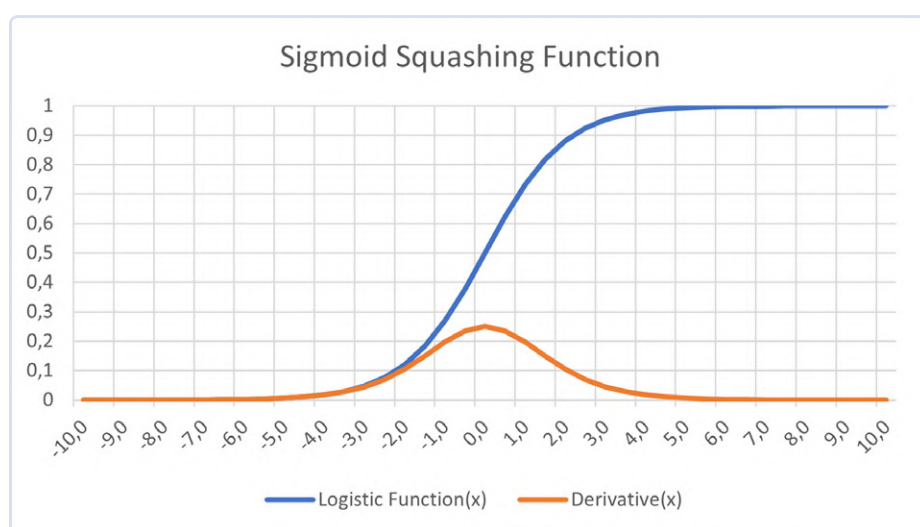


Figure 9. La fonction sigmoïde aide à la classification en poussant rapidement les valeurs d'entrée vers un résultat « probable » ou « improbable ».

combinés, répondent spécifiquement aux exigences de la tâche.

Régression logistique

Les premières expériences avec les neurones artificiels ont rencontré un obstacle notable. Ils étaient performants pour classer les informations pouvant être séparées en deux groupes distincts à l'aide d'une ligne droite, grâce à l'utilisation de fonctions linéaires, exprimées mathématiquement par $f(x) = ax + b$ si Cependant, ce type de fonctionnalité présente des limites lorsque les données à classer ne sont pas linéairement séparables, par exemple quand elles se regroupent de manière circulaire, situation où ces modèles linéaires échouent.

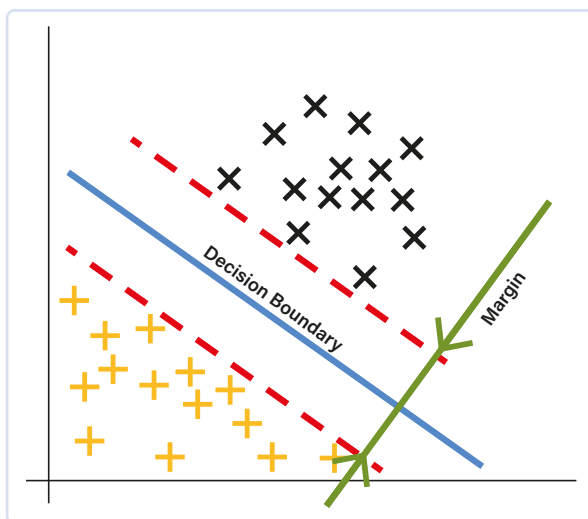
La régression logistique utilise des fonctions qui rapprochent rapidement les valeurs d'entrée de 0 ou de 1. Un exemple est la fonction sigmoïde, $s(z) = 1 \div (1 + \exp$

$(-z))$, qui renvoie 0,5 pour une entrée de 0,0, proche de 1 pour une entrée de 10, et proche de 0 pour une entrée de -10,0 (**figure 9**). Grâce à ces propriétés mathématiques, les réseaux de neurones peuvent, par de nombreuses répétitions, être entraînés à des tâches de classification relativement simples, comme celle de savoir si un capteur RGB se trouve devant la couleur violette (probable/improbable).

Machines à vecteurs de support (SVM)

Si vous jetez sur votre tapis des M&M rouges suivis de M&M bleus, vous constateriez probablement que vous pouvez placer une règle entre les deux groupes de couleurs. La règle sépare les deux groupes et son angle sur le tapis peut être représenté par un vecteur (une ligne avec une magnitude et une direction). C'est ainsi que fonctionnent, à la base, les machines

Figure 10. Avec les SVM, les données sont classées de part et d'autre d'une limite de décision, définie comme un vecteur, avec une marge (distance entre les groupes opposés les plus proches).



à vecteurs de support. Toutefois, dans un espace à deux dimensions, cette règle imaginaire est aussi large que possible pour maximiser la marge (distance) entre les deux groupes lorsque le vecteur est calculé (figure 10). Une fois déployé dans un modèle d'IA, nous pouvons désormais classer les M&Ms rouges et bleus.

À ce niveau superficiel, ce n'est pas différent d'une fonction linéaire. Les SVM deviennent intéressants lorsque les données comportent davantage de dimensions. Par exemple, supposons que vos données puissent être représentées sur les axes X, Y et Z. Dans ce cas, vous n'êtes pas obligé d'utiliser une fonction linéaire. Dans ce cas, vous ne cherchez plus une ligne pour séparer les classes, mais un hyperplan (un plan bidimensionnel dans un espace tridimensionnel). Ils conviennent également aux cas où certains de vos M&M rouges et bleus ont dérivé vers le groupe opposé. Les SVM nécessitent plus de calculs que la régression logistique, mais sont plus performants sur les données comportant de nombreuses dimensions et sur les ensembles de données non structurés.

Arbres de décision et forêts aléatoires

Les arbres de décision sont, à un niveau simple, similaires à la création d'organigrammes. Ils utilisent une stratégie de division et de conquête pour classer un ensemble de données en autant de groupes que nécessaire. Lorsqu'un utilisateur applique l'algorithme d'IA à sa tâche, il parcourt l'organigramme pour fournir une réponse. Appliqué à des images, il pourrait commencer par diviser l'ensemble de données en arrière-plans, puis en couleurs dominantes, et enfin en caractéristiques d'objet (oreilles, pattes, rayures).

La décision de diviser l'ensemble de

données à chaque nœud dépend de l'entropie des échantillons examinés. Par exemple, si de nombreuses images contiennent du rouge et que quelques-unes sont bleues, l'entropie de « contient du rouge » est élevée. L'ajout d'un nœud « L'image contient-elle du rouge ? » permet au modèle d'IA de traiter rapidement et efficacement les images rouges. Si l'ensemble de données est partagé à 50/50 sur la question rouge/bleu, il recherchera une classification alternative avec une entropie plus élevée.

Les arbres de décision posent un problème d'ajustement excessif et de biais. L'ajustement excessif est combattu en élaguant les nœuds sur les caractéristiques de faible importance au cours de l'apprentissage. Le surajustement et le biais peuvent être traités à l'aide de l'algorithme de la forêt aléatoire qui utilise en outre le bagging, pour réduire la variance de l'ensemble des données, et le caractère aléatoire.

Réseaux neuronaux convolutifs (CNN)

Pour les applications de vision artificielle et de reconnaissance d'objets, les CNN à apprentissage profond sont utilisés. Ces algorithmes multicouches commencent par utiliser la convolution pour trouver des caractéristiques (lignes droites, courbes, couleurs) dans l'image. Viennent ensuite des couches de mise en commun qui réduisent essentiellement l'échantillonnage des résultats de la convolution. Le réseau profond comprend généralement de nombreuses couches de convolution et de mise en commun, autant que nécessaire pour obtenir la classification souhaitée. La dernière couche est appelée couche entièrement connectée. C'est à ce stade que l'algorithme renvoie une classification.

L'un des avantages des CNN est qu'ils fonctionnent bien même si la position de

l'objet recherché se déplace par rapport aux exemples de l'objet dans les données d'apprentissage.

Réseaux neuronaux récurrents (RNN) et transformateurs

Les images sont des moments statiques dans le temps, mais certaines applications nécessitent de comprendre ce qui s'est passé avant, comme dans le cas de la parole et de l'écriture. Par exemple, si j'écris « John a construit un amplificateur. Il l'a fait fonctionner du premier coup », nous savons que « il » fait référence à John et que « il » fait référence à l'amplificateur. Les RNN sont beaucoup plus performants pour traiter ce type de tâches temporelles, notamment le traitement du langage naturel, la traduction et la reconnaissance de l'écriture manuscrite. La plupart des réseaux neuronaux transmettent simplement toutes leurs données à l'étape suivante au fur et à mesure qu'ils traitent les entrées. Les RNN s'écartent de cette approche en mettant en œuvre une « mémoire » dans certaines couches cachées, en conservant des informations pour des décisions futures, et même en anticipant les données futures, telles que les mots au-delà de ceux qui sont actuellement traités, pour prendre leurs décisions. Les transformateurs améliorent les RNN et ont rendu possibles de grands modèles linguistiques tels que ChatGPT. Ils utilisent des mécanismes d'attention plus performants lors de l'analyse du langage, en particulier lorsque des relations doivent être trouvées entre les mots de différentes phrases plutôt qu'entre les mots d'une même phrase.

Il n'y a pas que ChatGPT et Midjourney

L'IA a permis des avancées significatives, modifiant la manière dont nous abordons certaines tâches et rendant possibles des activités auparavant difficiles. En médecine, l'intelligence artificielle améliore significativement l'analyse des scanners médicaux et accélère le développement de nouveaux médicaments. Pour les entreprises, elle est capable d'analyser les journaux d'accès aux réseaux pour détecter les cyberattaques, anticiper les menaces, et même examiner le code source. Dans le domaine des systèmes embarqués, l'apprentissage automatique peut se substituer à des filtres analogiques complexes et à l'utilisation de



filtres numériques énergivores. Cependant, l'attention médiatique se concentre souvent sur les aspects les moins flatteurs de l'IA, comme les fausses vidéos génératives ou les erreurs de ChatGPT lors d'opérations simples. Bien que certaines inquiétudes et risques éthiques méritent une attention sérieuse, il ne faudrait pas pour autant renoncer aux nombreux bénéfices apportés par l'intelligence artificielle.

Comme toute technologie, l'IA peut être utilisée à des fins bénéfiques, malveillantes ou néfastes. Il serait judicieux de consacrer davantage de temps à comprendre le fonctionnement interne de cette techno-

logie, comme nous l'avons exploré ici, et de l'appliquer à des défis que les humains trouvent fastidieux, complexes, ou à des tâches sujettes aux erreurs humaines. Internet offre une abondance de tutoriels, guides et exemples sur la manière de construire vos propres modèles d'IA ou de les entraîner à traiter des ensembles de données qui résistaient à d'autres méthodes. Le moment est peut-être venu d'expérimenter avec l'IA et de relever les défis que vous envisagez. ◀

240564-04

À propos de l'auteur

Stuart Cording est à la fois ingénieur et journaliste, avec plus de 25 ans d'expérience dans l'industrie électronique. Il est spécialisé dans la création de contenu vidéo, mettant l'accent sur les analyses techniques approfondies et la compréhension des technologies. Il porte un intérêt particulier à la technologie elle-même, à son intégration dans les applications finales et aux perspectives d'évolution future. Vous pouvez découvrir de nombreux articles récents de Stuart Cording sur le site d'Elektor à l'adresse www.elektormagazine.com/cording.

Questions ou commentaires ?

Contactez Elektor (redaction@elektor.fr).

LIENS

- [1] S. Cording, "Understanding the Neurons in Neural Networks (Part 1): Artificial Neurons," [elektormagazine.com](https://www.elektormagazine.com/articles/neural-networks-part-1-artificial-neurons), May 2021: <https://www.elektormagazine.com/articles/neural-networks-part-1-artificial-neurons>
- [2] RGB in % for orange: <https://www.farb-tabelle.de/en/rgb2hex.htm?q=orange>
- [3] S. Cording, "Understanding the Neurons in Neural Networks (Part 3): Practical Neurons," [elektormagazine.com](https://www.elektormagazine.com/articles/neural-networks-part-3-practical-neurons), June 2021: <https://www.elektormagazine.com/articles/neural-networks-part-3-practical-neurons>
- [4] Izoelektro RAM-1: <https://www.ram-center.com/>
- [5] Edge Impulse, "The New AI- Powered Grid": https://cdn.prod.website-files.com/618cdeef45d18e4ef2fd85f3/621cef966699cbc24cdae67e_Smart-Grid-Monitoring.pdf
- [6] Edge Impulse: <https://edgeimpulse.com/>
- [7] B. Kopp, "Getting Started with Using Amazon Mechanical Turk to Label Data," U.S. Bureau of Labor Statistics, September 2019: <https://www.fcs.gov/assets/files/docs/gasp/Kopp.pdf>
- [8] D. J. Mankowitz, A. Michi, "AlphaDev discovers faster sorting algorithms," Google DeepMind, June 2023: <https://deepmind.google/discover/blog/alphadev-discovers-faster-sorting-algorithms>
- [9] Incorrect count of "r" characters in the word "strawberry": <https://community.openai.com/t/incorrect-count-of-r-characters-in-the-word-strawberry/829618>

Rejoignez notre communauté



www.elektormagazine.fr/community

elektor
design > share > earn

ordinateurs monocartes pour les projets d'IA

Aperçu et contexte

Saad Imtiaz (Elektor)

Il existe une multitude d'ordinateurs monocartes (SBC), mais choisir celui qui convient spécifiquement aux applications d'IA peut se révéler complexe. Dans cet article, nous examinons les facteurs clés à considérer lors du choix d'un SBC adapté à l'IA. Nous y présentons également une sélection des plateformes les plus efficaces actuellement disponibles. Ce guide vous aidera à trouver le matériel idéal pour vos projets d'IA.

L'intelligence artificielle (IA) est évoluée d'un concept futuriste à un élément essentiel de notre quotidien. De la reconnaissance vocale sur nos smartphones, aux algorithmes prédictifs dans les services de streaming, en passant par l'amélioration des diagnostics médicaux et la maintenance prédictive et du contrôle de la qualité dans l'industrie manufacturière, l'IA est partout. Au cœur de ces avancées se trouve la nécessité de disposer d'un matériel puissant et efficace, capable de répondre aux exigences de calcul élevées des algorithmes d'apprentissage automatique. Le choix du bon matériel est crucial pour la réussite de tout projet d'IA, car il influence directement l'efficacité, l'évolutivité et le succès de l'application.

Quelles sont les exigences du développement IA ?

Le développement d'applications IA requiert un matériel capable d'effectuer beaucoup de calculs simultanément. En effet, les algorithmes d'IA, en particulier les modèles d'apprentissage profond, impliquent de grands ensembles de données et des calculs complexes qui doivent être traités rapidement

et avec précision. Les principales exigences matérielles pour le développement de l'IA comprennent un processeur haute performance, un puissant GPU pour le traitement parallèle, une capacité importante de mémoire (RAM) et, de manière croissante, une unité de traitement neuronal (NPU) dédiée.

Un NPU est conçu pour accélérer le calcul des réseaux neuronaux, offrant une augmentation significative des performances par rapport aux CPU et GPU ordinaires. Avec l'évolution constante du domaine de l'IA, les NPU deviennent une caractéristique standard des plateformes matérielles modernes d'IA, offrant des capacités de traitement spécialisées qui sont cruciales pour répondre aux exigences des tâches d'IA.

Choisir le bon matériel pour vos projets d'IA/ML

Choisir le matériel approprié pour le développement de l'IA dépend de plusieurs facteurs, notamment la complexité des modèles d'IA, la taille des ensembles de données, la nécessité de traitement en temps réel et les contraintes de puissance spécifiques au projet. Par exemple, un projet axé sur l'IA en périphérie - où le traitement de l'IA est effectué localement sur un appareil - peut privilégier l'efficacité énergétique et un format compact plutôt qu'une puissance de calcul élevée. En revanche, les applications d'IA basées sur le cloud peuvent tirer parti de GPU et de NPU puissants, adaptés aux traitements massifs dans les centres de données.

Il est également essentiel de prendre en compte la compatibilité avec les frameworks d'IA tels que TensorFlow, PyTorch et TensorRT, ainsi que la disponibilité des outils de développement et le soutien de la communauté. L'objectif est de sélectionner une plateforme qui répond non seulement aux exigences techniques du projet, mais qui facilite également le développement et le déploiement.

Évolutions récentes dans le matériel dédié à l'IA

L'essor récent du matériel dédié à l'IA a marqué une évolution notable, avec l'intégration continue de NPU avec des CPU et des GPU. Cette intégration vise à optimiser le traitement des tâches liées à l'IA, car les NPU sont conçues pour exécuter des opérations essentielles pour les réseaux neuronaux, telles que les multiplications de matrices et les convolutions. (Voir l'encadré « **Comprendre les NPU** »).

Choisir une plateforme matérielle pour débuter en développement IA

À mesure que le domaine de l'IA évolue, le spectre de plateformes matérielles disponibles pour les développeurs s'élargit considérablement. Face à une telle diversité, il est essentiel de comprendre que si certaines plateformes excellent dans certains domaines, elles peuvent être moins performantes dans d'autres. Le choix du bon matériel dépend des exigences spécifiques de votre projet, que vous soyez novice en IA ou un expert en quête de solutions pour des applications complexes.

Bien que cet article ne puisse couvrir tous les SBC disponibles en raison de leur nombre, des ressources en ligne [1][2] existent pour vous permettre de filtrer et de sélectionner des SBC selon vos exigences spécifiques. Vous trouverez ci-dessous une liste de plateformes matérielles adaptées à divers projets d'IA et de ML, allant de la vision par ordinateur et du traitement du langage naturel à la robotique et à l'informatique en périphérie. Ces cartes ont été choisies pour leur capacité à gérer diverses applications, offrant une base solide pour le développement de l'IA à tous les niveaux d'expertise.

Le **tableau 1** présente une comparaison de ces cartes SBC, incluant d'autres cartes remarquables non détaillées ici, ce qui vous permet de comparer leurs performances, fonctionnalités et adéquation aux différentes applications d'IA.

Nvidia Jetson Orin Nano

Lorsqu'il s'agit de plateformes d'IA, Nvidia est souvent le premier nom qui nous vient à l'esprit. Depuis des années, Nvidia est à l'avant-garde du développement matériel pour l'IA, repoussant continuellement les limites avec ses puissants GPU et ses NPU spécialisés. Tout a commencé avec des GPU conçus initialement pour les jeux, qui sont rapidement devenus la clé de voûte de la recherche en IA grâce à leurs capacités de traitement parallèle exceptionnelles. Introduite en 2016, la plateforme Jetson de Nvidia a révolutionné l'IA et la robotique, en apportant un traitement puissant de l'IA à la périphérie - où elle est aujourd'hui utilisée dans les véhicules autonomes, les robots industriels et d'innombrables autres applications.

Comprendre les NPU : Le cerveau derrière le matériel d'IA

Une unité de traitement neuronal (NPU) est un processeur spécialisé conçu pour traiter les types de calculs spécifiques requis par les modèles d'IA, ce qui les rend beaucoup plus efficaces que les CPU et GPU traditionnels pour ces tâches. Les modèles d'IA, en particulier les réseaux neuronaux, s'appuient fortement sur des opérations telles que les multiplications de matrices, les convolutions et les fonctions d'activation, qui impliquent le traitement de grandes quantités de données en parallèle.

Par exemple, la multiplication matricielle est au cœur des réseaux neuronaux, où les données d'entrée sont multipliées par des matrices de poids sur plusieurs couches. Les NPU excellent dans ce domaine en exécutant ces opérations simultanément sur de nombreuses unités de traitement, ce qui accélère considérablement les calculs. De même, dans les réseaux neuronaux convolutifs (CNN), les convolutions et les opérations de mise en commun permettent d'extraire des caractéristiques des images, tâches que les NPU traitent avec une grande efficacité grâce à leurs capacités de traitement parallèle.

En outre, les NPU sont optimisées pour la nature répétitive et hautement parallèle de ces opérations, ce qui leur permet de traiter les données plus rapidement et avec une consommation d'énergie plus faible que les CPU à usage plus général ou même les GPU. Leurs performances sont souvent mesurées en TOPS (Tera Operations Per Second), ce qui indique leur capacité à effectuer des trillions d'opérations par seconde, un facteur essentiel pour la vitesse et l'efficacité des applications d'intelligence artificielle. En se concentrant sur ces types de calculs spécifiques, les NPU permettent un traitement plus rapide et plus économe en énergie, ce qui les rend essentielles pour le matériel d'IA moderne.

Cependant, si ces plateformes de niveau industriel sont impressionnantes, tous les projets ne nécessitent pas un supercalculateur. Parfois, un démarrage modeste suffit - comme entraîner un modèle d'IA à distinguer votre chat de celui de votre voisin. Examinons donc le Nvidia Jetson Orin Nano, une option puissante mais accessible qui pourrait marquer le départ de votre parcours de développement de l'IA vers des projets d'IA avancés.

Le Nvidia Jetson Orin Nano [3] est une plateforme d'IA compacte dotée d'une puissance considérable, ce qui la rend idéale pour une large gamme de projets d'IA et de ML (voir **figure 1**). Équipé d'un processeur ARM Cortex-A78AE à six cœurs et jusqu'à 8 Go de RAM LPDDR5, il offre jusqu'à 40 TOPS (*Tera Operations Per Second*) de performances pour l'IA, idéal pour des tâches telles que l'apprentissage profond, la vision par ordinateur et la robotique.

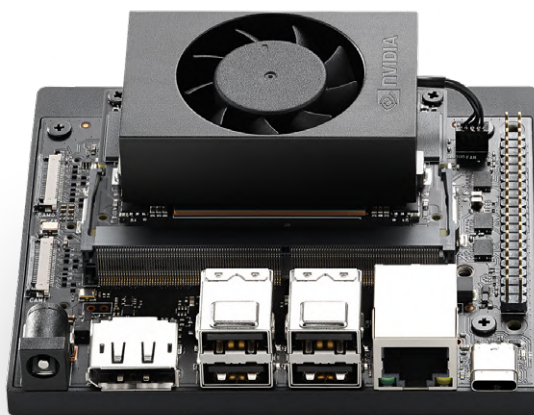


Figure 1.
Nvidia Jetson Orin Nano
(Source : Nvidia)

Tableau 1. Comparaison des SBC.

	Jetson Orion Nano 4 GB	Raspberry Pi 5	BeagleBone® AI-64	Rock 5C	
AI Accelerator	20 TOPS	11 TOPS (only with AI kit)	8 TOPS	6 TOPS	
PCIe	1 x4 + 3 x1 (PCIe Gen3, Root Port, & Endpoint)	1 lane Gen 2	4 x2 lane Gen 2	1 x1 lane Gen 2.1	
SoC	6-core Arm® Cortex®-A78AE v8.2 64-bit CPU 1.5MB L2 + 4MB L3 1.5 GHz	Broadcom BCM2712	Texas Instruments TDA4VM SoC	Rockchip RK3588S2	
Memory	4 GB 64-bit LPDDR5 - 34 GB/s	4 GB LPDDR4X	4 GB LPDDR4	4 GB LPDDR4X	
Storage	(Supports external NVMe)	-	16 GB eMMC	-	
Display	1x 4K30 multi-mode DP 1.2 (+MST)/eDP 1.4/HDMI 1.4	Dual 4k60	Mini Display Port 4k30 4 lane MIPI DSI	1x HDMI 2.1 8Kp60 1x MIPI DSI 1080p60	
GPIO	40-pin GPIO header	40-pin GPIO header	40-pin GPIO header	40-pin GPIO header	
Cellular	-	-	-		
Wi-Fi	-	802.11ac 2.4/5GHz	-	WiFi 6	
BLE	-	5.2	-	5.4	
USB Ports	3x USB 3.2 Gen2 (10 Gbps) 3x USB 2.0	2 x USB-C. 2 x USB-A Built in hub	2 x USB-A + 1 x USB-C	2 x USB-A -2.0 + 1 x USB-A 3.0 + 1 x USB-A 3.0 OTG	
Camera	Up to 4 cameras (8 via virtual channels) 8 lanes MIPI CSI-2 D-PHY 2.1 (up to 20 Gbps)	2 x 4-lane MIPI DSI	2 x 4-lane MIPI DSI	1x 4-lane MIPI CSI or 2x 2-lane MIPI CSI	
Ethernet	Gigabit Ethernet port	Gigabit Ethernet port	Gigabit Ethernet port	Gigabit Ethernet port	-

Intégré dans le vaste écosystème IA de Nvidia, notamment CUDA, cuDNN et TensorRT, l'Orin Nano est compatible avec les principaux frameworks tels que TensorFlow et PyTorch, offrant ainsi une grande flexibilité pour répondre à divers besoins de développement. Il est équipé d'options d'E/S à grande vitesse telles que USB 3.1, PCIe et Gigabit Ethernet, tout en

conservant l'efficacité énergétique, le rendant adapté à des applications d'Edge AI. Que vous soyez en train d'automatiser des processus, de construire des dispositifs intelligents ou de vous lancer dans la robotique, le Jetson Orin Nano fournit les performances robustes et la polyvalence nécessaires pour concrétiser vos projets d'IA. Avec un mélange de puissance et d'accessibilité, il constitue un excellent choix tant pour les débutants que pour les développeurs expérimentés.

Raspberry Pi 5

Lorsqu'il est question de choisir un matériel pour presque n'importe quel projet, le Raspberry Pi est souvent privilégié pour son coût abordable, sa grande polyvalence, et le vaste soutien de sa communauté. Le Raspberry Pi 5 [4] (voir **figure 2**) poursuit cette tradition ; équipé d'un processeur ARM Cortex-A76 à quatre cœurs et pouvant embarquer jusqu'à 8 GB de RAM, il est compatible avec un large éventail de

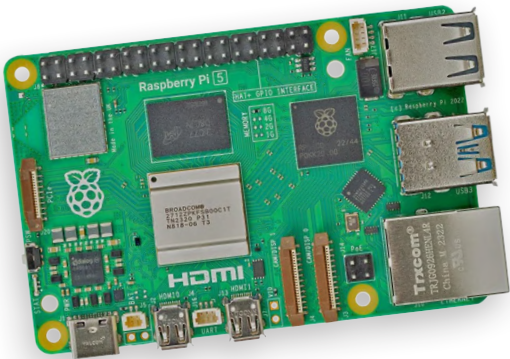


Figure 2. Raspberry Pi 5
(Source : Raspberry Pi)

Tachyon	Google Coral	Orange Pi 5	ASUS Tinker Board 2S	Alta AML-A311D-CC
12 TOPS	4 TOPS (int8); 2 TOPS per watt	6 TOPS	2.3 TOPS	5 TOPS
2 lanes Gen 3	-	PCIe 3.0 slot	PCIe Gen3 x2	Not specified
1 x Gold Plus @ 2.7 GHz +3xGold@2.4GHz +4xSilver@1.9GHz	NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F) @ 1.5GHz	Rockchip RK3588S	Dual 2.0 GHz Cortex-A72 and Quad 1.5 GHz Cortex-A53	Amlogic A311D SoC
4 GB LPDDR4X	1 or 4 GB LPDDR4	2 GB/4 GB DDR4	2 GB/4 GB LPDDR4	4 GB LPDDR4
64 GB	8 GB eMMC	Optional eMMC Module	16 GB eMMC storage	
Single 4K60 along with 4 lane MIPI DSI	39-pin FFC connector for MIPI-DSI display (4-lane) HDMI 2.0a (full size)	HDMI 2.0	HDMI 2.0, MIPI-DSI	HDMI 2.0, 4K
40-pin GPIO header	40-pin GPIO header	40-pin GPIO header	40-pin GPIO header	40-pin GPIO header
Sub-GHz 5G - CAT-18				
802.11ax 2.4/5/6GHz (WiFi 6)	Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5GHz)	802.11 a/b/g/n/ac	802.11 a/b/g/n/ac	
5.2	Bluetooth 4.2	Bluetooth 5.0	Bluetooth 5.0	
2 x USB-C (1 x port inc USB-C 3.1 with integrated Display Port)	Type-C OTG; Type-C power; Type-A 3.0 host; Micro-B serial console	USB 3.0, USB 2.0	1 x USB 3.2 Gen1 Type-C (OTG), 3 x USB 3.2 Gen1 Type-A, 1 x USB 2.0 Type-A	4 x USB 3.0
2 x 4-lane MIPI CSI	4-pin FFC connector for MIPI-CSI2 camera (4-lane)	MIPI-CSI interface	MIPI-CSI interface	Not specified
Gigabit Ethernet port	Gigabit Ethernet	Gigabit Ethernet	Gigabit Ethernet	Gigabit Ethernet

frameworks d'IA tels que TensorFlow Lite, OpenCV, et PyTorch. Cela le rend adapté à diverses tâches, y compris les projets de base de vision par ordinateur et d'apprentissage automatique.

Cependant, bien que le Raspberry Pi 5 soit puissant, il n'est pas spécifiquement optimisé pour les applications d'IA gourmandes en ressources. Pour les tâches plus exigeantes, le kit Raspberry Pi AI [5] avec l'accélérateur d'IA Hailo 8L change la donne, (voir **figure 3**). Le Hailo 8L ajoute 13 TOPS de puissance de traitement d'IA, permettant au Raspberry Pi de gérer plus efficacement des tâches d'IA complexes telles que la reconnaissance d'images en temps réel et l'apprentissage profond.

En conclusion, le Raspberry Pi 5 constitue un excellent point de départ pour le développement IA, en particulier lorsqu'il est associé au Hailo 8L pour des performances accrues, ce qui en fait un choix polyvalent et rentable pour les débutants et les développeurs avancés.

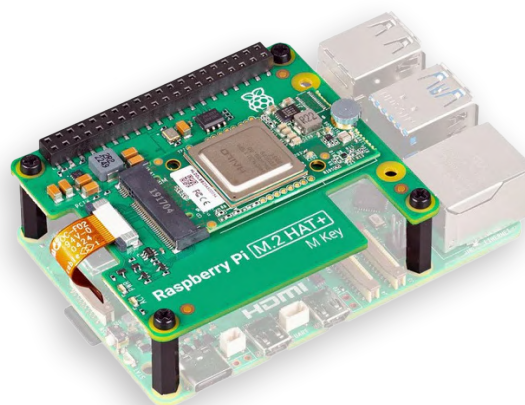


Figure 3.
Raspberry Pi AI Kit
(Source : Raspberry Pi)

Orange Pi 5

Les Orange Pi 5 et 5 Pro [6] sont des choix de plus en plus populaires dans les domaines de l'IA et de l'informatique en général, offrant une alternative puissante et abordable aux autres SBC. Les deux modèles sont

Figure 4. Orange Pi 5
(Source : Orange Pi)

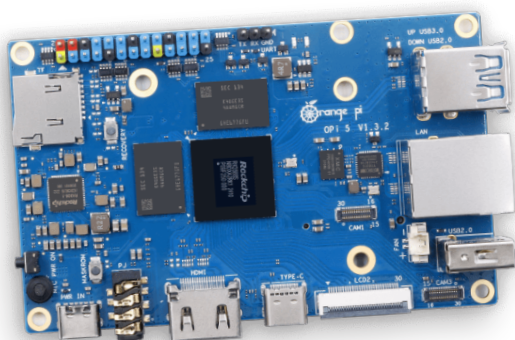
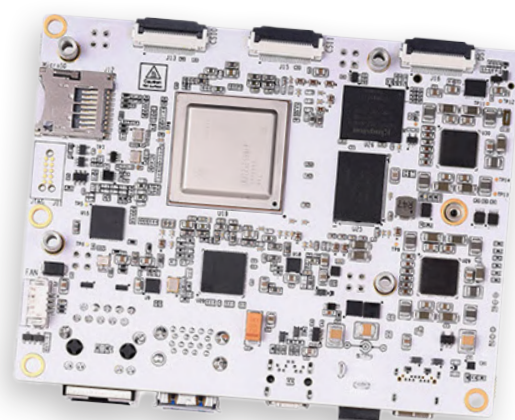


Figure 5. BeagleBone AI-64 (Source : Texas Instruments)



équipés du processeur Rockchip RK3588S, qui présente une configuration octa-core avec quatre cœurs ARM Cortex-A76 et quatre cœurs Cortex-A55. Cette configuration offre une puissance de traitement robuste, renforcée par jusqu'à 16 Go de RAM, ce qui rend ces cartes bien adaptées aux applications gourmandes en ressources, notamment l'IA et l'apprentissage automatique.

L'une des principales caractéristiques de la série Orange Pi 5 (**figure 4**) est son NPU intégré, qui fournit jusqu'à 6 TOPS de performances en matière d'IA. Ce NPU améliore considérablement la capacité de la carte à gérer des tâches d'IA telles que la détection d'objets, la reconnaissance d'images et le traitement du langage naturel, ce qui en fait un concurrent de choix pour les projets nécessitant un traitement de l'IA en temps réel. Les Orange Pi 5 et 5 Pro sont compatibles avec une vaste gamme de frameworks d'IA, notamment TensorFlow, PyTorch et OpenCV. Cette compatibilité étendue permet aux développeurs d'utiliser ces cartes pour diverses applications d'IA, de la vision par ordinateur à des modèles d'apprentissage automatique plus complexes. Le GPU Mali-G610 MP4 renforce encore leur capacité à gérer les tâches graphiques intensives. Une différence majeure entre le Pi 5 d'Orange et le 5 Pro est l'intégration exclusive du Wi-Fi et du Bluetooth dans la version Pro.

L'Orange Pi 5 Pro, avec ses options de RAM plus élevées et ses fonctionnalités améliorées, offre des performances et une flexibilité supplémentaires, ce qui le rend idéal pour les projets plus exigeants. Il comprend des options de connectivité avancées telles que HDMI 2.1, USB 3.0 et PCIe 3.0, permettant l'ajout de périphériques à haut débit et l'expansion des possibilités d'utilisation.

BeagleBone AI-64

Le BeagleBone AI-64 [7] est un ordinateur monocarte puissant conçu pour l'IA, afin de répondre aux exigences des projets d'IA embarquée et du calcul en temps réel (**figure 5**). Alimenté par le SoC TDA4VM de Texas Instruments, il est doté d'un CPU ARM Cortex-A72 à double cœur cadencé à 2,0 GHz, garantissant des performances robustes pour une variété de tâches. La carte AI-64 est équipée d'un accélérateur d'IA capable de fournir 8 TOPS, ce qui la rend bien adaptée aux applications exigeantes d'IA et d'apprentissage automatique.

Avec 4 Go de RAM LPDDR4 et 16 Go de stockage eMMC intégré, la BeagleBone AI-64 offre suffisamment de mémoire et de stockage pour la plupart des projets d'IA embarqués. La carte prend en charge Linux Yocto, un système d'exploitation flexible et personnalisable, idéal pour les développeurs nécessitant une adaptation précise de leur environnement système à des applications spécifiques. La connectivité est également un point fort, avec des ports USB-A doubles, USB-C, et Ethernet Gigabit, assurant une intégration aisée dans divers réseaux. La carte offre aussi une grande flexibilité d'extension grâce à son en-tête GPIO de 40 broches, permettant l'ajout facile de capteurs et de périphériques variés.

L'horloge en temps réel et les capacités d'alimentation via USB-PD de la BeagleBone AI-64 renforcent encore son adéquation avec les applications industrielles et IdO où la synchronisation précise et une alimentation fiable sont primordiales. Cependant, l'absence de Wifi et de BLE intégrés peut nécessiter des modules supplémentaires pour la connectivité sans fil, en fonction des besoins du projet. Globalement, le BeagleBone AI-64 est un choix judicieux pour les développeurs qui cherchent à mettre en œuvre des solutions d'IA avancées dans des environnements embarqués, en particulier lorsque le traitement en temps réel et l'intégration matérielle robuste sont essentiels.

BeagleY-AI

Le BeagleY-AI [8] est un autre choix remarquable au sein de la famille BeagleBoard, particulièrement adapté aux applications de vision et les traitements DSP intensifs, (voir figure 6). Elle est alimentée par le processeur de vision Texas Instruments AM67A, qui comprend un processeur ARM Cortex-A53 quadricœur

cadencé à 1,4 GHz, et deux DSP (processeurs de signaux numériques) C7x avec accélérateur de multiplication matricielle (MMA), capable de délivrer 4 TOPS de performance IA. Cette configuration est particulièrement bien adaptée aux tâches qui nécessitent un traitement d'image en temps réel, une vision par ordinateur et d'autres charges de travail d'IA qui bénéficient d'un traitement DSP spécialisé.

Le BeagleY-AI dispose de 4 Go de mémoire LPDDR4, garantissant des performances fluides pour les applications exigeantes. Elle inclut également plusieurs interfaces à haut débit, telles que PCIe Gen3, USB 3.1 et Gigabit Ethernet avec prise en charge PoE+, offrant d'excellentes options de connectivité pour les tâches gourmandes en données. La carte supporte l'affichage sur trois écrans simultanément via les interfaces microHDMI, MIPI-DSI et OLDI (LVDS), ce qui la rend idéale pour les configurations multi-écrans dans les applications de vision pilotées par l'IA.

Avec le Wi-Fi 6 et le Bluetooth 5.4, le BeagleY-AI offre une connectivité sans fil avancée, ce qui constitue un avantage significatif par rapport au BeagleBone AI-64. Il est donc particulièrement bien adapté aux projets qui nécessitent des performances réseau robustes et une communication à faible latence, comme les caméras intelligentes, les robots autonomes et d'autres appareils de périphérie.

Le BeagleY-AI offre également une large gamme d'options d'E/S, notamment le MIPI CSI pour les caméras, un connecteur d'extension à 40 broches et plusieurs ports USB, ce qui en fait une plateforme polyvalente pour les développeurs. De plus, avec un connecteur JTAG et un UART de console, les développeurs peuvent facilement déboguer et optimiser leurs applications.

La BeagleBone AI-64 est très performante pour les applications d'IA industrielles qui nécessitent un traitement en temps réel et une intégration matérielle robuste, tandis que la BeagleY-AI excelle dans les projets d'IA destinés à la vision, offrant des capacités DSP avancées et des options de connectivité supérieures. Les deux cartes constituent de puissantes plateformes pour le développement en IA, et le choix entre elles dépend des exigences spécifiques de votre projet.

Tachyon

Le Tachyon est un nouvel ordinateur monocarte (SBC) conçu par Particle [9], qui fait actuellement l'objet d'une campagne Kickstarter [10] au moment de la rédaction de cet article. Cette carte est prête à transformer le domaine de l'informatique portable et à distance en combinant le puissant SoC Qualcomm Snapdragon avec la connectivité 5G, l'accélération de l'IA et des options d'E/S étendues, le tout dans un facteur de forme compatible avec Raspberry Pi (voir **figure 7**

Le Tachyon est un nouvel ordinateur monocarte (SBC) conçu par Particle [9], qui fait actuellement l'objet d'une campagne Kickstarter [10] au moment où nous écrivons ces lignes. Cette carte est prête à redéfinir le paysage de l'informatique portable et à distance en combinant le puissant SoC Qualcomm Snapdragon avec la connectivité 5G, l'accélération de l'IA et des options d'E/S étendues, le tout dans un facteur de forme compatible avec Raspberry Pi, (voir **figure 7**). Au cœur du Tachyon se trouve le SoC Qualcomm QCM6490, un chipset également utilisé dans les smartphones de milieu à haut de gamme. Il contient un CPU Kryo octa-core (1× Gold Plus @ 2,7 GHz, 3× Gold @ 2,4 GHz, et 4× Silver @ 1,9 GHz), délivrant une puissance de traitement impressionnante. En complément du CPU, on trouve un GPU Qualcomm Adreno 643 et un DSP Qualcomm Hexagon 770 avec un accélérateur d'IA 12 TOPS, rendant cette carte parfaitement capable de gérer des tâches avancées d'IA/ML, comme la détection d'objets en temps réel et la classification vidéo/audio.

Tachyon est équipé de jusqu'à 8 Go de mémoire LPDDR4X et 64 Go de stockage intégré, garantissant un espace et une vitesse suffisants pour la plupart des applications. La carte supporte une variété de

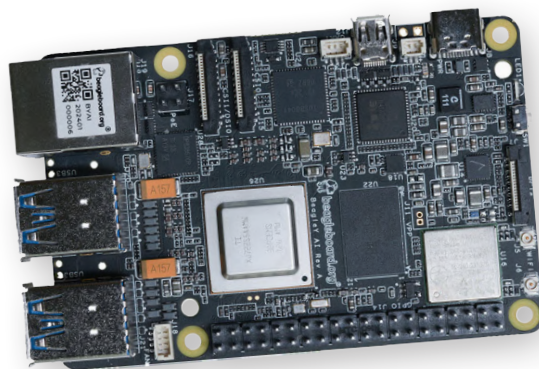


Figure 6. BeagleY AI (Source : Texas Instruments)



Figure 7. Tachyon (Source : Particle)

Le choix du matériel approprié pour le développement de l'IA dépend de plusieurs facteurs, tels que la complexité et les besoins en ressources des modèles d'IA, le volume et la nature des ensembles de données, la nécessité d'un traitement à faible latence ou en temps réel, ainsi que les contraintes de puissance et d'efficacité du projet.

systèmes d'exploitation, notamment Ubuntu 24.04, Yocto, Android, Qualcomm Linux et même Windows 11, offrant aux développeurs une flexibilité sans précédent. En termes de connectivité, la carte comprend deux ports USB-C (dont un supportant DisplayPort), Wi-Fi 6, Bluetooth 5.2, et des capacités cellulaires 5G sub-GHz, idéales pour les applications de calcul en périphérie et IoT nécessitant des connexions fiables et rapides.

Conçu pour une polyvalence maximale, Tachyon est doté d'un connecteur GPIO à 40 broches compatible avec les HAT Raspberry Pi, d'une double interface de caméra MIPI CSI et d'un support d'affichage 4K via MIPI DSI, ce qui le rend capable de supporter tout, des assistants IA à la vision par ordinateur, en passant par l'hébergement de médias et les jeux portables. Avec des antennes intégrées pour la 5G et le Wifi, Tachyon peut être déployé dans divers environnements, que ce soit à la maison, sur le terrain ou au sein d'applications d'entreprise.

En résumé, Tachyon est un SBC puissant et flexible qui apporte des performances de niveau smartphone à l'espace edge computing et IdO. Elle est conçue pour être une carte « passe-partout », capable de tout gérer, des projets personnels aux déploiements à grande échelle, ce qui en fait un excellent choix pour les développeurs qui cherchent à repousser les limites de ce qui est possible avec un SBC.

Rock 5C et 5C Lite

Les Rock 5C et Rock 5C Lite [11] sont de puissants ordinateurs monocartes (SBC) développés par Radxa, conçus pour les utilisateurs à la recherche d'un équilibre entre performances et prix abordable, comme le montre **la figure 8**. Les deux cartes sont construites autour des SoC Rockchip, le Rock 5C étant équipé du plus puissant RK3588S2 et le Rock 5C Lite du RK3582.

Le Rock 5C se distingue par son processeur octa-core (quatre ARM Cortex-A76 et quatre Cortex-A55), un GPU Mali G610MP4 et un NPU capable de fournir jusqu'à 6 TOPS de performances IA. Cette configuration lui permet d'accomplir des tâches allant des projets AI/ML aux applications multimédias, y compris le décodage vidéo en 8K. Il est également équipé du Wi-Fi 6, du Bluetooth 5.4, de plusieurs ports USB et d'un port Gigabit Ethernet, offrant ainsi de solides options de connectivité. Le Rock 5C prend en charge jusqu'à 32 Go de RAM LPDDR4x, ce qui en fait une plateforme polyvalente pour un large éventail de cas d'utilisation. D'autre part, le Rock 5C Lite est une version plus simplifiée, dotée d'un CPU Cortex-A76 à double cœur et d'un Cortex-A55 à quadruple cœur, mais sans GPU. Il comprend toujours un NPU doté d'une puissance de traitement de l'IA de 5 TOPS, ce qui en fait un choix solide pour les projets d'IA qui ne nécessitent pas de traitement graphique intense. La version Lite prend en charge les mêmes options de connectivité que le 5C, sans le GPU, ce qui en fait un choix efficace pour les applications axées sur l'IA où l'efficacité énergétique et la rentabilité sont essentielles.

Les deux cartes offrent des options de stockage extensibles via des emplacements eMMC et microSD et offrent des sorties d'affichage doubles avec des interfaces HDMI 2.1 et MIPI DSI. Elles sont également dotées d'options de refroidissement externe, ce qui est crucial pour maintenir les performances sous de lourdes charges. La Rock 5C est idéale pour les utilisateurs qui ont besoin de capacités multimédias et IA complètes, tandis que la Rock 5C Lite est conçue pour les applications axées sur l'IA où la puissance du GPU est moins critique. Les deux cartes constituent d'excellents choix en fonction des besoins spécifiques de votre projet.

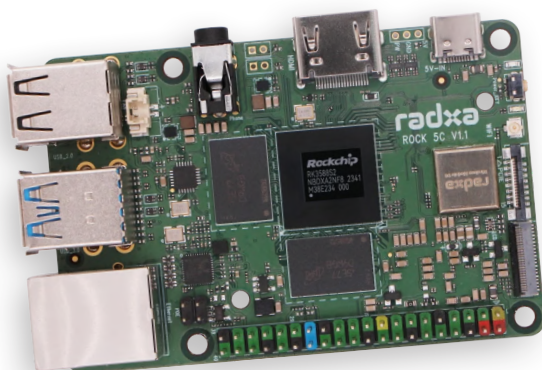


Figure 8. Rock 5C
(Source : Radxa)

Encore plus d'options

À mesure que l'intelligence artificielle continue de s'intégrer dans divers aspects de notre vie quotidienne, de la santé à la fabrication, l'importance de choisir le bon matériel pour développer des projets d'IA ne peut être sous-estimée. La réussite de tout projet d'IA est profondément liée aux capacités du matériel sous-jacent, qui doit gérer efficacement les calculs intensifs requis par les algorithmes d'apprentissage automatique.

Dans ce domaine en pleine évolution, le choix du bon ordinateur monocarte est crucial. Bien que cet article ait fourni une sélection de SBC polyvalents et puissants adaptés à une gamme d'applications d'IA - de la vision par ordinateur à la robotique - il existe de nombreuses autres options [12]. Chaque plateforme a ses propres avantages uniques, ce qui rend essentielle l'adaptation du matériel aux besoins spécifiques de votre projet. Alors que vous poursuivez votre parcours de développement de l'IA, n'oubliez pas que le bon matériel permet non seulement d'accélérer vos projets, mais aussi d'évoluer en fonction de vos ambitions, que vous soyez au début de votre aventure ou que vous cherchiez à repousser les frontières du possible avec l'IA. ◀

240473-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (saad.imtiaz@elektor.com) ou contactez Elektor (redaction@elektor.fr).



À propos de l'auteur

Saad Imtiaz, ingénieur senior chez Elektor, est spécialisé en mécatronique et possède une solide expérience dans les systèmes embarqués et le développement de produits. Tout au long de sa carrière, il a collaboré avec un large éventail d'entreprises, des startups novatrices aux multinationales bien établies, en pilotant des projets de prototypage et de développement à la pointe de la technologie. Avec un parcours significatif dans l'industrie aéronautique et à la tête d'une startup technologique, Saad apporte à Elektor une combinaison unique de compétences techniques et d'esprit entrepreneurial. Il contribue au développement de projets dans les domaines du logiciel et du matériel.



Produits

- > **Raspberry Pi 5 Ultimate Starter Kit (8 GB)**
www.elektor.fr/20721
- > **Raspberry Pi AI Kit**
www.elektor.fr/20879
- > **Google Coral USB Accelerator**
www.elektor.fr/19366



LIENS

- [1] The Single Board Computer Database: <https://hackerboards.com/>
- [2] Single Board Computer List : https://www.blackmagicboxes.com/?page_id=466
- [3] NVIDIA Jetson Orin:
<https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/>
- [4] Raspberry Pi 5: <https://www.raspberrypi.com/products/raspberry-pi-5/>
- [5] Raspberry Pi AI Kit: <https://www.raspberrypi.com/products/ai-kit/>
- [6] Orange Pi 5:
<http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/details/Orange-Pi-5.html>
- [7] BeagleBone AI-64: <https://www.beagleboard.org/boards/beaglebone-ai-64>
- [8] BeagleY-AI: <https://www.beagleboard.org/boards/beagle-y-ai>
- [9] Particle: <https://www.particle.io/>
- [10] Particle Tachyon: <https://www.kickstarter.com/projects/particle-iot/tachyon-powerful-5g-single-board-computer-w-ai-accelerator>
- [11] Radxa ROCK 5C: <https://radxa.com/products/rock5/5c/>
- [12] Brian Tristram Williams, "2024: An AI Odyssey - Desktop Versus Embedded Accelerators: A Look at Some Options," Elektor 9-10/2024: <https://elektormagazine.com/230181-H-01>

des données de capteurs aux modèles d'apprentissage automatique

Détection de gestes avec Edge Impulse et un accéléromètre

Koen Vervloesem (Belgique)

Nous sommes habitués à interagir avec nos appareils via des claviers, des souris ou des écrans tactiles. Toutefois, il peut être intéressant de rechercher des alternatives à ces interfaces classiques.

La reconnaissance des gestes représente une telle alternative prometteuse. Pour cela, nous pouvons utiliser Edge Impulse afin de développer un modèle d'apprentissage automatique qui détecte les gestes en exploitant les données provenant d'un accéléromètre.

L'une des méthodes de reconnaissance gestuelle utilise une caméra qui enregistre les mouvements de la main [1]. Toutefois, une méthode alternative consiste à intégrer un accéléromètre dans un appareil porté au poignet ou tenu en main, et il est possible d'utiliser l'apprentissage automatique pour identifier les gestes à partir des données de mouvement, réduisant ainsi le volume de données nécessaires au traitement.

Dans cet article, je vais créer un système de

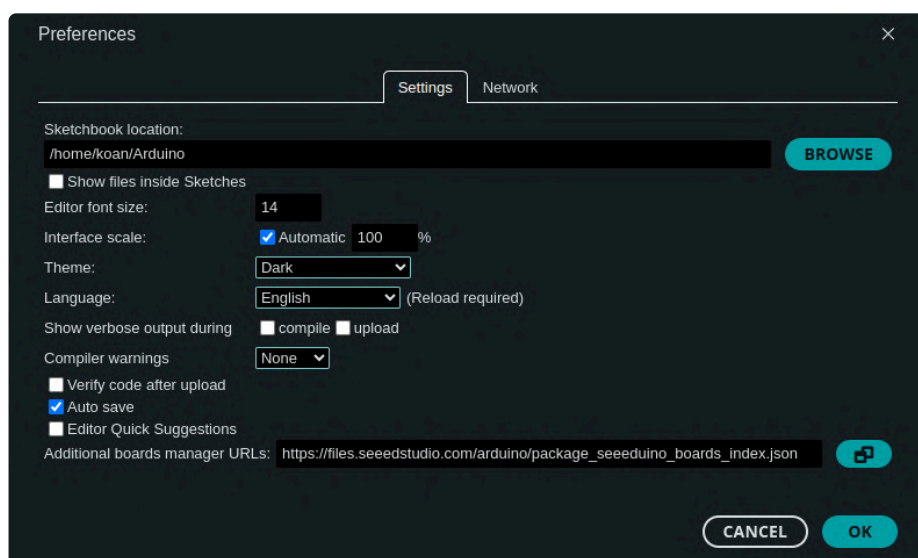


Figure 1. Ajouter l'URL des cartes Seeed à l'EDI Arduino.

reconnaissance gestuelle fonctionnant sur une carte de microcontrôleur XIAO nRF52840 Sense de Seeed Studio, équipée d'un accéléromètre. En utilisant l'EDI Arduino [2], je développerai d'abord un croquis Arduino pour lire les données du capteur d'accéléromètre. Je réaliserai les gestes spécifiques que je souhaite que le système reconnaisse et je transmettrai ces données à la plateforme Edge Impulse [3] pour entraîner le modèle.

Sur Edge Impulse, je créerai un modèle d'apprentissage automatique qui reconnaît les gestes. Ce modèle sera ensuite déployé sur le microcontrôleur en tant que modèle TensorFlow Lite qui peut être facilement utilisé comme bibliothèque Arduino. Vous pouvez ensuite vous appuyer sur ce modèle pour contrôler votre ordinateur par gestes.

Configuration de l'EDI Arduino

Tout d'abord, installez l'EDI Arduino, disponible pour les systèmes d'exploitation Windows, macOS et Linux. La carte Seeed nécessite un package de carte spécifique pour fonctionner avec l'EDI Arduino. Accédez à *Fichier / Préférences* et ajoutez l'URL https://files.seeedstudio.com/arduino/package_seeeduino_boards_index.json au champ *URL du gestionnaire de cartes supplémentaires* (figure 1). Cliquez sur *OK* pour charger les informations nécessaires sur le package de cartes depuis ce fichier en ligne.

Ensuite, ouvrez le *gestionnaire de cartes* en cliquant sur l'icône de la carte dans la barre latérale gauche et recherchez « seeed nrf ». Sélectionnez le paquet nommé Seeed nRF52 mbed-enabled Boards et cliquez sur *Installer*.

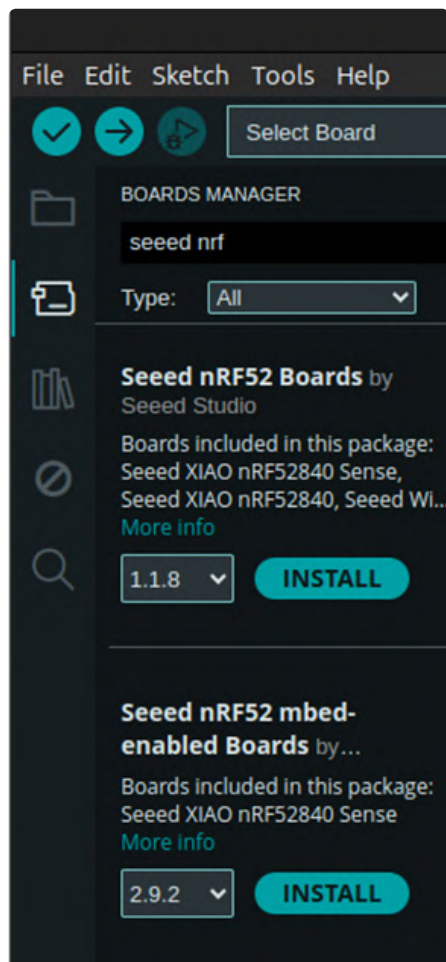


Figure 2. Installer le paquet Seeed nRF52 mbed-enabled Boards pour Arduino.

Ce paquet est optimisé pour les applications d'apprentissage automatique embarquées (figure 2). Après l'installation, connectez l'appareil au port USB de votre ordinateur, allez dans *Outils / Carte / Seeed nRF52 mbed-enabled Boards*, et choisissez *Seeed XIAO BLE Sense-nRF52840*. Ensuite, naviguez vers *Outils / Port* et sélectionnez le port série de votre appareil.

Pour vérifier que vous pouvez flasher un croquis Arduino sur l'appareil, ouvrez *Fichier / Exemples / 01.Basics / Blink*. Cela ouvre un croquis d'exemple qui fait simplement clignoter la LED intégrée. Cliquez sur l'icône *Téléverser* (la flèche orientée vers la droite) pour construire et flasher le croquis. Si tout se passe bien, la LED sur la carte commence à clignoter après environ 1 minute.

Essai de l'accéléromètre

Le nRF52840 Sense de Seeed Studio XIAO est équipé d'une unité de mesure inertielle (IMU) de haute précision à six axes LSM6DS3 qui inclut un accéléromètre à trois axes et un gyroscope à trois axes. Pour l'utiliser dans votre croquis Arduino, allez sur l'icône *Library*

Manager dans la barre latérale gauche et recherchez « seeed lsm ». Installez la bibliothèque *Seeed Arduino LSM6DS3*.

La bibliothèque propose un code d'exemple, naviguez donc vers *Fichier / Exemples / Seeed Arduino LSM6DS3 / HighLevelExample*. Compilez et exécutez ce code, puis ouvrez le *Serial Monitor* dans le coin supérieur droit de l'EDI Arduino pour visualiser les coordonnées X, Y et Z de l'accéléromètre et du gyroscope, ainsi que la température du capteur interne (figure 3). Déplacez maintenant la carte Seeed dans votre main et observez les valeurs changeantes.

Modification du code de l'accéléromètre

Je souhaite utiliser les composantes X, Y et Z de l'accéléromètre pour développer un modèle d'apprentissage automatique, mais l'appareil un format de transmission des données plus compact. De plus, au lieu d'envoyer un échantillon toutes les secondes, comme c'est le cas avec l'instruction `delay(1000);`, je veux établir une fréquence d'échantillonnage de 50 Hz pour capturer avec précision les

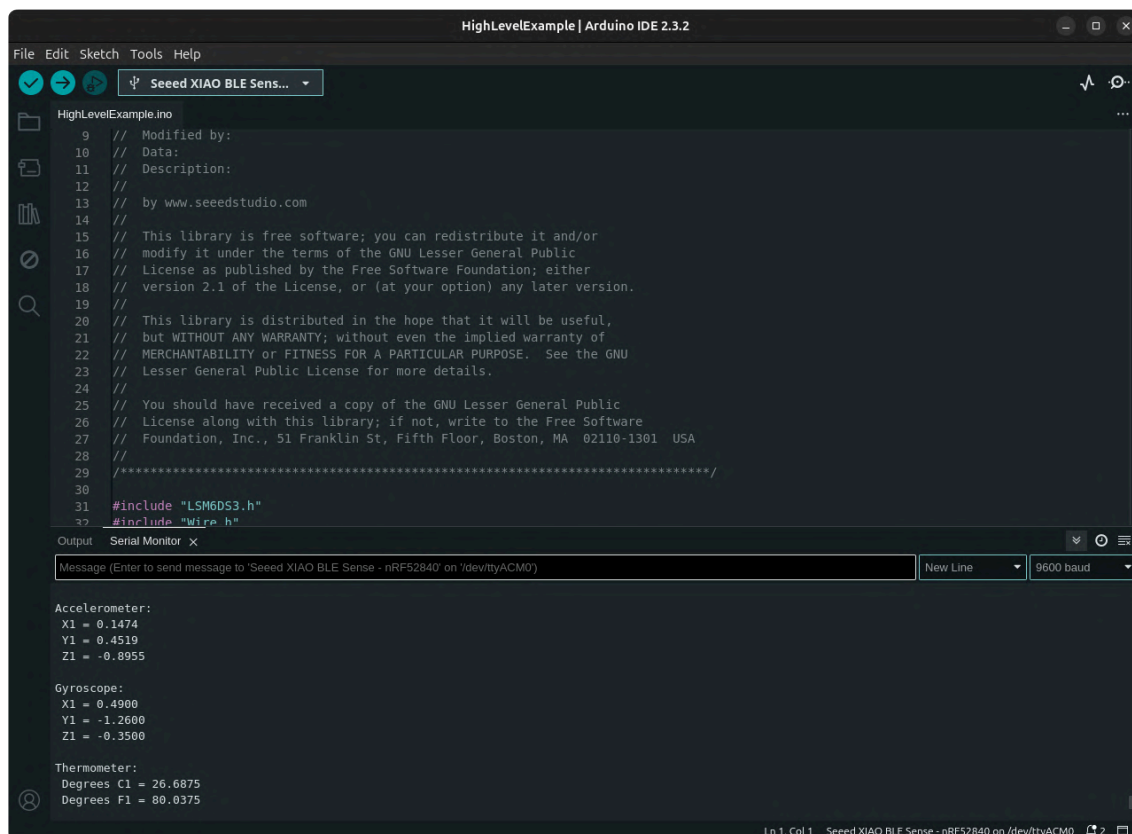


Figure 3. Les relevés de l'accéléromètre lors du déplacement de la carte Seeed.

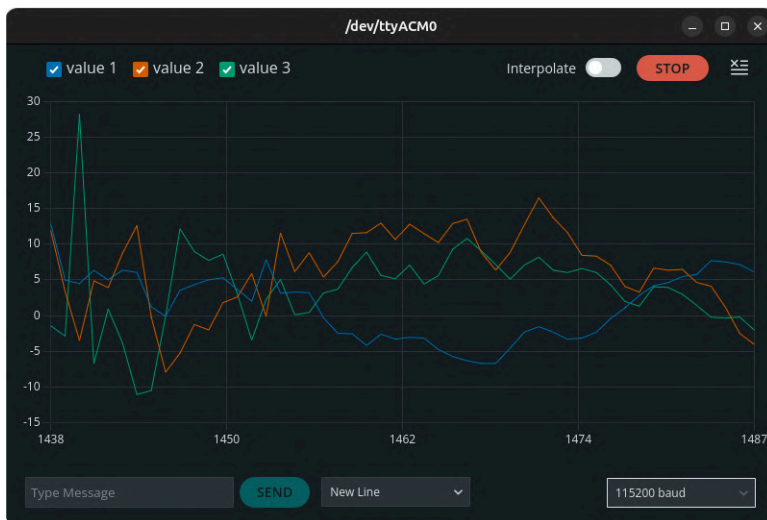


Figure 4. Le traceur série de l'EDI Arduino affiche les données de l'accéléromètre.

mouvements rapides de la main. Créez donc un nouveau croquis avec le code du **listage 1**, également disponible sur [4].

Exécutez ce code sur l'appareil et observez la sortie sur le moniteur série. Cela vous permettra de visualiser les variations de mouvement de façon beaucoup plus rapide, affichant les trois composantes d'accélération en m/s^2 . Si vous cliquez sur l'icône *Serial Plotter* à côté du *Serial Monitor*, vous pouvez même visualiser les trois tracées des composantes de l'accéléromètre (**figure 4**).

Gestes

Pour démarrer avec l'acquisition de données d'accéléromètre en vue de créer un ensemble de données d'entraînement, il est d'abord



Listage 1. Lecture des données de l'accéléromètre LSM6DS3.

```
// XIAO BLE Sense LSM6DS3 Accelerometer Raw Data

#include "LSM6DS3.h"
#include "Wire.h"

//Create a instance of class LSM6DS3
LSM6DS3 myIMU(I2C_MODE, 0x6A); //I2C device address 0x6A

#define CONVERT_G_TO_MS2 9.80665f
#define FREQUENCY_HZ 50
#define INTERVAL_MS (1000 / (FREQUENCY_HZ + 1))

static unsigned long last_interval_ms = 0;

void setup() {
  Serial.begin(115200);
  while (!Serial)
    ;

  if (myIMU.begin() != 0) {
    Serial.println("Device error");
  } else {
    Serial.println("Device OK!");
  }
}

void loop() {
  if (millis() > last_interval_ms + INTERVAL_MS) {
    last_interval_ms = millis();
    Serial.print(myIMU.readFloatAccelX() * CONVERT_G_TO_MS2, 4);
    Serial.print('\t');
    Serial.print(myIMU.readFloatAccelY() * CONVERT_G_TO_MS2, 4);
    Serial.print('\t');
    Serial.println(myIMU.readFloatAccelZ() * CONVERT_G_TO_MS2, 4);
  }
}
```

Figure 5. Créer un nouveau projet dans Edge Impulse..

nécessaire de s'inscrire sur Edge Impulse. Commencez par créer un compte, qui offre un compte gratuit dans le cadre de son plan communautaire pour les étudiants, les universités et les développeurs individuels [5]. Dans votre tableau de bord, cliquez sur *Créer un nouveau projet*, donnez un nom au projet, et choisissez de rendre le projet public ou privé (**figure 5**).

Ensuite, il est essentiel de définir quels gestes

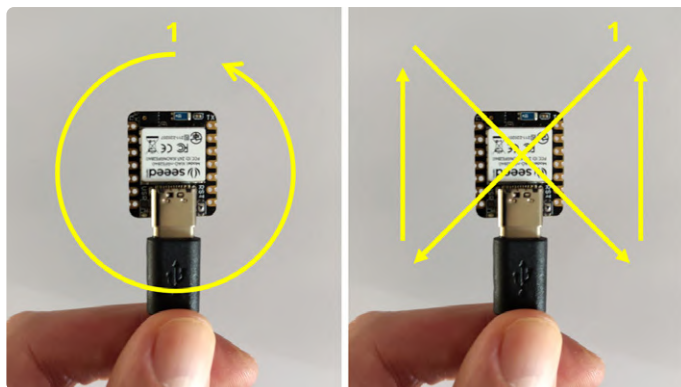


Figure 6. Les deux gestes que nous voulons détecter avec l'accéléromètre intégré de la carte.

vous souhaitez que votre modèle reconnaisse. Prenez, par exemple, les gestes formant un cercle et une croix. Pour enregistrer geste circulaire, effectuez un mouvement circulaire avec votre dispositif. Pour une croix, déplacez la carte dans un plan imaginaire allant d'en haut à gauche à en bas à droite, puis d'en haut à droite à en bas à gauche, et retournez à la position de départ. (Figure 6).

Nous allons maintenant créer un jeu de données avec au moins 50 échantillons pour chaque mouvement, et 50 échantillons pour les mouvements inconnus (aléatoires). Assurez-vous de maintenir la carte de manière constante à travers tous les échantillons, et de réaliser les mouvements dans une direction cohérente, par exemple dans le sens horaire ou antihoraire pour le cercle, et en respectant une durée approximative similaire pour chaque échantillon (par exemple 2,5 s).

Transmission des données à Edge Impulse

Nous disposons déjà du script permettant d'enregistrer les valeurs de l'accéléromètre ; il nous suffit de transmettre ces données à Edge Impulse pendant nos mouvements d'entraînement, puis de diviser les données en durées de 2,5 s. Pour ce faire, nous pouvons utiliser le Data Forwarder d'Edge Impulse, qui fait partie du logiciel Edge Impulse CLI [6]. Commencez par installer les prérequis sur votre ordinateur, à savoir Python 3 et Node.js. Par la suite, installez les outils CLI nécessaires avec la commande :

```
npm install -g edge-impulse-cli
```

Lancez ensuite le Data Forwarder d'Edge Impulse avec :

```
edge-impulse-data-forwarder
```

Lorsque vous exécutez cette commande, vous serez invité à entrer votre nom d'utilisateur ou votre adresse électronique et votre mot de passe pour Edge Impulse. Le programme se

connecte ensuite à votre appareil via l'interface série USB et vous demande de spécifier le projet auquel les données doivent être associées.

Le transfert de données détecte la fréquence des données, ainsi que les trois axes, et vous invite à nommer les axes. Nommez-les x, y, et z, séparés par des virgules. Ensuite, nommez l'appareil, après quoi le programme affiche l'URL de la section *Data acquisition* du projet (figure 7).

Enregistrement des données de l'entraînement

Dans la section *Collect data* à droite, vous devriez voir votre appareil connecté. Le nombre d'axes et la fréquence devraient déjà être corrects, mais nous devons modifier la durée d'échantillonnage par défaut qui est de 10 s. Nous utiliserons des échantillons de 2,5 s, mais nous répéterons le même geste plusieurs fois dans chaque enregistrement, puis nous diviserons l'enregistrement avec

```
/bin/edge-impulse-data-forwarder
Edge Impulse data forwarder v1.27.1
? What is your user name or e-mail address (edgeimpulse.com)? koen@vervloesen.eu
? What is your password? [hidden]
Endpoints:
  Websocket: wss://remote-ngmt.edgeimpulse.com
  API: https://studio.edgeimpulse.com
  Ingestion: https://ingestion.edgeimpulse.com

[SER] Connecting to /dev/ttyACM0
[SER] Serial is connected (D1:4A:E5:69:16:66:06:9E)
[WS] Connecting to wss://remote-ngmt.edgeimpulse.com
[WS] Connected to wss://remote-ngmt.edgeimpulse.com

? To which project do you want to connect this device? Koen Vervloesen / gestures
[SER] Detecting data frequency...
[SER] Detected data frequency: 50Hz
? 3 sensor axes detected (example values: [2.3115,-6.3171,7.1498]). What do you want to call them? Separate the names with ',': x,y,z
? What name do you want to give this device? xiao
[WS] Device "xiao" is now connected to project "gestures". To connect to another project, run "edge-impulse-data-forwarder --clean".
[WS] Go to https://studio.edgeimpulse.com/studio/509218/acquisition/training to build your machine learning model!
```

Figure 7. Le Edge Impulse Data Forwarder envoie les données de l'accéléromètre de votre microcontrôleur vers le cloud pour l'entraînement.

un intervalle de 2,5 secondes entre chaque répétition. Pour atteindre 50 répétitions, nous avons besoin d'un enregistrement de 250 s. Comme il est difficile de se concentrer aussi longtemps sur la réalisation de mouvements précis, nous enregistrerons 5 échantillons de 50 s, chacun composé de 10 répétitions. Définissez donc une durée d'échantillon de 50 000 ms, et n'oubliez pas d'attribuer une étiquette à chaque geste, par exemple *circle*. Maintenez la planche stable et cliquez sur *Start sampling*. Effectuez un geste circulaire ne dépassant pas 2,5 secondes, à répéter toutes les 5 secondes jusqu'à ce que l'échantillonnage s'arrête. Répétez cette procédure cinq fois. Ensuite, changer l'étiquette pour le deuxième nom de classe *cross*, enregistrer cinq échantillons de 50 s en faisant ce geste (croix). Enfin, enregistrez des échantillons avec l'étiquette *unknown* en faisant des mouvements aléatoires ou en posant la carte sur une table. Après cela, vous devriez avoir recueilli plus de 12 minutes de données (figure 8).

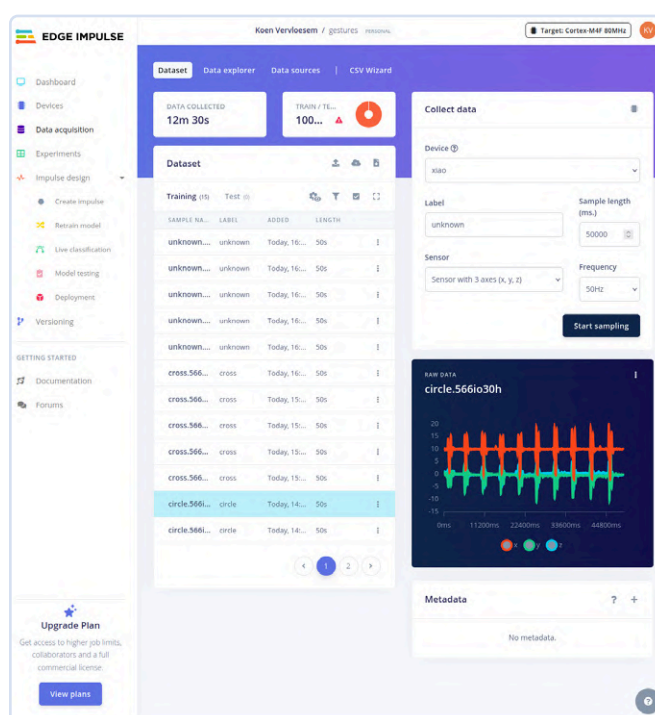


Figure 8. Nous venons d'enregistrer 12 minutes de relevés d'accéléromètre.

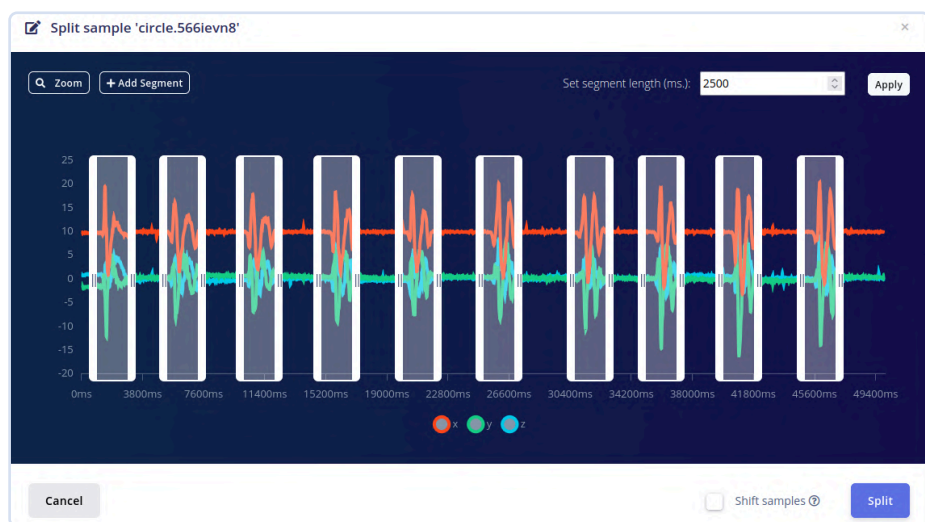


Figure 9. Diviser les enregistrements de l'accéléromètre en 10 échantillons de gestes

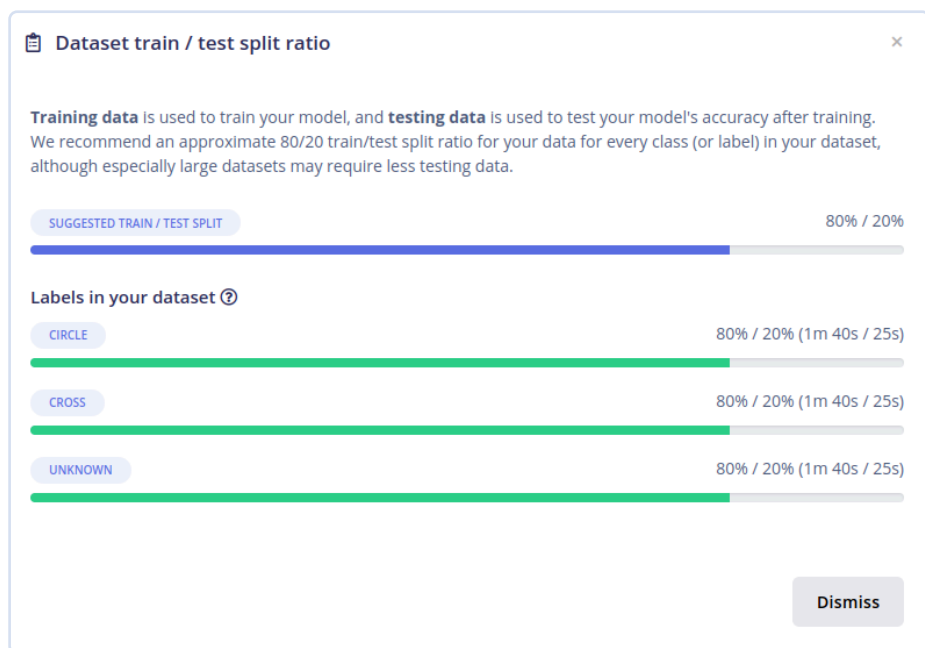


Figure 10. Il s'agit d'une répartition équilibrée entre entraînement et test.

Pour chaque échantillon enregistré, cliquez sur les trois points situés à côté, puis sélectionnez *Split sample*. Pour *Set segment length (ms.)*, entrez 2500 et cliquez sur *Appliquer*. Si vos gestes ne dépassent pas 2,5 s, Edge Impulse les segmente automatiquement. Cliquez sur *Diviser* pour extraire les gestes individuels en tant qu'échantillons. Pour chaque échantillon original, vous obtiendrez 10 échantillons de 3 s maintenant (figure 9). Le nombre d'échantillons inconnus peut varier en fonction de vos mouvements aléatoires. Vous pouvez ajouter des échantillons supplémentaires en cliquant sur *Add Segment*, par

exemple pour inclure des échantillons avec la carte posée à plat. Divisez tous les échantillons enregistrés en conséquence. Si la reconnaissance automatique des segments par Edge Impulse est erronée, vous pouvez toujours ajuster manuellement les segments. À la fin, vous devriez avoir 150 échantillons au total, ce qui donne un peu plus de 6 minutes de données.

Accédez à l'option *Train / Test split* signalée par un point d'exclamation, puis cliquez sur *Perform train / test split*. Chacune de vos étiquettes devrait maintenant être divisée en environ 80 % de données d'entraînement et

20 % de données de test (figure 10). Vous pouvez encore déplacer des échantillons des données d'entraînement vers les données de test ou l'inverse en cliquant sur les trois points d'un échantillon et en choisissant *Move to test set* ou *Move to training set* pour remédier à un déséquilibre.

Création du modèle

Cliquez sur *Create impulse* sous *Impulse design*. Vous devez d'abord configurer votre périphérique cible. Bien que le XIAO nRF52840 Sense ne soit pas listé, vous pouvez sélectionner le *Nordic nRF52840 DK (Cortex-M4F 64MHz)*, qui a le même processeur. Donnez un nom à votre appareil et cliquez sur *Save*. Une impulsion est une série de blocs traitant vos données depuis la série temporelle brute jusqu'aux caractéristiques de sortie. Dans le premier bloc, configurez la taille de la fenêtre à 2 500 ms et l'augmentation de la fenêtre à 400 ms. Cliquez ensuite sur *Add a processing block* et choisissez *Spectral analysis*. Ce bloc identifie les caractéristiques du signal dans le domaine des fréquences. Ensuite, cliquez sur *Add a learning block* et choisissez *Classification*. Ce bloc utilise les caractéristiques spectrales pour détecter si le signal correspond à un cercle, à une croix ou à un geste inconnu. Enfin, cliquez sur *Save impulse* (figure 11).

Cliquez ensuite sur *Spectral features*. Sous *Filter*, cliquez sur *Type* et choisissez *low* pour ajouter un filtre passe-bas. Réglez la fréquence de coupure à 10 Hz pour éliminer les signaux de basse fréquence et réglez l'ordre à 2 pour utiliser un filtre de Butterworth du second ordre. Vous pourrez modifier ces valeurs ultérieurement. Sous *Analyse*, réglez la longueur de la FFT sur 128 fréquences de sortie. Cliquez ensuite sur *Save parameters*. Vous accédez alors à l'onglet *Generate features*. Cliquez sur *Generate features*. Les résultats s'affichent dans le *Feature explorer*. Si les caractéristiques sont clairement différentes pour différents gestes, vous devriez voir les couleurs de chaque geste regroupées (figure 12).

Cliquez ensuite sur *Classify*. L'*architecture du réseau neuronal* présente une couche d'entrée, deux couches denses et une couche de sortie avec les trois classes. Cliquez sur *Add an extra layer*, choisissez *Dropout*, et réglez le taux d'abandon à 0,2. Cela améliore la précision et vous pouvez ajuster le taux ultérieurement pour trouver la valeur optimale.

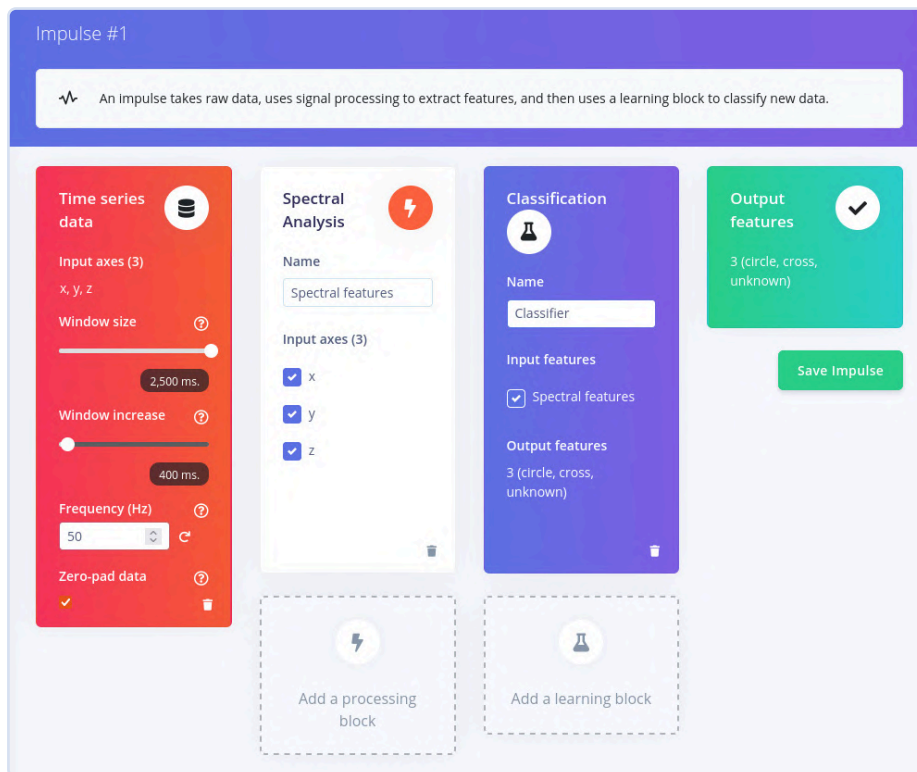


Figure 11. Créez un *impulse* pour classer vos données.

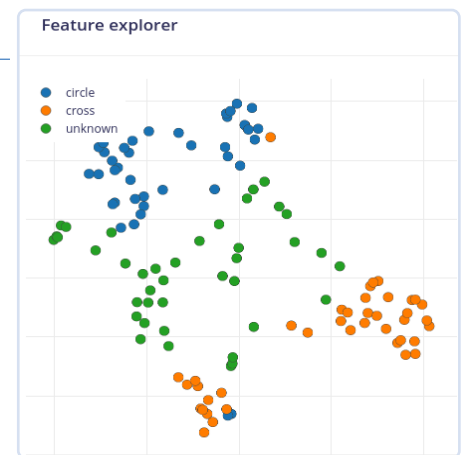


Figure 12. L'explorateur de caractéristiques montre une distinction prometteuse entre les trois classes.

Déplacez la couche entre les deux couches denses. Augmentez le nombre de cycles d'entraînement à 60, puis cliquez sur **Save & train**.

Dans notre cas, tous les cercles et toutes les croix ont été correctement identifiés, mais 14 % des mouvements inconnus ont été mal classés en tant que cercles (**figure 13**). Expérimentez avec différents taux d'abandon ou d'autres paramètres, entraînez à nouveau le modèle et observez l'impact sur la précision. Notez qu'Edge Impulse affiche par défaut la version quantifiée (optimisée) du modèle. Cliquez dessus pour passer à la version *non optimisée (float32)* et vérifiez si la précision est meilleure. Dans notre cas, la précision était moindre, avec 18 % des cercles incorrectement reconnus comme des croix.

Test du modèle

Naviguez jusqu'à **Live classification**. Réglez la longueur de l'échantillon à 3 000, cliquez sur **Start sampling**, et faites un geste avec votre carte. Le modèle traitera les données en temps réel et affichera ses prédictions, indiquant l'horodatage ainsi que la probabilité associée à chaque geste détecté.

Edge Impulse a formé le modèle en se basant uniquement sur les données d'apprentissage, vous devez donc encore le tester sur l'ensemble de données de test. Allez dans **Model testing** et cliquez sur **Classify all**. Cela exécutera le modèle sur tous les échantillons des données de test et comparera la sortie aux étiquettes attribuées.

Si la précision de ce test de modèle est significativement inférieure à celle obtenue sur les données d'apprentissage, cela indique un surajustement potentiel. Vous pouvez essayer de modifier certains paramètres dans l'étape d'apprentissage du modèle ou rassembler de nouveaux échantillons pour y remédier.

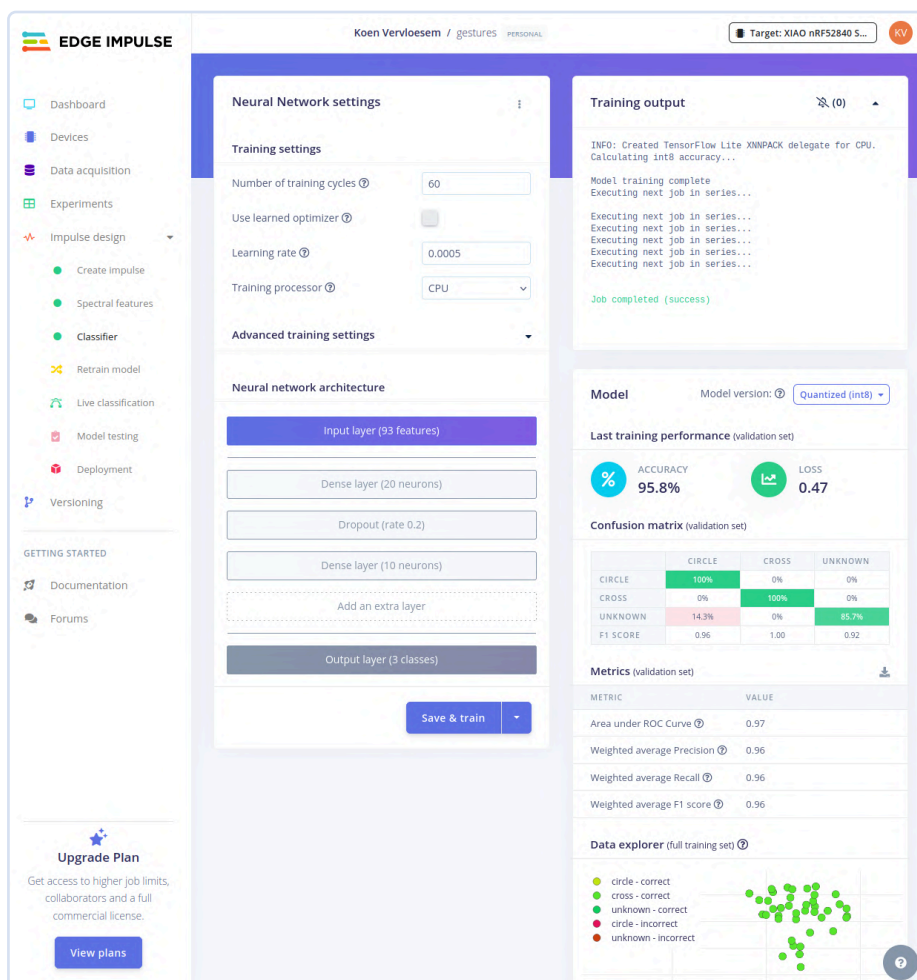


Figure 13. Le modèle entraîné atteint une précision de 95,8 %.

Déploiement sur votre appareil

La dernière étape consiste à déployer votre modèle d'apprentissage automatique sur votre carte. Dans l'onglet *Deployment*, sélectionnez *Arduino library* en la recherchant dans le champ de texte. Choisissez la quantification, puis cliquez sur *Build* pour générer un fichier ZIP contenant la bibliothèque Arduino que vous pouvez installer sur votre ordinateur. Lancez l'EDI Arduino et allez dans *Sketch / Include Library / Add .ZIP Library...* et sélectionnez le fichier ZIP téléchargé. Dans mon exemple, avec un projet intitulé *gestures*, le fichier s'appelle *ei-gestures-arduino-1.0.1.zip*. Vous trouverez ensuite des exemples de code sous *File / Examples / gestures_inferencing*. Cependant, ces exemples sont conçus pour d'autres cartes que la XIAO nRF52840 Sense. Pour cette dernière, vous pouvez exécuter le croquis Arduino présenté dans le **listage 2**, basé sur les exemples d'Edge Impulse pour d'autres cartes.

Notez que le fichier d'en-tête `gestures_inferencing.h` au début fait référence à la bibliothèque que vous avez installée avec le modèle d'apprentissage automatique intégré. Dans la fonction `loop`, le code s'exécute en continu et gère l'acquisition des données depuis l'accéléromètre, stocke ces données dans une mémoire tampon, convertit la mémoire tampon brute en un signal traitable, puis exécute le classificateur sur le signal. Enfin, la fonction affiche les prédictions

```
Output Serial Monitor X
Message (Enter to send message to 'Seed XIAO BLE Sense - nRF52840' on '/dev/ttyACM0')
circle: 0.39302
cross: 0.39083
unknown: 0.21555
Classified as: circle
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.31808
cross: 0.57633
unknown: 0.10559
Classified as: cross
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.41435
cross: 0.50746
unknown: 0.07819
Classified as: cross
Sampling...
Predictions (DSP: 8 ms., Classification: 0 ms., Anomaly: 0 ms.):
circle: 0.20718
cross: 0.38553
unknown: 0.40729
Classified as: unknown
Sampling...
```

Figure 14. Détection en direct des gestes sur le capteur Seeed XIAO nRF52840.

pour chaque classe, ainsi que la classe ayant la probabilité la plus élevée. Compilez et téléchargez ce sketch sur votre carte via l'IDE Arduino, puis réalisez des gestes en tenant la carte pour observer les prédictions de gestes en temps réel dans la sortie du moniteur série (figure 14).

Pour aller plus loin

Utiliser Edge Impulse pour développer un modèle d'apprentissage automatique qui détecte des gestes via des données d'accéléromètre est relativement simple. Ce modèle peut être déployé sous forme de bibliothèque Arduino compatible avec plusieurs cartes de développement. Vous pouvez également personnaliser le sketch de base sur un Seeed XIAO nRF52840 Sense pour réaliser différentes actions en fonction des gestes reconnus. Par exemple, vous pouvez configurer

la carte pour qu'elle fonctionne comme un clavier USB, envoyant des commandes à votre ordinateur pour gérer des applications telles qu'un lecteur multimédia. [K](#)

240449-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (koen@vervolessem.eu), ou contactez Elektor (redaction@elektor.fr).



Produits

> **Raspberry Pi AI Kit**
www.elektor.fr/20879

LIENS

- [1] Antonio Aloisio, « télécommande IR universelle basée sur l'IA », édition spéciale IA 2024 d'Elektor : <https://www.elektormagazine.fr/240433-04>
- [2] EDI Arduino : <https://www.arduino.cc/en/software>
- [3] Edge Impulse : <https://edgeimpulse.com>
- [4] Téléchargements : <https://www.elektormagazine.fr/240449-04>
- [5] Edge Impulse plans : <https://edgeimpulse.com/pricing>
- [6] Edge Impulse CLI : <https://docs.edgeimpulse.com/docs/tools/edge-impulse-cli>



Listage 2. Inférence des gestes à partir des données de l'accéléromètre LSM6DS3.

```
#include <gestures_inferencing.h>
#include <LSM6DS3.h>
#include <Wire.h>

#define CONVERT_G_TO_MS2 9.80665f
```

```

#define MAX_ACCEPTED_RANGE 2.0f

static bool debug_nn = false; // Set this to true to see
                                // e.g. features generated from the raw signal
LSM6DS3 myIMU(I2C_MODE, 0x6A);

void setup()
{
    Serial.begin(115200);
    Serial.println("Edge Impulse Inferencing Demo");

    if (!myIMU.begin()) {
        ei_printf("Failed to initialize IMU!\r\n");
    }
    else {
        ei_printf("IMU initialized\r\n");
    }

    if (EI_CLASSIFIER_RAW_SAMPLES_PER_FRAME != 3) {
        ei_printf("ERR: EI_CLASSIFIER_RAW_SAMPLES_PER_FRAME
                    should be equal to 3 (the 3 sensor axes)\n");
        return;
    }
}

/**
 * @brief Return the sign of the number
 *
 * @param number
 * @return int 1 if positive (or 0) -1 if negative
 */
float ei_get_sign(float number) {
    return (number >= 0.0) ? 1.0 : -1.0;
}

/**
 * @brief      Get data and run inferencing
 *
 * @param[in]  debug  Get debug info if true
 */
void loop()
{
    ei_printf("Sampling...\n");

    // Allocate a buffer here for the values we'll read from the IMU
    float buffer[EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE] = { 0 };

    for (size_t ix = 0; ix < EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE; ix += 3) {
        // Determine the next tick (and then sleep later)
        uint64_t next_tick = micros() + (EI_CLASSIFIER_INTERVAL_MS * 1000);

        buffer[ix] = myIMU.readFloatAccelX();
        buffer[ix+1] = myIMU.readFloatAccelY();
        buffer[ix+2] = myIMU.readFloatAccelZ();
    }
}

```



```

    for (int i = 0; i < 3; i++) {
        if (fabs(buffer[ix + i]) > MAX_ACCEPTED_RANGE) {
            buffer[ix + i] = ei_get_sign(buffer[ix + i]) * MAX_ACCEPTED_RANGE;
        }
    }

    buffer[ix + 0] *= CONVERT_G_TO_MS2;
    buffer[ix + 1] *= CONVERT_G_TO_MS2;
    buffer[ix + 2] *= CONVERT_G_TO_MS2;

    delayMicroseconds(next_tick - micros());
}

// Turn the raw buffer in a signal which we can the classify
signal_t signal;
int err = numpy::signal_from_buffer(buffer,
                                     EI_CLASSIFIER_DSP_INPUT_FRAME_SIZE, &signal);
if (err != 0) {
    ei_printf("Failed to create signal from buffer (%d)\n", err);
    return;
}

// Run the classifier
ei_impulse_result_t result = { 0 };

err = run_classifier(&signal, &result, debug_nn);
if (err != EI_IMPULSE_OK) {
    ei_printf("ERR: Failed to run classifier (%d)\n", err);
    return;
}

// print the predictions
ei_printf("Predictions ");
ei_printf("(DSP: %d ms., Classification: %d ms., Anomaly: %d ms.)",
           result.timing.dsp, result.timing.classification, result.timing.anomaly);
ei_printf(": \n");
int max_value_index = 0;
for (size_t ix = 0; ix < EI_CLASSIFIER_LABEL_COUNT; ix++) {
    ei_printf("    %s: %.5f\n", result.classification[ix].label,
              result.classification[ix].value);
    if (result.classification[ix].value >=
        result.classification[max_value_index].value) {
        max_value_index = ix;
    }
}
#if EI_CLASSIFIER_HAS_ANOMALY == 1
    ei_printf("    anomaly score: %.3f\n", result.anomaly);
#endif

    ei_printf("Classified as: %s\n",
              result.classification[max_value_index].label);
}

```

créez un neurone d'intégration- et-tir avec fuite



Intelligence artificielle sans logiciel

Clemens Valens (Elektor)

Les réseaux neuronaux artificiels constituent la base de l'intelligence artificielle. Ils sont généralement implémentés dans des logiciels, mais leur efficacité énergétique est relativement faible. Les réseaux neuronaux biologiques sont beaucoup plus performants. Dans cet article, nous présentons un circuit qui simule un neurone biologique, vous permettant de construire et d'expérimenter votre propre réseau neuronal.

Dessiner une moustache sur une affiche dans le métro est rarement perçu comme un acte d'intelligence. En revanche, lorsqu'un filtre Instagram ajoute des oreilles de chien à votre image cela est souvent perçu comme une prouesse technologique. Cela souligne combien la perception de l'intelligence, qu'elle soit artificielle ou non, est empreinte de subjectivité.

D'autre part, ce qui n'est pas subjectif, c'est l'énorme quantité d'énergie requise pour accomplir des tâches jugées intelligentes. Des millions, voire des milliards de transistors, sont sollicités pour créer des oreilles de chien virtuelles, consommant ainsi une quantité considérable d'énergie. En comparaison, chaque être vivant, qu'il s'agisse d'une huître ou même de brins d'herbe, réalise des opérations bien plus complexes que les systèmes d'IA les plus avancés, et ce, avec une efficacité énergétique bien supérieure.

Pour les animaux, quelques neurones intelligemment interconnectés et un peu de nourriture suffisent à accomplir des merveilles.

Réseaux neuronaux

Les neurones sont généralement considérés comme les éléments fondamentaux du cerveau, et donc de l'intelligence. Lorsqu'ils sont interconnectés, ils forment un réseau neuronal (**figure 1**). Les réseaux neuronaux artificiels (RNA) ont été efficacement appliqués à la reconnaissance vocale, à l'analyse d'images et au contrôle adaptatif dans le cadre de l'apprentissage automatique et de l'intelligence artificielle. [1]

Au sein d'un réseau neuronal, qu'il soit artificiel ou biologique, un neurone reçoit des signaux de neurones connectés. Il traite ces signaux, et s'il y a un résultat, il le transmet à d'autres neurones. Chaque signal d'entrée est pondéré. Les poids sont ajustables, permettant ainsi au neurone de s'adapter et au réseau de d'apprendre. Le résultat (instantané) est déterminé par une fonction non linéaire appliquée à la somme des entrées.

En général, les neurones artificiels fonctionnent avec des données numériques car ils sont principalement implémentés dans des logiciels. Les neurones biologiques utilisent des signaux électriques et chimiques. L'avantage des neurones logiciels est leur facilité de production en grande quantité et la simplicité avec laquelle un réseau peut être modifié. Cependant, il existe une tendance croissante à implémenter des neurones directement sur du matériel, en utilisant des

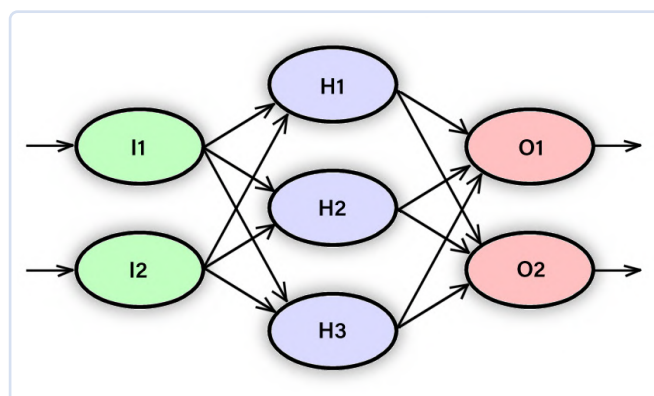


Figure 1. Un réseau neuronal artificiel simple avec deux neurones d'entrée (I1 & I2), une couche cachée de trois neurones (H1-3), et deux neurones de sortie (O1 & O2). Une drosophile possède environ 100 000 neurones.

technologies telles que la logique programmable des FPGA. Ces neurones sont de petits circuits construits avec des transistors. Ces neurones artificiels tentent d'imiter les signaux électriques utilisés par les neurones biologiques. Le neurone présenté dans cet article est un tel circuit.

Les neurones biologiques ne génèrent pas des niveaux de signal continus, mais produisent plutôt des impulsions courtes (**figure 2**). La fréquence et la durée des impulsions sont variables, et il est supposé que l'information est véhiculée dans le temps séparant deux impulsions et dans la largeur de l'impulsion. Ce type de communication pourrait être assimilé à une modulation de position d'impulsion (PPM). Les neurones qui produisent de telles impulsions sont appelés neurones à impulsions (SN) ; un réseau composé de tels neurones est appelé réseau neuronal à impulsions (SNN).

Neurone à impulsions

Un neurone à impulsions peut être assimilé à un pluviomètre à auget basculant qui mesure les précipitations (**figure 3**). Les impulsions entrantes (les gouttes de pluie) remplissent le seau, et lorsque le seau est plein, il bascule et se vide, générant ainsi une impulsion. À

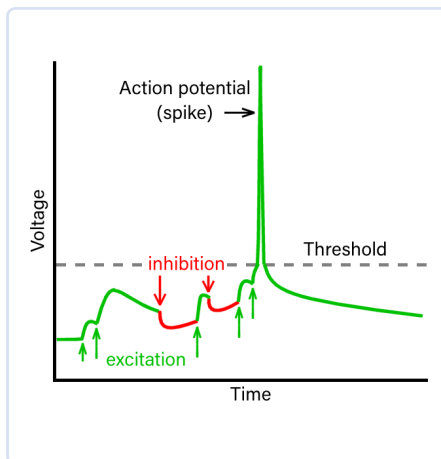


Figure 2. Les signaux d'un neurone à pointes (source : Alan Woodruff / QBI).

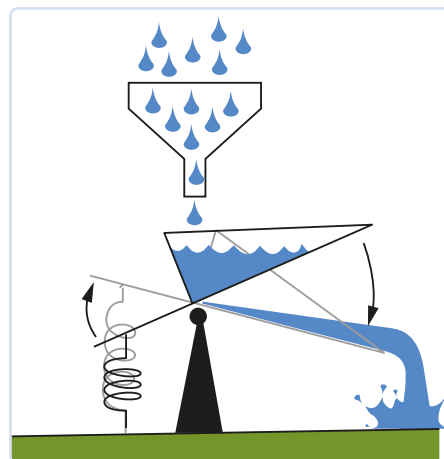


Figure 3. Un pluviomètre à auget basculant.

l'instar du pluviomètre, le SN ajoute des poids aux gouttes de pluie entrantes et peut contrôler la largeur de l'impulsion émise. En outre, le SN permet au seau de fuir, ce qui ajoute une dimension temporelle. Si les impulsions d'entrée arrivent trop lentement, le seau ne se remplira jamais et le SN ne se déclenchera jamais. En d'autres termes, le SN a la capacité d'oublier les stimuli précédents.

L'équivalent électronique du pluviomètre à auget basculant est l'oscillateur de relaxa-

tion. Dans un tel circuit, un condensateur est chargé et lorsqu'un certain seuil est atteint, un transistor ou un autre interrupteur contrôlé électroniquement décharge le condensateur, ce qui produit une impulsion. Le transistor unijonction (programmable) (UJT ou PUT) serait parfait pour construire un SN. Cependant, ces dispositifs sont aujourd'hui difficiles à obtenir à des prix raisonnables. Heureusement, il est possible de fabriquer un PUT en utilisant deux transistors bipolaires, une solution. Il est

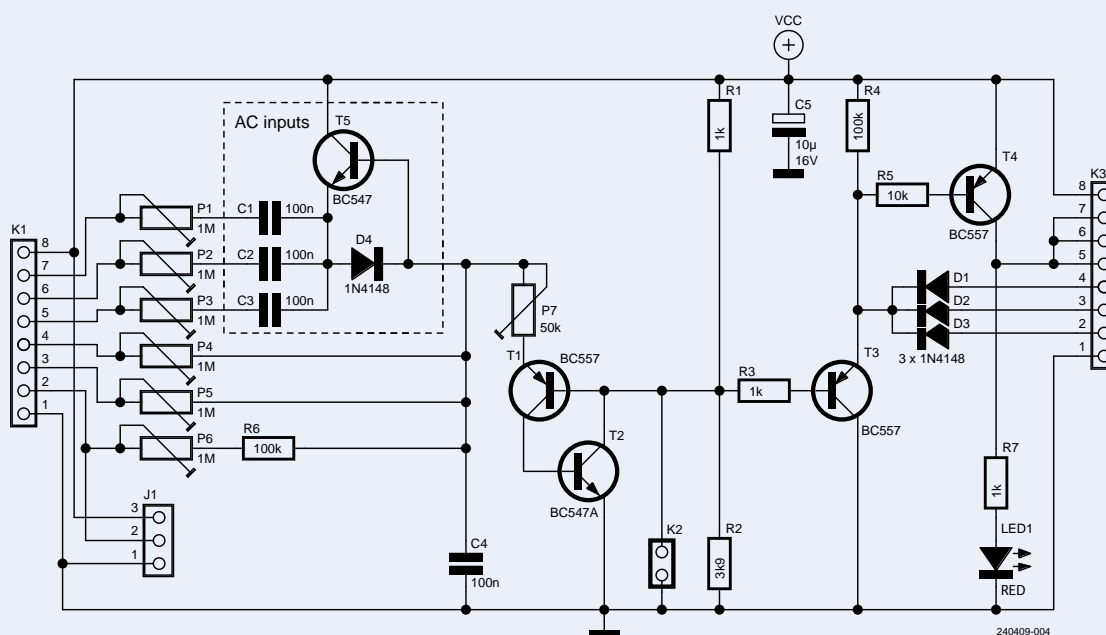


Figure 4. Ce neurone leaky integrate-and-fire spiking est très flexible et configurable. Pas aussi puissant que le Loihi d'Intel ou le TrueNorth d'IBM, il constitue un point de départ amusant.

à noter que pour implémenter un SN(N) dans un FPGA (ou dans du silicium, comme le Loihi d'Intel ou le TrueNorth d'IBM), les MOSFET restent les commutateurs de choix.

Circuit

Le circuit de notre neurone à impulsion est présenté dans la **figure 4**. Il est un peu plus complexe qu'un simple oscillateur à relaxation, mais cela est dû aux possibilités de connexion du neurone.

Le condensateur C4 joue le rôle du seau, tandis que les transistors T1 et T2 forment le commutateur qui permet de vider ce seau. Le niveau qui fait déclencher le commutateur est déterminée par les résistances R1 et R2. Le seau est rempli par les impulsions arrivant via D4, qui combine les signaux appliqués aux trois entrées supérieures des neurones (stimuli) (K1, broches 7, 6 et 5).

C1/C2/C3, D4, T5 et C4 forment une pompe à diodes avec trois entrées (C1, C2 et C3). Chaque impulsion entrante charge l'un des condensateurs C1, C2 ou C3. D4 transfère cette charge à C4, ce qui a pour effet d'augmenter la tension ainsi sa tension. T5, un émetteur suiveur, copie cette tension sur l'anode de D4, ce qui permet d'additionner les impulsions. Par conséquent, la tension sur C4 augmente à chaque impulsion. Sur un oscilloscope, la tension sur C4 ressemble à un escalier.

Lorsque la tension sur C4 devient supérieure à la tension de base de T1 (la tension réglée avec R1 et R2 plus sa chute de diode B-E), T1 commence à conduire. Un courant commence à circuler de C4 à travers P7 vers l'émetteur de T1 et de son collecteur dans la base de T2, jusqu'à la masse. Cela fait en sorte que T2 conduit également, faisant chuter la tension de base de T1 en dessous de celle de son émetteur. T1 est maintenant passant et C4 peut se décharger rapidement. Le seau bascule ainsi. Lorsque le courant de décharge de C4 devient suffisamment faible, T2 cesse de conduire et la base de T1 remonte au niveau de tension défini par R1 et R2. Comme la tension sur C4 est presque nulle, la tension de base de T1 est maintenant beaucoup plus élevée que sa tension d'émetteur et T1 se bloque. Le circuit revient à son état initial, prêt pour un nouveau cycle de remplissage.

Cette configuration particulière de T1 et T2 est parfois appelée thyristor transistorisé, populaire dans les années 1970. Il est crucial

que T2 n'ait pas une valeur bêta (h_{FE}) trop élevée, pour éviter une conduction continue même quand le courant de décharge est presque nul. Un transistor de type A ou B est idéal pour cette application, mais un transistor de type C pourrait être trop sensible. Le BC547A est recommandé pour T2, tandis que pour T5, un autre BC547, la valeur de bêta est moins critique, ce qui autorise l'utilisation d'un BC547A, simplifiant ainsi la nomenclature.

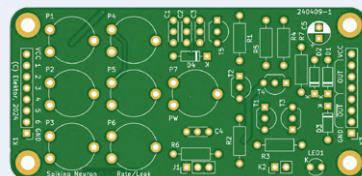
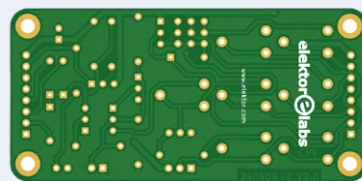
Entrées et sorties

Lorsque le condensateur C4 est en cours de charge, le signal au niveau du nœud formé par les résistances R1 et R2 est élevé ; lorsqu'il se décharge, ce signal devient bas. Le potentiomètre P7 permet de contrôler la durée de cette décharge. Une valeur basse sur P7 résulte en une impulsion active-basse courte sur R1/R2. Le point de test K2 permet de mesurer cette condition.

Le signal présent sur R1/R2 est tamponné par le transistor PNP T3. Les diodes D1, D2, et D3 génèrent trois sorties actives basses pour le neurone. Lorsque T3 est bloqué, ces sorties sont relevées par la résistance R4, et tirées vers le bas quand T3 est conducteur. De même, le blocage de T3 entraîne celui de T4, désactivant ainsi les sorties actives hautes du neurone (K3, aux broches 5, 6, et 7). Lorsque T3 conduit, T4 conduit également, activant ces sorties. La LED1 s'illumine pour signaler une impulsion, indiquant que les sorties sont à basse impédance lorsqu'une impulsion est émise, et à haute impédance dans le cas contraire. Cela évite que deux neurones



Liste des composants



Résistances

R1, R3, R7 = 1 kΩ
R2 = 3,9 kΩ
R5 = 10 kΩ
R4, R6 = 100 kΩ

Condensateurs

C1, C2, C3, C4 = 100 nF, pas de 2,5 mm
ou de 5 mm.
5 mm
C5 = 10 μF, 16 V, pas de 2,5 mm

Semi-conducteurs

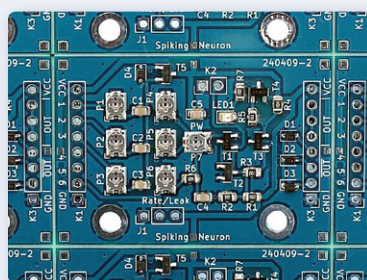
D1, D2, D3, D4 = 1N4148
T1, T3, T4 = BC557
T2, T5 = BC547A

Divers

JP1 = connecteur à 3 voies, pas de 2,54 mm
K1 = socket à 8 broches, pas de 2,54 mm
K2 = connecteur à 2 voies, pas de 2,54 mm
K3 = connecteur à 8 broches,
au pas de 2,54 mm

Essais avec des CMS

Pour optimiser l'expérimentation avec un réseau neuronal, j'ai conçu un modèle réduit de neurone en utilisant des composants montés en surface (CMS).



J'ai demandé au fabricant de circuits imprimés de produire un panneau 3 x 3 de ces neurones, avec des rainures en V (voir l'image ci-dessous). En incluant des plots de distribution de l'alimentation autour de la carte, ils peuvent tous être connectés facilement à l'alimentation avec des ponts de fils ou des résistances de 0 Ω. En utilisant du fil fin, il est rapide de câbler un réseau comme le montre la figure 1, et même de l'étendre avec un neurone d'entrée et de sortie supplémentaire (comme je l'ai fait). Pour faciliter les choses, il est recommandé de procéder à un câblage systématique afin de mémoriser aisément les connexions des sorties aux entrées.

s'influencent mutuellement lorsqu'ils pilotent le même neurone.

Les deux types de sorties permettent au neurone soit de stimuler soit d'inhiber un autre neurone. Cette fonctionnalité est plus claire lorsqu'on examine les entrées du neurone. Il existe également deux types d'entrées. Les trois entrées supérieures de type « AC » se connectent à la pompe à diodes déjà mentionnée. Les trois entrées inférieures de type « DC » se connectent chacune à C4 via un potentiomètre de poids, permettant de contrôler directement la charge du condensateur par chargement ou déchargement.

L'entrée inférieure (K1, broche 2) est équipée d'un cavalier J1 pour simplifier les connexions standard. Lorsque les broches 1 et 2 de J1 sont court-circuitées, le potentiomètre P6 offre une fuite réglable. En court-circuitant les broches 2 et 3, P6 charge C4, provoquant l'oscillation du neurone. La résistance R6 assure que cette oscillation est possible sur toute la gamme de P6.

Pour transformer les entrées AC en DC, il faut remplacer C1, C2, et C3 par une diode (avec l'anode connectée au potentiomètre), substituer D4 par un pont de fils, et ne pas monter T5. L'installation des trois diodes est nécessaire pour éviter que les entrées des différents neurones s'interfèrent mutuellement. Remplacer C1, C2, et C3 par un pont de fils créerait davantage d'entrées de fuite.

Le circuit n'est pas critique en termes d'alimentation et peut fonctionner avec des tensions variant de 3,5 V à 15 V ou plus, selon l'adaptation de C5. La consommation de courant est principalement déterminée par la LED et reste inférieure à 1 mA pour une alimentation de 5.

Caractéristiques

En résumé, nous pouvons maintenant dresser la liste suivante des caractéristiques de notre neurone à intégration-et-tir avec fuite :

- 3 entrées stimulus/AC avec des poids individuels
- 3 entrées d'inhibition/fuite/DC avec des poids individuels
- 3 sorties de stimulus (actif haut)
- 3 sorties d'inhibition/fuite (actif bas)
- Largeur d'impulsion de sortie réglable (« intensité »)
- Générateur d'impulsions à fonctionnement libre

Tester le réseau

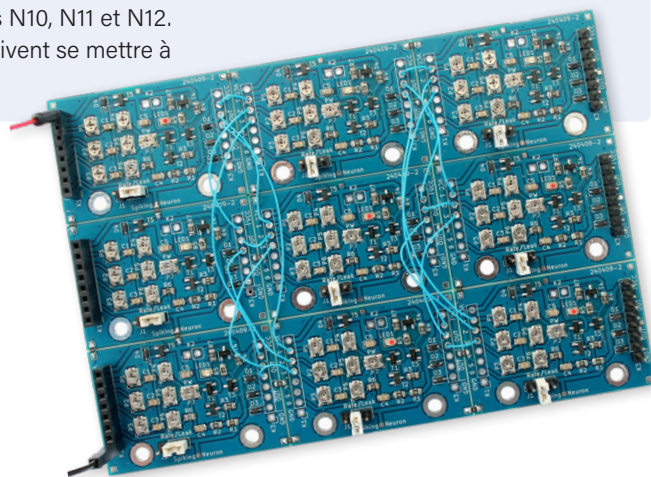
Voici une bonne méthode pour vérifier si le réseau fonctionne comme il se doit. Tournez les potentiomètres P1 à P6 de chaque nœud à fond dans le sens inverse des aiguilles d'une montre, c'est-à-dire réglez-les à leur valeur maximale. Mettez P7 de chaque nœud en position médiane. Sur le neurone situé dans le coin supérieur gauche du panneau, appelons-le N00 (Nrow,column), placez un cavalier sur les broches 2 et 3 de J1. Cela permet de connecter P6 à VCC. Alimenter le panneau. La LED de N00 commence à clignoter. Le taux peut être réglé avec P6 et l'intensité avec P7. Ajustez ensuite les potentiomètres de poids correspondants (P1 pour moi) sur les neurones connectés N10, N11 et N12. Leurs LED doivent commencer à clignoter. Plus la valeur du potentiomètre de poids est petite, plus la LED clignote rapidement.

Retirez le cavalier de N00 et placez-le sur N01, le neurone situé en dessous. Comme précédemment, la LED commence à clignoter, et vous devriez pouvoir ajuster la vitesse de clignotement (P6) et l'intensité (P7) de la LED. Comme précédemment, ajustez le potentiomètre de poids (P2 dans mon cas) des neurones connectés N10, N11 et N12. A nouveau, leurs LED doivent se mettre à clignoter.

Répétez cette procédure pour les autres neurones. Évidemment, les neurones de sortie n'étant pas connectés à d'autres, il n'est pas nécessaire de régler leurs poids. Si toutes les LEDs peuvent être contrôlées comme décrit ci-dessus, cela signifie que la partie stimulus du réseau est opérationnelle. Pour tester la partie fuite/inhibition de chaque nœud, oscillez les neurones d'entrée en plaçant un cavalier sur les broches 2 et 3 de J1. Puis, créez un court-circuit entre les broches 1 et 2 de JP1 sur un autre nœud, ce qui devrait modifier significativement le clignotement de sa LED.

Il ne vous reste plus qu'à entraîner le réseau pour une tâche spécifique. Je laisse cet exercice au lecteur, car je n'en ai pas encore exploré les possibilités.

Vous pouvez observer mon réseau neuronal 3x3 en action. Il semble que les diodes clignent au hasard, mais bien sûr, ce n'est pas le cas. Le réseau exécute une tâche complexe que je ne saisis pas encore totalement.



- LED pour visualiser le feu des neurones
- Tension d'alimentation : 3,5 V à 15 V

Créer un réseau de neurones à impulsions

Vous aurez probablement besoin d'au moins sept neurones pour un petit réseau neuronal capable d'accomplir une tâche très simple (Figure 1). Dans un tel réseau, il y a deux

neurones d'entrée, deux neurones de sortie et trois neurones dans la couche dite cachée. La raison pour laquelle il en est ainsi et son fonctionnement sortent du cadre de cet article. Consultez Wikipedia et ChatGPT pour plus d'informations.

Pour configurer notre petit réseau, les neurones d'entrée sont connectés à des capteurs comme des photorésistances ou

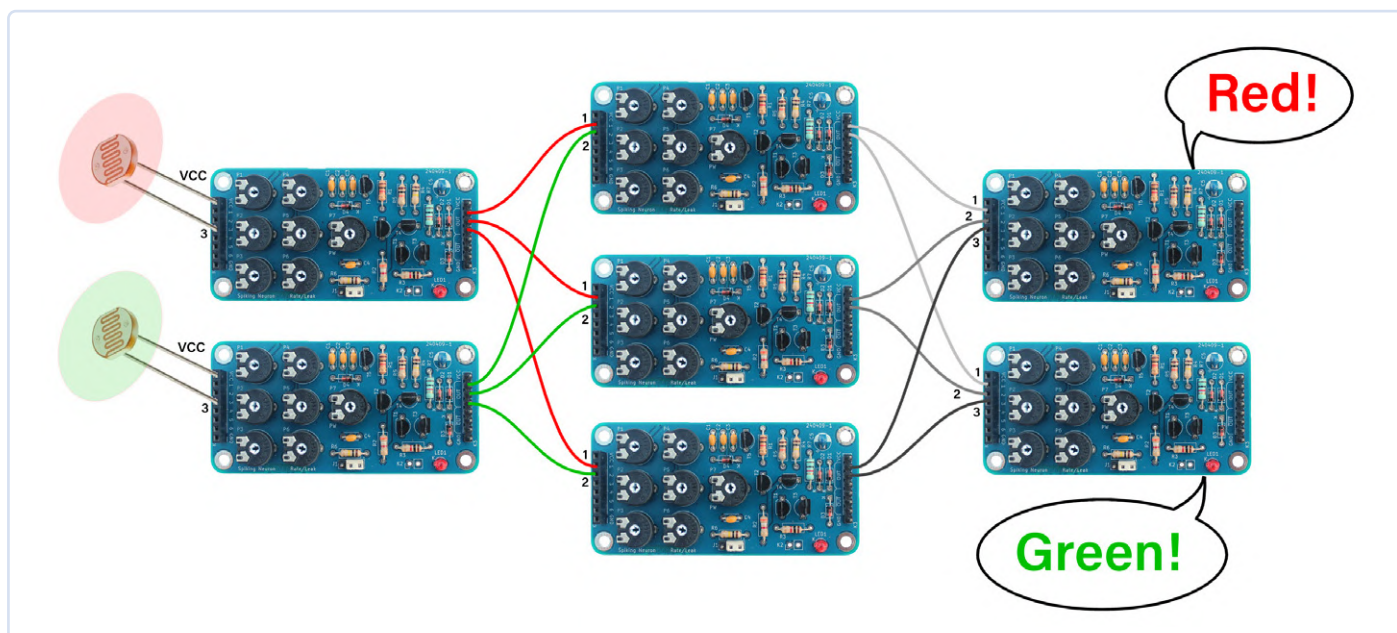


Figure 5. Lorsqu'il est configuré (entraîné) correctement, ce simple réseau neuronal peut faire la distinction entre la lumière rouge et la lumière verte et détecter par exemple l'état d'un feu de circulation. Très astucieux. Pensez-vous vraiment que vous pouvez accomplir une tâche aussi complexe avec moins de composants ? Notez que les connexions de l'alimentation électrique ont été omises pour plus de clarté.

des thermistances, ou à tout autre dispositif capable de convertir une grandeur physique en tension ou en courant. Ces capteurs, une fois connectés à une des entrées de courant continu du neurone, peuvent réguler le taux de stimulation du neurone. Les signaux de sortie des deux neurones d'entrée sont alors dirigés vers les trois neurones de la couche cachée, tandis que les sorties de cette dernière alimentent les neurones de sortie, comme montré dans la **figure 5**.

Il suffit maintenant de régler les potentiomètres de tous les neurones pour que le réseau génère le résultat souhaité. Certes, dans le réseau de la figure 5, cela implique de régler une vingtaine de potentiomètres et, honnêtement, je ne sais pas exactement comment procéder, mais je suis convaincu que c'est faisable. En quelque sorte. C'est ce qui rend ce genre de projet excitant : il laisse beaucoup de place à l'expérimentation.

En ce qui concerne l'expérimentation, les valeurs des résistances, potentiomètres et condensateurs à gauche du thyristor transistorisé T1/T2 (jusqu'à P7 inclus) peuvent être modifiées. N'hésitez pas à essayer d'autres valeurs. Je recommande toutefois de conserver des taux des impulsions relativement élevés, pour une meilleure visibilité lors de l'utilisation d'un oscilloscope. Les taux des impulsions dans les neurones biologiques sont plutôt faibles, souvent inférieurs à 100 Hz. Les fichiers de conception du neurone à impulsions (traversants et SMT) peuvent être téléchargés ici [1].

240409-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (clemens.valens@elektor.com), ou contactez Elektor (redaction@elektor.fr).



Produits

> Get Started with the MAX78000FTHR Bundle
www.elektor.fr/19798



FEATURED TOPIC

Visit our **Embedded & AI** page for articles, projects, news, and videos.

www.elektormagazine.com/
embedded-ai



LIENS

[1] S. Cording, "Understanding the Neurons in Neural Networks (Part 1): Artificial Neurons," [elektormagazine.com: https://www.elektormagazine.com/articles/neural-networks-part-1-artificial-neurons](https://www.elektormagazine.com/articles/neural-networks-part-1-artificial-neurons)

[2] This project at Elektor Labs: <https://www.elektormagazine.com/labs/spiking-neural-network-node>

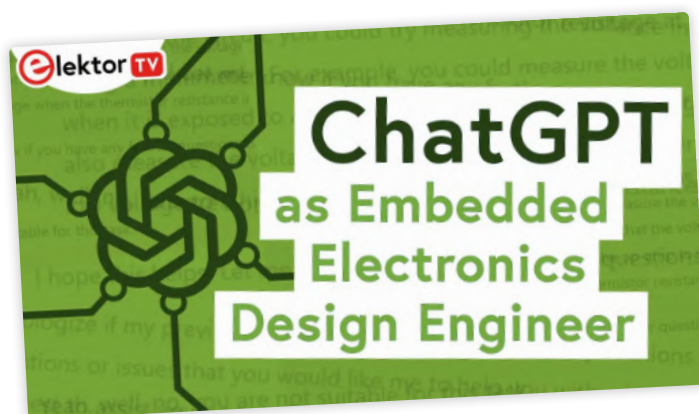
ChatGPT

pour la conception électronique

GPT-4o fait-il mieux ?

Clemens Valens (Elektor)

Au début de l'année 2023, j'ai voulu savoir si ChatGPT pouvait servir d'outil pour la conception électronique. J'ai interagi avec la version « Feb 13 », construite sur l'architecture GPT-3. Bien que ses compétences linguistiques m'aient profondément impressionné, ses aptitudes en ingénierie ne m'ont pas convaincu. Depuis, GPT-3 a évolué pour devenir GPT-4o. Est-il désormais à la hauteur ? Il est temps de le découvrir.



Watch Clemens's original video (https://youtu.be/pM1n_q7q670) from February 2023, when he tested ChatGPT as a tool for designing electronic circuits and embedded software.



Lors de ma première conversation avec ChatGPT, j'en ai rapidement conclu qu'il n'était pas encore prêt pour les tâches complexes que j'avais en tête. Vous pouvez en voir un exemple dans une vidéo sur Elektor TV [1]. Cette limite est compréhensible, car il faut souvent plusieurs années d'apprentissage aux humains pour maîtriser l'électronique. Cependant, un problème plus préoccupant est rapidement apparu : ChatGPT passait sans hésitation du sens à l'absurde, au point de mentir ouvertement. À partir de là, j'ai décidé de prendre tout ce qu'il disait avec beaucoup plus de précaution.

ChatGPT va-t-il remplacer Wikipedia ?

Au cours des derniers mois, j'ai continué à interagir avec ChatGPT [2] et j'ai constaté des améliorations significatives. J'ai même réalisé que j'avais progressivement arrêté d'utiliser Wikipédia, mon site web de référence depuis toujours. ChatGPT intègre probablement une grande partie des contenus de Wikipédia, mais sa capacité à les synthétiser rend les recherches bien plus rapides et efficaces.

Déboguons un circuit

L'un des exercices que j'ai tenté avec GPT-3 consistait à déboguer un circuit simple, illustré à la **figure 1**. Il s'agit d'un détecteur de circuit ouvert. La LED est censée rester éteinte lorsque J1 est court-circuité et s'allumer lorsque J1 est ouvert. Cependant, une erreur s'est glissée dans le schéma : le transistor Q1 est spécifié comme étant PNP au lieu de NPN attendu dans ce cas, ce qui empêche le circuit de fonctionner correctement. En réalité, si vous reproduisez le circuit de la figure 1 sur une plaque à essai, vous constaterez qu'il fonctionne, mais à l'envers : la LED s'allume lorsque J1 est court-circuité et s'éteint lorsque J1 est ouvert. Ce comportement s'explique par le fait que J1 court-circuite la base et l'émetteur, transformant Q1 en une simple diode.

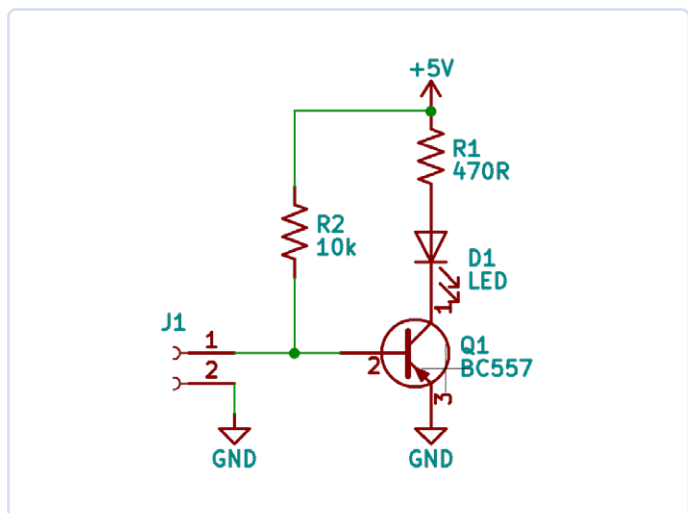


Figure 1. Un détecteur de circuit ouvert simple mais défectueux. ChatGPT peut-il trouver l'erreur ? Le pouvez-vous ?

En sollicitant ChatGPT pour identifier et résoudre ce problème, j'ai rencontré une difficulté majeure : lui transmettre correctement le schéma du circuit. À la question de savoir comment procéder, il m'a conseillé de charger une image du schéma sur GitHub, par exemple, puis de partager le lien. C'est ce que j'ai fait. Voici la réponse que ChatGPT m'a donnée :

Je vois, merci d'avoir partagé le schéma. J'ai examiné le schéma et cela ressemble à un simple circuit de détection en circuit ouvert. Voici quelques réflexions et questions initiales :

- > Le schéma indique une pile de 9 V comme source d'alimentation. Est-ce l'alimentation que vous utilisez pour le circuit ?
- > Le circuit utilise un ampli-op 741 comme comparateur. Y a-t-il une raison spécifique pour laquelle vous avez choisi cet ampli-op ?
- > Le circuit utilise une thermistance comme élément de détection pour identifier un circuit ouvert. Comment la thermistance est-elle utilisée dans le circuit et quel est le comportement attendu lorsqu'un circuit ouvert est détecté ?
- > Le schéma montre deux LED, une verte et une rouge. Comment ces LED sont-elles utilisées dans le circuit et quel est le comportement attendu lorsqu'un circuit ouvert est détecté ?

D'après le schéma, il est possible que le circuit ne fonctionne pas comme prévu si les valeurs ou les connexions des composants sont incorrectes. Pour faciliter le débogage, pourriez-vous fournir plus d'informations sur les problèmes ou les comportements inattendus que vous rencontrez avec le circuit ?

Il est évident que ChatGPT n'a pas réellement analysé mon circuit, mais s'est contenté de restituer une réponse issue de sa mémoire, évoquant un montage similaire : un détecteur de température haute/basse illustré à la **figure 2**. Malgré plusieurs tentatives, je n'ai pas réussi à lui faire comprendre mon schéma. J'ai donc fini

par abandonner. Si ChatGPT est incapable de lire les schémas, il ne peut ni les analyser ni les corriger.

Recherche de défauts avec GPT-4o

La version GPT-4o de ChatGPT a évolué de plusieurs façons par rapport à GPT-3 [3]. Non seulement il traite mieux le langage naturel et vous avertit de ses limites, mais son interface utilisateur s'est également améliorée. Vous pouvez maintenant télécharger des fichiers en plus de taper du texte, ce qui lui permet de travailler avec des images et d'autres documents. J'ai utilisé cette fonctionnalité avec succès, par exemple, pour traduire des scans de documents chinois en anglais. Cela montrait que ChatGPT analysait réellement les documents fournis. J'ai donc décidé d'essayer l'exercice de débogage de schéma précédent avec GPT-4o. Après avoir posé les mêmes questions préliminaires qu'avec GPT-3, j'ai téléchargé le fichier plutôt que de le diriger vers GitHub. Voici sa réponse (y compris le formatage). Notez que j'ai omis les conseils de débogage, car GPT-4o est bien plus verbeux que GPT-3.

Je vois que vous avez téléchargé un schéma intitulé « Open Circuit Detector ». D'après les composants listés, il semble s'agir d'un circuit simple impliquant une LED, un transistor (BC557), des résistances et un connecteur (J1).

Voici une brève analyse du circuit :

Composants :

1. **D1** : LED
2. **Q1** : Transistor PNP BC557
3. **R1** : Résistance de 470 Ω (probablement une résistance de limitation de courant pour la LED)
4. **R2** : Résistance de 10k Ω (probablement une résistance de base pour le transistor)
5. **J1** : Connecteur

Description du circuit :

- > La LED (D1) est connectée en série avec la résistance R1 et le collecteur du transistor PNP (Q1).

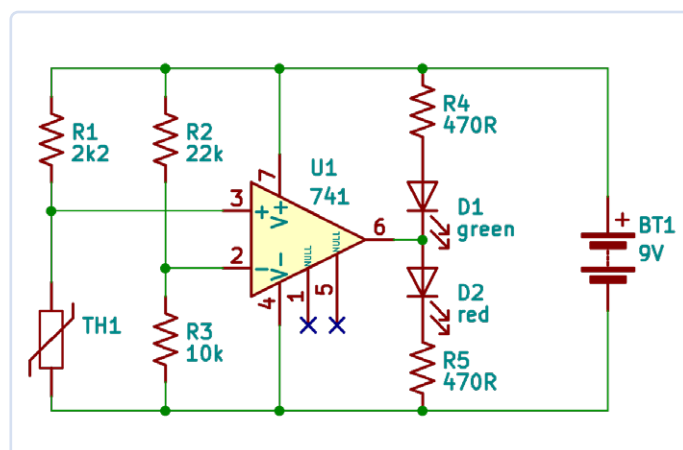


Figure 2. Est-ce que c'est ce que GPT-3 a cru voir ?

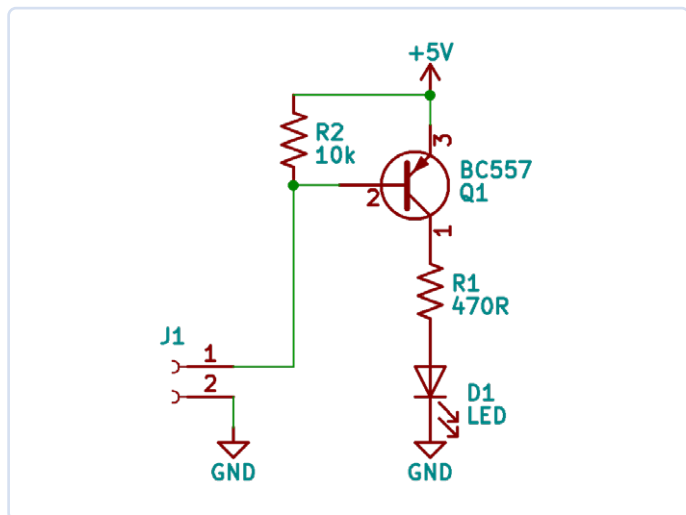


Figure 3. Apparemment, GPT-4o a interprété le circuit comme suit.

- L'émetteur de Q1 est connecté à l'alimentation +5V.
- La base de Q1 est connectée à la jonction de R2 et à l'une des broches du connecteur (J1).

➤

Fonctionnement attendu :

- Lorsque la base du transistor Q1 est tirée vers le bas (proche de la masse), Q1 devrait s'allumer, permettant au courant de circuler de l'émetteur au collecteur et d'allumer la LED.
- S'il y a un circuit ouvert ou une absence de connexion au niveau de J1, le transistor ne s'allumera pas et la LED restera éteinte.

[J'ai laissé de côté les conseils de dépannage]

Faites-moi savoir si vous avez besoin d'aide pour une partie spécifique du circuit ou si vous rencontrez un problème particulier !

La plupart des éléments ci-dessus sont corrects, à l'exception d'un détail : l'émetteur du transistor est connecté à +5 V au lieu d'être relié à la terre. Il semble qu'il ait interprété le circuit comme le montre la **figure 3**. La description par ChatGPT du fonctionnement attendu du circuit le confirme. Après avoir signalé son erreur, il a fourni l'analyse ci-dessous.

Merci de nous l'avoir signalée ! Corrigons l'analyse :

Description corrigée du circuit :

- Q1 (BC557) est un transistor PNP, et dans cette configuration corrigée :
 - Émetteur: Connecté à la **masse**.
 - Collecteur: Connecté à une extrémité de la LED à travers la résistance **R1** (470Ω), qui est ensuite connectée à l'**alimentation +5V**.
 - Base: Connectée à la jonction de **R2** (10kΩ) et de l'une des broches du connecteur **J1**.

Fonctionnement corrigé :

- Dans un transistor PNP comme le BC557, le courant circule de l'émetteur vers le collecteur lorsque la base est plus négative (tension plus faible) par rapport à l'émetteur.

- L'émetteur étant mis à la terre, le circuit allume la LED lorsque la base du transistor est à une tension supérieure à celle de l'émetteur, c'est-à-dire lorsqu'elle est tirée vers le bas par rapport à l'émetteur, le transistor s'éteint et la LED s'éteint.
- Comportement attendu :
- LED allumée : La LED doit s'allumer lorsque la tension à la base de Q1 est suffisamment élevée (proche de +5V), ce qui signifie que le transistor est conducteur.
- LED éteinte : Si la base de Q1 est ramenée à la masse (ou très proche de la masse), le transistor cesse de conduire et la LED s'éteint.

[...]

Si vous rencontrez toujours des problèmes, veuillez fournir plus de détails sur ce que fait le circuit par rapport à ce que vous attendez de lui, et je pourrai vous aider davantage !

Maintenant, la description du circuit correspond au schéma téléchargé. En revanche, l'« opération corrigée » est contradictoire. D'abord, il explique comment fonctionne un transistor PNP, puis il continue avec ce qui semble être un transistor NPN. Le paragraphe « Comportement attendu » le confirme. En informant ChatGPT de cette contradiction, il a répondu (en supprimant à nouveau les explications verbeuses) :

Vous avez raison, et je m'excuse pour la confusion dans mes explications précédentes. Clarifions le comportement correct pour un transistor PNP comme le BC557.

[...]

Si l'émetteur du BC557 est connecté à la masse, le transistor ne se comportera pas comme prévu pour allumer une LED, à moins que le circuit ne soit conçu d'une manière très spécifique. Normalement, l'émetteur d'un transistor PNP doit être connecté à une tension positive (par exemple, +5V), avec le collecteur connecté à la charge (LED et résistance), et la base doit être tirée vers le bas pour allumer le transistor.

[...]

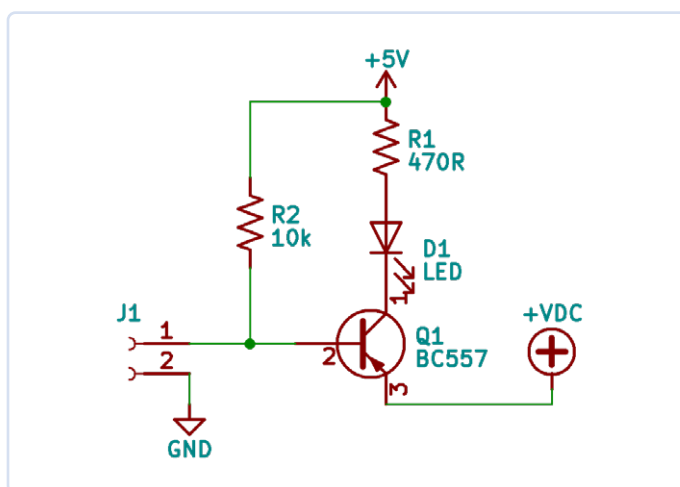


Figure 4. ChatGPT suggère de connecter l'émetteur à une tension positive, ce qui est dangereux pour Q1.

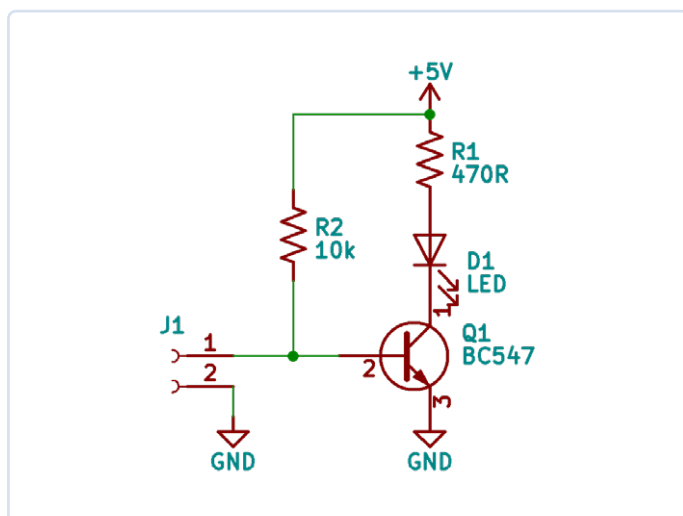


Figure 5. Le circuit que nous cherchions. ChatGPT ne l'a pas proposé.

Si vous corrigez la connexion de l'émetteur à une tension positive, la description originale de la façon dont le circuit devrait fonctionner (la LED s'allume lorsque la base est basse par rapport à l'émetteur) restera vraie.

Cette fois, ChatGPT signale une possible erreur dans le circuit. Cependant, la solution qu'il propose (**figure 4**) est incomplète, car la LED doit être inversée pour que cela fonctionne. De plus, la tension positive doit être d'au moins 5,6 V pour que le transistor Q1 commence à conduire. En outre, si J1 est court-circuité dans le circuit de la figure 4, le transistor risque de griller, car le courant de base n'est limité que par l'alimentation.

En suggérant à ChatGPT de remplacer Q1 par un transistor NPN (**figure 5**), il reprend :

En résumé, le remplacement du BC557 par un BC547 inversera la logique du circuit. Avec le BC547, la LED s'allume lorsque le socle est tiré vers le haut, et elle s'éteint lorsque le socle est bas ou ouvert.

Malheureusement, ChatGPT n'a toujours pas compris que J1 est censé être ouvert ou court-circuité. En le signalant, j'ai obtenu cette réponse (abrégée) :

Résumé de l'opération :

- > J1 court-circuité (circuit fermé) : La LED est **éteinte**, ce qui indique que le circuit est court-circuité.
- > J1 ouvert (circuit ouvert) : Le voyant **est allumé**, ce qui indique que le circuit est ouvert.
- > Mise en œuvre pratique :

- > Résistance de tirage: Si J1 est ouvert, le socle du BC547 peut flotter à moins qu'il n'y ait un pull-up assez fort vers le +5V. Vous pouvez utiliser le pull-up interne du microcontrôleur (si applicable) ou ajouter une résistance de pull-up externe pour s'assurer que le socle est tiré vers le haut lorsque J1 est ouvert.
- > Résistance de tirage vers le bas: Si vous voulez vous assurer que le socle est bien relié à la terre lorsque J1 est court-circuitée, vous pouvez ajouter une petite résistance d'excursion vers le bas au socle, mais ce n'est généralement pas nécessaire si J1 est une connexion solide à la terre.

Un microcontrôleur est apparu de nulle part et ChatGPT semble avoir oublié R2. Ceci a mis fin à mon expérience.

L'IA a encore un long chemin à parcourir

Les résultats ci-dessus nous permettent de conclure que GPT-4o peut effectivement traiter des images, mais son interprétation semble souvent influencée par ses connaissances préalables. Ce biais est en fait assez humain. Dans ce cas particulier, ChatGPT s'attendait clairement au circuit de la figure 3 plutôt qu'à celui de la figure 1. Toutefois, bien qu'il ait remarqué que Q1 était de type PNP, il a traité le circuit comme s'il s'agissait d'un transistor NPN. Un humain, lui, n'aurait jamais commis cette erreur.

Lorsqu'il est confronté à des données inattendues, ChatGPT produit parfois des résultats incohérents. Cela m'a rappelé l'ordinateur d'échecs de mon frère à la fin des années 80. Si vous jouiez de manière conventionnelle, il était difficile à battre. Mais dès que vous faisiez un mouvement inattendu, il perdait le fil et vous le battiez systématiquement. ChatGPT fonctionne de manière similaire. Il est impressionnant lorsqu'il s'agit de traiter le langage naturel, et c'est un excellent moteur de recherche et une interface pour Wikipédia. Cependant, en matière de résolution de problèmes complexes, il n'est pas plus performant qu'un ordinateur d'échecs des années 80. En somme, l'intelligence artificielle a encore un long chemin à parcourir. ◀

240410-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (clemens.valens@elektor.com) ou contactez Elektor (redaction@elektor.fr).



Produit

- > **Practical Electronics Crash Course (Bundle)**
www.elektor.fr/20824

LIENS

- [1] Video: ChatGPT as Embedded Electronics Design Engineer, Elektor TV: https://youtu.be/pM1n_q7q670?si=PxL3VePdo8ytAUyjd
- [2] OpenAI: <https://openai.com/>
- [3] ChatGPT 4: <https://openai.com/index/gpt-4/>

intégrer l'IA périphérique avec l'ESP32-P4

Anant Gupta, Sun Xiangyu, et Xie Wei (Espressif)

L'ESP32-P4 est le dernier-né de la famille des SoC haute performance d'Espressif, conçu pour révolutionner le secteur des systèmes embarqués. Cette puissante puce est dotée de fonctionnalités avancées, notamment de solides capacités d'intelligence artificielle, ce qui en fait un choix idéal pour les développeurs désireux de concevoir des appareils connectés et intelligents. Dans cet article, nous explorerons en détail les capacités d'IA de l'ESP32-P4, les bibliothèques d'IA qu'elle prend en charge et présenterons un exemple concret d'application qui illustre son immense potentiel.

L'ESP32-P4 d'Espressif est un microcontrôleur à double cœur basé sur un processeur RISC-V capable d'atteindre une vitesse d'horloge allant jusqu'à 400 MHz. Comme le montre le schéma fonctionnel de la **figure 1**, l'ESP32-P4 dispose d'un sous-système de mémoire très flexible et adaptable, doté de 768 Ko de SRAM intégrée, 8 Ko de RAM TCM à latence nulle, et une PSRAM extensible. La puce intègre également une large gamme de périphériques, notamment SPI, I2S, I2C, LED PWM, MCPWM, RMT, ADC, UART et TWAI. Il offre également des fonctionnalités avancées pour les interfaces homme-machine telles que MIPI-CSI avec ISP intégré, MIPI-DSI, 14 entrées tactiles

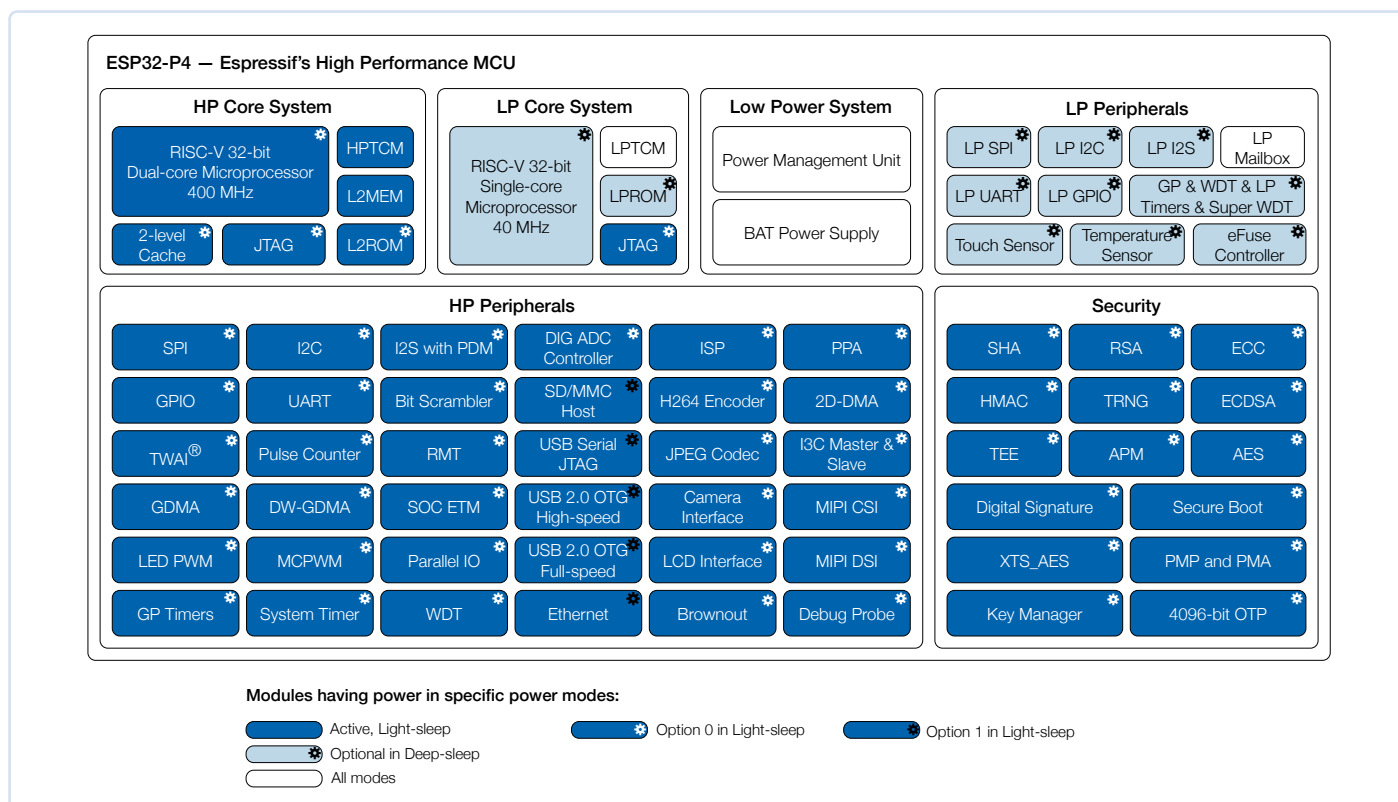


Figure 1. Schéma fonctionnel de l'ESP32-P4.

capacitives, etc. Il est équipé d'accélérateurs matériels notamment l'accélérateur de traitement de pixels (PPA), l'encodage H.264 et le 2D-DMA. Grâce à ses hautes performances et sa connectivité E/S complète, l'ESP32-P4 promet de transformer radicalement le domaine des systèmes embarqués.

Capacités IA de l'ESP32-P4

L'ESP32-P4 est conçue pour prendre en charge une vaste gamme d'applications d'IA, des modèles simples d'apprentissage automatique aux algorithmes complexes d'apprentissage profond. Les capacités d'IA de la puce reposent sur son processeur RISC-V à double cœur, qui constitue une base solide pour le traitement de l'IA. L'ESP32-P4 est équipée d'un ensemble d'accélérateurs matériels et d'instructions optimisées pour l'IA, qui facilitent les calculs complexes, ce qui en fait un choix particulièrement adapté aux applications exigeant un traitement d'IA de haute performance.

L'une des principales caractéristiques des capacités d'IA de l'ESP32-P4 est sa compatibilité avec l'extension *Xai* du jeu d'instructions RISC-V. L'extension *Xai* fournit une gamme d'instructions spécifiquement conçues pour les applications d'IA et d'apprentissage automatique, notamment les :

- **Instructions vectorielles** : elle permettent à la puce de traiter de grands vecteurs de données, une fonctionnalité essentielle pour de nombreux algorithmes d'IA et d'apprentissage automatique.
- **Instructions matricielles** : elles permettent à la puce d'effectuer des opérations sur de grandes matrices de données, ce qui est une exigence commune à de nombreux algorithmes d'apprentissage profond.

L'extension *Xai* offre également une gamme d'autres instructions spécifiquement conçues pour les applications d'IA et d'apprentissage automatique, notamment des instructions pour la manipulation et le déplacement de données vectorielles alignées et non alignées, l'arrondi configurable et les modes de saturation.

ESP-SR : cadre de reconnaissance vocale

ESP-SR est un environnement complet destiné aux développeurs souhaitant intégrer la reconnaissance vocale dans leurs applications grâce à l'IA. Ce cadre inclut divers modules, notamment le traitement audio frontal, la détection des mots de réveil et la reconnaissance des commandes vocales. Avec ESP-SR, les développeurs peuvent créer des applications capables de reconnaître et de répondre aux commandes vocales, ce qui en fait un choix idéal pour des applications telles que la domotique intelligente et les assistants vocaux (**figure 2**).

Le cadre ESP-SR offre plusieurs fonctionnalités clés :

- **Traitement audio frontal (algorithmes AFE)** : ce module offre un ensemble d'API pour le traitement audio, incluant la suppression du bruit profond, l'annulation d'écho et la séparation aveugle des sources.
- **Détection des mots de réveil (WakeNet)** : ce module permet aux développeurs de détecter les mots de réveil, tels que « Alexa » ou « OK Google », ou tout autre mot de réveil personnalisé, et de déclencher des commandes personnalisées.
- **Reconnaissance des commandes vocales (Mul)** : ce module

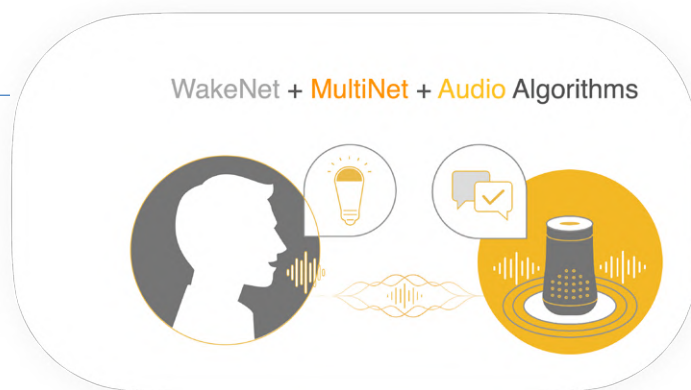


Figure 2. ESP-Speech Recognition.

propose un ensemble d'API pour la reconnaissance vocale, permettant aux développeurs de créer des applications capables de reconnaître et de répondre aux commandes vocales.

ESP-DL : bibliothèque d'apprentissage profond

ESP-DL est une bibliothèque d'apprentissage profond qui propose des API pour l'inférence de réseaux neuronaux, le traitement d'images et les opérations mathématiques (**figure 3**). Avec ESP-DL, les développeurs peuvent déployer leurs modèles d'apprentissage profond sur l'ESP32-P4, ouvrant ainsi la porte à un large éventail d'applications utilisant l'IA.

La bibliothèque ESP-DL offre de nombreuses fonctionnalités :

➤ Format de modèle standard ESP-DL

Ce format binaire utilisé pour stocker le graphique du modèle, les poids et d'autres informations. Ce format est similaire au format de modèle ONNX, tout en substituant les Protobuf par des FlatBuffers, ce qui rend nos modèles plus légers et permet une désérialisation directe. Donc, l'accès aux données sérialisées ne nécessite pas de les copier dans une partie distincte de la mémoire. L'accès aux données devient significativement plus rapide par rapport aux formats exigeant une manipulation plus complexe, tels que les Protobuf.

➤ Implémentation efficace et précise des opérateurs

Grâce à des directives d'IA, nous avons implémenté efficacement les opérateurs IA courants, tels que Conv2d, Pool2D, Gemm, Add, Mul, etc. Parallèlement, nous avons corrigé les erreurs de précision des opérateurs 8 bits de versions antérieures. Nous avons implémenté l'opérateur PyTorch afin de garantir que les résultats obtenus par notre outil de quantification soient cohérents avec les résultats obtenus sur ESP-DL.

➤ Planificateur de mémoire statique

Le planificateur de mémoire statique évalue la taille de mémoire maximale requise et les décalages de mémoire de chaque variable en fonction de la séquence topologique des opérateurs. Cela permet d'éviter la surcharge de temps et la fragmentation potentielle de la mémoire causée par l'allocation de la mémoire pendant l'exécution du modèle. Nous avons conçu un nouveau planificateur de mémoire statique pour la structure de mémoire RAM/PSRAM interne commune. Étant donné que la RAM interne a une vitesse d'accès plus rapide mais une capacité limitée, nous fournissons une API qui permet aux utilisateurs de personnaliser la taille de la RAM interne que le modèle peut utiliser. Le planificateur de mémoire répartit automatiquement les différentes couches dans les zones de mémoire optimales de la mémoire en fonction de la taille de la RAM interne spécifiée par l'utilisateur, maximisant ainsi l'efficacité de l'exécution tout en minimisant l'utilisation de la mémoire.

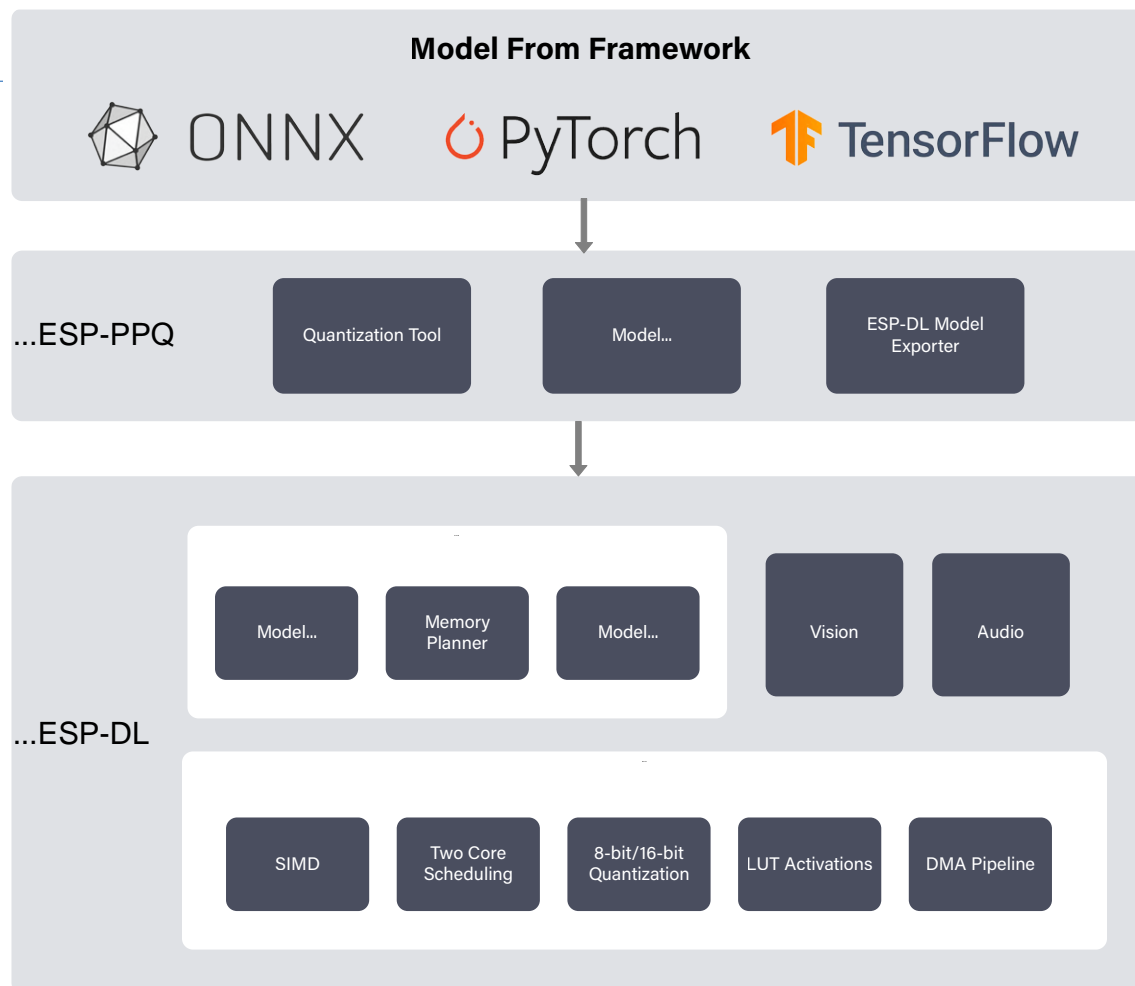


Figure 3. Framework ESP-DL.

► Planification double cœur

Nous avons ajouté une planification automatique à double cœur pour certains opérateurs à grande intensité de calcul, permettant ainsi aux modèles d'utiliser pleinement la puissance de calcul des architectures à double cœur. Actuellement, Conv2D et DepthwiseConv2D prennent en charge cette planification. Voici quelques-uns de nos résultats expérimentaux : pour les couches à forte intensité de calcul, l'utilisation d'un double cœur peut réduire le temps de moitié.

Cas d'utilisation : classification d'images

L'ESP32-P4, grâce à ses capacités avancées d'intelligence artificielle, se révèle être une plateforme idéale pour la classification d'images. Grâce à la bibliothèque ESP-DL, les développeurs peuvent implémenter des modèles d'apprentissage profond capables de détecter et de classer différents objets à partir d'images. Cette application présente un large éventail d'utilisations potentielles, des systèmes de sécurité à la domotique intelligente. Grâce à la haute performance de l'ESP32-P4, ses capacités d'IA et à sa connectivité E/S étendue, les développeurs peuvent créer un système de classification d'images à la fois précis et efficace.

Pour implémenter la classification d'images sur l'ESP32-P4, les développeurs peuvent utiliser la bibliothèque ESP-DL pour déployer des modèles d'apprentissage profond pré-entraînés comme le MobileNet V2. Le modèle est entraîné sur un ensemble de données d'images. Il est donc capable de détecter et de classer différentes images. Grâce au moteur PPA et 2D-DMA de l'ESP32-P4, il est possible d'accélérer significativement le traitement et la gestion du flux vidéo, permettant à la puce d'effectuer des traitements complexes d'IA en temps réel. Nous utiliserons la carte ESP32-P4-Function-EV-Board illustrée à

la **figure 4**. Cette carte de développement multimédia est basée sur la puce ESP32-P4. et est compatible avec USB2.0, MIPI-CSI, MIPI-DSI et plusieurs autres périphériques. Grâce à toutes ces caractéristiques exceptionnelles, cette carte constitue un choix idéal pour le développement de produits audio et vidéo connectés, à faible coût, à hautes performances et à faible consommation d'énergie.

Portage de MobileNet V2 sur ESP32-P4 avec ESP-DL

MobileNet V2 est un modèle d'apprentissage profond léger, adapté aux performances sur les appareils mobiles et embarqués. Il est largement utilisé dans divers domaines, tels que la classification d'images, la détection d'objets, la reconnaissance faciale, entre autres. Utilisons MobileNet V2 comme exemple pour comprendre comment déployer le modèle avec ESP-DL.

1. **Préparer un modèle pré-entraîné** : nous pouvons directement obtenir un modèle pré-entraîné via l'interface PyTorch et exporter le modèle vers un fichier ONNX pour le processus de quantification.

```
torchvision.models.mobilenet.  
mobilenet_v2(pretrained=True)  
torch.onnx.export(  
    model=model,  
    args=tuple(  
        [  
            torch.zeros(  
                size=[1] + input_shape[1:],  
                device=self.device_str,
```

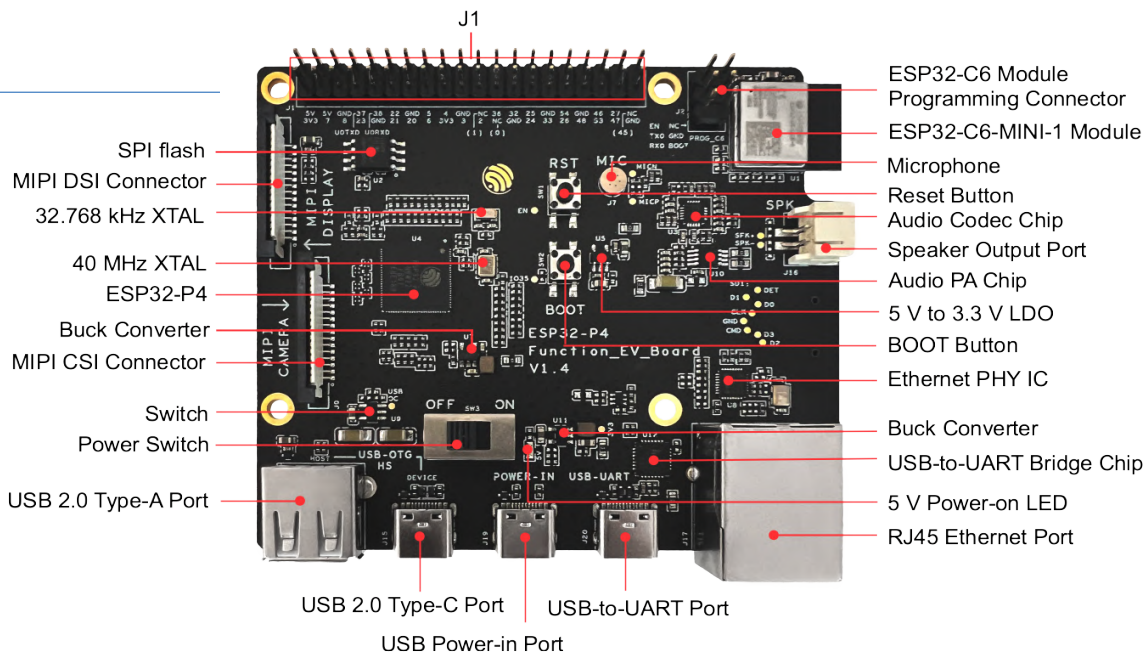


Figure 4. ESP32-P4-Function-EV-Board.

```

        dtype=self.input_dtype,
    )
    for input_shape in self.input_shape
    ]
),
f=orig_onnx_path,
opset_version=11,
do_constant_folding=True,
)

```

2. **Préparer un jeu de données de calibration et quantifier le modèle :** sur la base de l'outil de quantification ESP-PPQ, la fonction encapsulée `quantize_model_wrapper` permet de charger le fichier ONNX de l'étape 1 et d'effectuer la quantification à partir du jeu de données de calibration. Une fois la quantification terminée, exportez le modèle pour obtenir un fichier avec le suffixe *espdll*.

```

ppq_graph, executor = quantize_model_wrapper(
    onnx_import_file=orig_onnx_path,
    calib_data_loader=self.calib_iter,
    calib_steps=self.calib_steps,
    input_shape=self.input_shape,
    platform=self.platform,
    input_dtype=self.input_dtype,
    dispatching_override=None,
    dispatching_method="conservative",
    collate_fn=collate_fn,
    device=self.device_str,
    verbose=1,
)

```

```

PFL.Exporter(platform=self.platform).export(
    file_path=export_onnx_path,
    graph=ppq_graph,
    config_path=export_config_path,
    modelVersion=self.model_version,
    valuesForTest=valuesForTest,
)

```

3. **Déploiement du modèle :** chargez le fichier de modèle *espdll* dans

la mémoire flash. Lors du chargement avec ESP-DL, celui-ci alloue automatiquement la mémoire nécessaire aux processus intermédiaires. À ce stade, transmettez les données d'entrée à l'interface `Model::run`. Les résultats de l'inférence finale peuvent être récupérés grâce à la bibliothèque d'opérateurs d'accélération d'ESP-DL.

```

Model *model = new Model("model",
    fbs::MODEL_LOCATION_IN_FLASH_PARTITION);
std::map graph_test_inputs = get_graph_test_inputs(model);
model->run(graph_test_inputs);
std::map outputs = model->get_outputs();

```

L'avenir de l'IA embarquée avec l'ESP32-P4

L'ESP32-P4 se distingue comme un puissant SoC prêt à révolutionner le domaine des systèmes embarqués. Doté de capacités d'IA avancées, d'une connectivité E/S complète, et supportant des bibliothèques d'IA telles que ESP-SR et ESP-DL, ce SoC est parfaitement adapté aux développeurs désireux de concevoir des dispositifs intelligents et interconnectés. Le cas d'utilisation de la classification d'images avec MobileNet V2 illustre le potentiel de l'ESP32-P4 pour exploiter les capacités d'IA et comment un modèle pré-entraîné peut être efficacement porté via le framework ESP-DL, démontrant comment la puce peut être utilisée pour créer des systèmes précis et efficaces alimentés par l'IA. À mesure que le domaine de l'IA continue de se développer, l'ESP32-P4 est destiné à jouer un rôle essentiel dans l'évolution future des systèmes embarqués. ◀

240568-04

Questions ou commentaires ?

Contactez Elektor (redaction@elektor.fr).



produits

- **Elektor Special: Espressif Guest-Edited Edition 2023**
www.elektor.com/EP-0526
- **Elektor Special: Espressif Guest-Edited Edition (PDF)**
www.elektor.com/ED-0526

fonctions vocales avec le Raspberry Pi Zero

Donner de la voix à votre projet

Somnath Bera (Inde)

Explorons la reconnaissance vocale (STT) et la synthèse vocale (TTS) sur le Raspberry Pi Zero, avec différentes options logicielles, des instructions d'installation et des applications potentielles pour des projets intégrant le traitement audio et la commande vocale.

La synthèse vocale (TTS) et la conversion de la parole en texte (STT) sont les deux principales fonctions vocales que nous allons explorer avec le Raspberry Pi Zero, un ordinateur monocarte équipé d'un processeur performant cadencé à 1 GHz, qui peut être surcadencé à 1,3 GHz avec un dissipateur thermique en métal pour un refroidissement passif (pour un surcadencage jusqu'à 1,7 GHz, un refroidissement actif est nécessaire). En termes de performances, cette carte offre des performances comparables au Raspberry Pi 4, bien qu'elle dispose d'une mémoire limitée à 512 MB. J'ai testé plusieurs logiciels de TTS et STT sur cette carte, et je présente les résultats détaillés ci-dessous, ainsi que des idées de projets pour inspirer les lecteurs.

Surcadencage

Comme le Pi Zero dispose d'une RAM limitée (seulement 512 MB), le surcadencage fournit une augmentation significative des performances, ce qui est particulièrement utile pour les tâches de TTS et STT. Ce gain en puissance et en efficacité vient toutefois au prix d'une augmentation potentielle de la chaleur et d'une possible instabilité du CPU. Sur le Raspberry Pi Zero, l'overclocking est simple. En ajoutant seulement deux lignes au fichier `/boot/config.txt`, en sauvegardant et en redémarrant, vous pouvez immédiatement bénéficier d'une hausse des performances. Il est essentiel de surveiller la température du CPU ; avec un bon dissipateur thermique en métal, il est possible de surcadencer en toute sécurité

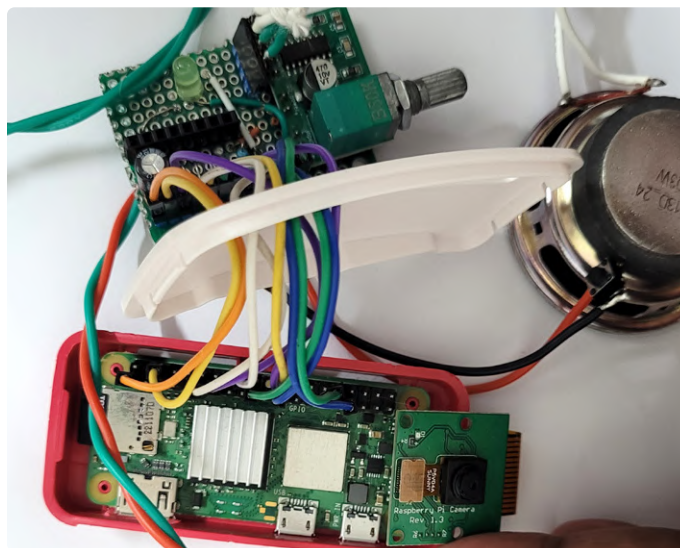


Figure 1. Gros plan sur le Raspberry Pi Zero avec son dissipateur thermique métallique.

jusqu'à 1,3 GHz, en maintenant une température de 65°C à 70°C sans compromettre la stabilité. Pour des fréquences supérieures, comme jusqu'à 1,7 GHz, un refroidissement actif devient nécessaire. Référez-vous à la **figure 1** qui montre le dissipateur métallique que j'ai utilisé

Éditez le fichier `config.txt`, par exemple avec l'éditeur Nano :

```
sudo nano /boot/config.txt
```

et ajoutez ces deux lignes :

```
arm_freq=1300
over_voltage=2
```

Si vous souhaitez essayer l'overclocking jusqu'à 1,7 GHz, augmentez le dernier paramètre à 4 ou 5. Puis enregistrez et redémarrez pour appliquer les modifications. Pour surveiller la température du processeur, vous pouvez utiliser la commande `vcgencmd measure_temp`.

Sortie son sur le Raspberry Pi Zero

Le Raspberry Pi Zero ne dispose pas de connecteur intégré pour la sortie audio. Pour obtenir du son, vous devez connecter un périphérique HDMI, tel qu'un téléviseur. Cependant, il existe des adaptateurs HDMI vers AV qui peuvent séparer le signal audio. Vous pouvez également utiliser deux des broches GPIO compatibles PWM (12, 13, 18, 19) pour activer la sortie audio en ajoutant une seule ligne au fichier `/boot/config.txt`. Ouvrez le fichier avec `sudo nano /boot/config.txt`, puis ajoutez la ligne suivante :

```
dtoverlay=audremap,pins_18_13
```

Sauvegardez le fichier et redémarrez ; la sortie stéréo sur les broches 18 et 13 est maintenant activée.

Ces broches émettent des signaux carrés modulés en largeur d'impulsion de niveau logique (3,3 V), qui ne conviennent pas pour alimenter directement des haut-parleurs, mais qui peuvent être suffisants pour une petite paire d'écouteurs. Vous pouvez utiliser un module d'amplification audio, tel que le PAM8403. J'ai réalisé le montage présenté dans la **figure 2**. Les signaux droit et gauche sont filtrés par des filtres RC passe-bas, puis couplés en courant alternatif aux entrées de l'ampli via les deux condensateurs de 10 µF. Vous êtes maintenant prêt à profiter d'une sortie stéréo de qualité ! Vous pouvez voir le circuit en fonctionnement sur [1], où j'utilise le Raspberry Pi Zero pour diffuser une chanson bouddhiste. Le montage complet est illustré dans la **figure 3**.

Logiciel de synthèse vocale

Examinons quelques-unes des options offertes par les logiciels de synthèse vocale.

> **eSpeak / eSpeak-ng** : eSpeak et sa nouvelle version eSpeak-ng (nouvelle génération) sont des logiciels de synthèse vocale open-source matures, disponibles pour les systèmes basés sur Linux (comme Debian) et Mac OS. Ces

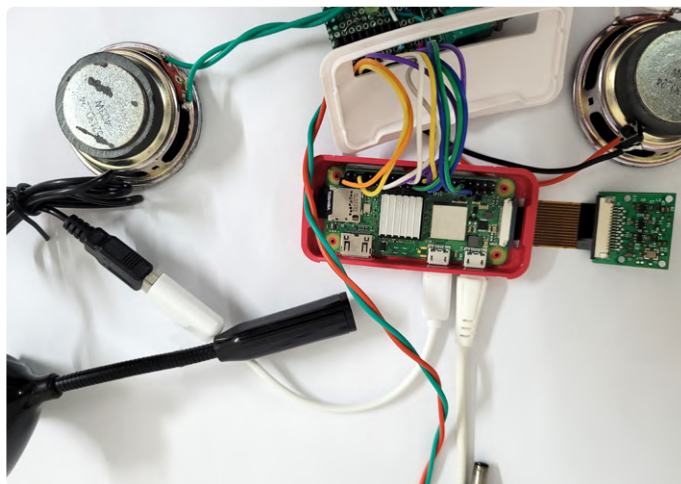


Figure 3. Configuration complète avec microphone, amplificateur et haut-parleurs.

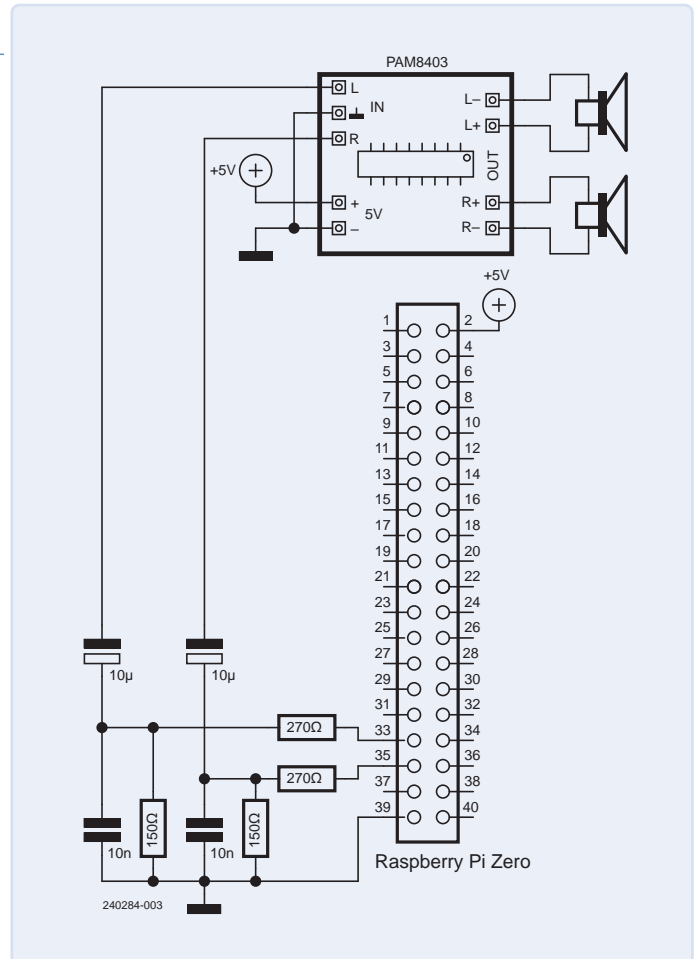


Figure 2. Schéma du projet

programmes sont faciles à installer et peuvent être utilisés immédiatement. Pour installer eSpeak-ng sur le Pi Zero, il suffit de suivre les instructions fournies, en utilisant les commandes :

```
sudo apt-get install espeak-ng
espeak-ng -f temp.txt
espeak-ng "Hello Readers, welcome to this project"
```

Vous devriez alors entendre la phrase « Hello Readers... ». Outre l'anglais, eSpeak-ng supporte plusieurs langues et voix. `espeak-ng -v en+f2 "Input text"` émettra une voix féminine. Pour plus de détails, consultez le manuel avec `man espeak-ng`.

> **Google Text-To-Speech (gTTS)** : gTTS offre des voix beaucoup plus naturelles que eSpeak ou eSpeak-ng. Je l'utilise souvent dans mes projets d'IA. Vous pouvez l'installer :

```
pip install gtts playsound
```

Voir le **listage 1** pour un exemple de code Python pour démarrer.

Entrée vocale sur le Pi Zero

La sortie vocale fonctionne parfaitement jusqu'à présent. Pour l'entrée vocale, un microphone USB économique (environ 2 \$ chez robu.in) est parfaitement fonctionnel et prêt à l'emploi. Après l'avoir



Listage 1. Utilisation de la synthèse vocale de Google.

```
from gtts import gTTS
import subprocess

def text_to_speech(text):
    # Use gTTS to convert text to speech and save to an audio file
    tts = gTTS(text=text, lang="en")
    tts.save("output.mp3")
    # Use mpg123 to play the audio file
    subprocess.run(["mpg123", "-q", "output.mp3"])

if __name__ == "__main__":
    while True:
        # Get the string to speak
        question = "Please ask your question:"
        text_to_speech(question)
```

branché, la commande `lsusb` vous permettra de vérifier que le microphone est détecté. Voici ce que j'obtiens avec cette commande dans mon exemple :

```
Bus 001 Device 002: ID 4c4a:4155 Jieli Technology
UACDemoV1.0
Bus 001 Device 001: ID 1d6b:0002 Linux Foundation 2.0
root hub
```

Vous aurez peut-être besoin d'un câble OTG pour faire l'interface entre le microUSB (côté Pi Zero) et l'USB (microphone). Maintenant que le microphone est connecté, pour le tester complètement, essayez cette commande dans un terminal :

```
arecord -f S16_LE -r 33100 -d 10 -c 1 output.wav
```

Les paramètres sont expliqués ci-dessous.

- `-f S16_LE` : spécifie le format audio. `S16_LE` signifie signé 16 bits little-endian.
- `-r 33100` : Définit le taux d'échantillonnage à 33 100 Hz.
- `-d 10` : spécifie la durée de l'enregistrement en secondes (par exemple, 10 s dans cet exemple).
- `-c 1` : Définit le nombre de canaux audio à 1 (mono). Pour les enregistrements stéréo, utilisez `-c 2`.
- `output.wav` : spécifie le nom du fichier de sortie (dans ce cas, `output.wav`).

Après avoir exécuté cette commande, `arecord` capturera l'audio à un taux d'échantillonnage de 33 100 Hz pendant 10 s et l'enregistrera en tant que fichier `output.wav` dans le répertoire actuel. Vous pouvez modifier des paramètres tels que la durée (`-d`), le nom du fichier de sortie et d'autres paramètres en fonction de vos besoins spécifiques.

Pour lire l'audio enregistré, exécutez `aplay output.wav`. Veillez à régler la position du microphone et à l'éloigner du haut-parleur pour garantir des enregistrements audio nets.

Jusqu'ici, tout fonctionne bien ! Vous avez configuré votre source audio, installé le logiciel de synthèse vocale sur le Pi Zero et essayé avec succès l'enregistrement audio et la conversion du texte en parole avec eSpeak, eSpeak-ng ou gTTS. Passons maintenant à la conversion de la parole en texte.

Logiciels de transcription de la parole

Google offre un service payant pour la transcription de la parole nécessitant une connexion internet. En outre, pour la synthèse vocale, Google propose également une API en ligne qui est gratuite jusqu'à une certaine limite, après quoi un compte payant est nécessaire.

Dans le cadre de ce projet, sur le Pi Zero, il est préférable d'utiliser un logiciel de synthèse vocale hors ligne, à condition que la parole soit suffisamment claire. Bien qu'il existe une fonction de Google (voir ci-dessous), mais il y a aussi deux excellents logiciels libres disponibles pour la conversion parole-texte : `spchcat` et `Pocketsphinx`. Alors que `spchcat` est basé sur une ligne de commande, `Pocketsphinx` est une solution basée sur une bibliothèque. Les deux logiciels peuvent être utilisés avec Python en utilisant différentes méthodes. Le module `subprocess` est utilisé pour appeler `spchcat`, tandis que le module `pocketsphinx` peut être installé directement dans Python avec `pip`. Voici la procédure d'installation de ces deux outils performants.

Spchcat

`Spchcat` est un logiciel open-source qui utilise le modèle de reconnaissance TensorFlow. Il est actuellement compatible avec Linux et Raspberry Pi et prend en charge 46 modèles de langues, y compris des langues indiennes telle que le tamoul et le bengali, ma langue

maternelle. Vous pouvez télécharger le logiciel et les instructions depuis [2].

Spchcat est un outil en ligne de commande qui traite l'audio à partir de fichiers WAV. C'est un programme volumineux, environ 1,2 Go, qui ne peut pas être directement installé sur un Pi Zero. Pour contourner cette limitation, suivez les étapes suivantes :

1. Téléchargement : commencez par télécharger le logiciel *spchcat* depuis son dépôt GitHub sur votre ordinateur.
2. Transfert : utilisez *scp* (*secure copy*) pour transférer le fichier téléchargé sur la carte SD de votre Pi Zero.
3. Installation : comme le Pi Zero ne peut pas gérer l'installation, retirez la carte SD et insérez-la dans un Raspberry Pi B+ ou un Raspberry Pi 4.
4. Installation sur le Pi 4 : double-cliquez sur le logiciel téléchargé sur le Raspberry Pi 4. Le processus d'installation prendra environ 25 minutes.
5. Dernière étape : après l'installation, remplacez la carte SD sur votre Pi Zero. Le logiciel *spchcat* est maintenant prêt à être utilisé sur votre Pi Zero.

Grâce à cette méthode, vous pouvez surmonter les contraintes de capacité du Pi Zero.

Après l'installation, vous pouvez commencer à capturer de l'audio via le microphone par défaut en lançant le logiciel. Les résultats s'afficheront directement dans le terminal :

spchcat

Si vous ne disposez pas de microphone ou si vous souhaitez transcrire de l'audio provenant d'un autre programme, vous pouvez utiliser l'argument `--source` avec la valeur `system`. Cette option tentera d'écouter l'audio joué par votre machine, y compris les vidéos ou les chansons, et de transcrire toute parole détectée :

```
spchcat --source=system
```

Lancez une vidéo YouTube ou jouez un fichier audio sur votre système. Vous verrez la transcription vocale apparaître dans la fenêtre du terminal au fur et à mesure de la lecture.

Pour traiter un fichier WAV avec *spchcat*, utilisez la commande suivante :

```
spchcat /home/bera/myaudio.wav
```

Après le traitement, le texte transcrit s'affichera dans le terminal. Pour choisir la langue de transcription, utilisez l'argument `--language`, par exemple pour l'allemand ou l'anglais, utilisez :

```
spchcat --language=de_DE  
spchcat --language=en_US
```

Attention, *spchcat* est gourmand en ressources et peut surchauffer le processeur. La surveillance de la température du processeur est



Liste des composants

La plupart de ces composants, à l'exception du Raspberry Pi Zero, sont disponibles à des prix très modérés sur les sites Internet chinois courants.

- Raspberry Pi Zero 2 W
- Module d'amplification PAM8403
- Haut-parleurs de 4 ohms (2 unités)
- Microphone USB
- Câble USB OTG

cruciale. Il est essentiel de surveiller la température du CPU et d'utiliser un dissipateur thermique de haute qualité sur le Pi Zero pour gérer efficacement la chaleur produite lors de son fonctionnement. Vous pouvez tester *spchcat* avec des fichiers WAV d'exemple sans avoir à les créer vous-même. Téléchargez des fichiers WAV depuis [3] et utilisez-les pour évaluer les résultats de la transcription avec votre Pi Zero. Ils vous aideront à juger la qualité d'enregistrement requise pour une transcription précise.

Pour enregistrer la transcription produite par la commande dans un fichier, la méthode typique sous Linux est d'utiliser le caractère `>` :

```
spchcat your-audio-file.wav > /tmp/transcript.txt
```

Enfin, pour utiliser *spchcat* dans un script Python, vous devez utiliser le module *subprocess*. Assurez-vous que votre environnement utilise Python 3.7 ou plus. Si *subprocess* n'est pas préinstallé, vous pouvez l'installer avec `pip install subprocess`.

Sur le Pi Zero, *spchcat* est capable de traiter jusqu'à 4 minutes d'enregistrement et de transcription de fichiers WAV. Sur un Raspberry Pi 4, il peut effectuer des transcriptions en continu à partir de sources telles que des vidéos YouTube. La qualité de l'enregistrement est cruciale pour obtenir des résultats précis. Placer le microphone près de la source sonore améliore généralement la précision de la transcription.

PocketSphinx

PocketSphinx est un moteur de reconnaissance vocale léger et open-source développé par l'université Carnegie Mellon. Il traite l'audio mono canal 16 bits PCM provenant de l'entrée standard ou de fichiers et tente de reconnaître la parole en utilisant des modèles acoustiques et linguistiques par défaut. Plus simple et moins complexe que *Spchcat*, *PocketSphinx*, il supporte plusieurs langues, dont l'anglais, le chinois, le français, l'espagnol, l'allemand et le russe. Cependant, je ne l'ai testé qu'avec l'anglais. *PocketSphinx* peut convertir des fichiers WAV en texte directement, mais ce processus peut s'avérer lent lorsqu'il est exécuté avec Python. Pour installer *PocketSphinx* ainsi que le modèle de langue anglaise, utilisez les commandes suivantes :

```
pip install pocketsphinx  
sudo apt-get install -y python3-pocketsphinx  
pocketsphinx-en-us
```


En Python, la classe `LiveSpeech` du module `PocketSphinx` permet la reconnaissance vocale. La valeur du time-out, ici fixée à 1,0, peut être ajustée selon les besoins. Voici un programme Python très simple pour tester `PocketSphinx` :

```
from pocketsphinx import LiveSpeech
speech = LiveSpeech()
speech = LiveSpeech(silence_timeout=1.0)
print("Listening...")
for phrase in speech:
    print(f"Transcript: {phrase}")
```

La transcription en temps réel est possible à la fois sur le Pi Zero et le Raspberry Pi 4. Sur le Pi Zero, il arrive que `PocketSphinx` omette certains mots, alors qu'il fonctionne parfaitement sur le Raspberry Pi 4. J'ai trouvé que la qualité de la transcription offerte par `PocketSphinx` est souvent inférieure à celle de `spkchat`.

Google Speech-to-Text

Google Speech-to-Text (STT) supporte 125 langues et est très efficace, ce qui en fait une excellente option pour les appareils à capacité limitée, comme le Raspberry Pi Zero. Pour utiliser Google STT,

installez la bibliothèque Python `SpeechRecognition`, qui permet de l'utiliser :

```
pip install SpeechRecognition
```

Un programme d'exemple simple d'utilisation de cette bibliothèque, avec Google Speech-To-Text, est présenté dans le **listage 2**.

La bibliothèque `SpeechRecognition` a l'avantage de supporter plusieurs moteurs de reconnaissance vocal en plus de celui de Google. Par exemple, Pour utiliser `PocketSphinx`, vous pouvez réutiliser le code du listage 2 en remplaçant `text = recognizer.recognize_google(audio)` par `text = recognizer.recognize_sphinx(audio)` à la ligne 15.

Le **listage 3** montre comment utiliser cette bibliothèque avec le moteur de reconnaissance vocale de Google pour transcrire du texte à partir d'un fichier. Je vous recommande vivement de le tester ! Bien entendu, il est possible de combiner l'enregistrement audio et la transcription dans un même script Python. Par exemple, le **listage 4** illustre comment enregistrer de l'audio pendant dix secondes tout en posant une question, puis affiche la transcription textuelle de cette question.

J'ai utilisé avec succès Speech-to-Text (STT) et Text-to-Speech (gTTS)



Listage 2. Utilisation de la bibliothèque `SpeechRecognition` avec Python.

```
import speech_recognition as sr
def speech_to_text():
    # Initialize the recognizer
    recognizer = sr.Recognizer()
    # Use the default microphone as the audio source
    with sr.Microphone() as source:
        print("Speak something...")
        audio = recognizer.listen(source)
    try:
        print("Recognizing...")
        # Use Google Web Speech API to perform speech recognition
        text = recognizer.recognize_google(audio)
        return text
    except sr.UnknownValueError:
        print("Could not understand audio")
        return ""
    except sr.RequestError as e:
        #print(f"Error: ")
        return ""
if __name__ == "__main__":
    while True:
        # Call the speech_to_text function to get text from speech
        result = speech_to_text()
        print(f"You said: ")
```



Listage 3. Reconnaissance vocale depuis un fichier audio.

```
import argparse
import speech_recognition as sr

def speech_to_text(audio_file):
    # Initialize the recognizer
    recognizer = sr.Recognizer()
    # Load the audio file
    with sr.AudioFile(audio_file) as source:
        audio_data = recognizer.record(source)
    try:
        # Use Google Speech Recognition for speech-to-text
        text = recognizer.recognize_google(audio_data)
        return text
    except sr.UnknownValueError:
        print("Google Speech Recognition could not understand the audio")
        return ""
    except sr.RequestError as e:
        print(f"Could not request results from Google Speech Recognition service; ")
        return ""

if __name__ == "__main__":
    # Create argument parser
    parser = argparse.ArgumentParser(description="Convert audio file to text using Google Speech Recognition")
    parser.add_argument("audio_file", help="Path to the audio file")
    parser.add_argument("destination_file", help="Path to save the text output")
    # Parse command-line arguments
    args = parser.parse_args()
    # Call the speech_to_text function to convert speech to text
    result = speech_to_text(args.audio_file)
    if result:
        # Save the recognized text to the destination file
        with open(args.destination_file, "w") as f:
            f.write(result)
        print(f"Recognized text saved to ")
    else:
        print("Speech recognition failed.")
```

de Google dans divers projets avec le Raspberry Pi Zero, y compris dans des projets basés sur des microprocesseurs comme le *terminal ESP32 ChatGPT* [4][5], où il génère des réponses vocales via un haut-parleur I2S.

Futurs cas d'utilisation ?

La conversion du texte en parole et inversement est une tâche importante. Les grandes entreprises de médias s'appuient sur des logiciels coûteux et des ordinateurs performants pour effectuer ces tâches en temps réel pendant les diffusions en direct. Bien que l'utilisation d'un Raspberry Pi Zero puisse sembler simpliste, cette carte offre des avantages notables, notamment sa taille compacte,

sa capacité à être alimentée avec diverses alimentations 5 V et sa consommation d'énergie réduite.

Voici quelques applications potentielles pour la transcription vocale sur le Pi Zero que je recommande :

- **Encyclopédie parlante** : un terminal mains libres où vous posez une question, la confirmez, et le Pi Zero récupère les réponses d'un service tel que ChatGPT, puis les énonce à haute voix. Tout cela peut être réalisé avec un Pi Zero. En savoir plus : [6].
- **Sortie vocale pour les projets d'IA** : dans les projets d'IA, comme ceux que j'ai développés avec Edge Impulse pour la



Listage 4. Reconnaissance vocale et enregistrement audio.

```
import subprocess
import speech_recognition as sr

def record_audio(file_name, duration=10):
    # Record audio from USB microphone using arecord
    command = ["arecord", "-D", "plughw:1,0", "-f", "S16_LE", "-r", "16000", "-d", str(duration), file_name]
    subprocess.run(command)

def speech_to_text(file_name):
    recognizer = sr.Recognizer()
    # Load the audio file for recognition
    with sr.AudioFile(file_name) as source:
        audio_data = recognizer.record(source)
    text = recognizer.recognize_google(audio_data)
    return text.lower() # Convert to lowercase for easier comparison

if __name__ == "__main__":
    while True:
        # Record audio for question
        record_audio("question.wav", duration=10) # Increased duration for longer responses
        speech_text = speech_to_text("question.wav")
        print(speech_text)
```

classification d'objets, les résultats de la classification (par exemple, un nombre ou un nom) peuvent être verbalisés par l'appareil grâce à eSpeak ou eSpeak-ng, vous libérant ainsi de la nécessité d'utiliser un écran. Voir un exemple sous [7].

- **Terminal pour personnes aveugles** : un système permettant à une personne malvoyante de poser des questions, le Raspberry Pi recherchant les réponses sur divers sites, y compris OpenAI, pour fournir des réponses vocales.
- **Robotique** : idéal pour les jouets robotisés à commande vocale, en particulier pour les marchés internationaux, en convertissant les commandes en langue étrangère en anglais pour un traitement ultérieur.
- **Lecteurs de codes-barres portables à commande vocale** : utiles dans les supermarchés, où l'appareil lit les codes-barres et annonce les résultats par via des haut-parleurs. Les capacités multilingues d'eSpeak-ng pourraient être un grand particulièrement bénéfiques.
- **Assistant vocal pour les malentendants** : un assistant vocal portable qui fournit une transcription en temps réel des phrases prononcées pour les personnes malentendantes.
- **Assistant de voyage** : facilite la communication dans les pays étrangers en traduisant et en vocalisant les interactions avec les habitants, comme les chauffeurs de taxi ou les serveurs.
- **Lecteur de musique à interaction vocale** : un lecteur

de musique portable qui réagit aux commandes vocales, similaire à Alexa d'Amazon.

Ce ne sont là que quelques exemples parmi tant d'autres possibles. Quelles autres applications impliquant la parole pourriez-vous envisager avec des nano-ordinateurs comme le Raspberry Pi ?

240284-04



À propos de l'auteur

Somnath Bera, ingénieur en mécanique diplômé du Jalpaiguri Govt. Engineering College en Inde, occupe le poste de directeur général chez NTPC, le plus grand producteur d'électricité du pays. Passionné par l'électronique, il a réalisé plus de 60 projets novateurs sur Elektor Labs, dont plus de 10 ont été publiés dans le magazine Elektor. Ses projets sont principalement axés sur des solutions aux enjeux de gestion des déchets et des ressources naturelles. Somnath aime exploiter des approches et des plateformes innovantes telles qu'Arduino, Raspberry Pi et ESP32, intégrant divers capteurs et systèmes sans fil, afin de concevoir des solutions efficaces et économiques.

Questions ou commentaires ?

Envoyez un courriel à l'auteur (berasomnath@gmail.com), ou contactez Elektor (redaction@elektor.fr).



Produits

- > Raspberry Pi Zero 2 WH (avec connecteurs)
www.elektor.fr/20952
- > John Allwork, *Programming Voice-controlled IoT Applications with Alexa and Raspberry Pi* (Elektor, 2023)
www.elektor.fr/20400

LIENS

- [1] Exemple de fichier audio joué sur Raspberry Pi Zero : <https://youtu.be/YXC7VIVIX9c>
- [2] Aide et codes de fichier pour spchcat : <https://github.com/petewarden/spchcat>
- [3] Fichiers WAV d'exemple à tester avec spchcat : https://www.voiptroubleshooter.com/open_speech/
- [4] « Terminal ESP32-ChatGPT » sur Elektor Labs : <https://www.elektormagazine.fr/labs/esp32-chatgpt-terminal>
- [5] Somnath Bera, « ESP32 ChatGPT Terminal », édition spéciale IA 2024 d'Elektor : <http://www.elektormagazine.fr/230536-04>
- [6] "Raspberry Pi Zero Talking Encyclopedia" on Elektor Labs: <https://www.elektormagazine.com/labs/raspberry-pi-zero-encyclopedia>
- [7] Exemple : Transcription de texte dans les projets d'IA : <https://www.youtube.com/watch?v=C5QrCl4XIJU>

Ils nous font confiance, n'est-ce pas ?

elektor.fr

Réactivité d'elektor
Tout va plus vite chez elektor, expédition, suivi des commandes, communications !!!
Un véritable service à la carte !!!
Bravo !!!
Date of experience: June 04, 2024

Prix très compétitifs et livraison...
Prix très compétitifs et livraison rapide
Date of experience: May 11, 2024

Aucun problème
Aucun problème. Délai de livraison court et respecté. Suivi de livraison détaillé. Contenu correspondant à la commande. État et fonctionnement des produits parfait. Aucun défaut. Je recommande.
Date of experience: May 25, 2024

Produits disponibles
Produits disponibles, bonne description, envoi rapide et soigneusement emballé, nickel.
Date of experience: May 10, 2024

Nous aimons l'électronique et les projets, et nous faisons tout notre possible pour répondre aux besoins de nos clients.
Le magasin Elektor : **Jamais cher, toujours surprenant**

Elektor Store

Reviews 365 • Excellent



VERIFIED COMPANY

Consultez d'autres avis sur notre page Trustpilot : www.elektor.com/TP/fr



Vous pouvez également vous faire votre propre opinion en visitant notre Elektor Store, www.elektor.fr



le rôle croissant de l'IA périphérique

Une tendance qui structure l'avenir

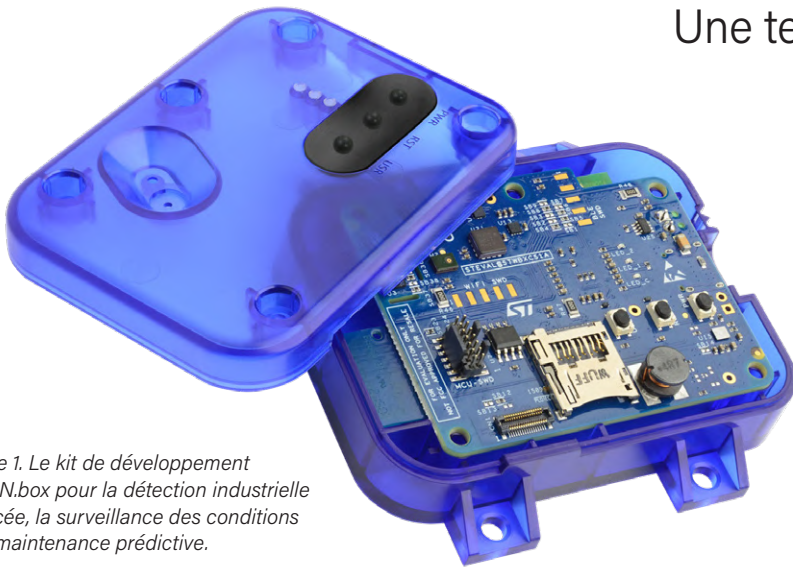


Figure 1. Le kit de développement STWIN.box pour la détection industrielle avancée, la surveillance des conditions et la maintenance prédictive.

Vianney Taufour (STMicroelectronics)

Au rythme de l'évolution de la technologie, l'intelligence artificielle (IA) franchit les limites de l'informatique en nuage (Cloud) et s'installe dans le monde des appareils périphériques. L'IA périphérique (Edge AI) représente une évolution majeure vers une intelligence décentralisée et en temps réel sur des dispositifs tels que les capteurs, les moteurs, les caméras et les véhicules. En traitant les données localement, l'IA périphérique réduit la latence, améliore la confidentialité et permet une prise de décisions en temps réel, essentielle pour de nombreuses applications industrielles et grand public.

Cette tendance croissante est due à la convergence des systèmes embarqués, des algorithmes d'apprentissage automatique et d'un large éventail de capteurs. De la détection d'anomalies dans les moteurs industriels à la reconnaissance faciale dans les systèmes de sécurité, la créativité requise pour extraire des informations significatives de ces capteurs est à l'origine

de certaines des applications d'IA les plus innovantes dans les systèmes embarqués. Alors que les ingénieurs et les développeurs s'efforcent d'exploiter de nouvelles possibilités grâce à l'IA, la polyvalence de l'Edge AI devient de plus en plus évidente. Examinons de plus près comment cette tendance se manifeste dans diverses applications du monde réel.

Cas d'utilisation de l'Edge AI dans le monde réel

Détection des anomalies dans les performances des moteurs

L'Edge AI permet de détecter en temps réel les anomalies dans les moteurs fonctionnant à des vitesses variables. En utilisant des capteurs de vibrations, tels que le capteur de vibrations numérique à 3 axes à bande passante ultra-large et à faible bruit utilisé dans notre kit de développement STWIN.box (figure 1), les modèles d'IA peuvent identifier les écarts de performance susceptibles de signaler des problèmes mécaniques potentiels. Ceci est inestimable dans les secteurs où la détection précoce des anomalies peut prévenir des pannes coûteuses et améliorer l'efficacité des programmes de maintenance. Ce cas d'utilisation est également intéressant parce que l'apprentissage de l'algorithme de détection des anomalies peut être repris sur l'appareil lui-même pour s'adapter au schéma de vibration exact et fournir la meilleure précision possible pour chaque machine. Imaginez une usine avec des centaines de machines où vous pouvez affiner le modèle pour chacune d'elles ! Pour en savoir plus, consultez [1].

Détection et classification des défauts des moteurs

Dans l'industrie, les moteurs sont des composants essentiels qui entraînent tout, des tapis roulants aux bras robotisés. L'Edge AI appliquée aux moteurs permet de détecter les défauts en temps réel en analysant des données telles que la consommation de courant ou les vibrations. Des algorithmes avancés peuvent même classer le type de défaut, qu'il s'agisse d'un déséquilibre, d'un désalignement ou d'une usure mécanique, permettant ainsi des interventions précises

et opportunes. Pour plus d'informations, voir [2].

Détection des arcs électriques dans les systèmes à courant continu

Les installations solaires et autres systèmes électriques à courant continu (CC) sont confrontés à des défis spécifiques, dont l'un est le risque d'arc électrique. En appliquant l'Edge AI, ces systèmes peuvent détecter ce risque en temps réel, améliorant ainsi la sécurité et évitant les incendies ou les dommages coûteux. Cette application met en évidence le rôle croissant de l'IA dans l'amélioration de la fiabilité des infrastructures d'énergies renouvelables. L'utilisation de l'IA dans ce cas apporte une meilleure précision que l'approche algorithmique existante, ce qui est crucial pour éviter les faux positifs et réduire les temps d'arrêt. Pour en savoir plus, voir [3].

Capteurs de pression virtuels pour améliorer la sécurité des vélos à assistance électrique

Les capteurs traditionnels peuvent être remplacés ou complétés par des capteurs virtuels pilotés par l'IA qui déduisent la pression ou la force dans des systèmes complexes. Par exemple, dans les vélos à assistance électrique, l'Edge AI peut améliorer la sécurité en prédisant les changements de pression à partir de la consommation de courant par le moteur. Cela permet de créer des systèmes de contrôle plus intelligents qui optimisent les performances gratuitement, tout en garantissant la sécurité du cycliste. Pour plus d'informations, voir [4].

Détection du type de plancher pour les aspirateurs intelligents

Dans les appareils domestiques intelligents tels que les aspirateurs, l'Edge AI est utilisée pour reconnaître les différents types de planchers (tapis, bois dur ou carrelage) et ajuster les modes de nettoyage en conséquence. L'algorithme d'IA est capable de distinguer un sol dur d'un sol mou parmi 40 types d'échantillons différents avec une grande précision en utilisant un simple capteur de type temps de vol (ToF). Pour plus d'informations, voir [5].

Classification des panneaux de signalisation dans les véhicules autonomes

Les systèmes de conduite autonome

s'appuient sur une reconnaissance précise des objets pour naviguer sur les routes en toute sécurité. La classification des panneaux de signalisation au moyen de l'Edge AI en est un élément clé. Les modèles d'IA embarqués traitent les données des caméras pour identifier les panneaux en temps réel, ce qui permet aux véhicules de réagir à l'évolution de l'état des routes et des réglementations sans dépendre du cloud. Cette application, qui traite le signal d'une caméra RVB, est un exemple d'application de vision par ordinateur, longtemps réservée à des machines puissantes, et désormais disponible sur microcontrôleurs. Pour plus d'informations, voir [6].

Identification faciale et systèmes de sécurité

Dans les applications de sécurité, l'Edge AI permet la reconnaissance et l'identification faciales en temps réel en se passant de l'accès permanent au nuage. Cela est particulièrement important pour la protection de la vie privée et des données, car les informations biométriques sensibles peuvent être traitées et stockées localement. Des banques aux villes intelligentes, l'Edge AI améliore les systèmes de sécurité, offrant une reconnaissance plus rapide et plus précise sans compromettre la vie privée de l'utilisateur. Pour en savoir plus, consultez le site [7].

L'importance de la créativité dans le déploiement de l'IA

Si les cas d'utilisation de l'Edge AI démontrent son vaste potentiel (figure 2),

ils mettent également en évidence un aspect essentiel mais souvent négligé : la créativité. L'application de l'IA aux systèmes embarqués ne se limite pas à l'usage de modèles ou d'algorithmes prêts à l'emploi. Elle exige des développeurs et des ingénieurs qu'ils pensent différemment et qu'ils fassent preuve de créativité pour résoudre les problèmes d'une manière hors des possibilités des méthodes traditionnelles. La véritable puissance de l'Edge AI ne réside pas seulement dans sa capacité à automatiser des tâches, mais dans la manière dont elle permet de trouver des solutions innovantes en réinventant la manière dont les données sont collectées, traitées et utilisées.

Dans le monde des systèmes embarqués, des contraintes telles qu'une puissance de traitement, une mémoire et une efficacité énergétique limitées posent des défis importants. Cependant, ces contraintes inspirent également une résolution créative des problèmes. Les ingénieurs doivent aller au-delà du traitement traditionnel des données des capteurs et explorer comment divers flux de données peuvent être combinés, déduits ou même synthétisés pour produire des résultats utiles. Par exemple, les modèles d'IA doivent souvent être optimisés pour s'adapter aux contraintes strictes des appareils embarqués sans sacrifier la précision ou l'efficacité de leurs prédictions.

Prenons l'exemple des vélos à assistance électrique de Panasonic [4]. Dans ce cas, au lieu d'installer des capteurs de pression traditionnels pour mesurer la pression des



Figure 2. Le vaste potentiel de l'Edge AI pour une variété d'applications.

NanoEdge AI Studio simplifies edge AI development workflow

Deployment of NanoEdge AI Studio libraries, the market reference AutoML tool, is completely free for unlimited quantities on any STM32

... and available on any Arm® Cortex®-M MCU*

NANOEDGE AI STUDIO

The best combination for given data: ML model, hyperparameters, and preprocessing

On-device learning capability to fine-tune a deployed solution without retraining

Bring your own data approach: no need to create edge AI models



*under a special license agreement

7

Figure 3. NanoEdge AI Studio offre une interface simple pour développer des projets d'apprentissage machine.

pneus, ce qui ajouterait de la complexité et des coûts, les ingénieurs ont eu recours à une approche créative en utilisant l'IA. En exploitant les données de mesure du courant du moteur, ils ont conçu un capteur de pression virtuel - une solution qui déduit la pression du pneu en analysant le courant consommé par le moteur. À mesure que le moteur réagit aux sollicitations du conducteur et à l'état de la route, les variations subtiles de la consommation de courant peuvent être utilisées pour prédire la pression des pneus. Cette utilisation créative de l'IA permet non seulement de réduire la complexité du matériel, mais aussi d'améliorer les performances globales et la sécurité du vélo, tout en utilisant les données déjà saisies par le système. Cet exemple illustre parfaitement la manière dont la créativité dans le déploiement de l'IA peut conduire à des solutions innovantes et rentables. Sortant des sentiers battus, les ingénieurs ont reformulé le problème - la surveillance de la pression des pneus - afin que les données existantes fournissent de nouvelles informations. Au lieu de s'appuyer sur des capteurs coûteux ou spécialisés, ils ont utilisé les données existantes du moteur pour faire surgir une toute nouvelle fonction. C'est ce genre d'ingéniosité qui donne à l'intelligence artificielle toute sa puissance et sa polyvalence.

Réinventer la résolution de problèmes dans les systèmes d'IA embarqués

L'utilisation créative de l'IA dans les systèmes embarqués exige également que les ingénieurs envisagent la résolu-

tion des problèmes de manière globale, en tenant compte non seulement des aspects techniques, mais aussi des implications pratiques du déploiement de l'Edge AI. Par exemple :

Réutilisation des capteurs existants

Comme dans le cas de la bicyclette Panasonic, l'IA peut souvent donner aux capteurs existants un nouvel objectif. Les ingénieurs peuvent utiliser des capteurs de vibrations, conçus à l'origine pour surveiller l'usure mécanique, pour détecter également d'autres paramètres critiques tels que des conditions environnementales ou des anomalies de performance inattendues.

Combiner diverses sources de données

En combinant de manière créative des données provenant de plusieurs capteurs (comme le son, les vibrations et la température), il est possible de dresser un tableau plus complet de l'état de santé d'un système. Les modèles d'IA entraînés avec cette combinaison de données peuvent détecter des subtilités susceptibles d'échapper à un capteur isolé.

Déduire des informations cachées

L'IA peut déduire des informations qui ne sont pas directement mesurables. Par exemple, plutôt que de mesurer directement l'état d'une route ou d'un sol, les modèles d'IA peuvent utiliser les vibrations d'un aspirateur ou de la suspension d'une voiture pour déterminer les types de surface ou les niveaux d'usure. Cette capacité à extraire des informations cachées à partir de sources de données indirectes permet aux développeurs de résoudre

des problèmes d'une manière nouvelle et passionnante.

Optimiser l'efficacité énergétique

L'énergie est une ressource précieuse pour tout ce qui est alimenté par batterie, comme les appareils portables ou intelligents. La créativité dans le déploiement de l'IA consiste à concevoir des modèles qui fournissent des résultats précis en consommant un minimum d'énergie. Cela signifie souvent qu'il faut trouver des moyens astucieux de réduire la complexité des modèles d'IA ou d'effectuer les calculs avec le moins d'énergie possible.

Dans tous ces cas, les développeurs doivent dépasser le paradigme traditionnel de la simple automatisation des tâches et considérer l'IA comme un outil permettant de découvrir de nouvelles idées et de relever des défis complexes de manière plus efficace. Cela nécessite un changement de mentalité, car l'objectif n'est pas simplement d'appliquer l'IA, mais de repenser la manière dont l'IA peut reformuler et simplifier le problème lui-même.

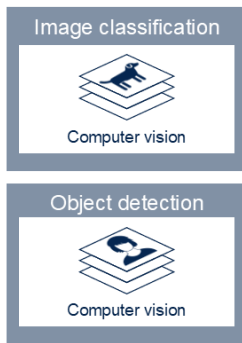
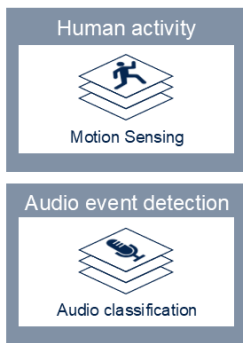
Comment STMicroelectronics structure le futur de l'Edge AI

STMicroelectronics est à l'avant-garde de ce mouvement de promotion de l'Edge AI, en fournissant des solutions complètes qui répondent aux différents besoins des ingénieurs et des développeurs dans tous les secteurs d'activité. Avec Edge AI Suite [8], ST propose une large gamme d'outils logiciels adaptés à différents profils d'utilisateurs et niveaux d'expertise, ainsi qu'une vaste collection d'études de cas [9] pour inspirer les développeurs.

Start with edge AI optimized models STM32 model zoo



A collection of application-oriented models optimized for STM32



Hosted on GitHub



Model training scripts

- Scripts to generate and validate



Application code example

- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation



9

Figure 4. Le STM32 Model Zoo contient tous les modèles dont vous avez besoin pour commencer.

Pour les novices en matière d'IA ou de systèmes embarqués, NanoEdge AI Studio (**figure 3**) [10] offre une plateforme intuitive de bout en bout qui permet aux utilisateurs d'importer facilement leurs propres données, d'entraîner des modèles et de les déployer sur des dispositifs embarqués. Cette solution est idéale pour les développeurs qui cherchent à mettre en œuvre rapidement des systèmes de détection d'anomalies, de classification des défauts ou de surveillance de l'environnement basés sur l'IA.

D'autre part, STM32Cube.AI [11] et ST Edge AI Developer Cloud [12] sont conçus pour des utilisateurs plus avancés qui ont besoin d'une plus grande souplesse et d'une meilleure personnalisation. Ils donnent accès à un zoo de modèles (**figure 4**) [13] et à des outils avancés pour la conception, l'entraînement et le déploiement de modèles personnalisés. Qu'il s'agisse de la classification des panneaux de signalisa-

tion dans les véhicules autonomes ou de la reconnaissance faciale dans les systèmes de sécurité, STM32Cube.AI permet aux développeurs d'atteindre les limites de ce qui est faisable avec l'Edge AI.

La vision de ST en matière d'Edge AI repose sur la fourniture de solutions évolutives et adaptables qui répondent aux besoins des ingénieurs et des développeurs travaillant à la pointe de l'innovation. En associant l'expertise matérielle à des logiciels pour embarquer de l'IA, ST contribue à ouvrir de nouvelles perspectives dans des secteurs allant des énergies renouvelables aux maisons intelligentes, en passant par les transports et l'automatisation industrielle. Alors que l'Edge AI continue de gagner en importance, l'utilisation créative des capteurs, l'application de divers algorithmes et l'adaptabilité de solutions telles que celles de STMicroelectronics seront essentielles pour façonner l'avenir des appareils connectés intelligents. ◀

Vf : Helmut Müller — 240620-04



À propos de l'auteur

Vianney Taufour est responsable de l'équipe AI Solutions Developer Engagement chez STMicroelectronics à Grenoble, en France, et possède plus de 25 ans d'expérience dans l'industrie des semi-conducteurs. Il possède une solide expérience en matière de communication, d'image de marque, de marketing et d'électronique, ayant occupé divers postes en recherche et développement, en marketing et en communications d'entreprise. Vianney est titulaire d'un Mastère en génie électrique et microélectronique de l'Institut National des Sciences Appliquées (INSA) de Lyon.

LIENS

- [1] Détection des anomalies dans les performances des moteurs : <https://tinyurl.com/anomaly-detection-case>
- [2] Détection et classification des défauts des moteurs : <https://tinyurl.com/fault-detection-case>
- [3] Détection des arcs en courant continu dans les systèmes électriques (AFCI) : <https://tinyurl.com/arc-fault-detection>
- [4] Capteurs de pression virtuels pour améliorer la sécurité des vélos à assistance électrique : <https://tinyurl.com/e-bikes-panasonic>
- [5] Détection du type de plancher pour les aspirateurs intelligents : <https://tinyurl.com/floor-type-detection>
- [6] Classification des panneaux de signalisation pour les véhicules autonomes : <https://tinyurl.com/traffic-sign-classification>
- [7] Identification faciale et systèmes de sécurité : <https://tinyurl.com/face-identification>
- [8] Edge AI Suite : <https://www.st.com/st-edge-ai-suite>
- [9] ST Edge AI - Études de cas : https://www.st.com/content/st_com/en/st-edge-ai-suite/case-studies.html
- [10] NEAI Studio : <https://stm32ai.st.com/nanoedge-ai/>
- [11] STM32 Cube.AI : <https://stm32ai.st.com/stm32-cube-ai/>
- [12] ST Edge AI Developer Cloud : <https://stm32ai.st.com/st-edge-ai-developer-cloud/>
- [13] STM32 AI model zoo : <https://stm32ai.st.com/model-zoo/>



exploiter la puissance de l'IA en périphérie

Un entretien avec François de Rochebouët de STMicroelectronics

Questions de l'équipe éditoriale d'Elektor

Découvrez comment les solutions innovantes d'IA pour les microcontrôleurs révolutionnent les systèmes embarqués et transforment les industries grâce à des prises de décision en temps réel et économes en énergie. François de Rochebouët (responsable marketing pour les solutions d'IA, chez STMicroelectronics) évoque le rôle essentiel de l'IA en périphérie (Edge AI) dans la construction d'un avenir durable.

Elektor : Merci François de nous avoir rejoints. Pourriez-vous commencer par vous présenter, et nous parler un peu de votre cheminement dans le domaine de l'IA et de l'innovation ?

François de Rochebouët : J'ai toujours été passionné par l'innovation et les limites du possible, en particulier dans le domaine des systèmes embarqués et de l'IA. En tant que cofondateur de Cartesiam, nous avons été les premiers à proposer des solutions d'IA pour les microcontrôleurs. Aujourd'hui, chez STMicroelectronics, je dirige le marketing de l'IA pour les microcontrôleurs STM32. Ma vision est claire : l'IA en périphérie (Edge AI) est cruciale pour l'avenir. Nous avons besoin d'une IA à la périphérie, plus proche de l'endroit où les données sont générées, car les solutions basées sur l'informatique en cloud ne seront pas en mesure de gérer l'énorme quantité de données, générées par notre monde de plus en plus connecté. L'IA durable doit traiter les données localement, ce qui permet d'économiser de l'énergie et d'améliorer la réactivité.

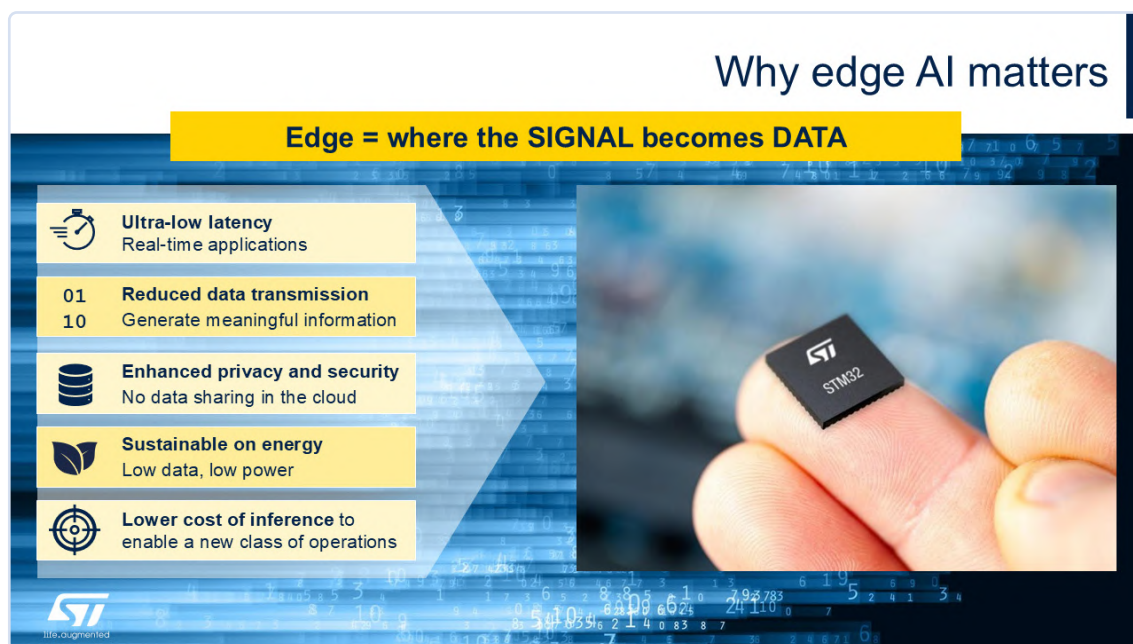


Figure 1. Les avantages de l'IA en périphérie.

Elektor : Pourquoi l'Edge AI est-elle essentielle à la construction d'un avenir durable ?

François : La quantité de données que nous générons aujourd'hui est considérable, et ne va qu'augmenter de manière exponentielle. Le recours à l'IA en nuage (cloud) est synonyme de consommation d'énergie élevée, de problèmes de bande passante et de latence. L'Edge AI (**figure 1**) traite les données localement, à la source, réduisant ainsi la nécessité d'un transfert constant de données vers le cloud. Cela permet non seulement de réduire la consommation d'énergie, mais aussi de prendre des décisions plus rapides et plus efficaces. Pour assurer la durabilité, nous avons besoin de solutions d'intelligence artificielle capables de fonctionner dans des environnements à faible consommation d'énergie, et c'est là où les systèmes embarqués et les microcontrôleurs sont excellents.

Elektor : Quels sont les principaux défis auxquels les développeurs sont confrontés lorsqu'ils conçoivent des solutions d'IA en périphérie ?

François : L'un des principaux défis consiste à changer la façon dont les développeurs envisagent la résolution des problèmes. Dans les systèmes conventionnels, les développeurs s'appuient sur des capteurs pour obtenir des informations directes, mais ils changent d'approche avec l'IA. Elle nous permet de déduire des informations complexes à partir de données apparemment sans rapport. Cela nécessite un nouvel état d'esprit. Les développeurs sont également confrontés à des contraintes telles que la limitation de la puissance de traitement, de la mémoire et de l'efficacité énergétique. Ils doivent donc optimiser les modèles d'IA pour qu'ils s'adaptent à ces paramètres contraignants, tout en garantissant la précision et la fiabilité.

Elektor : Pouvez-vous donner des exemples de la façon dont l'IA embarquée a contribué à résoudre les problèmes des clients de manière créative ?

François : Les vélos à assistance électrique de Panasonic en sont un exemple. Au lieu d'utiliser des capteurs de pression traditionnels pour la surveillance des pneus, ils ont exploité les données de mesure de l'intensité du courant du moteur. En appliquant l'IA, ils ont créé un capteur de pression virtuel qui déduit la pression des pneus à partir des données existantes. Cela a permis de réduire la complexité et le coût du matériel tout en améliorant les performances. Un autre exemple est la détection d'anomalies dans les moteurs fonctionnant à des vitesses différentes. En utilisant l'IA, nous avons aidé nos clients à prévoir les défaillances à un stade avancé et à minimiser les temps d'arrêt, optimisant ainsi le calendrier de maintenances.

Elektor : Comment STMicroelectronics relève-t-elle le défi de concevoir des solutions d'IA en périphérie, avec des ressources limitées ?

François : Nous pensons que c'est la combinaison de ressources matérielles de haute performance et d'un écosystème logiciel adapté qui fait la différence. Sur le plan matériel, nous disposons des microcontrôleurs STM32 [1], leaders dans leur catégorie, qui offrent l'efficacité énergétique et les performances nécessaires à l'Edge AI. Côté logiciel, nous avons rassemblé nos outils et ressources dans la ST Edge AI Suite [2] (**figure 2**), qui rationalise le parcours du développeur, du concept au déploiement, et offre une large collec-

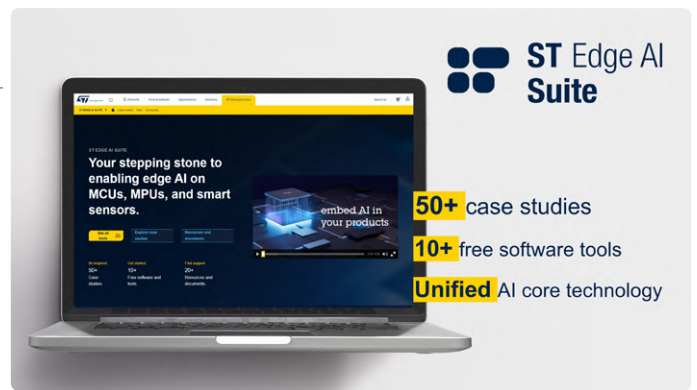


Figure 2. La ST Edge AI Suite, un ensemble d'outils permettant d'intégrer des fonctions d'intelligence artificielle dans les systèmes embarqués.

tion d'études de cas pour l'inspiration. De plus, la STM32 Developer Zone [3] offre un support, des tutoriels et des ressources complets, aidant les développeurs à créer des solutions innovantes avec les STM32. Cette combinaison de matériel et de logiciel permet aux développeurs de relever les défis et de créer de puissantes applications d'IA, quel que soit leur niveau de connaissances.

Elektor : Comment STMicroelectronics soutient-elle les développeurs ayant des niveaux de compétences différents dans le domaine de l'Edge AI ?

François : Chez ST, nous comprenons que les développeurs viennent d'horizons différents, c'est pourquoi nous avons adapté nos outils pour répondre à leurs besoins. Pour les nouveaux venus dans le domaine de l'IA, nous proposons NEAI Studio [4], une plateforme intuitive pour créer et déployer des modèles d'IA avec un minimum d'effort. NEAI Studio supporte une approche BYOD (Bring Your Own Device), avec une solution AutoML générant directement des bibliothèques, optimisées d'algorithmes d'apprentissage automatique les plus adaptés, en fonction des ensembles de données fournis par les utilisateurs.

Pour les utilisateurs avancés, STM32Cube.AI [5], ST Edge AI Core [6] et ST Edge AI Developer Cloud [7] fournissent des solutions d'optimisation et de benchmarking avancées. Les développeurs peuvent également bénéficier de scripts pour réentraîner des modèles à partir du « zoo de modèles » [8] ou de leurs propres modèles, en adoptant l'approche BYOM (Bring Your Own Model). Ce support complet permet aux développeurs d'intégrer leurs modèles pré-entraînés dans l'écosystème, en vue d'une optimisation et d'un déploiement plus poussés. De cette manière, que vous soyez un maker, un ingénieur chevronné dans les systèmes embarqués ou un data scientist, nous avons des outils qui correspondent à votre niveau d'expertise, et vous permettent de tirer le maximum de votre productivité et de votre innovation.

Elektor : Quel est le rôle du matériel dans les cas d'utilisation plus complexes de l'IA, comme la vision par ordinateur ?

François : Le matériel joue un rôle important, d'autant plus que les cas d'utilisation de l'IA deviennent de plus en plus complexes. Par exemple, la vision par ordinateur nécessite une puissance de calcul et une mémoire importantes. C'est là qu'intervient le matériel spécialisé, comme notre ST Neural-ART accelerator, une unité de traitement neuronal (NPU) conçue pour les microcontrôleurs STM32. Il est sans précédent en termes de puissance de calcul et d'efficacité énergétique, permettant aux développeurs d'implémenter des algorithmes d'IA plus exigeants, tels que la détection d'objets en temps réel, tout en conservant une faible consommation d'énergie.

From DMIPS to TOPS, the paradigm shift Opening a new range of embedded AI applications

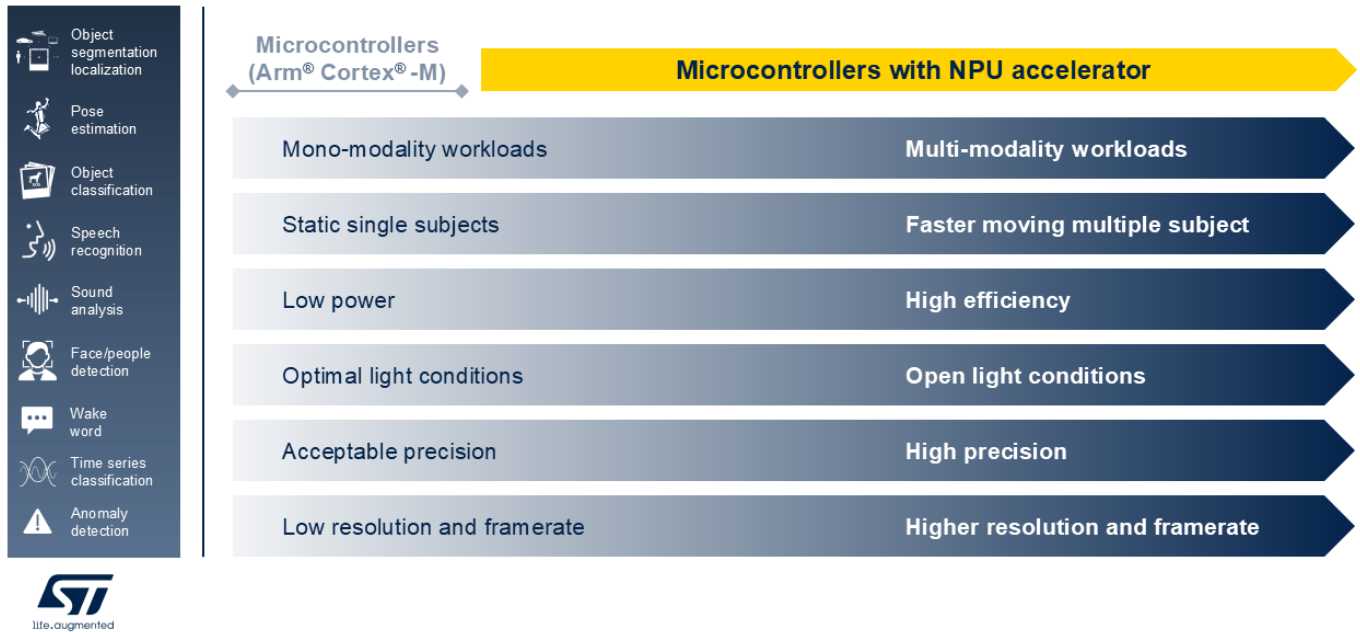


Figure 3. Le NPU accelerator offre une plus grande puissance de calcul pour les tâches d'intelligence artificielle, tout en maintenant une faible consommation d'énergie.

Elektor : Pouvez-vous nous en dire plus sur le ST Neural-ART accelerator ?

François : Le ST Neural-ART accelerator change la donne pour les applications d'IA en périphérie. Il s'agit d'un accélérateur matériel (NPU) conçu spécifiquement pour les réseaux neuronaux, qui offre un gain de performance de 600 fois supérieur environ par rapport à notre microcontrôleur STM32H7, tout en maintenant la consommation d'énergie à un niveau minimum (**figure 3**). Cet équilibre est essentiel pour les cas d'utilisation de l'IA dans les appareils alimentés par batterie, tels que les vêtements, les caméras intelligentes ou les capteurs industriels. Le NPU permet aux développeurs de déployer des modèles d'IA plus complexes, comme ceux nécessaires à la vision par ordinateur ou à la prise de décision en temps réel, directement sur des microcontrôleurs à faible consommation. Nous venons de publier un livre blanc pour en savoir plus à ce sujet et nous invitons les gens à le lire pour en savoir plus.

Elektor : Comment STMicroelectronics envisage-t-elle l'évolution de l'Edge AI ?

François : L'avenir de l'IA en périphérie est incroyablement prometteur. Nous allons voir plus d'appareils intelligents prendre des

décisions en temps réel, avec moins de dépendance à l'égard de l'infrastructure en cloud. Ce sera la clé pour construire des systèmes plus durables et plus réactifs, en particulier dans des domaines tels que les villes intelligentes, la santé et l'automatisation industrielle. Chez ST, nous nous concentrons sur la fourniture d'outils, de matériel et d'écosystèmes qui permettront aux développeurs d'innover et de repousser les limites de ce qui est possible avec l'IA en périphérie.

Elektor : Finalement, que pouvons-nous attendre de STMicroelectronics lors du prochain STM32 Summit ?

François : Le STM32 Summit [9], qui se tiendra le 10 décembre 2024, sera une formidable occasion pour les développeurs et les ingénieurs d'en savoir plus sur notre vision de l'Edge AI. Nous dévoilerons de nouveaux développements passionnants, notamment un aperçu de notre future feuille de route en matière d'IA, et de la manière dont nous soutenons davantage la communauté avec des outils et des ressources. J'invite tout le monde à se joindre à l'émission, pour en savoir plus sur notre stratégie Edge AI, et pour être aux premières loges pour voir comment ST stimule l'innovation dans ce domaine. ◀

VF : Laurent Rauber — 240621-04

LIENS

- [1] Les microcontrôleurs STM32 : <https://www.st.com/stm32>
- [2] ST Edge AI suite : <https://www.st.com/st-edge-ai-suite>
- [3] STM32 Developer Zone : https://www.st.com/content/st_com/en/stm32-mcu-developer-zone.html
- [4] NEAI Studio : <https://stm32ai.st.com/nanoedge-ai/>
- [5] STM32 Cube.AI : <https://stm32ai.st.com/stm32-cube-ai/>
- [6] ST Edge AI Core : <https://www.st.com/en/development-tools/stedgeai-core.html>
- [7] ST Edge AI Developer Cloud : <https://stm32ai.st.com/st-edge-ai-developer-cloud/>
- [8] STM32 AI model zoo : <https://stm32ai.st.com/model-zoo/>
- [9] STM32 Summit : <https://www.st.com/stm32-summit>

horloge en VHDL réalisée avec ChatGPT

Jean-François Simon (Elektor)

Jusqu'à présent, je n'avais jamais utilisé de FPGA, ni les langages Verilog ou VHDL. Découvrez dans cet article comment ChatGPT m'a aidé à créer mon premier projet basé sur un FPGA en partant de zéro.

Il y a quelque temps, je me suis procuré une carte DE10-Lite du fabricant Terasic (**figure 1**), contenant un FPGA de la série Intel MAX 10, quelques afficheurs 7-segments, des interrupteurs à glissière, des boutons poussoirs avec leur circuit anti-rebond, des LED, etc. En somme, une carte sympathique et bien équipée pour débuter, mais que j'avais un peu délaissée. L'article de Theo Mulder publié récemment dans Elektor [1] m'y a fait repenser. Cette édition spéciale IA est l'occasion parfaite pour me lancer ! J'ai choisi d'utiliser le langage VHDL ; reportez-vous à l'encadré dédié dans [1] pour vous aider à choisir entre VHDL et Verilog.

Parmi les projets de magazines d'électronique qui apportent toujours, décennie après décennie, une inexplicable satisfaction, les horloges figurent en bonne place. Il y a un côté rassurant à observer l'écoulement régulier des secondes et le clignotement des LED rouges. La DE10-Lite, d'une taille compacte d'environ 8 cm x 10 cm, contient justement un afficheur 7 segments à 6 chiffres et une petite protection en plexiglas. La situation est donc idéale pour faire une horloge à LED ! Evidemment, il est un peu exagéré d'utiliser un composant en boîtier BGA à 484 broches contenant 50 000 éléments logiques pour fabriquer l'équivalent d'une pendule de salle de bains à 5 euros. Mais laissons cela de côté, l'important est d'apprendre et de s'amuser ! Si vous connaissez d'autres exemples amusants d'abus de moyens, faites-le nous savoir dans le courrier des lecteurs.

Petit avertissement

Cet article contient quelques points de détail, mais ne pourra pas être un tutoriel de bout en bout. Ce serait beaucoup trop long et ennuyeux.



Figure 1. La carte de développement DE10-Lite.

De plus, même si vous tapez *exactement* les mêmes « prompts » que moi dans ChatGPT, je ne peux pas garantir que les réponses que vous obtiendrez seront les mêmes que celles que j'ai reçues, selon que vous utilisez la version gratuite ou payante, selon le modèle (ChatGPT 4, 4o, 4o mini), etc. Voyons plutôt cela comme un retour d'expérience. Je vous invite à essayer vous-même d'utiliser une IA comme ChatGPT pour vos projets ! Il n'y a pas de raison de se limiter aux FPGA, ou à la carte DE10-Lite en particulier, ou à l'utilisation de Quartus, ou au langage VHDL. Suivez votre inspiration et soyez persévérant ! Ce type de projet n'avance qu'en résolvant successivement des problèmes plus ou moins grands.

Premiers pas

On dit souvent que la première chose à faire pour se familiariser avec le matériel et le logiciel est de faire clignoter une LED. Mais, sur un FPGA, il n'y a pas de fonction prête à l'emploi comme `delay()` sur Arduino. La première étape sera donc encore plus modeste : allumer une des LED de la carte lorsque l'on appuie sur l'un des boutons de la carte. Je me rends donc sur le site ChatGPT et lui donne le *prompt* (instruction) suivant :

Ecris-moi un code VHDL très simple pour allumer une LED lorsqu'un bouton est pressé sur une carte FPGA DE10-lite.

La réponse de ChatGPT est présentée dans le **Listing 1**. Pour pouvoir implémenter ce programme dans le FPGA, le chemin est encore long. Voici un aperçu des étapes à suivre. Tout d'abord, téléchargez et installez Quartus Prime Lite [2]. Créez un nouveau projet vide en cliquant sur *New Project Wizard*. Lors de la création du projet, donnez-lui un nom (j'ai choisi « test1 ») et, dans la liste déroulante *Devices*, sélectionnez le FPGA que nous allons utiliser en particulier : le 10M50DAF484C7G. Ensuite, créez un nouveau fichier VHDL (*File* → *New*) puis copiez et collez le code dans ce fichier.

Ensuite, il faut indiquer à Quartus les broches d'entrées-sorties que nous souhaitons utiliser. Je recommande de télécharger le *System CD* fourni par Terasic [3], qui contient de la documentation et des fichiers d'exemple. Le *User Manual* sera d'une grande aide, et



Listage 1. ChatGPT a généré ce code simple pour allumer une LED.

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;

entity LED_Button is
    Port (
        KEY : in std_logic; -- Input button
        LED : out std_logic -- Output LED
    );
end LED_Button;

architecture Behavioral of LED_Button is
begin
    -- Process to control LED
    process(KEY)
    begin
        if (KEY = '0') then -- Active low button (pressed)
            LED <= '1';      -- Turn on LED
        else
            LED <= '0';      -- Turn off LED
        end if;
    end process;
end Behavioral;
```

indique notamment la correspondance entre les numéros de broches et les composants de la carte. J'utilise ici le bouton *KEY0*, connecté à la broche *PIN_B8* du FPGA, et la LED *LEDR0* (*PIN_A8*). La façon la plus simple de procéder est d'utiliser le *Pin Planner*, dans le menu *Assignments*. Avant cela, il faut indiquer au système quels sont les signaux présents dans le projet. Une *Compilation* complète (a. dans la **Figure 2**) conviendrait pour cela, mais elle est assez lente. Cliquez donc sur *Start Analysis & Elaboration* (b.), une sorte de compilation partielle, qui fera l'affaire et est plus rapide.

La première erreur ne se fait pas attendre : *Error (12007): Top-level design entity "test1" is undefined*. Par défaut, Quartus donne à l'entité de plus haut niveau le même nom que le nom du projet

(ici « test1 »). Or, le code fourni par ChatGPT décrit une entité nommée *LED_Button*. Dans le menu *Assignments* → *Settings* → *General* il est possible de spécifier le nom de l'entité de plus haut niveau ; entrez *LED_Button* au lieu de *test1*. Ensuite, l'analyse et l'élaboration peuvent s'exécuter correctement. Le *Pin Planner* est ensuite utilisé pour spécifier les emplacements *PIN_B8* pour *KEY* et *PIN_A8* pour *LED*. Enfin, effectuez une compilation complète (cliquez sur le triangle bleu ou CTRL-L).

Programmation du FPGA

Pour charger la conception fraîchement compilée dans le FPGA, utiliser le *Programmer* (dans le menu *Tools*). Après la programmation, on peut voir que toutes les LED de la carte, à l'exception de *LEDR0*, sont allumées. Un appui sur *KEY0* allume cette dernière. Victoire ! Pour ceux de nos lecteurs qui voudraient tenter l'expérience, je recommande la vidéo [4], qui donne des informations utiles pour ces premiers pas. L'auteur de la vidéo utilise un FPGA Cyclone IV au lieu d'un MAX 10, pensez donc à adapter ses instructions à votre cas précis.

Les neuf autres LED (*LEDR1* à *LEDR9*) restent allumées car, par défaut dans Quartus, le paramètre suivant est activé : *Reserve all unused pins as inputs tri-stated with weak pull-up*. Ces résistances de tirage limitent le courant à une valeur très faible, mais suffisante pour allumer les LED. Ce paramètre peut être modifié dans le menu *Assignments* → *Device* → *Device and Pin Options* → *Unused Pins*. Par exemple, il peut être réglé sur *As output driving ground* pour forcer un état bas et éteindre les LED. Après recompilation et reprogrammation, elles sont effectivement éteintes, mais certains segments de l'afficheur à 7 segments sont maintenant allumés. En fait, ces segments sont actifs à l'état bas, comme le précise la documentation. Une autre approche consisterait à forcer les LED ou les segments à 1 ou 0 dans le code. Nous verrons cela plus loin.

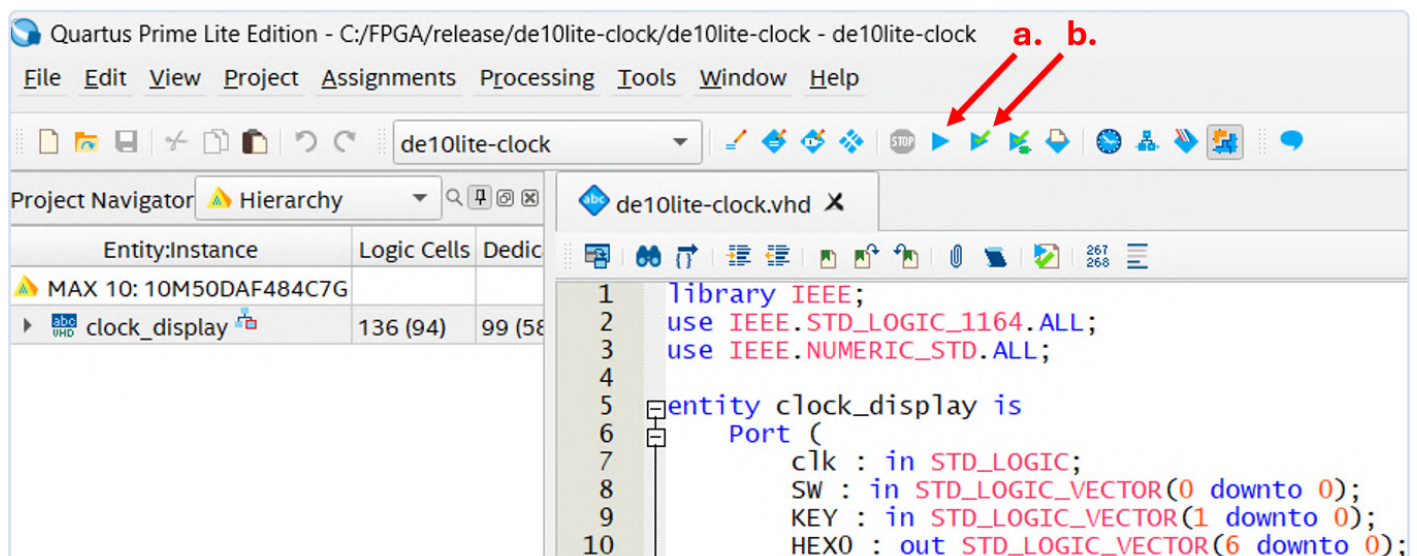


Figure 2. Analyse, élaboration et compilation dans Quartus.



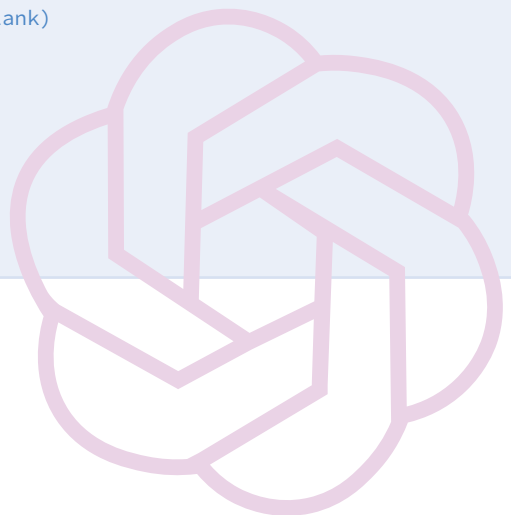
Listage 2. Commande de l'afficheur à 7 segments.

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
use IEEE.STD_LOGIC_ARITH.ALL;
use IEEE.STD_LOGIC_UNSIGNED.ALL;

entity seven_segment_display is
    Port (
        clk : in STD_LOGIC; -- Clock signal
        digit : in STD_LOGIC_VECTOR(3 downto 0); -- 4-bit input for the digit (0-9)
        segments : out STD_LOGIC_VECTOR(6 downto 0) -- Outputs for the 7-segment display
    );
end seven_segment_display;

architecture RTL of seven_segment_display is
    signal seg_reg : STD_LOGIC_VECTOR(6 downto 0); -- Register for storing segment data
begin
    -- Sequential logic to update the display register on every clock edge
    process(clk)
    begin
        if rising_edge(clk) then
            case digit is
                when "0000" => -- Digit 0
                    seg_reg <= "0000001"; -- a,b,c,d,e,f are on, g is off
                when "0001" => -- Digit 1
                    seg_reg <= "1001111"; -- b,c are on
                when "0010" => -- Digit 2
                    seg_reg <= "0010010"; -- a,b,d,e,g are on
                when "0011" => -- Digit 3
                    seg_reg <= "0000110"; -- a,b,c,d,g are on
                when "0100" => -- Digit 4
                    seg_reg <= "1001100"; -- b,c,f,g are on
                when "0101" => -- Digit 5
                    seg_reg <= "0100100"; -- a,c,d,f,g are on
                when "0110" => -- Digit 6
                    seg_reg <= "0100000"; -- a,c,d,e,f,g are on
                when "0111" => -- Digit 7
                    seg_reg <= "0001111"; -- a,b,c are on
                when "1000" => -- Digit 8
                    seg_reg <= "0000000"; -- All segments are on
                when "1001" => -- Digit 9
                    seg_reg <= "0000100"; -- a,b,c,d,f,g are on
                when others =>
                    seg_reg <= "1111111"; -- All segments off (blank)
            end case;
        end if;
    end process;

    -- Combinational logic to drive the segments
    segments <= seg_reg;
end RTL;
```





Listage 3. Entité de plus haut niveau pour afficher un chiffre.

```
library IEEE;
use IEEE.STD_LOGIC_1164.ALL;
use IEEE.STD_LOGIC_ARITH.ALL;
use IEEE.STD_LOGIC_UNSIGNED.ALL;

entity top_level is
  Port (
    clk : in STD_LOGIC; -- Clock input
    segments : out STD_LOGIC_VECTOR(6 downto 0) -- Outputs to the 7-segment display
  );
end top_level;

architecture Behavioral of top_level is
  signal digit_signal : STD_LOGIC_VECTOR(3 downto 0); -- 4-bit digit signal
begin
  -- Instantiate the seven_segment_display module
  U1 : entity work.seven_segment_display
    Port map (
      clk => clk,
      digit => digit_signal,
      segments => segments
    );

  -- Set the digit to "4"
  digit_signal <= "0100"; -- Binary for 4
end Behavioral;
```

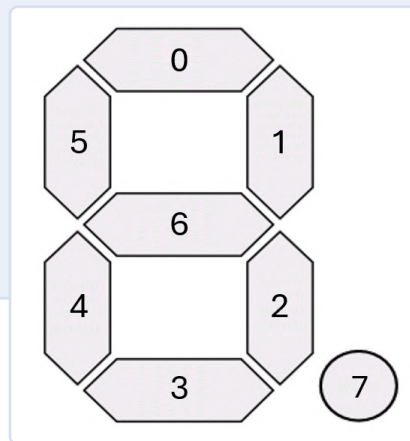


Figure 3. Adressage des différents segments de l'écran.

Quelques chiffres

Voyons maintenant comment afficher des chiffres sur l'afficheur à 7 segments. J'ai saisi le prompt suivant :

Ecris-moi un code VHDL capable d'écrire n'importe quel chiffre de 0 à 9 sur l'afficheur 7 segments intégré d'une carte de développement DE10-lite.

La réponse de ChatGPT comprend du code, ainsi que des explications que j'omets ici faute de place ; le code est visible dans le **listage 2**. Pour faciliter la réutilisation de ce code, je l'ai copié-collé dans un nouveau fichier VHDL appelé (par exemple) *7segment.vhd*. N'oubliez pas de cliquer sur *Add Current File to Project* dans le menu *Project*. Le code fourni est générique, et n'inclut pas le nécessaire pour afficher un chiffre en particulier afin de vérifier le fonctionnement. Je complète donc ma demande : *J'ai collé ce code dans un fichier 7segment.vhd. Montre-moi quel code utiliser dans l'entité de plus haut niveau pour afficher le chiffre « 4 »*. La réponse est donnée dans le **listage 3**.

Voyons cela. L'entité de plus haut niveau, *top_level*, possède une entrée *clk* pour le signal d'horloge, qui est transmis à l'entité *seven_segment_display*. Le signal *digit_signal* est utilisé pour contenir la valeur du chiffre, encodé en binaire sur 4 bits. Pour le chiffre 4, l'équivalent binaire est 0100. L'entité *seven_segment_display* est instanciée et « mappée » avec les signaux de niveau supérieur (*clk*, *digit_signal* et *segments*). Le *mapping* consiste à relier les signaux entre eux. Par exemple, *digit => digit_signal* connecte le signal de niveau supérieur *digit_signal* au port d'entrée *digit* de l'entité *seven_segment_display*, et permet ainsi au premier de contrôler le second. Les signaux *clk* et *segments* sont

également mappés de la même manière. Enfin, le nombre à afficher est défini à l'aide de l'instruction *digit_signal <= "0100"*. Notez la différence entre l'opérateur d'affectation *<=* et l'opérateur d'association *=>*, utilisé notamment pour le *mapping* des signaux.

Avant de pouvoir contempler le résultat, il faut ajouter des *Pin Constraints*, afin d'obliger Quartus à utiliser les broches adéquates, reliées aux segments de l'afficheur. Les segments sont généralement marqués d'une lettre allant de « a » à « g », mais Terasic a choisi de les numéroter de 0 à 6 (**figure 3**). Le manuel d'utilisation nous donne la correspondance des broches ; un extrait est présenté dans le **tableau 1**. Pour cet exemple d'affichage à un chiffre, sept connexions sont nécessaires, il serait donc envisageable d'utiliser le *Pin Planner* à cette fin. Cependant, pour l'horloge finale, nous aurons besoin des six chiffres et des points décimaux, soit 48 connexions.

Il est donc plus pratique de procéder autrement. Chaque contrainte de broche correspond à deux lignes dans le *Fichier de Configuration Quartus* (fichier QSF) : une ligne *set_instance_assignment*, qui indique à Quartus le type d'E/S, et une ligne *set_location_assignment*, qui indique l'affectation de la broche elle-même. Il est beaucoup plus simple de copier ces contraintes de broches à partir d'un fichier QSF existant. Ouvrir le fichier *DE10_LITE_Default.qsf* dans le dossier *Demonstrations/Default* du CD Système. Copier les 96 lignes dédiées à l'affichage 7 segments :

Tableau 1. Brochage de l'afficheur.

Signal Name	FPGA Pin No.	Description	I/O Standard
HEX00	PIN_C14	Seven Segment Digit0[0]	3.3-V LVTTL
HEX01	PIN_E15	Seven Segment Digit0[1]	3.3-V LVTTL
HEX02	PIN_C15	Seven Segment Digit0[2]	3.3-V LVTTL
HEX03	PIN_C16	Seven Segment Digit0[3]	3.3-V LVTTL
HEX04	PIN_E16	Seven Segment Digit0[4]	3.3-V LVTTL
HEX05	PIN_D17	Seven Segment Digit0[5]	3.3-V LVTTL
HEX06	PIN_C17	Seven Segment Digit0[6]	3.3-V LVTTL
HEX07	PIN_D15	Seven Segment Digit0[7], DP	3.3-V LVTTL
HEX10	PIN_C18	Seven Segment Digit1[0]	3.3-V LVTTL
HEX11	PIN_D18	Seven Segment Digit1[1]	3.3-V LVTTL
HEX12	PIN_E18	Seven Segment Digit1[2]	3.3-V LVTTL
etc.	etc.	etc.	etc.

```
set_instance_assignment -name IO_STANDARD "3.3-V LVTTL"
-to HEX0[0]
(...)
(...)
set_location_assignment PIN_L19 -to HEX5[7]
```

Collez ces lignes dans le fichier QSF de votre projet en cours. Le code des listages 2 et 3 nécessite un signal d'horloge. Copiez donc aussi le bloc de 6 lignes dédié à l'horloge interne du FPGA, indiqué par l'en-tête `# CLOCK`. Pour que cela fonctionne, vous devrez renommer `clk` en `MAX10_CLK1_50` et `segments` en `HEX0` dans le fichier VHDL principal (l'entité de plus haut niveau) :

```
entity top_level is
Port (
MAX10_CLK1_50 : in STD_LOGIC;
HEX0 : out STD_LOGIC_VECTOR(6 downto 0)
);
end top_level;
```

Il faut également renommer les mêmes signaux au moment de l'instanciation de notre affichage à 7 segments afin de faire correspondre correctement les signaux d'une entité à l'autre :

```
U1 : entity work.seven_segment_display
Port map (
clk => MAX10_CLK1_50,
digit => digit_signal,
segments => HEX0
);
```

Enfin, il est temps de procéder à une nouvelle compilation. En principe, le code devrait compiler sans erreurs. Si ce n'est pas le cas, signalez ces erreurs à ChatGPT pour qu'il vous aide à trouver le problème. Après la reprogrammation, on obtient l'affichage de la **figure 4**. Ce n'est pas un « 4 » ! Que s'est-il passé ? Inspectons le code responsable de la génération des signaux pour les segments :

```
case digit is
when "0000" => -- Digit 0
seg_reg <= "0000001"; -- a,b,c,d,e,f are on, g is off
```

Dépannage

Les segments sont actifs à l'état bas ; le commentaire après le double tiret implique que, selon ChatGPT, la série de bits est dans l'ordre suivant : « abcdefg ». Or, selon le *User Manual*, le segment « 0 » (ou « a ») correspond au signal HEX0[0]. En VHDL, les bits sont numérotés en commençant par la gauche, le bit le plus à gauche ayant l'indice le plus élevé et le bit le plus à droite ayant l'indice le plus bas. En d'autres termes, HEX0[0], ou « a », devrait se trouver tout à droite : « gfedcba ». Il faut donc soit en informer ChatGPT et le laisser corriger le problème, soit inverser manuellement toutes les chaînes de bits de gauche à droite : 0000001 devient 1000000, 1001111 devient 1111001, etc. Une fois cela fait, après une nouvelle recompilation et reprogrammation, le chiffre 4 apparaît correctement.

Comme vous le voyez, c'est un peu laborieux. Mais il ne faut pas se décourager, il y a eu du progrès. Nous disposons maintenant d'une entité d'affichage permettant d'afficher le chiffre de notre choix sur le

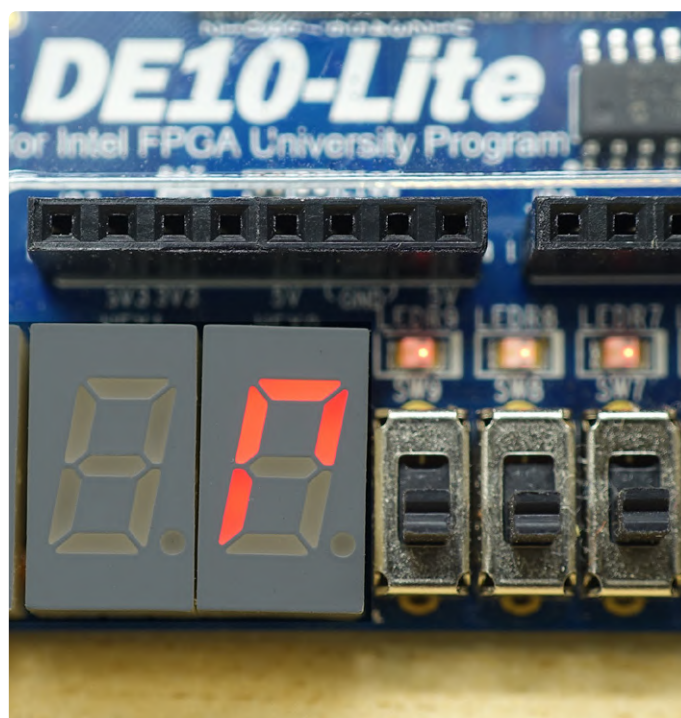


Figure 4. Essai d'affichage d'un chiffre...

Quelques problèmes

A différents moments durant le projet, ChatGPT a rencontré les problèmes suivants :

- Génération de code utilisant des fonctions sans les avoir correctement importées, ce qui entraînait des erreurs telles que « object 'to_unsigned' is used but not declared. »
- Déclarer un process tel que `process(clk_1hz, reset)` et utiliser par la suite des instructions comme `rising_edge(KEY(0))` à l'intérieur de celui-ci, ce qui provoque une erreur, car KEY ne figure pas dans la liste de sensibilité (la première série de parenthèses dans la déclaration).
- Générer des verrous (latches). Je vous laisserai chercher pourquoi les latches sont particulièrement problématiques dans les projets à FPGA, mais en résumé, ils peuvent conserver un état sans aucune entrée d'horloge, ce qui les rend imprévisibles. Quartus refuse de compiler du code qui génère des verrous et produit des messages d'erreur.
- Tenter de modification des signaux à partir de deux processus différents, ce qui entraîne des messages d'erreur.
- Difficultés pour gérer deux horloges distinctes – où l'horloge principale (rapide) est utilisée pour détecter les pressions sur les boutons avec une certaine réactivité, tandis que l'horloge 1 Hz ne fait qu'incrémenter les compteurs de temps. ChatGPT a parfois confondu les deux, ce qui a entraîné un affichage incohérent, ou bien une très mauvaise réactivité à l'appui sur les boutons.
- Création de processus sensibles à deux horloges en même temps. Quartus a du mal à déterminer quel front d'horloge doit déclencher les événements.
- Répondre « Pour résoudre ce problème, nous allons utiliser un flag (drapeau) », et fournir ensuite un code qui n'implémente pas réellement de flag. Également, donner des explications incorrectes sur certains comportements.
- Avoir des problèmes de logique. Exemple : l'horloge se réinitialise à 00:00 après 19:59:59 au lieu de 23:59:59.
- Difficultés à résoudre les erreurs de logique lorsqu'elles sont présentes dans plusieurs parties du code à la fois. ChatGPT a tendance à ne corriger l'erreur qu'à un seul endroit, ce qui ne résout pas le problème entièrement et conduit à des cycles répétitifs.

Résoudre tous ces problèmes s'est avéré être une tâche difficile, mais assez amusante et très instructive !

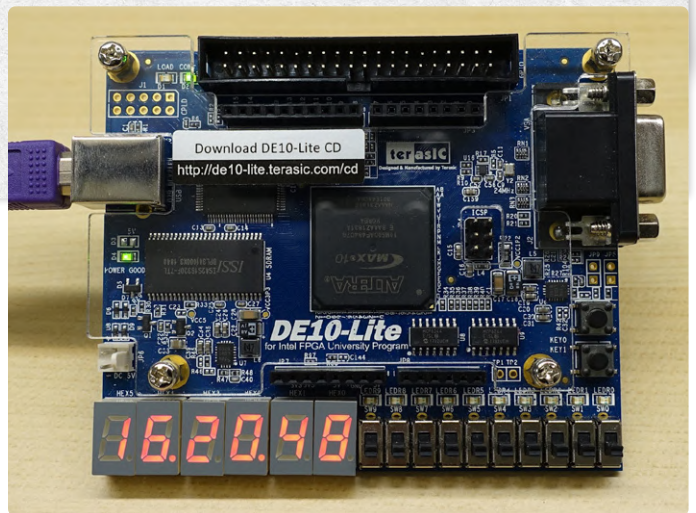


Figure 5. L'horloge terminée.

premier emplacement de l'afficheur. Les *pin constraints* sont en place permettant d'utiliser aisément les cinq autres chiffres. À présent, il n'y a plus qu'à continuer à dialoguer avec ChatGPT pour avancer ! Personnellement, j'aime bien procéder par étapes, je lui ai donc demandé de procéder par petites évolutions, afin d'être certain de bien comprendre. Cependant, il est tout à fait possible de lui faire votre demande finale (ici, une horloge comptant les heures, minutes et secondes sur l'afficheur 7-segments à 6 chiffres d'une carte DE10-Lite) dès le début et de tout faire d'un coup. Cela serait très certainement bien plus rapide.

Une vraie horloge ?

Voici les étapes que j'ai suivies par la suite, sous la forme des *prompts* donnés à ChatGPT :

- *Adapte le code pour afficher « 123456 » sur un afficheur à sept segments de 6 chiffres. Les registres d'affichage sont HEX0 à HEX5, un pour chaque chiffre.*
- *Excellent, merci. Maintenant, écris le code VHDL pour faire une horloge utilisant cet affichage et indiquant l'heure au format hh:mm:ss.*
- *Maintenant, donne-moi le code que je dois ajouter pour pouvoir régler l'horloge. Je voudrais que l'appui sur KEY[0] incrémente les minutes et KEY[1] incrémente les heures. Pas besoin d'intégrer un anti-rebond dans le code, il y a un circuit dédié sur la carte pour cela.*

Après avoir passé un certain temps à chasser les bugs (voir l'encadré **Quelques problèmes**), en dialoguant avec ChatGPT, et aussi en cherchant sur Google, un code fonctionnel pour une horloge complète a pu être réalisé. Le code final est trop long pour être reproduit ici, je vous invite à le consulter en ligne [5]. L'horloge terminée (**figure 5**) peut être vue en action dans la vidéo [6].

D'abord, l'entité `clock_display` est déclarée, ainsi que toutes ses entrées et sorties. Ensuite, dans le bloc `architecture`, les signaux (analogues aux variables en programmation logicielle) nécessaires à son fonctionnement sont créés : ceux-ci servent à stocker l'heure courante et l'état des boutons. Un premier `process` est utilisé pour créer une horloge de 1 Hz à partir de l'horloge de 50 MHz embarquée sur la DE10-Lite, en comptant jusqu'à 49.999.999 ; lorsque cette valeur est atteinte, le signal `clk_1hz` passe à l'état haut pendant un cycle d'horloge, puis revient à l'état bas.

Un second `process` se charge d'incrémenter les signaux pour les secondes, les minutes et les heures, ainsi que de détecter les pressions sur les boutons. L'heure courante est incrémentée toutes les secondes et les afficheurs à sept segments affichent chaque chiffre à l'aide d'une instance distincte de l'entité `seven_segment_display`. Chaque instance reçoit le chiffre à afficher (sur 4 bits) à son entrée et pilote les segments correspondants pour représenter ce chiffre. Par ailleurs, les deux boutons KEY[0] et KEY[1] sont utilisés pour régler l'heure manuellement. KEY[0] incrémente les minutes et KEY[1] incrémente les heures. L'horloge est remise à zéro lorsque l'interrupteur SW[0] est actionné. Les points décimaux sur l'écran sont utilisés pour séparer les heures, les minutes et les secondes. Ceux des chiffres 0, 1, 3 et 5



sont forcés à 1 afin de les maintenir éteints tandis que les autres sont synchronisés avec le signal d'horloge le plus lent avec les instructions `dpX <= clk_separators`. De cette manière, tous les segments de l'affichage sont explicitement utilisés ; cela vous permet de choisir l'option *Reserve all unused pins as output driving ground* pour éteindre les LED *LEDR0* à *LEDR9* sans perturber l'affichage.


Notez que, par défaut, lorsque le FPGA est programmé, le design est stocké dans la mémoire volatile du FPGA. A la prochaine mise sous tension, le programme de démonstration Terasic est rechargé à partir de la mémoire Flash interne du FPGA. Pour que celui-ci démarre l'horloge directement à chaque mise sous tension, vous devez indiquer au *Programmer* d'utiliser le fichier POF au lieu du fichier SOF ; la programmation est légèrement plus longue.

Au final, l'horloge fonctionne très bien. Au démarrage, si l'interrupteur en bas à droite est en position basse, elle commence immédiatement à compter les secondes. Pour la mettre à l'heure, actionner l'interrupteur pour la maintenir à 00:00:00. Puis, en utilisant une autre horloge comme référence, actionnez à nouveau l'interrupteur au début de la minute suivante pour commencer à compter les secondes. Les secondes sont maintenant réglées. Appuyez ensuite sur les touches KEY0 et KEY1 pour régler respectivement les minutes et les heures.

Une approche intéressante

Ce projet utilisant ChatGPT pour concevoir une horloge en VHDL a été une expérience fascinante. Il s'agit d'une approche nouvelle par rapport à la méthode traditionnelle consistant à apprendre les bases à l'aide de livres et d'autres ressources et à commencer à programmer en partant de zéro. Avec ChatGPT, vous pouvez commencer avec un code globalement fonctionnel, qui ne contient généralement aucune erreur de syntaxe, ce qui permet d'obtenir des résultats visibles rapidement – un grand avantage pour les impatientes. Cependant, une attention particulière est requise pour résoudre les erreurs de logique et les autres défis qui se présentent en cours de route. Bien entendu, cette méthode n'est pas destinée à remplacer totalement l'apprentissage traditionnel, mais elle le complète efficacement. Les ressources listées à la fin de [1] m'ont été bien utiles. ChatGPT (ou toute autre IA) est un atout supplémentaire dans votre boîte à outils. Il peut présenter assez clairement des informations complexes, ce qui est très utile lorsque les recherches traditionnelles sur internet s'avèrent décevantes (faute de connaître le bon mot-clé par exemple). Il m'a également bien aidé pour analyser du code, en particulier du code non commenté trouvé sur Internet, et pour m'expliquer les algorithmes sous-jacents.

Ceci ne prétend pas être la meilleure façon de faire une horloge ; ce genre de projet peut être abordé de multiples façons et je vous

encourage à essayer vos propres solutions. D'autres outils utilisant l'IA, comme Claude [7], ainsi que des assistants intégrés à votre EDI peuvent être utilisés. Telle qu'elle est présentée ici, l'horloge perd l'heure lorsqu'on la met hors tension. L'ajout d'une batterie de secours constituera un petit exercice pour nos lecteurs. Et maintenant, quoi d'autre ? Pour poursuivre l'apprentissage, il me reste à entreprendre d'autres projets à FPGA. Une étape importante sera d'utiliser la simulation, car il s'agit d'une partie essentielle du développement sur FPGA que je n'ai pas abordée ici. Un avantage inhérent aux FPGA est la portabilité du code : je pourrais ainsi essayer de transférer cette horloge sur une carte plus petite et moins coûteuse, comme la iCE40HX1K-EVB d'Olimex. Cela impliquerait de concevoir un petit PCB pour ajouter des afficheurs à 7 segments, et apprendre les bases du multiplexage en VHDL, car ladite carte n'a pas d'afficheurs intégrés ni même assez de broches d'E/S pour contrôler tous les segments directement. Est-ce bien nécessaire ? Sans doute pas, mais instructif, sûrement ! 

240602-04

Questions ou commentaires ?

Envoyez un courriel à l'auteur (jean-francois.simon@elektor.com), ou contactez Elektor (redaction@elektor.fr).



À propos de l'auteur

Jean-François Simon a une passion de longue date pour l'électronique et s'intéresse à des sujets aussi variés que la conception de circuits, le test et la mesure, le prototypage, la rétro-ingénierie, la radio logicielle, etc. Il aime créer, modifier et améliorer ses outils et autres systèmes. Il aime également la mécanique, l'usinage et tout ce qui est technique. https://x.com/JFS_Elektor



produits

- > **Alchitry Au FPGA Development Board (Xilinx Artix 7)**
www.elektor.fr/19641
- > **Carte de développement FPGA iCEBreaker**
www.elektor.fr/20625

LIENS

- [1] T. Mulder, « FPGA pour les débutants », Elektor 3-4/2024 : <https://www.elektormagazine.fr/230067-04>
- [2] Téléchargement de Quartus Prime Lite : <https://tinyurl.com/downloadquartus>
- [3] System CD de la DE10-Lite : <https://de10-lite.terasic.com/cd>
- [4] Tutoriel : faire clignoter une LED avec un FPGA, étape par étape (vidéo) : <https://www.youtube.com/watch?v=JJ3XkNcLdx8>
- [5] Code source sur GitHub : <https://github.com/jfpsimon/de10lite-clock>
- [6] L'horloge en action (vidéo) : <https://www.youtube.com/watch?v=vdxtk3s0xEA>
- [7] Claude AI : <https://claude.ai>

l'impact réel de l'IA

Sayash Kapoor à propos
des « faux miracles de l'IA »
et plus encore



Sayash Kapoor (doctorant, Université de Princeton)

C. J. Abate (Elektor)

Dans cet entretien, Sayash Kapoor, ingénieur logiciel de l'université de Princeton et ancien employé de Facebook, évoque l'impact sociétal de l'IA et les risques réels qu'elle pose aujourd'hui, tels que les déplacements de main-d'œuvre et la diffusion d'informations erronées. Il explique également comment l'IA peut améliorer la prise de décision humaine et partage les enseignements de son livre *AI Snake Oil*, qui remet en question de nombreuses affirmations exagérées concernant l'intelligence artificielle.

Sayash Kapoor : Avant de fréquenter l'université de Princeton, j'ai travaillé chez Facebook en tant qu'ingénieur logiciel, où j'ai développé l'IA pour la modération de contenu.

C. J. : Vos recherches à l'université de Princeton portent sur l'impact sociétal de l'intelligence artificielle. Quand et comment avez-vous commencé à vous intéresser à ce sujet ?

Sayash : Lorsque j'étais chez Facebook, j'ai pu constater, de visu, l'impact de l'IA sur la société. Je

me suis particulièrement intéressé à la manière dont l'IA peut causer des dommages dans le monde réel, mais aussi à la manière dont elle peut être utilisée pour améliorer les résultats sociétaux. J'ai développé un intérêt particulier pour l'influence des politiques publiques sur l'IA, en voyant comment des régulations comme le RGPD guidaient le fonctionnement des grandes entreprises telles que Facebook.

C. J. : L'IA représente-t-elle un risque existentiel pressant qui devrait préoccuper tous les ingénieurs et les étudiants en EE/ECE ? Ou s'agit-il simplement d'un outil pratique qui, bien que nécessitant une réflexion approfondie, est en fin de compte une force positive ?

Sayash : Au cours des dernières années, nous avons entendu de nombreuses affirmations selon lesquelles l'IA allait anéantir l'humanité. Dans notre livre *AI Snake Oil*, nous consacrons un chapitre entier à cet argument. Nous examinons les nombreuses failles sous-jacentes aux affirmations selon lesquelles l'IA mettra fin à l'humanité et, sur cette base, nous concluons que, même si nous devons prendre au sérieux les risques liés à l'IA, les discussions actuelles sur les menaces existentielles de l'IA surestiment largement l'impact réel de l'IA. Dans le même temps, je ne veux pas minimiser les risques bien réels que l'IA pose aujourd'hui, y compris le déplacement de la main-d'œuvre, la dépendance excessive à l'égard d'informations erronées et les menaces pour la sécurité.

C. J. : Comment envisagez-vous l'avenir des systèmes pilotés par l'IA en termes d'amélioration de la prise de décision humaine dans des tâches complexes ? Quels sont les principaux défis à relever pour créer des modèles d'IA non seulement performants, mais aussi conformes aux valeurs et aux attentes humaines dans diverses applications du monde réel ?

Sayash : L'IA est déjà utilisée pour résoudre de nombreuses tâches complexes. L'un des aspects intéressants de l'histoire de l'IA est que dès qu'une tâche peut être résolue de manière fiable, nous cessons de l'appeler IA. Par exemple, il y a encore quelques décennies, la vérification orthographique était considérée comme un problème difficile à résoudre. Aujourd'hui, des outils comme la correction automatique de l'orthographe se sont estompés en arrière-plan et sont devenus partie intégrante de la vie de tous les jours.

J'espère voir davantage d'outils de ce type fonctionner de manière suffisamment fiable pour être relégués à l'arrière-plan. Nous voyons déjà des outils d'IA capables de modifier une grande partie du travail intellectuel. Ils sont utilisés pour la découverte automatisée de médicaments et comme assistants de codage. Le plus grand défi pour rendre ces outils utiles à un grand nombre de personnes est peut-être d'accroître leur fiabilité, car les systèmes d'IA générative d'aujourd'hui ont un comportement aléatoire et les utilisateurs ne disposent pas vraiment de références pour travailler avec eux. Par exemple, les chatbots peuvent souvent « halluciner », c'est-à-dire fabriquer des réponses incorrectes aux questions des utilisateurs. Si nous parvenons à réduire les erreurs et les hallucinations, l'IA générative pourra être beaucoup plus utile pour les applications conséquentes de l'IA.

C. J. : Avec Arvind Narayanan, vous avez écrit un livre intitulé *AI Snake Oil*. Qu'est-ce qui vous a poussé à écrire ce livre et qu'est-ce qui en fait un ouvrage incontournable ?

Sayash : L'une des plus grandes sources de confusion au sujet de l'IA aujourd'hui est le fait que l'IA est un terme générique. Il est utilisé pour désigner de nombreux types de technologies distinctes qui n'ont pas grand-chose à voir les unes avec les autres. Dans le livre *AI Snake Oil*, nous faisons la distinction entre les différents types d'IA pour souligner où l'IA fonctionne bien, quels types d'IA n'ont pas beaucoup progressé et comment nous pouvons nous attendre à ce que l'avenir de ces types d'IA se concrétise.

Nous expliquons comment l'IA générative, utilisée pour générer du texte, des images, du son, etc., a fait d'énormes progrès au cours de la dernière décennie - bien qu'il y ait certains risques que nous anticipons,

Qu'est-ce que p(doom) ? Qu'est-ce que les membres de la communauté Elektor devraient savoir à ce sujet ?



Drafts

✦ P(doom) est un terme utilisé dans le domaine de la sécurité de l'IA pour décrire la probabilité de résultats catastrophiques résultant de l'intelligence artificielle (IA). Cela peut aller d'une grave perturbation sociétale à l'extinction de l'humanité... Il est important pour la communauté Elektor, en tant qu'experts en technologie et en innovation, d'être conscients des risques potentiels associés à l'IA et de contribuer aux discussions et aux solutions qui peuvent aider à atténuer ces risques.

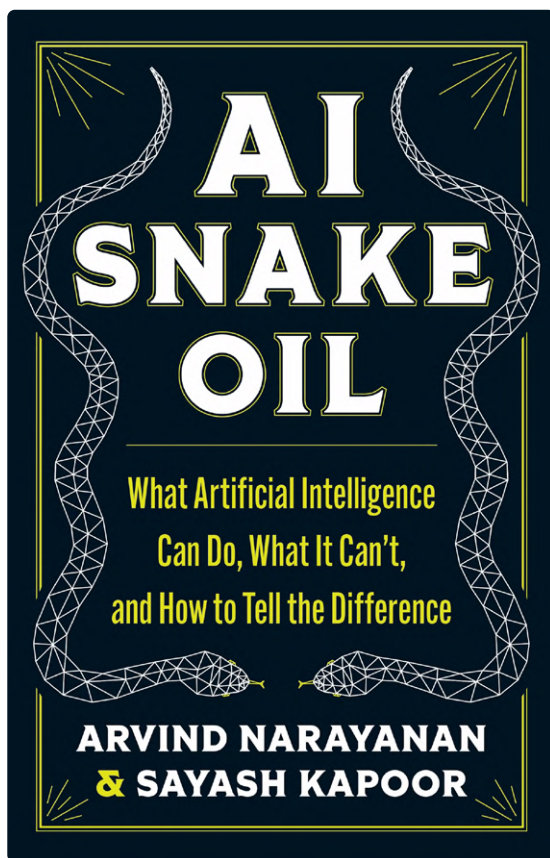
Enter a prompt for Gemini



tels que le déplacement de la main-d'œuvre et la dépendance excessive à l'égard de chatbots incorrects. D'autre part, l'IA prédictive est utilisée pour prédire l'avenir des individus et, sur cette base, prendre des décisions conséquentes à leur sujet - par exemple, si un accusé doit être libéré avant son procès ou si un candidat à un emploi doit être invité à un entretien.



*Si nous parvenons à
réduire les erreurs et les
hallucinations,
l'IA générative pourra être
beaucoup plus utile pour
les applications conséquentes
de l'IA.*



Source : Sayash Kapoor

Nos recherches nous ont permis de constater que l'IA prédictive est loin de fonctionner aussi bien que le prétendent ses concepteurs, et qu'elle ne s'est pas améliorée au cours des dernières décennies. Pour tous ceux qui cherchent à utiliser l'IA dans leur vie ou qui prennent des décisions concernant l'achat d'IA pour leurs institutions, nous espérons fournir les connaissances de base sur ce qui fait que l'IA fonctionne, où elle échoue, et comment faire la différence.

C. J. : Le terme « snake oil » implique une tromperie. Pour que les choses soient claires : qui sont les trompeurs ? Les programmeurs et les développeurs ? Les entreprises qui exploitent les solutions d'IA ? Les consommateurs ? Sommes-nous tous en train de nous tromper sur l'IA ?

Sayash : Nous discutons des nombreuses sources d'engouement pour l'IA : Les chercheurs en IA, les entreprises et les journalistes qui traitent de l'IA. Lorsque les chercheurs qui travaillent sur l'IA ne parviennent pas à évaluer correctement leurs modèles, cela conduit à des affirmations exagérées sur la façon dont l'IA peut fonctionner dans le monde réel. Dans

nos travaux antérieurs, nous avons mis en évidence une crise de reproductibilité dans la recherche scientifique qui utilise l'apprentissage automatique.

Les entreprises font également du battage médiatique lorsqu'elles vendent des produits qui ne fonctionnent pas aussi bien qu'elles le prétendent. Ces dernières années, nous avons vu de nombreuses affirmations concernant des produits d'IA qui ne fonctionnent pas - et souvent ne peuvent pas fonctionner - et pourtant les entreprises sont prêtes à faire des affirmations exagérées pour vendre leurs produits. Dans certains cas, les entreprises vendent des produits commercialisés comme étant de l'IA alors qu'en réalité, c'est un humain qui prend les décisions en coulisses.

Enfin, l'engouement pour l'IA est également dû aux journalistes qui donnent une image erronée du fonctionnement de l'IA. Nous avons recueilli une série de pièges dans le journalisme sur l'IA et constaté que les articles sur l'IA, même dans les médias respectés, souffrent de ces pièges. Par exemple, les articles utilisent des images de robots pour parler d'applications de l'IA qui n'ont rien à voir avec la robotique, ce qui donne aux lecteurs une idée trompeuse des progrès de l'IA.

C. J. : Parlez à nos lecteurs de $p(\text{doom})$. Qu'est-ce que c'est ? S'agit-il d'une mesure utile ?

Sayash : $P(\text{doom})$ ou la probabilité de la catastrophe due à l'IA est une mesure utilisée dans la communauté de la sécurité de l'IA pour articuler le risque existentiel de l'IA - la probabilité que l'IA nous anéantisse tous. Elle est devenue populaire comme moyen de quantifier à quel point quelqu'un pense que le risque existentiel de l'IA est élevé. En règle générale, les estimations de probabilité sont justifiées de trois manières : inductive (basée sur des données passées), déductive (basée sur des théories ou modèles empiriquement vérifiés sur le monde) ou subjective (basée sur des suppositions de prévisionnistes).

Malheureusement, aucune des méthodes ci-dessus ne fonctionne pour justifier $p(\text{doom})$. Nous ne pouvons pas faire d'estimations inductives parce qu'il n'existe pas de « classe de référence » d'événements similaires à partir desquels nous pourrions extrapoler le risque existentiel de l'IA. De même, nous ne disposons pas de théories et de modèles du monde établis qui nous permettent d'anticiper le risque existentiel de l'IA. Nous nous contentons donc d'estimations subjectives, qui sont par nature spéculatives et qui surestiment le risque d'événements peu probables.

Ce qui est inquiétant, c'est que les estimations de $p(\text{doom})$ sont prises au sérieux par les décideurs politiques. Contrairement à la plupart des autres données qu'ils utilisent pour élaborer leurs politiques,



Si nous parvenons à réduire les erreurs et les hallucinations, l'IA générative pourra être beaucoup plus utile pour les applications conséquentes de l'IA.

les estimations de p(doom) sont intrinsèquement spéculatives. Si les particuliers et les entreprises peuvent, bien entendu, utiliser toutes les informations dont ils disposent pour prendre des décisions, les gouvernements et les décideurs politiques doivent appliquer des normes plus strictes en matière de preuves.

C. J. : En septembre 2023, vous avez contribué à organiser un atelier sur les modèles fondamentaux responsables et ouverts. À l'époque, quelle était votre position sur la question de savoir si l'IA devait être ouverte ou fermée ? Avez-vous la même opinion aujourd'hui ?

Sayash : Lorsque nous avons organisé l'atelier, je voulais mieux comprendre les arguments concernant les risques liés aux modèles fondamentaux ouverts. L'un des principaux résultats de l'atelier a été un document de recherche dans lequel nous nous sommes penchés sur cette question. Nous avons constaté que de nombreuses allégations concernant les effets néfastes des modèles d'IA diffusés ouvertement n'étaient pas vraiment justifiées.

Par exemple, au cours de l'été 2023, un groupe de chercheurs du MIT a publié des articles affirmant que les modèles fondamentaux ouverts pouvaient aider les attaquants à créer des armes biologiques. Mais nous avons également constaté que les mêmes informations disponibles dans ces modèles fondamentaux ouverts étaient également disponibles sur Wikipédia. Dans l'article, nous avons introduit le concept de « risque marginal » afin de préciser si le risque provient des modèles fondamentaux ouverts ou s'il existe indépendamment de la publication des modèles.

Il s'agit essentiellement de comparer le risque des modèles ouverts à la fois avec celui des technologies existantes ainsi qu'avec celui des modèles fondamentaux fermés. Depuis, des recherches ont montré que le risque marginal des modèles fondamentaux ouverts est faible — notamment dans des domaines comme la création d'armes biologiques.

C. J. : Y a-t-il quelque chose que vous savez aujourd'hui et que vous ne saviez pas il y a un an qui vous effraie à propos de l'IA et de son potentiel ? Et inversement, y a-t-il quelque chose que vous savez aujourd'hui et que

vous ne saviez pas il y a un an qui vous enthousiasme à propos de l'IA et de son potentiel ?

Sayash : Je suis enthousiasmé par le potentiel de l'IA pour les travailleurs du savoir. Je pense que la plupart des travailleurs du savoir peuvent utiliser l'IA de manière utile. Bien que des problèmes tels que le manque de fiabilité et les hallucinations doivent être résolus, l'IA générative représente en fin de compte un grand pas en avant dans nos capacités technologiques. En même temps, je suis préoccupé par l'utilisation croissante de l'IA générative pour des deepfakes non consentis, notamment des nudes. Nous avons vu des cas de personnes — principalement des femmes — être ciblées par des deepfakes générés par IA.

C. J. : Pensez-vous que les départements EE/ECE prennent la sécurité de l'IA au sérieux ? Ou diriez-vous qu'il y a un manque de sensibilisation et d'éducation à ce sujet ?

Sayash : Nous avons constaté un grand intérêt pour la sécurité de l'IA dans tous les départements. Certains des plus grands spécialistes de la sécurité de l'IA à Princeton, comme le professeur Prateek Mittal, font partie du département ECE. Je suis sûr qu'à l'avenir, l'intérêt sera encore plus grand. ◀

240555-04

Questions ou commentaires ?

Contactez Elektor (redaction@elektor.com).



les dernières nouveautés de BeagleBoard

BeagleY-AI, BeagleV-Fire, BeagleMod, BeaglePlay et
BeagleConnect Freedom

contribué par BeagleBoard

La fondation *BeagleBoard.org* est une organisation à but non lucratif et un leader dans le matériel open-source, fournissant des cartes compatibles avec Linux conçues pour les *maker*, les éducateurs et les professionnels. Le dynamisme de la communauté de BeagleBoard.org garantit aux utilisateurs l'accès à une assistance étendue, à des tutoriels et à des ressources pour le développement de logiciels. Dans cet article, cinq plateformes de matériel libre sont présentées, pour des applications telles que l'IA, l'IdO, l'automatisation, et bien d'autres choses encore.

BeagleMod : une solution informatique embarquée puissante pour les applications industrielles

BeagleMod™ (**figure 1**) est un module informatique embarqué compact et performant, spécialement conçu pour les applications industrielles et les automatismes [1]. Entièrement compatible avec la norme Open Standard Modules™ (OSM) V1.1, BeagleMod est boosté par le processeur TI AM62x et intègre du matériel de pointe pour un éventail large de cas d'utilisation, notamment l'interface homme-machine (IHM), l'automatisation robotique et l'équipement médical (voir l'encadré « **BeagleMod : aperçu des caractéristiques techniques** »).

Principales caractéristiques de BeagleMod

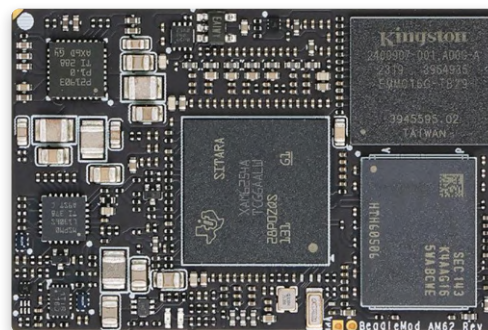
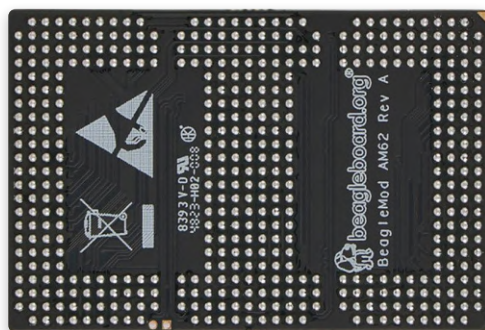
1. Processeur :

Au cœur de BeagleMod se trouve un processeur Quad 64 bits Arm® Cortex®-A53, qui garantit un traitement à grande vitesse et la capacité de gérer plusieurs tâches simultanément. Il est donc idéal pour les systèmes nécessitant un traitement robuste des données et des performances multitâches.

2. Mémoire et stockage :

Le BeagleMod est doté d'une mémoire vive DDR4 de 2 Go, qui peut être étendue à 4 Go pour répondre

Figure 1. BeagleMod™ est un module compact et hautes performances alimenté par le processeur TI AM62x, idéal pour les applications d'automatisation industrielle, d'IHM et d'équipement médical.



aux besoins d'applications plus exigeantes. Le module comprend également 16 Go de mémoire eMMC pour une conservation fiable des données, ainsi qu'une mémoire EEPROM de 32 Kbit pour stocker les paramètres et configurations essentiels.

3. Alimentation :

Le module fonctionne avec une seule alimentation en courant continu de 5 V, à travers les soudures FTGA (*Fused Tin Grid Array*), ce qui le rend économe en énergie et facile à intégrer dans les systèmes à faible consommation d'énergie.

4. Dimensions et facteur de forme :

Mesurant seulement 30 mm × 45 mm, le BeagleMod est suffisamment compact pour être utilisé dans des environnements où l'espace est restreint, tout en offrant des fonctionnalités étendues grâce à son empreinte OSM Size-M avec 476 contacts.

5. Conditions d'utilisation :

Le module est conçu pour les environnements standard et difficiles. Il fonctionne de manière efficace à des températures allant de 0°C à +70°C, avec une version industrielle optionnelle supportant des températures allant de -40°C à +85°C. Il présente également une tolérance à l'humidité de 10 % à 95 %.

6. Normes et certifications :

La carte BeagleMod est certifiée selon plusieurs normes, notamment FCC, CE, REACH et RoHS, ce qui garantit la conformité avec les réglementations mondiales et les exigences de sécurité environnementale.

Support d'interface complet

BeagleMod prend en charge une large gamme d'interfaces, ce qui en fait une solution flexible pour les applications industrielles, automobiles et autres systèmes embarqués :

- JTAG pour le débogage et le développement.
- 5 ports UART pour la communication série.
- 2 interfaces RGMII pour la connectivité Ethernet 10/100/1000 Mbit.
- 2 interfaces SDIO (4 bits) pour les cartes mémoire et autres périphériques.
- 6 sorties PWM pour le contrôle de moteurs, de LEDs ou d'autres dispositifs.
- 2 interfaces SPI, y compris OSPI (8 bits) pour une communication rapide avec des capteurs, des mémoires de stockage ou d'autres microcontrôleurs.
- 2 interfaces I²S pour le transfert de données audio.
- 2 bus CAN pour la communication automobile.
- 2 ports USB 2.0 pour les périphériques.
- 3 interfaces I²C pour la connexion à divers capteurs et dispositifs.
- 1 interface CSI (4 voies) pour les caméras.
- 1 interface RGB (18 bits) pour les écrans.
- 1 OLDI (2×4 voies LVDS) pour les écrans haute résolution.
- GPIOs pour le contrôle des entrées et sorties personnalisées.

BeagleMod : Aperçu des caractéristiques techniques

Composant	Spécification
Processeur	TI AM62 Quad-Core A53
Mémoire RAM	2 Go DDR4, extensible à 4 Go
Stockage	16 Go eMMC, 32 Kbit EEPROM
Alimentation	5 V DC via OSM FTGA
Facteur de forme	30 mm × 45 mm, taille OSM-M
Température de fonctionnement	0°C à +70°C, -40°C à +85°C (optionnel)
Humidité de fonctionnement	10 %~95 % HR
Certifications	FCC, CE, REACH, RoHS JTAG, UART, RGMII, SDIO, PWM, SPI, I2S, CAN, USB 2.0, I2C, CSI, RGB, OLDI, GPIO
Interfaces	JTAG, UART, RGMII, SDIO, PWM, SPI, I2S, CAN, USB 2.0, I2C, CSI, RGB, OLDI, GPIO
Fonctionnalités supplémentaires	OSM FTGA, GPIO, diverses interfaces de contrôle

Les applications de la carte BeagleMod

Le haut niveau d'intégration de BeagleMod, associé à sa taille compacte et à sa puissance de traitement, en fait un outil idéal pour une grande variété d'applications :

➤ Interface homme-machine (IHM) :

Le module peut être intégré dans des systèmes d'interface utilisateur pour l'automatisation industrielle, offrant un contrôle souple et réactif des machines, des systèmes industriels et des panneaux de contrôle.

➤ L'automatisation robotique :

Le puissant processeur Cortex-A53, associé aux interfaces GPIO, PWM et I²C en temps réel, permet un contrôle précis des bras robotiques, des machines autonomes et des systèmes d'automatisation industriels.

➤ Maison intelligente et économies d'énergie :

Les interfaces polyvalentes et la faible consommation d'énergie du BeagleMod le rendent idéal pour les applications domestiques intelligentes, notamment les systèmes d'éclairage intelligents, les commandes HVAC et les dispositifs de gestion de l'énergie.

➤ Matériel médical :

L'intégration de plusieurs interfaces UART, SPI et I²S permet d'utiliser le BeagleMod dans des outils de diagnostic médical, des systèmes de surveillance des patients, comme d'autres équipements médicaux sensibles nécessitant un traitement fiable des données en temps réel.

Kit de personnalisation et d'évaluation

BeagleMod offre une grande flexibilité pour des configurations et des extensions personnalisées afin de répondre à des besoins spécifiques. Seeed Studio fournit des services de personnalisation, permettant aux utilisateurs d'adapter BeagleMod à leurs besoins particuliers. Le kit d'évaluation BeagleMod comprend le module de base et les outils de développement nécessaires pour démarrer rapidement.

BeagleY-AI : une plateforme informatique de pointe pour l'IA à haute performance

BeagleY-AI [2], développé par la Fondation BeagleBoard.org, est un puissant ordinateur monocarte conçu pour l'informatique de pointe axée sur l'IA (figure 2). Au cœur de cet ensemble, le processeur Texas Instruments AM67A offre un mélange exceptionnel de performances et d'évolutivité, ciblant les applications qui exigent un traitement en temps réel pour l'IA, la vision et la robotique (voir l'encadré « **BeagleY-AI : aperçu des caractéristiques techniques** »).

Les caractéristiques principales

1. La puissance de traitement :

Le SoC AM67A intègre un processeur Quad 64-bit Arm Cortex-A53 à 1,4 GHz, ainsi que deux DSP C7x pour les charges de travail d'IA, chacun offrant 4 TOPS de calcul, ce qui le rend idéal pour l'apprentissage profond et les applications basées sur la vision. Ces processeurs sont

pris en charge par un accélérateur de multiplication matricielle (MMA), ce qui favorise une inférence plus rapide des modèles d'IA et permet de créer des systèmes intelligents et améliorés par l'IA.

2. Mémoire et stockage :

Équipé de 4 Go de RAM LPDDR4, la carte BeagleY-AI assure une gestion efficace des modèles d'IA gourmands en mémoire et du traitement des données en temps réel. Elle dispose également d'un emplacement pour carte microSD avec support UHS-1, permettant un transfert de données à grande vitesse et un démarrage fiable. La carte offre un espace de stockage suffisant pour l'exécution d'algorithmes d'IA complexes et le traitement de grandes quantités de données de capteurs.

3. Connectivité et extension :

Avec 4 ports USB 3.0, un port USB 2.0 Type-C et Gigabit Ethernet, la carte BeagleY-AI assure une connectivité transparente aux périphériques et une communication réseau à haut débit. L'interface PCIe Gen3 prend en charge l'intégration de périphériques haute performance, élargissant l'utilité de la carte pour les applications nécessitant un calcul d'IA avancé et une mise en réseau. L'en-tête GPIO à 40 broches est compatible avec les HAT Raspberry Pi existants, ce qui permet aux utilisateurs de créer et de déployer rapidement des applications d'IA de pointe avec des capteurs, des actionneurs et des options d'E/S supplémentaires.

4. Capacités en matière d'IA et de vision :

Conçu pour les applications d'intelligence artificielle et de vision, BeagleY-AI intègre 2 interfaces de caméra MIPI-CSI, capables de connecter plusieurs caméras haute résolution pour des tâches telles que l'analyse vidéo en temps réel, la vision artificielle et la robotique autonome. Le GPU et les accélérateurs vidéo intégrés garantissent que le streaming vidéo et les tâches de reconnaissance d'images sont traités efficacement sans trop solliciter les cœurs du CPU.

5. L'alimentation et sa gestion :

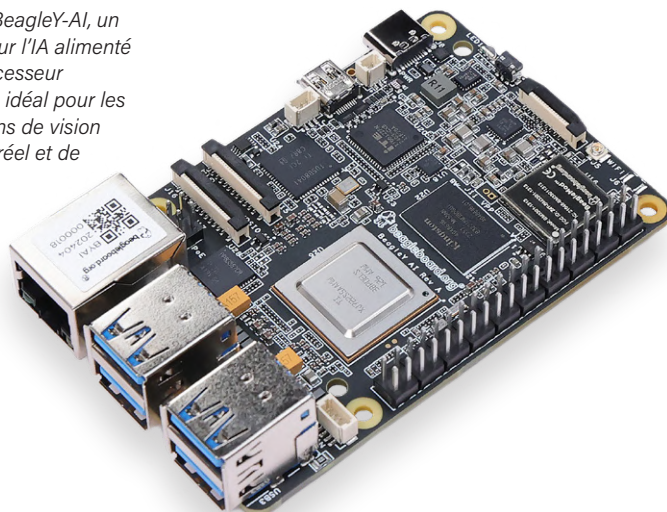
La carte est alimentée par une alimentation 5 V/3 A DC via USB-C, ce qui la rend compatible avec un large éventail de sources d'alimentation. Elle inclut une broche pour ventilateur avec contrôle PWM, garantissant que les charges de travail IA haute performance sont soutenues avec une gestion de la température adaptée.

Applications d'IA en périphérie

BeagleY-AI est parfaitement adapté à une variété d'applications en périphérie de réseau pilotées par l'IA :

- **Interfaces homme-machine (IHM) :** système de contrôle et de retour d'information en temps réels.
- **Automatisation industrielle :** surveillance basée sur l'IA, maintenance prédictive et automatisation des procédés.
- **Robotique :** navigation autonome, détection d'objets et manipulation à l'aide de modèles d'IA avancés.
- **Vidéosurveillance et analyse :** détection d'objets, reconnaissance faciale et analyse de modèles à haute performance dans les systèmes de sécurité.
- **Matériel médical :** applications de l'IA de périphérie

Figure 2. BeagleY-AI, un SBC axé sur l'IA alimenté par le processeur TI AM67A, idéal pour les applications de vision en temps réel et de robotique.



BeagleY-AI : Aperçu des caractéristiques techniques

Composants	Spécification
Processeur	TI AM67A, Quad 64 bits Cortex-A53 @1,4 GHz
Accélérateurs IA	Double DSP C7x avec 4 TOPS chacun
Mémoire RAM	4 Go LPDDR4
Wi-Fi	802.11ax (Wi-Fi 6)
Bluetooth	Bluetooth 5.4 (BLE)
Ports USB	4 x USB 3.0, 1 x USB 2.0 Type-C
Ethernet	Gigabit Ethernet avec PoE+
Caméra	2 x interfaces MIPI CSI à 4 voies
Affichage	HDMI, OLDI et DSI MIPI
Stockage	Emplacement pour carte microSD avec support UHS-1
Extension	En-tête GPIO à 40 broches
Alimentation	5 V/3 A via USB-C
Connecteur de ventilateur	En-tête à 4 broches contrôlé par PWM

pour la surveillance des patients, les diagnostics et l'automatisation des soins de santé.

Une plateforme d'IA prête pour l'avenir

BeagleY-AI est conçu pour être une plateforme prête pour l'avenir, permettant l'intégration transparente d'applications d'IA à la périphérie, tout en maintenant la flexibilité avec de multiples options d'E/S et le support de périphériques. Son architecture évolutive lui permet de gérer les tâches les plus exigeantes dans des secteurs tels que la santé, la sécurité et l'automatisation. Un article sur la carte BeagleBoardY-AI rédigé par Brian T. Williams, rédacteur en chef d'Elektor, est disponible dans cette édition et dans les archives d'Elektor (www.elektormagazine.fr/240527-04).

BeagleV-Fire : carte de développement RISC-V et FPGA pour l'edge computing

BeagleV-Fire [3] est un ordinateur monocarte (SBC) révolutionnaire (figure 3) développé par la Fondation BeagleBoard.org, alimenté par le SoC Microchip PolarFire® MPFS025T. Il intègre des cœurs RISC-V et une matrice FPGA, offrant aux développeurs la flexibilité d'explorer le potentiel de l'architecture RISC-V et de la logique programmable. Idéal pour l'IA, la robotique et l'automatisation industrielle, BeagleV-Fire fournit le matériel nécessaire pour alimenter les applications de *edge computing* en temps réel (voir l'encadré « BeagleV-Fire : aperçu des caractéristiques techniques »).

Principales caractéristiques

1. Processeur :

- Le cœur de la BeagleV-Fire est le SoC MPFS025T, qui comprend :
- > 4 cœurs RISC-V 64 bits (RV64GC) pour le traitement des applications
 - > 1 noyau d'amorçage RISC-V 64 bits (RV64IMAC) pour la gestion du système
 - > Une matrice FPGA programmable, permettant aux développeurs d'accélérer les tâches matérielles personnalisées telles que le traitement des signaux, la cryptographie ou l'inférence IA.

2. Mémoire et stockage :

- > 2 Go de RAM LPDDR4 à 1866 MHz garantissent la multitâche et un traitement efficace des données.
- > Mémoire flash eMMC de 16 Go offrant un espace de stockage embarqué suffisant pour les distributions Linux et les applications, et un espace de stockage supplémentaire est disponible via l'emplacement microSD.

3. Connectivité et interfaces :

- > Gigabit Ethernet avec Realtek RTL8211F PHY pour une mise en réseau à grande vitesse.
- > Un port USB-C pour l'alimentation de la carte, la programmation et le transfert de données.

BeagleV-Fire : Aperçu des caractéristiques techniques

Composant	Spécification
Processeur	PolarFire® MPFS025T SoC avec 4x cœurs RISC-V, 1x cœur de démarrage
Mémoire RAM	2 Go LPDDR4
Stockage	16 Go eMMC, emplacement microSD
Ethernet	Gigabit Ethernet, Realtek RTL8211F PHY
Wi-Fi	Emplacement M.2 Key E pour 2,4 GHz/5 GHz
Ports USB	USB-C pour alimentation et données
Caméra	Interface caméra CSI (22 broches)
Extension	En-têtes P8 & P9 cape, connecteur SYZYGY
Alimentation	USB-C (5 V/3 A)
Débogage	JTAG, en-têtes UART
LEDs	12x LEDs programmables
Dimensions	86,38 × 54,61 mm, PCB à 12 couches, conforme RoHS

- > Un emplacement M.2 Key E pour l'ajout de modules Wi-Fi, permettant aux développeurs d'intégrer des communications sans fil, telles que le Wi-Fi 2,4 GHz/5 GHz.
- > Connecteur de caméra CSI, compatible avec le BeagleBone AI-64, le BeagleV-Ahead et le Raspberry Pi Zero/CM4, permettant l'entrée d'images et de vidéos de haute qualité pour les applications basées sur la vision.

4. Connecteurs d'extension :

Le BeagleV-Fire conserve les connecteurs familiers P8 et P9, offrant une compatibilité avec les connecteurs du BeagleBone Black, ce qui permet une intégration facile avec les modules d'extension existants. De plus, le connecteur haute vitesse SYZYGY permet d'accéder à des périphériques haute performance pour une extension matérielle avancée.

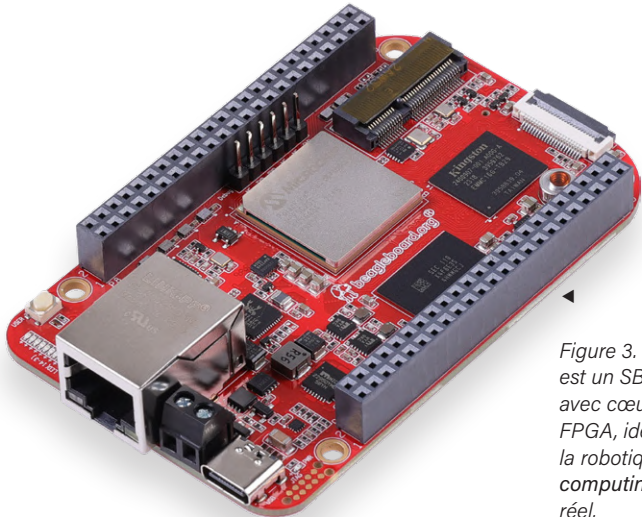


Figure 3. BeagleV-Fire est un SBC polyvalent avec cœurs RISC-V et FPGA, idéal pour l'IA, la robotique et l'edge computing en temps réel.

5. Débogage et développement :

La carte est équipée de :

- > Connecteurs de débogage JTAG et UART pour le débogage et la surveillance du matériel de bas niveau.
- > Leds et boutons programmables par l'utilisateur pour améliorer les expériences de développement et le retour d'information sur le système.

Applications idéales

- > Edge IA et automatisation industrielle :

Avec ses capacités de traitement en temps réel et à son FPGA programmable, BeagleV-Fire est idéal pour des tâches telles que la maintenance prédictive, le contrôle qualité alimenté par l'IA et l'automatisation des usines intelligentes. Sa faible consommation d'énergie le rend adapté aux applications d'IA en périphérie, permettant un traitement en temps réel sans dépendre d'une infrastructure cloud.

- > Robotique et systèmes de vision :

La prise en charge de la caméra CSI intégrée, ainsi que la capacité de la carte à s'interfacer avec des capteurs et des actionneurs via les connecteurs P8/P9, la rendent bien adaptée à la robotique guidée par la vision, aux systèmes autonomes et à la détection d'objets basée sur l'intelligence artificielle.

- > Exploration du matériel Open-Source :

Les développeurs, les bricoleurs et les chercheurs peuvent explorer l'architecture RISC-V de pointe tout en utilisant le FPGA pour accélérer des tâches telles que le cryptage des données, l'apprentissage en profondeur et le traitement des données en temps réel. La disponibilité de ressources open-source étendues, associée à la richesse des E/S de la carte, en fait un excellent choix pour l'expérimentation matérielle et le prototypage.

Mise en route et outils de développement

BeagleV-Fire est soutenu par un écosystème de développement complet. Les développeurs peuvent rapidement flasher les derniers logiciels du FPGA (*gateway*) sous Linux sans outil supplémentaire. *Gateway* peut également être flashé en utilisant FlashPro5/6 et Libero SoC Design Suite, ce qui facilite la conception et le débogage FPGA. Flasher Linux est aussi simple que de flasher une sdCard avec des outils comme Balena etcher. La carte prend également en charge Cockpit, SSH et l'accès au port série pour le développement et la gestion du système, ce qui permet aux développeurs de se plonger facilement dans leurs projets.

BeagleV-Fire offre une opportunité unique de travailler avec un SoC RISC-V associé à un FPGA dans un SBC compact et abordable. Que vous exploriez l'IA à la périphérie, que vous développiez des systèmes autonomes ou que vous créiez des applications avec accélérations matérielles personnalisées, BeagleV-Fire offre une plateforme puissante et flexible pour stimuler l'innovation. Pour plus d'informations sur le BeagleV-Fire et pour

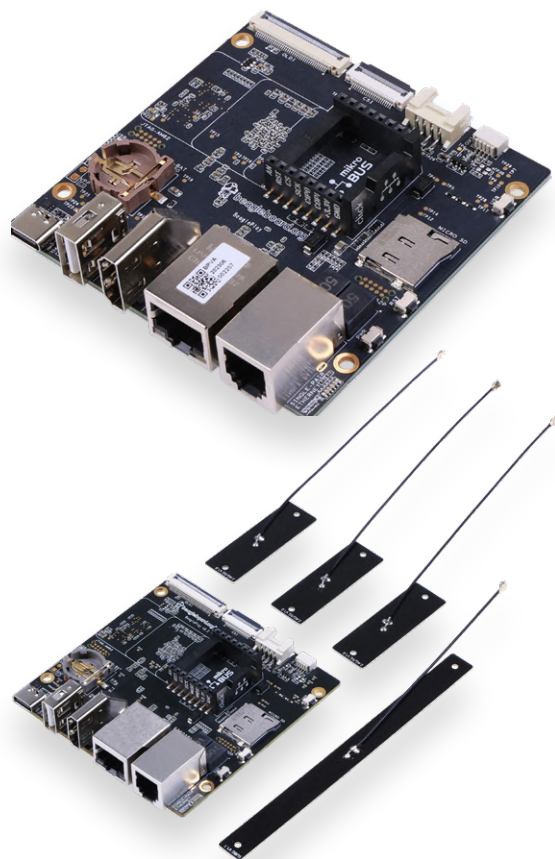
accéder à des tutoriels, veuillez consulter la documentation officielle de BeagleBoard.org [4].

BeaglePlay et BeagleConnect Freedom : un duo puissant pour le développement de l'IdO

BeaglePlay [5] et BeagleConnect Freedom [6] (**figures 4 et 5**) sont deux cartes complémentaires de la Fondation BeagleBoard.org, conçues pour simplifier le développement de l'IdO en fournissant un matériel puissant et open-source, avec une intégration facile des capteurs et de la connectivité.

BeaglePlay : une passerelle IdO à la connectivité inégalée

BeaglePlay est un ordinateur monocarte doté du processeur quadricœur Texas Instruments AM6254, offrant une grande puissance de calcul. Il excelle dans la création d'une passerelle IdO ou d'un hub d'automatisation grâce à sa large gamme d'options de connectivité. Que vous ayez besoin de Wi-Fi (2,4 GHz/5 GHz), de Gigabit Ethernet, d'Ethernet à paire unique ou d'un réseau sans fil Sub-1 GHz, BeaglePlay facilite la connexion des capteurs et des périphériques IdO grâce à des interfaces intégrées telles que mikroBUS, Grove et QWIIC. De plus, il prend en charge le HDMI pour les écrans et le CSI pour les caméras, ce qui le rend polyvalent pour les solutions IdO industrielles et domestiques.



▲ Figure 4. BeaglePlay, un SBC de passerelle IdO avec le processeur TI AM6254, offre de nombreuses options de connectivité pour la création de hubs IdO polyvalents.



Figure 5. BeagleConnect Freedom, une plateforme sans fil avec le microcontrôleur TI CC1352P7, prend en charge les nœuds IdO longue portée avec le protocole BLE et Sub-1 GHz.

Caractéristiques principales :

- Processeur : AM6254 (quadricœur Arm Cortex-A53)
- Mémoire : 2 Go DDR4, 16 Go de stockage eMMC
- Connectivité : Wi-Fi, Ethernet, BLE, Sub-1 GHz, Ethernet à paire unique
- Extension : mikroBUS, Grove, QWIIC, CSI, etc.

Grâce à son architecture plug-and-play et à sa facilité d'utilisation, BeaglePlay permet d'ajouter facilement des capteurs, des actionneurs et des indicateurs. Sa compatibilité avec des milliers de modules d'extension prêts à l'emploi simplifie le prototypage et le développement de systèmes complets.

BeagleConnect Freedom : l'IdO sans fil en toute simplicité

BeagleConnect Freedom est une plateforme matérielle ouverte, sans fil, construite autour du microcontrôleur TI CC1352P7, ce qui en fait un complément idéal à BeaglePlay pour créer des nœuds IdO sans fil et à faible consommation. Elle prend en charge les protocoles sans fil Bluetooth Low Energy (BLE) à 2,4 GHz et Sub-1 GHz, ce qui permet aux nœuds de capteurs de communiquer sur de longues distances. Avec ses deux prises mikro-BUS, ses capteurs intégrés et ses fonctions de gestion de charge de batterie, il est prêt pour le prototypage rapide de l'IdO.

Principales caractéristiques :

- Processeur : CC1352P7 (Arm Cortex-M4 à 48 MHz)
- Sans fil : BLE, Sub-1 GHz IEEE 802.15.4

- Extension : 2 prises mikroBUS pour capteurs et actionneurs
- Alimentation : support de batterie LiPo avec capacités de charge

BeagleConnect Freedom est parfait pour construire des nœuds IdO qui communiquent sans fil avec BeaglePlay, formant ainsi un réseau robuste d'appareils connectés.

Comment utiliser BeaglePlay et BeagleConnect Freedom ensemble

Lorsqu'ils sont utilisés ensemble, BeaglePlay agit comme un hub central ou une passerelle, tandis que BeagleConnect Freedom sert de nœuds de capteurs sans fil. Par exemple, vous pouvez utiliser BeaglePlay comme processeur central pour collecter les données de plusieurs BeagleConnect Freedom répartis sur une large zone. Cette combinaison vous permet de mettre en place un réseau maillé de capteurs sans vous soucier du câblage, ce qui le rend idéal pour les maisons intelligentes, l'automatisation industrielle ou la surveillance environnementale à grande échelle.

Voici comment vous pouvez combiner leurs avantages :

1. Configurez BeaglePlay en tant que hub IdO, en le connectant à votre réseau via Wi-Fi ou Ethernet.
2. Déployez des nœuds BeagleConnect Freedom dans votre maison, votre usine ou votre environnement pour transmettre sans fil des données de capteurs (comme la température ou l'humidité) à BeaglePlay.
3. Utilisez Zephyr RTOS sur BeagleConnect Freedom pour programmer et gérer les nœuds de capteurs.
4. Surveillez ou contrôlez le réseau grâce au traitement puissant de BeaglePlay et aux intégrations dans le cloud.

La combinaison de la connectivité élevée de BeaglePlay et des capacités sans fil à faible consommation de BeagleConnect Freedom crée un écosystème IdO flexible, évolutif et efficace pour les amateurs comme pour les professionnels.

Avec ces outils, le développement IdO n'a jamais été aussi accessible ! Que vous soyez un développeur chevronné ou débutant, BeaglePlay et BeagleConnect Freedom vous offrent toute la flexibilité matérielle et logicielle dont vous avez besoin. ◀

VF : Laurent Rauber — 240630-04

LIENS

- [1] BeagleMod : <https://www.beagleboard.org/boards/beaglemod>
- [2] BeagleY-AI : <https://www.beagleboard.org/boards/beagle-y-ai>
- [3] BeagleV-Fire : <https://www.beagleboard.org/boards/beagle-v-fire>
- [4] Documentation officielle de BeagleBoard.org : <https://docs.beagleboard.org/>
- [5] BeaglePlay : <https://www.beagleboard.org/boards/beagleplay>
- [6] BeagleConnect Freedom : <https://www.beagleboard.org/boards/beagleconnect-freedom>



détection des moustiques

avec Arduino Nicla Vision et des données open source

Jose Bagur (UVG), João Vitor Freitas da Costa (UERJ), Silvia A. Sotelo López (UPB),
Peter Ing (Arm SAAIA) et David J. Cuartielles Ruiz (Arduino/Université de Malmö)

L'équipe s'est confrontée à la problématique critique liée à la prolifération des moustiques et à leur rôle dans la propagation de maladies. Elle a donc développé un système de détection qui utilise le son pour identifier les moustiques et activer les dispositifs répulsifs. Face à l'augmentation alarmante des cas de dengue en Amérique, le projet vise à offrir une solution efficace pour la détection des moustiques en conditions réelles. Découvrez comment l'association de la technologie et de tinyML peut favoriser le développement durable !





Durant l'été 2024, le CIPT a collaboré avec l'Université de Harvard, Barnard College, UNIFEI, IBM Research Brazil et Arduino pour organiser le *Workshop on TinyML for Sustainable Development* à São Paulo, Brésil. L'atelier a réuni des personnes venant principalement des Amériques et passionnées par le développement de systèmes intelligents en périphérie. Pendant une semaine, les participants ont eu l'occasion de travailler avec des experts et d'acquérir des connaissances sur des sujets variés tels que la collecte de données, la formation des appareils Edge, les LLM *DIY*, l'apprentissage avec tinyML et les exemples d'application des technologies. Chaque participant était également invité à présenter ses propres travaux dans le domaine.

Au cours de cette activité de cinq jours, nous avons eu l'occasion de nouer des liens, d'assister à une série de conférences et de séances pratiques, de visiter l'université de São Paulo et son usine de fabrication de circuits imprimés, ainsi que de collaborer en équipe pour développer diverses solutions techniques. Nous, les auteurs de cet article, avons décidé de collaborer à la recherche de solutions potentielles aux populations croissantes de différents types de moustiques, vecteurs de maladies considérées comme éradiquées. Dans certaines régions, il y a plusieurs cycles de reproduction par saison, ce qui est inédit. Chaque été, les zones touchées par les maladies transmises par les insectes, telles que la malaria, s'élargissent. Nous avons eu l'idée de concevoir un système capable de détecter la présence de certains types de moustiques par l'analyse sonore de leurs battements d'ailes et de répondre par l'émission d'un brouillard répulsif, pour éloigner ces nuisibles.

L'augmentation du nombre de cas de dengue dans les Amériques est la principale motivation de ce projet. En 2024, 11 517 728 cas ont été enregistrés, soit une augmentation alarmante de 230 % par rapport au nombre de cas signalés en 2023 (rapport de l'OPS, septembre 2024 [1]).

Pour reproduire ce projet, vous aurez besoin de :

- Un Arduino Nicla Vision [2]. Nous nous concentrerons uniquement sur les capacités de détection sonore de l'appareil.

- Le jeu de données qui contient des sons de moustiques qui permettront de détecter leur présence.
- Un compte sur Edge Impulse, la plateforme en ligne qui créera le moteur d'inférence après l'avoir alimenté avec le jeu de données mentionné.
- Une carte relais pouvant être déclenchée par un signal de 3,3 V (directement depuis le Nicla Vision).
- Un piézo-humidificateur
- Des fils de connexion
- Un fer à souder (le Nicla Vision est livré avec des pastilles cranelées, mais n'a pas de broches pour connecter quoi que ce soit. Nous avons dû souder quelques fils pour que la carte puisse contrôler le reste du dispositif).

Arduino a fait don des cartes nécessaires à l'organisation de l'atelier et Edge Impulse propose des comptes gratuits pour les makers, qui peuvent créer le code pour les cartes sur la base de l'ensemble de données. Cependant, nous avons dû acheter localement la carte relais et l'humidificateur. São Paulo compte plusieurs magasins où il est possible de se procurer des composants électroniques du commerce.

Contexte

Les moustiques ont une durée de vie courte mais intense. Plusieurs espèces existantes sont très résistantes aux conditions climatiques, et leurs œufs fécondés peuvent survivre dans des environnements très hostiles jusqu'à ce que les niveaux de température et d'humidité adéquats facilitent leur éclosion. Les moustiques ont été largement étudiés par les scientifiques. Ils se nourrissent du sang d'autres animaux, ce qui les rend très aptes à transmettre des maladies grâce au mélange de sangs différents qu'ils transportent dans leur système. Certaines expériences sur l'ADN (comme Prometeo de Microsoft) ont démontré que le sang d'un seul moustique pouvait contenir des traces d'ADN de 3 ou 4 espèces différentes de mammifères, incluant les humains, les chiens, les vaches et les moutons. Cette capacité à passer rapidement d'un hôte à l'autre rend la transmission de maladies rapide.

Il est possible de distinguer le type et le sexe des moustiques en observant la taille

de leurs ailes. Cela signifie également qu'il est possible de classer ces insectes selon la signature sonore qu'ils produisent, due à leurs ailes mécaniquement distinctes. Bien que cela soit très difficile pour les humains, un système tinyML correctement entraîné pourrait y parvenir. C'est dans ce contexte que l'accès à une vaste base de données, contenant des milliers d'enregistrements sonores de moustiques, devient utile. Elle pourrait servir à développer un moteur d'inférence, soit une minuscule machine d'intelligence artificielle capable de distinguer les sons des moustiques.

Un défi inattendu rencontré lors de notre projet est qu'au Brésil, deux types différents de courant alternatif coexistent : 110 et 230 V. Lors de l'achat des composants, nous avons également acheté un fer à souder bon marché, aucun de nous n'ayant pensé à en apporter un. Nous devons souder des fils de liaison aux broches de la carte Nicla Vision afin de la connecter à la carte relais. Nous avons réalisé ensuite que notre fer à souder ne fonctionne pas sous 110 V AC. Malgré nos tentatives, le manque de puissance rendait impossible son utilisation, et nous avons dû abandonner. Finalement, nous avons dû nous résoudre à utiliser la prise de la machine à raser de notre chambre d'hôtel pour réaliser les soudures, ajoutant ainsi une anecdote mémorable à notre projet. À l'avenir, nous nous souviendrons toujours d'emporter un fer à souder USB-C !

Idée et réalisation

Lors de l'atelier, nous avons conçu une preuve de concept pour un dispositif capable de détecter la présence de moustiques et de produire un brouillard qui pourrait contenir différents types d'essences (telles que, la citronnelle) pour nous avertir de la présence du moustique ou simplement le faire fuir. L'idée était de vérifier si les outils permettaient de construire rapidement ce dispositif intelligent. En termes de technologie, l'Arduino Nicla Vision est équipé d'une caméra couleur, d'un microphone numérique, d'un IMU, d'un processeur double cœur (Arm Cortex M7 + M4), des capacités de communication (Wifi, BLE), d'un port d'extension I²C, d'une batterie/un connecteur d'alimentation externe, et de plusieurs d'E/S numériques et d'entrées analogiques. Pour notre projet,

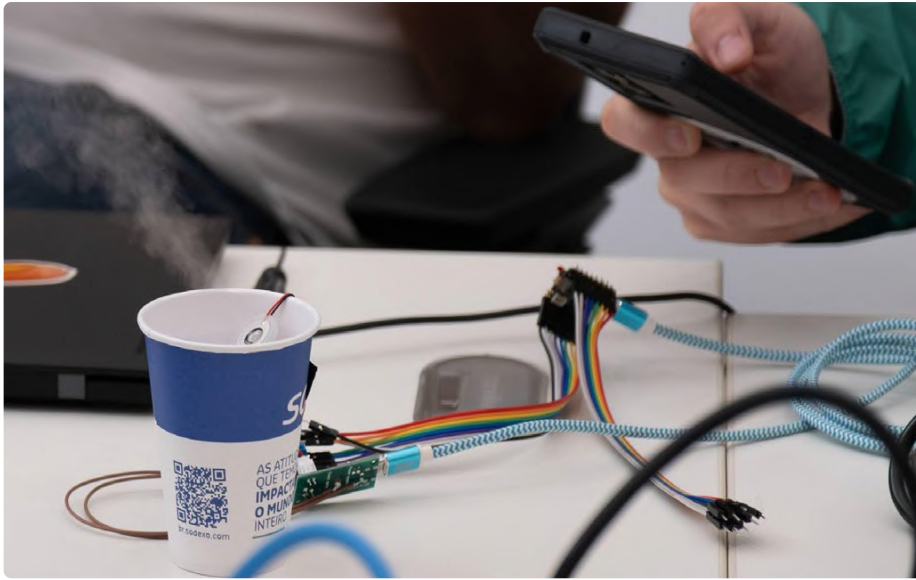


Figure 1. Évaluation en direct de l'humidificateur basé sur tinyML, à l'aide d'un enregistrement de l'ensemble de test lors de la clôture du Workshop on TinyML for Sustainable Development.

nous avons uniquement utilisé le microphone et le microcontrôleur Cortex M7, ainsi qu'une broche numérique pour contrôler l'humidificateur. Les autres fonctionnalités n'étaient pas nécessaires pour cette application. Tout le code (moteur d'inférence, communication série pour la surveillance, contrôle des broches numériques et collecte de données par le microphone) était exécuté avec un seul processeur, laissant le second libre pour d'autres tâches.

Pour générer le brouillard, nous avons utilisé un humidificateur non (figure 1). Ce dernier était équipé d'un élément piézoélectrique, activé par un oscillateur et un amplificateur, avec un interrupteur tactile offrant trois réglages : arrêt, demi-puissance et pleine puissance. Nous avons essayé deux méthodes différentes pour le faire fonctionner. La première, qui consistait à couper un des fils du piézo et à le connecter via un relais, s'est avérée infructueuse car l'impédance du relais perturbait le fonctionnement du piézo, empêchant la projection de la brume. La seconde méthode consistait à utiliser le relais pour actionner l'interrupteur tactile de l'humidificateur. Cela signifiait également que nous devions cliquer deux fois sur le relais pour activer l'humidificateur, et une seule fois pour le désactiver. Cette solution s'est révélée efficace, le système fonctionnant parfaitement sans aucun problème de rebond.

Méthodologie

Pour détecter efficacement la présence de moustiques et activer un répulsif approprié, nous avons opté pour l'utilisation d'un modèle d'apprentissage profond déployé sur l'Arduino Nicla Vision. Cette méthode permet à l'appareil d'« apprendre » et d'identifier les sons émis par les moustiques, offrant ainsi une

solution plus flexible et adaptative par rapport aux méthodes algorithmiques ordinaires. Ces approches conventionnelles deviennent souvent trop complexes dans les environnements réels, tandis que l'apprentissage profond excelle en s'adaptant à la variabilité et aux subtilités des données, garantissant ainsi une détection plus fiable.

tinyML a révolutionné les applications à la périphérie en permettant de déployer des modèles d'apprentissage profond tels que les réseaux neuronaux convolutifs sur les CPU que l'on trouve dans les microcontrôleurs, au lieu de recourir à des GPU gourmands en énergie.

Le Nicla Vision est doté d'un microcontrôleur à double cœur, combinant le puissant Arm Cortex M7 avec le cœur Arm Cortex M4. Cette configuration permet un équilibre entre un traitement puissant et une faible consommation d'énergie, ce qui la rend idéale pour les applications basées sur la vision qui

nécessitent à la fois une efficacité de calcul et une faible consommation énergétique, tout en restant en ligne avec les principes de durabilité.

Créer le modèle avec Edge Impulse

Edge Impulse a été le pionnier des applications tinyML en fournissant aux développeurs embarqués une plateforme MLOps qui facilite le processus de collecte et d'utilisation des données de capteurs pour construire et déployer des modèles sur des microcontrôleurs et des environnements Edge. Le processus général avec Edge Impulse est (voir figure 2 :

1. **Collecte des données** : Utiliser des capteurs pour collecter des données de l'environnement, ou importer des ensembles de données existants, à utiliser pour entraîner le modèle.

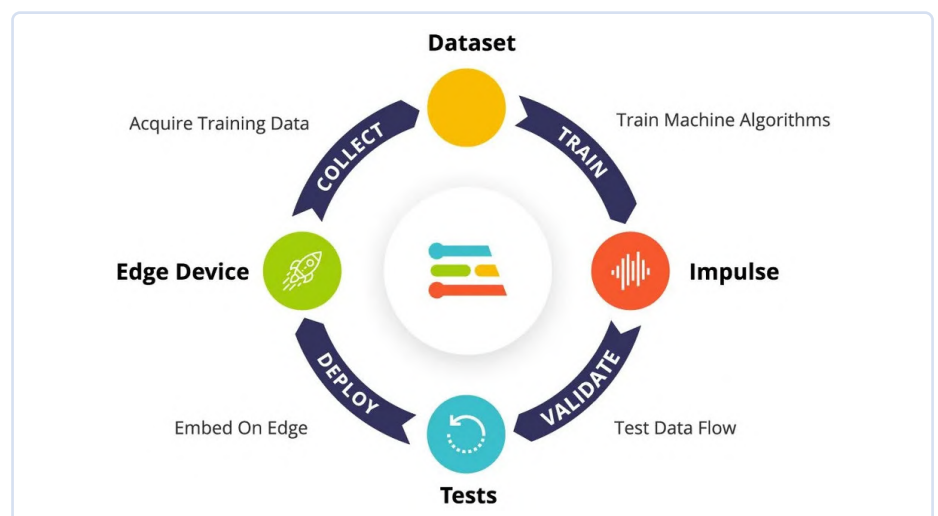


Figure 2. Le processus général de création de modèles avec Edge Impulse. (Source : Edge Impulse)

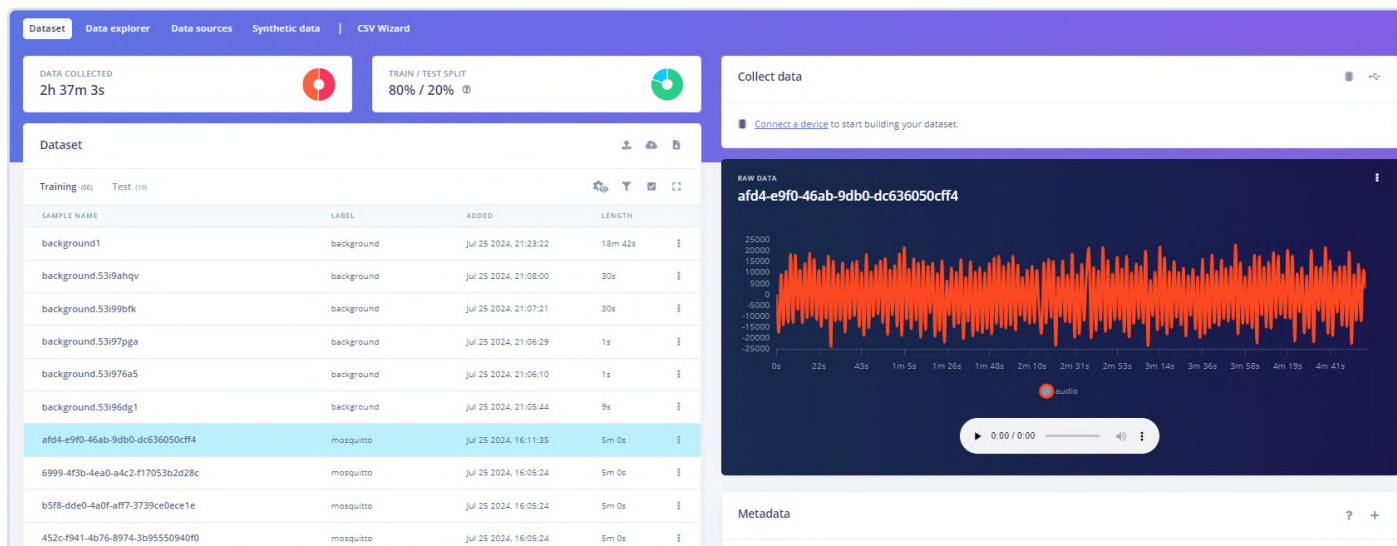


Figure 3. Travailler avec des ensembles de données en ligne et une capture audio en direct avec Edge Impulse.

2. **Prétraitement des données** : Exploiter les outils d'Edge Impulse pour étiqueter, prétraiter et préparer les données, les rendant ainsi prêtes pour l'apprentissage des modèles ML.
3. **Entraînement des modèles** : Créer et entraîner des modèles d'apprentissage automatique en utilisant un catalogue d'architectures préconçues ou personnalisées, basées sur les données collectées.
4. **Déploiement des modèles** : Installer ces modèles sur des appareils Edge, avec des déploiements optimisés et une intégration simplifiée adaptée à Arduino, y compris le Nicla Vision, pour une analyse et des actions en temps réel.

Pour développer un prototype fonctionnel et réaliser une démonstration efficace lors de l'atelier, nous avons besoin d'un moyen de démontrer le concept tout en le rendant possible à utiliser dans un cadre réel. Nous avons atteint cet objectif en combinant un ensemble de données de sons de moustiques réels avec quelques sons synthétisés dérivés d'enregistrements trouvés en ligne. De plus, nous avons capturé le bruit ambiant (**figure 3**), directement avec l'Arduino Nicla Vision à un taux d'échantillonnage de 8 KHz. Nous avons collecté 2 heures et 37 minutes de données au total, qu'Edge Impulse a automatiquement

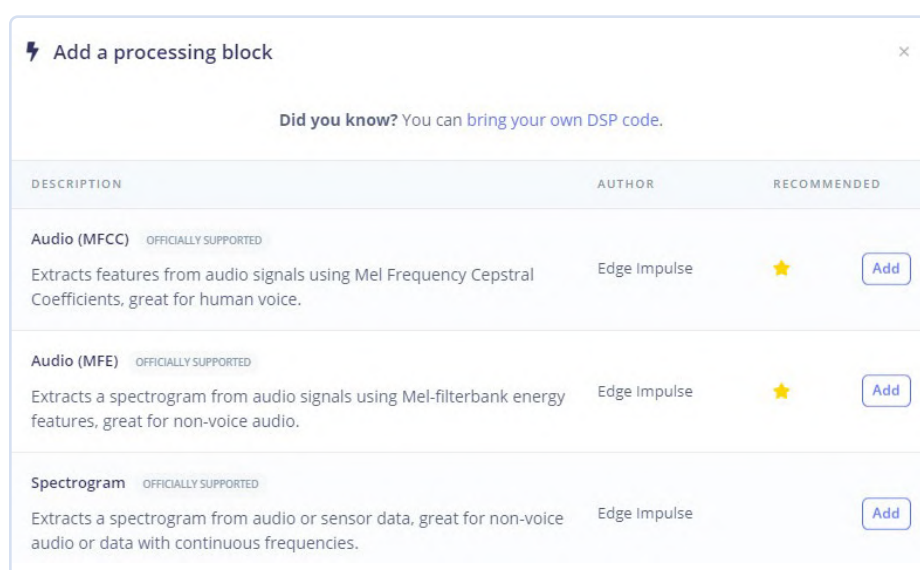


Figure 4. Edge Impulse propose des blocs DSP préconstruits adaptés à divers types de traitement audio.

divisées en 80% pour l'entraînement et 20% pour les tests. Cette méthode illustre bien comment Edge Impulse gère à la fois les ensembles de données en ligne et la capture en direct de données audio avec des cartes prises en charge telles que la Nicla Vision. En utilisant le bruit de fond de la salle, nous avons pu ajuster le modèle pour notre environnement de test, mettant en évidence l'adaptabilité de l'apprentissage automatique.

Cette étape était cruciale car les données initiales avaient été recueillies dans divers environnements intérieurs et extérieurs, avec des niveaux de bruit de fond très différents de ceux rencontrés lors de notre atelier. La capacité du modèle à s'adapter à ces nouvelles conditions a mis en évidence la capacité de l'apprentissage automatique à s'adapter à des environnements variés. Pour résoudre le problème de la détection de

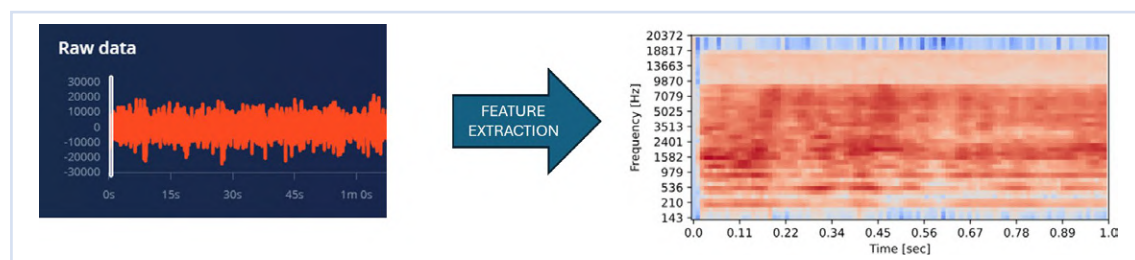



Figure 5. Le Mel-filterbank Energy Spectrogram convertit les sons de moustiques en une image 2D, en utilisant un modèle de classificateur similaire à la classification d'images.

 **Configure your target device and application budget** ✕

Target device

Define your target device requirements to inform model optimizations and performance calculations. No device yet? Use the default settings which you can change at any time.

Target device

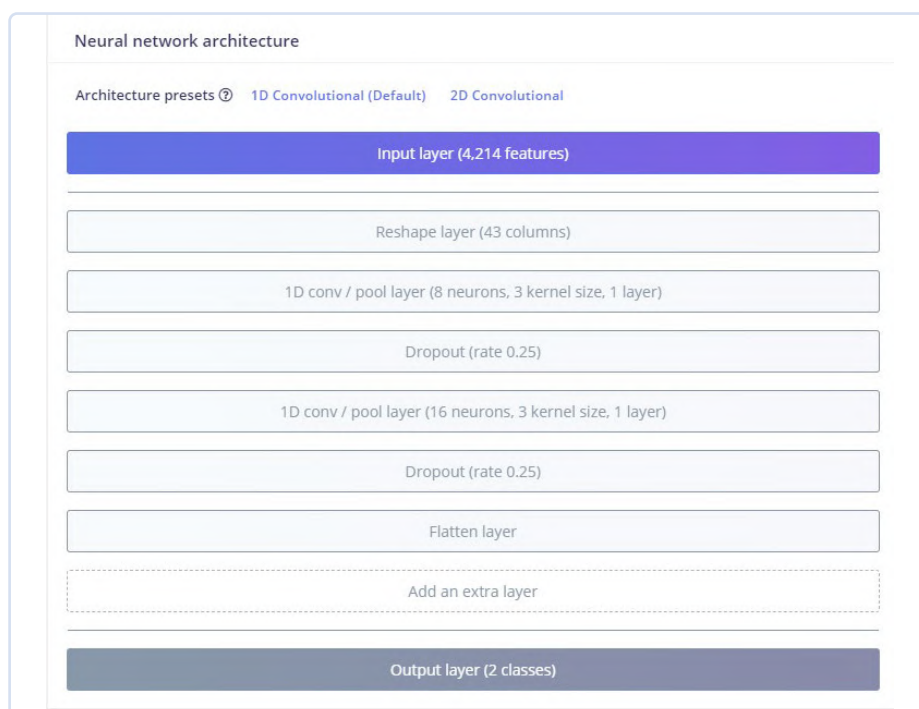
Arduino Nicla Vision (Cortex-M7 480MHz) ▼

Processor family

Cortex-M ▼

Clock rate ?

480 | MHz



deux classes distinctes de sons, nous avons choisi un modèle *Classifier* pour effectuer une classification binaire entre le bruit de fond et les moustiques. Ce type de modèle n'est pas entraîné directement sur l'entrée audio brute, mais plutôt sur des caractéristiques extraites grâce à des techniques de traitement numérique du signal (DSP) est utilisée pour dériver ces caractéristiques et les introduire dans le modèle. Edge Impulse fournit des blocs DSP préconstruits adaptés à divers types de traitement audio (**figure 4**) et optimisés pour le déploiement sur microcontrôleurs de la série Arduino Pro.

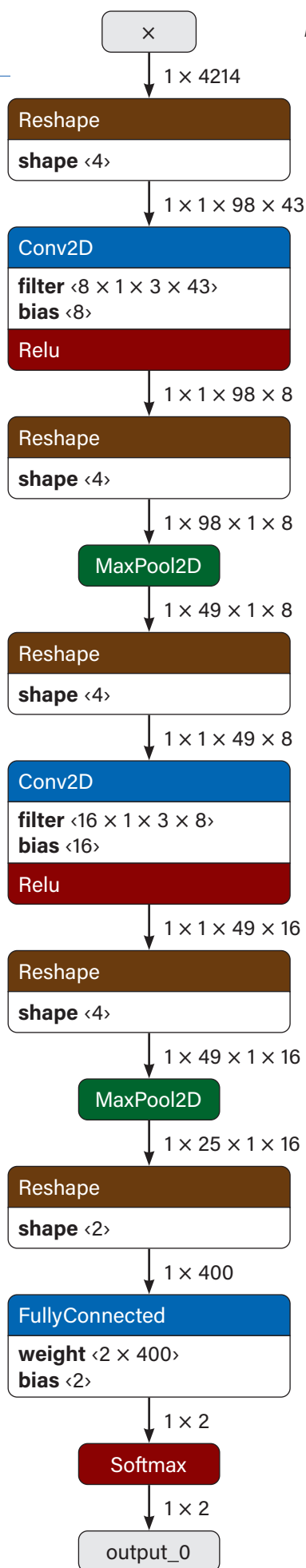


Figure 9. Schéma du réseau simple de type feed-forward.

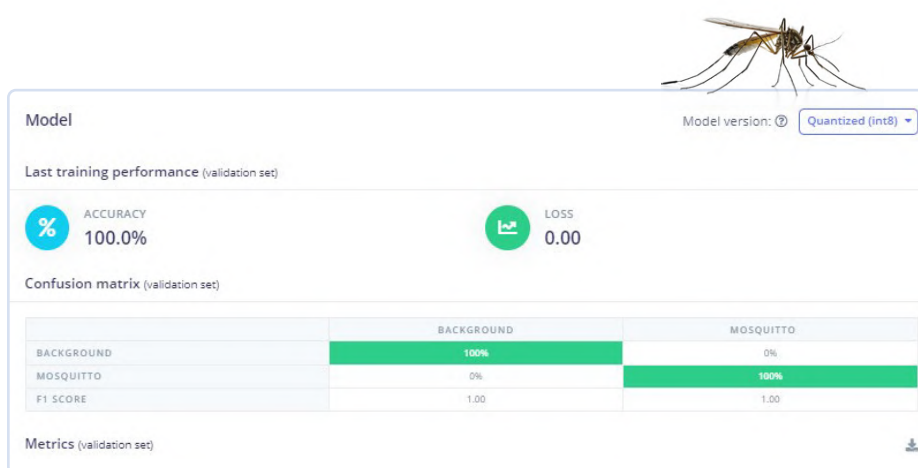


Figure 10. Le modèle a atteint une précision de 100 % et un score F1 de 1,00, classant avec succès à la fois le bruit de fond et les sons de moustiques de l'ensemble de test.

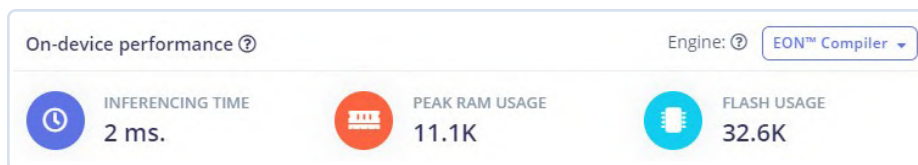


Figure 11. Les optimisations du compilateur EON d'Edge Impulse ont réduit l'utilisation de la RAM pendant l'exécution du modèle.



Figure 12. Le bloc DSP s'exécute en temps réel, avec une latence totale de 183 ms par inférence sur le Nicla Vision.

L'architecture du modèle par défaut fournie par Edge Impulse consiste en 2 couches de convolution 1D. Nous avons décidé d'opter pour ce mode (figure 8), car il s'est avéré idéal pour notre application. Graphiquement, ce modèle est un simple réseau feed-forward, comme présenté à la figure 9.

Edge Impulse en collaboration avec Arduino, a facilité et simplifié le processus de construction de modèles, offrant aux ingénieurs un accès simplifié aux résultats - en ligne avec la philosophie d'Arduino. Le résultat a été un modèle qui a fonctionné avec une précision de 100 % (figure 10) et un score F1 de 1,00 avec un ensemble de données relativement petit, classant avec précision à la fois le bruit de fond et les sons de moustiques de l'ensemble de test.

Les optimisations fournies par le compilateur EON d'Edge Impulse ont également réduit l'utilisation de la RAM du modèle à seulement 11,1 K lors de l'exécution et l'utilisation de la ROM à 32,6 K pour le stockage des poids et

du graphe d'exécution, comme le montre la figure 11.

En exécutant notre modèle sur la vision Nicla, nous pouvons atteindre une latence totale de 183 ms par inférence (figure 12). Cela offre également assez de temps pour que la machine à états commande l'actionneur, détail que nous aborderons dans la section suivante. Avec 2 MB de Flash et 1 MB de RAM partagés entre les cœurs, le Nicla Vision gère aisément ces opérations, ce qui ouvre des perspectives pour l'ajout d'autres fonctionnalités futures. Cette partie centrale du problème étant résolue, nous avons procédé à la construction finale.

Réalisation

La réalisation était assez simple en termes de connexions : nous avons utilisé l'Arduino Nicla Vision connecté à un ordinateur via USB pour alimenter la carte et surveiller les résultats du moteur d'inférence, une carte de relais abordable activée par des signaux de

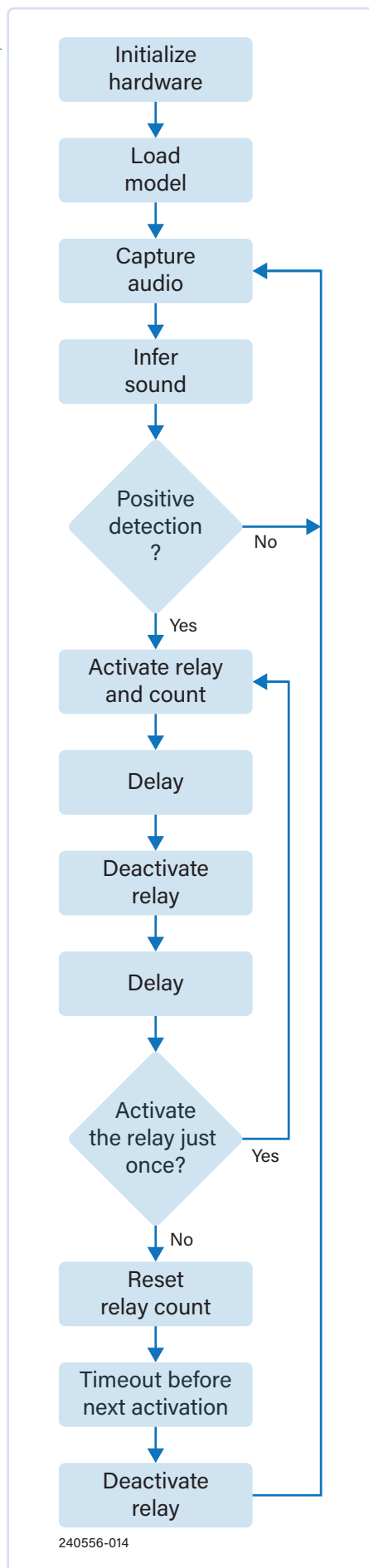


Figure 13. Organigramme illustrant le processus.

3,3 V, et un humidificateur. Le diagramme de la **figure 13** illustre cette configuration. Le code (lien de téléchargement [3]) a été généré à l'aide de Edge Impulse. Le processus peut être expliqué comme suit :

➤ Dans Edge Impulse, la base de données de sons de moustiques [4] a été téléchargée, y compris les enregistrements de bruit de fond, dans le but de construire un classificateur binaire pour détecter la présence de moustiques. Les signaux ont été enregistrés à une fréquence d'échantillonnage de 44 100 Hz. Une fenêtre d'échantillonnage

de 1000 ms a été configurée avec un décalage de fenêtre de 500 ms.

- Pour le prétraitement, le bloc Audio (MFE) (EdgeImpulse) a été utilisé, qui extrait des caractéristiques du Mel Spectrogram en intégrant les domaines temporel et fréquentiel, en mettant en évidence les basses fréquences qui sont les plus reconnaissables pour l'oreille humaine.
- Un réseau neuronal convolucional 1D a été construit pour la classification binaire pendant la formation, avec un taux d'abandon de 0,25.



Figure 14. Le modèle a atteint une précision de 100 % et un score F1 parfait de 1,00, avec des groupes bien séparés pour les sons de moustiques et les sons de fond, ce qui le rend prêt pour une utilisation pratique.

Les résultats de l'entraînement de notre modèle sont incroyablement encourageants (**figure 14**). Nous avons atteint une précision de 100%, ce qui signifie que le modèle a parfaitement identifié chaque son de moustique par rapport au bruit de fond dans l'ensemble de validation, sans aucune erreur. La valeur de perte est de 0,00, indiquant aucune erreur significative dans les prédictions du modèle. Les mesures telles que la précision, le rappel et le score F1 ont toutes atteint une valeur parfaite de 1,00, ce qui montre que le modèle est non seulement précis, mais aussi constamment fiable. L'explorateur visuel de données en bas de page le confirme, en montrant des groupes clairs et bien séparés pour les bruits de moustiques et les bruits de fond, confirmant la capacité du modèle à distinguer les deux avec une grande précision. Ceci est crucial pour une détection efficace des moustiques dans le monde réel avec l'Arduino Nicla Vision. Globalement, cette performance suggère que notre modèle est prêt à être déployé dans des contextes pratiques, où il peut aider à détecter efficacement les moustiques sur la base de leurs signatures sonores.


Vous pouvez en savoir plus sur le processus de programmation de l'Arduino Nicla Vision

avec Edge Impulse sur le site Web Arduino Docs [5]. Le code provenant d'Edge Impulse s'exécute comme thread principal dans le processeur, tandis que tout le reste du code dépend des décisions du moteur d'inférence. L'humidificateur ne se déclenchera qu'en cas de détection positive de moustiques. Une minuterie sera activée pour stopper l'humidificateur après un certain temps, et une seconde minuterie empêchera l'appareil de fonctionner en continu si le son persiste. Cette approche vise à éviter une diffusion excessive de citronnelle dans l'air en peu de temps.

Comme notre focus était sur la détection, nous avons activé l'appareil avec tout son ressemblant à celui d'un moustique. Nous avons réalisé des tests avec différents échantillons présents dans notre base de données, tous couronnés de succès.

Perspectives

La démonstration de faisabilité que nous présentons offre une solution rapide et économique pour la détection des moustiques. Ce même ensemble de données pourrait également servir à entraîner le système à reconnaître des espèces spécifiques de moustiques ou même à distinguer leur sexe, sachant que

seules les femelles sont vectrices de maladies, les mâles ne se nourrissant pas de sang. Il y a donc encore des améliorations à apporter au moteur d'inférence pour affiner la classification. Nous invitons les lecteurs d'Elektor à expérimenter et à partager leurs optimisations, contribuant ainsi à enrichir la communauté des développeurs et passionnés de tinyML ! 

240556-04

Questions ou commentaires ?

Contactez Elektor (redaction@elektor.fr).



Produits

- > **Arduino Pro Nicla Vision**
www.elektor.fr/20152
- > **Elektor Special: Guest-Edited by Arduino**
www.elektor.com/EP-0518
- > **Elektor Special: Arduino Guest Edition 2022 PDF**
www.elektor.com/ED-0518



LIENS

- [1] PAHO Report, September 2024 : <https://tinyurl.com/PAHO-report>
- [2] Arduino Nicla Vision : <https://www.arduino.cc/pro/hardware-product-nicla-vision/>
- [3] Téléchargement du code : https://github.com/dcuartielles/ictp_brazil_24
- [4] Mosquito sound database: <https://github.com/HumBug-Mosquito/ZooniverseData>
- [5] S. Romero, "Image Classification with Edge Impulse", Arduino Docs, September 2024 : <https://docs.arduino.cc/tutorials/nicla-vision/image-classification/>

L'IA d'aujourd'hui et de demain : les idées d'Espressif, d'Arduino et de SparkFun

Compilé par l'équipe de contenu d'Elektor

À mesure que la technologie de l'IA évolue, son intégration dans les systèmes logiciels et matériels devient de plus en plus complexe. Nous avons demandé à des ingénieurs d'Espressif, d'Arduino et de SparkFun - trois entreprises qui ont participé à la rédaction du magazine Elektor - de nous faire part de leurs réflexions. Nous leur avons posé la question suivante : comment envisagez-vous le rôle de l'IA dans l'évolution des solutions de votre entreprise, et anticipez-vous des défis qui pourraient survenir avec le progrès de la technologie de l'IA ?

Espressif

L'IA a le potentiel d'améliorer considérablement notre processus de développement, en améliorant l'efficacité, la précision et l'innovation. En intégrant l'IA dans nos flux de travail, nous pouvons rationaliser les différentes étapes de la conception des produits, de la logistique et du support client, ce qui permet d'obtenir des résultats plus rapides et plus fiables. Lors de la phase de conception, l'IA peut automatiser des tâches répétitives telles que la génération de code, les tests et le débogage, ce qui permet aux ingénieurs de se concentrer sur des aspects plus complexes et plus créatifs de la conception du produit. Cela permet non seulement d'accélérer le développement, mais aussi d'améliorer la qualité globale du produit final. Les capacités de l'IA en matière de simulation et de modélisation sont inestimables pendant la phase de vérification. En créant des simulations sophistiquées, nous pouvons tester différents scénarios de conception et optimiser les composants matériels avant la construction des prototypes physiques. Cela permet de réduire le nombre d'itérations et de minimiser le risque de défauts de conception coûteux, ce qui se traduit par une utilisation plus efficace des ressources et un délai de mise sur le marché plus court. L'IA peut également optimiser la logistique de la chaîne d'approvisionnement, en veillant à ce que les matériaux et les composants soient livrés juste à temps, en réduisant les coûts de stockage et en améliorant l'efficacité de la production. Elle peut analyser les données des processus de fabrication pour identifier les facteurs influençant le rendement des puces, permettant des ajustements pour améliorer l'efficacité de la production et réduire les coûts.

Anant Gupta

(responsable du marketing technique, Espressif)



L'IoT est un domaine très diversifié où les exigences des clients varient, ce qui entraîne un large éventail de demandes d'assistance. Nous pensons que l'IA peut contribuer de manière significative à répondre à ces demandes. Par exemple, dans le domaine du traitement du langage naturel (NLP), les chatbots alimentés par l'IA peuvent fournir une assistance 24 heures sur 24 et 7 jours sur 7, en répondant efficacement aux demandes des clients et en résolvant les problèmes. En tirant parti de la documentation open-source et des contributions de la communauté, l'IA peut gérer efficacement ces diverses demandes, en garantissant des réponses précises et rapides dans différents domaines et en améliorant l'ensemble de l'assistance à la clientèle.

Cependant, l'intégration de l'IA dans nos processus nécessite une réflexion approfondie. Nous devons nous assurer que l'IA est utilisée de manière éthique et responsable, en tenant compte des biais potentiels et en garantissant la confidentialité des données. En outre, nous devons investir dans la formation et le développement afin de doter notre équipe des compétences nécessaires pour exploiter efficacement les outils d'IA.

En adoptant l'IA et en relevant ses défis, nous pouvons positionner notre entreprise à la pointe de l'innovation et fournir des solutions de haute qualité qui répondent aux besoins évolutifs de nos clients.



David Cuartielles
(Co-fondateur, Arduino)

Arduino

Arduino participe à la transformation de l'IA de deux manières différentes. Tout d'abord, nous pensons que les microcontrôleurs (MCU) joueront un rôle crucial dans l'IA embarquée car ils offriront la possibilité de faire fonctionner des systèmes décentralisés à faible consommation d'énergie. Les MCU permettent de traiter l'apprentissage automatique dans des applications de petite envergure, de manière peu coûteuse et économe en énergie. L'IA embarquée, en tant que paradigme de calcul, est en accord avec la vision d'Arduino d'une électronique plus durable. De plus, il s'agit d'un meilleur moyen de faciliter l'accès à la technologie elle-même, puisqu'elle ne nécessite pas de connexion pour fonctionner, une ressource rare dans les endroits où cette technologie pourrait être la plus utile. C'est de cette manière que nous voulons contribuer aux Objectifs de Développement Durable des Nations Unies, en offrant une alternative viable à l'IA qui pourrait fonctionner n'importe où avec une très faible consommation d'énergie, ce qui lui permettrait de fonctionner avec des batteries pendant des années, ou avec de l'énergie solaire/éolienne pendant des décennies. Pour l'entraînement des dispositifs Arduino, nous soutenons actuellement l'approche de TinyML, qui consiste à effectuer un apprentissage externe, mais nous avons également

expérimenté les bibliothèques d'apprentissage intégré de Fraunhofer et d'autres outils similaires. Nous nous attendons à ce que de nouvelles alternatives d'apprentissage interne apparaissent, car les logiciels mûrissent dans ce domaine et nous comprenons de mieux en mieux comment optimiser les réseaux neuronaux pour des contextes d'utilisation spécifiques.

D'autre part, nous expérimentons et utilisons l'IA en interne pour les processus quotidiens chez d'Arduino. Outre les outils classiques tels que CoPilot pour assister le codage, ou l'IA générative pour créer des listes de tutoriels, nous utilisons des outils de traduction assistée par l'IA depuis des années. Tout notre contenu est sur Github, où nous connectons un outil tiers qui propose des traductions supervisées par des humains. Cet outil offre des traductions basées sur l'expérience acquise lors de nos interactions avec l'IA, nous permettant ainsi d'effectuer des traductions qui respectent notre terminologie propre au monde de l'électronique.

Aujourd'hui, nous vivons une transition vers cette nouvelle façon de rendre les objets intelligents. Pour ceux qui souhaitent développer une technologie de capteurs intelligents utilisant des caméras, microphones, capteurs IMU à 6 axes, etc., Arduino propose une gamme complète de cartes permettant d'expérimenter, d'entraîner des modèles de ML et de déployer des systèmes rapidement. Des cartes Nicla, à faible consommation et de petite taille, aux PLCs robustes, toutes nos cartes peuvent être programmées avec les outils de TensorFlow ou du framework, plus convivial, de Edge Impulse.

SparkFun

Je crois - et je ne pense pas être le seul - que l'"IA" est le terme le plus galvaudé au monde à l'heure actuelle. En effet, des études récentes montrent que les consommateurs se détournent des produits et services qui vantent leurs capacités « d'IA ». Ce mixeur est-il vraiment une IA ? Mon mixeur doit-il vraiment être doté d'une IA ? Si je fabrique un petit robot mignon qui passe la journée à zigzaguer dans mon salon en évitant habilement tous les obstacles sur son chemin, est-ce une démonstration d'intelligence artificielle ? Ou s'agit-il simplement d'une douzaine de lignes de code bien écrites et d'un capteur de proximité ?

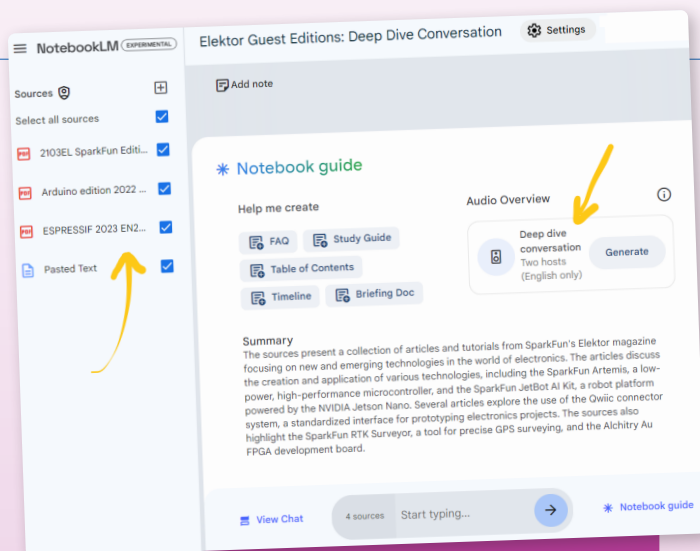
L'IA n'est pas nouvelle - ses racines remontent aux années 1950, lorsque Allen Newell, J.C. Shaw et Herbert Simon ont mis au point le Logic Theorist. Cependant, il a récemment connu une ascension fulgurante dans la conscience des gens, propulsé principalement par des programmes graphiques comme Midjourney, et aux interpréteurs de texte comme OpenAI. Comme le dit l'expression, il aura fallu 70 ans pour que l'IA devienne un succès instantané ! Cette technologie révolutionnaire va transformer notre manière d'accomplir nombre de nos tâches quotidiennes. Comme c'est toujours le cas avec une telle avancée technologique, de nombreuses craintes l'entourent, et elles ne sont pas, forcément, toutes infondées. Parmi les plus grandes peurs : une avalanche de fausses informations, les biais des IA, une disparition massive des emplois et, bien sûr, une prise de contrôle mondiale totale si l'IA décidait qu'elle doit éliminer la plus grande menace pour les humains, à savoir... les humains eux-mêmes. Les deux premiers exemples sont déjà visibles.

Rob Reynolds
(Technologue créatif,
SparkFun Electronics)



Les fausses informations qui semblent provenir de sources légitimes, en version imprimée comme en audio et vidéo, sont omniprésentes sur le web. De même, des exemples de biais dans les modèles d'IA, allant du subtil au franchement alarmant, commencent à se multiplier.

En ce qui concerne les pertes massives d'emplois, les précédents grands progrès technologiques ont montré que ce n'est généralement pas le cas. Ce qui s'est produit par le passé — et ce que je pense que nous pouvons anticiper aujourd'hui —, c'est que les métiers évolueront, que certains postes seront redéfinis, mais que de nouvelles technologies créent également de nouveaux rôles. Cela demandera peut-être à beaucoup de travailleurs d'acquiescer de nouvelles compétences ou de s'adapter à de nouvelles fonctions. Cependant, pour de nombreuses entreprises, il peut être plus judicieux de former leurs employés actuels aux nouvelles technologies, plutôt que d'embaucher massivement de nouveaux employés. Et dans des secteurs comme le service client, qui sera probablement fortement affecté par l'IA, certaines entreprises essaient déjà de remplacer la majorité de ces postes par des chatbots, tandis que les plus avisées utilisent l'IA pour les tâches répétitives, libérant leurs équipes pour qu'elles se concentrent sur l'amélioration de l'expérience client. Les études montrent que les consommateurs de la génération Z sont prêts à abandonner une entreprise après une seule mauvaise expérience, ce qui rend cette approche bien plus avantageuse.



L'édition spéciale IA : Conversation en profondeur avec l'IA

Au cours des dernières années, Elektor a collaboré avec des "rédacteurs invités" remarquables pour publier des éditions spéciales du magazine Elektor : Sparkfun a été l'invité d'un numéro en 2021, Arduino en 2022 et Espressif en 2023. Lors de la préparation de cet article, les rédacteurs d'Elektor ont téléchargé les versions PDF des trois premières éditions spéciales du magazine Elektor ainsi que des informations supplémentaires sur chaque édition, sur NotebookLM de Google. Suite à cela, une "conversation approfondie" en anglais a été générée. Toutefois, le résultat a été quelque peu décevant. Il semble que le programme n'ait pas réussi à analyser correctement le contenu des trois publications. Bien que les PDF aient été téléchargés plusieurs fois et que le processus ait été réessayé, les conversations audio qui en ont résulté, bien qu'amusantes et intéressantes, n'ont pas saisi ou reflété l'intégralité du contenu des trois éditions rédigées en collaboration avec des invités. Des idées et des contenus clés manquaient ou étaient résumés de manière inexacte, ce qui entraînait des lacunes dans la profondeur et la valeur globale de la conversation. Bien que l'IA ait fourni un chat intéressant qui semblait réaliste, elle n'avait pas une compréhension complète de certains détails techniques et de certaines nuances éditoriales. Cela a mis en évidence certaines limites du programme à évaluer avec précision trois magazines épais au contenu technique complexe et spécialisé. Malgré cela, l'outil a démontré un grand potentiel pour générer des discussions créatives, bien qu'il faille encore l'affiner pour garantir une couverture complète de tous les documents téléchargés dans le système. Nous reconnaissons également que nous aurions peut-être pu fournir un contenu et un contexte supplémentaires. Nous continuerons à expérimenter !

Les éditions spéciales du magazine Elektor

En 2021, Elektor a publié la première édition du magazine Elektor collaboration avec un invité. En collaboration avec notre premier invité spécial, SparkFun, nous avons créé et publié un magazine riche en tutoriels approfondis, de projets électroniques utilisant les produits SparkFun, de contenu exclusif et d'idées du fondateur et des équipes de SparkFun.

Fort du succès de cette édition SparkFun, nous avons invité en 2022 nos amis d'Arduino, à mettre en avant les capacités avancées de leur plateforme. Ensemble, nous avons élaboré une édition spéciale de 140 pages d'Elektor, enrichie par des contributions d'experts d'Arduino. Ce numéro spécial a bénéficié à la fois à la communauté professionnelle d'Elektor, incluant maker, ingénieurs et universitaires, ainsi qu'à Arduino.

En 2023, notre collaboration s'est tournée vers Espressif, explorant les futurs possibles offerts par les technologies d'intelligence artificielle générative, comme ChatGPT. Cette édition a examiné comment ces technologies pourraient transformer les industries et comment Espressif envisage de les utiliser pour innover, enrichie par des perspectives de notre communauté et de nos partenaires. Les magazines rédigés en collaboration avec les invités ont été publiés (en version imprimée et numérique) et commercialisés en anglais, en allemand, en néerlandais et en français. Cette initiative a permis à nos invités et à Elektor de présenter leurs marques sur de nouveaux marchés et de promouvoir nos produits auprès de nouveaux publics. Les retours ont été exceptionnels, chaque édition touchant des centaines de milliers de lecteurs passionnés.



240517-04

LIENS

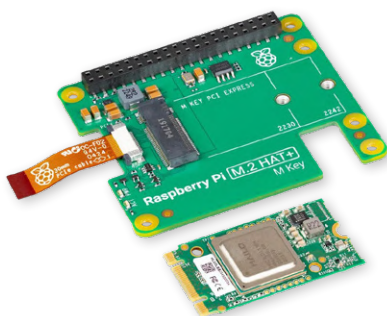
- [1] Espressif Guest-Edited edition of ElektorMag (2023): <https://elektormagazine.com/elektor-espressif>
- [2] Arduino Guest-Edited edition of ElektorMag (2022): <https://elektormagazine.com/elektor-arduino>
- [3] SparkFun Guest-Edited edition of ElektorMag (2021): <https://elektormagazine.com/magazine/elektor-171>

Mettez la main sur du matériel d'IA de pointe !

Kit Raspberry Pi AI

Le Raspberry Pi AI Kit comprend le HAT M.2 pour Raspberry Pi et un module d'accélération IA Hailo, destiné à être utilisé avec le Raspberry Pi 5. Il offre un moyen accessible, économique et écoénergétique d'intégrer une IA performante. Découvrez des applications telles que le contrôle de processus, la sécurité, l'automatisation domestique et la robotique !

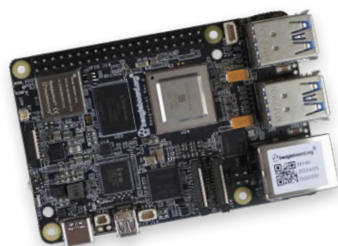
www.elektor.fr/20879



Unitree Go2 Pro Robot quadrupède

L'Unitree Go2 est un robot quadrupède conçu pour la recherche et le développement de systèmes autonomes dans les domaines de l'interaction homme-robot (HRI), SLAM et le transport. Grâce à ses quatre jambes et à ses 12 degrés de liberté (12DOF), ce robot peut s'adapter à divers types de terrains.

www.elektor.fr/20357



BeagleY-AI SBC avec GPU, DSP et accélérateurs d'IA

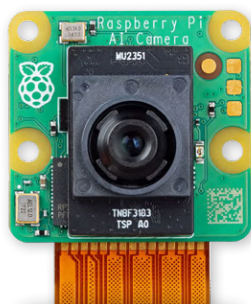
Le BeagleY-AI est un ordinateur monocarte 64 bits quadricœur, puissant, open-source et abordable, équipé d'un GPU, d'un DSP et d'accélérateurs de vision/apprentissage profond, conçu pour les développeurs et les créateurs. Les utilisateurs peuvent profiter des images logicielles Debian Linux fournies par BeagleBoard.org, qui incluent un environnement de développement intégré.

www.elektor.fr/20991

Raspberry Pi AI Camera

La Raspberry Pi AI Camera est un module caméra compact basé sur le capteur de vision intelligent Sony IMX500. L'IMX500 combine un capteur d'image CMOS de 12 MP avec une accélération d'inférence intégrée pour divers modèles de réseaux neuronaux, permettant aux utilisateurs de développer des applications IA avancées basées sur la vision sans nécessiter d'accélérateur séparé.

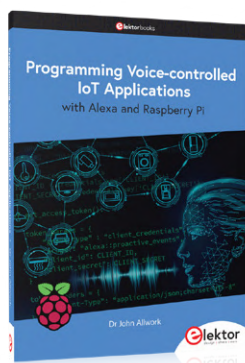
www.elektor.fr/20953



Kit de développement Waveshare Jetson Orin Nano AI

Ce kit de développement informatique pour l'IA en périphérie est basé sur le module Jetson Orin Nano et propose une grande variété d'interfaces périphériques telles que M.2, DP, USB, etc. Le kit est également livré avec une carte réseau sans fil AW-CB375NF préinstallée, qui prend en charge Bluetooth 5.0 et le Wi-Fi double bande, avec deux antennes PCB supplémentaires, pour assurer une connexion réseau sans fil rapide et fiable ainsi qu'une communication Bluetooth.

www.elektor.fr/20762



Programming Voice-controlled IoT Applications with Alexa and Raspberry Pi

Ce livre est divisé en deux parties : la création de compétences Alexa et la conception de dispositifs IoT (Internet des objets) utilisant un Raspberry Pi. Il couvre des sujets tels que le développement de compétences Alexa, les achats intégrés, et l'utilisation d'AWS Lambda, ainsi que la création d'appareils domestiques intelligents contrôlés par Alexa. Les lecteurs apprennent également la messagerie MQTT, la création de notifications proactives, et la transformation d'un Raspberry Pi en un appareil Alexa autonome.

www.elektor.fr/20400

chronologie de l'intelligence artificielle

L'équipe rédactionnelle d'Elektor a utilisé ChatGPT et Gemini pour concevoir une frise chronologique détaillée retraçant l'histoire de l'IA, spécialement destinée aux ingénieurs en électronique, aux fabricants de composants électroniques et aux étudiants en ingénierie. Cette chronologie met l'accent sur les avancées clés en matière d'algorithmes, de matériel et d'applications d'IA pertinents pour ces audiences.

**Années
1940-
1950**

premiers concepts et bases

1943 Modèle de neurones de McCulloch et Pitts: Modèle mathématique de neurones artificiels, jetant les bases des réseaux neuronaux.

1950 « Computing Machinery and Intelligence » d'Alan Turing. Turing présente le test de Turing, proposant un moyen de déterminer si une machine peut avoir un comportement intelligent indiscernable de celui d'un humain.



1956 Conférence de Dartmouth : John McCarthy et d'autres inventent le terme « intelligence artificielle » et proposent des idées fondamentales pour l'apprentissage automatique et l'IA symbolique.

**Années
1960**

les débuts de la recherche en IA

1961 Unimate : Le premier robot industriel, Unimate, est présenté et commence à travailler sur les chaînes de montage de General Motors, illustrant l'automatisation de la fabrication.

1966 ELIZA : Joseph Weizenbaum développe ELIZA, un programme qui imite la conversation humaine, démontrant ainsi les premières capacités de traitement du langage naturel.

1969 Shakey le robot : Développé à Stanford, Shakey est le premier robot mobile polyvalent capable de percevoir et de raisonner sur son environnement.

**Années
1970**

l'hiver de l'IA et des systèmes experts

1970 Langage de programmation Prolog : un langage de programmation logique crucial pour la recherche en IA, en particulier pour les systèmes experts.

1972 Robot humanoïde WABOT-1 : L'université de Waseda, à Tokyo, développe un robot humanoïde de taille réelle, démontrant les progrès de la robotique et de la technologie des capteurs.

**1974-
1980**

Premier hiver de l'IA : Les coupes budgétaires et les attentes trop ambitieuses entraînent une réduction des activités de recherche, ce qui marque le premier hiver de l'IA.



**Années
1980**

Essor de l'apprentissage automatique et des réseaux neuronaux

1980 Essor des systèmes experts : des systèmes tels que XCON pour la configuration des systèmes informatiques remportent un succès commercial, suscitant l'intérêt pour les applications de l'IA.

1986 ● Algorithme de rétropropagation : David Rumelhart, Geoffrey Hinton et Ronald Williams popularisent l'algorithme de rétropropagation pour l'apprentissage des réseaux neuronaux.

Années

1990 ● **IA et puissance de calcul**

1992 ● TD-Gammon : Gerald Tesauro développe un programme de jeu de backgammon qui apprend à jouer à un niveau élevé grâce à l'apprentissage par renforcement.

1997 ● Deep Blue contre Garry Kasparov : Deep Blue d'IBM bat le champion du monde d'échecs Garry Kasparov, démontrant le potentiel de l'IA dans la résolution de problèmes complexes.

Années

2000 ● **L'IA dans la robotique et les applications grand public**

2000 ● ASIMO : Honda dévoile ASIMO, un robot humanoïde avancé capable de marcher et d'interagir avec les humains.

2005 ● Stanley : Le véhicule autonome de Stanford utilise des algorithmes d'IA pour traiter les données des capteurs, planifier les itinéraires et contrôler sa direction, son accélération et son freinage.

2006 ● Création du terme « Deep Learning » (apprentissage profond) : Geoffrey Hinton et son équipe introduisent les techniques d'apprentissage profond, ce qui permet des avancées en matière de vision par ordinateur et de reconnaissance vocale.

Années

2010 ● **Essor de l'IA et de l'apprentissage profond**

2011 ● IBM Watson gagne « Jeopardy ! » : Watson bat des champions humains dans l'émission « Jeopardy ! », démontrant un traitement avancé du langage naturel et une recherche de connaissances.

2012 ● AlexNet remporte la compétition ImageNet : AlexNet révolutionne la vision par ordinateur en remportant la compétition ImageNet grâce à l'utilisation de CNN profonds.

2015 ● Création d'OpenAI: OpenAI est créée pour promouvoir et développer l'IA conviviale.

2016 ● AlphaGo bat Lee Sedol : AlphaGo de DeepMind bat le champion de Go Lee Sedol, mettant en évidence la puissance de l'apprentissage par renforcement profond et de la recherche arborescente de Monte Carlo.

Années

2020 ● **L'IA dans l'industrie et la vie quotidienne**

2020 ● Publication du GPT-3 : Le modèle de langage peut générer un texte comparable à celui d'un humain, démontrant ainsi le potentiel des transformateurs à grande échelle dans le traitement du langage naturel.



2021 ● GPU A100 de NVIDIA : Le GPU A100 marque une avancée significative dans l'accélération matérielle des applications d'apprentissage profond.

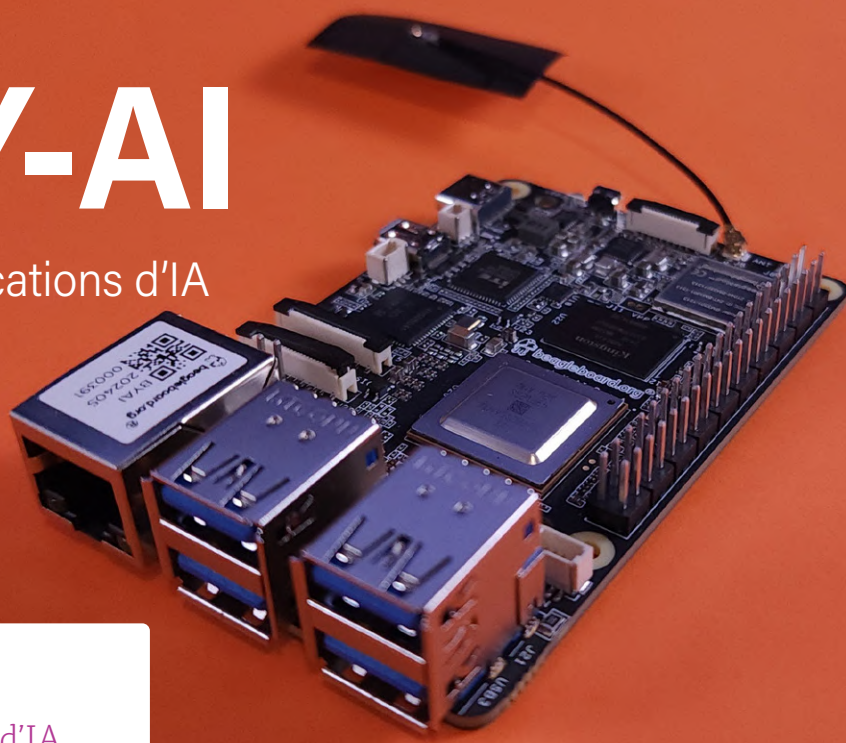
2022 ● DALL-E et la synthèse d'images : DALL-E est capable de générer des images à partir de descriptions textuelles.

2023 ● Assistants IA : L'adoption généralisée d'agents conversationnels alimentés par l'IA améliore l'interaction homme-machine dans diverses applications. Google Gemini a été officiellement annoncé le 6 décembre.

Image Source: Adobe Stock
Image Chat GPT source: Amika Studio - stock.adobe.com

BeagleY-AI

Le dernier SBC pour les applications d'IA



Brian Tristam Williams (Elektor)

Le BeagleY-AI, avec ses capacités d'IA et sa conception open-source, est-il le choix idéal pour votre prochain projet ? Continuez votre lecture pour évaluer les atouts et les contraintes de ce SBC à la fois compact et puissant.

Le BeagleY-AI est le dernier né de BeagleBoard.org dans le monde des ordinateurs mono-cartes, ciblant cette fois le domaine en pleine croissance de l'IA et de l'apprentissage automatique. Compact, open-source et doté d'un hardware spécialisé, le BeagleY-AI se positionne comme un outil polyvalent pour les développeurs et chercheurs qui ont besoin d'une plateforme fiable pour les tâches liées à l'IA. Mais est-il à la hauteur de l'engouement qu'il suscite ? Voyons cela de plus près.

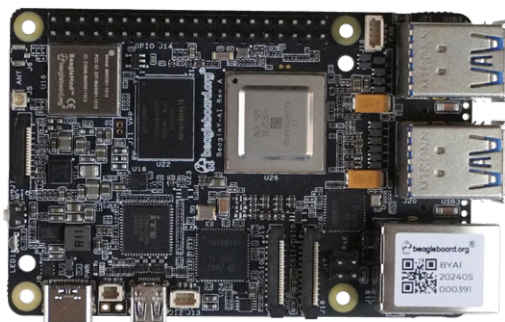


Figure 1. BeagleY-AI, vue de dessus.

Matériel et conception

On trouve au cœur du BeagleY-AI le processeur Texas Instruments AM67A, une UC Cortex-A53 quadricœur cadencée à 1,4 GHz. Cette UC peut sembler modeste comparée aux vitesses d'horloge plus élevées que l'on trouve dans d'autres SBC, mais elle a été développée pour des applications industrielles et automobiles. Texas Instruments produit des SoC (System on a Chip) conçus pour une fiabilité à toute épreuve, avec des capacités de commande en temps réel pilotées par les cœurs MCU Cortex-R5 et l'accélération de l'IA par deux DSP C7x. Ces accélérateurs sur puce permettent au BeagleY-AI d'offrir de puissantes performances en IA et de gérer des tâches précises et critiques, même dans des environnements exigeants.

Pour moi, la valeur réside dans le matériel accolé à la puce. L'AM67A est complété par deux DSP C7x C7x avec accélérateurs de multiplication de matrice (MMA), capables de fournir jusqu'à 4 TOPS (Téra-opérations par seconde) en performances d'IA. Il convient donc aux applications d'apprentissage en profondeur, telles que la reconnaissance d'images et la détection d'objets en temps réel.

Il est important de noter que le chiffre de 4 TOPS se réfère spécifiquement aux opérations sur 8 bits, qui sont couramment utilisées dans les tâches d'inférence de l'IA pour optimiser la vitesse et l'efficacité. Pour plus de précision, avec des opérations sur 16 ou 32 bits, les performances en TOPS seraient moindres.

La carte (**figure 1**) est bien équipée en termes de

connectivité, avec quatre port USB 3.0, un port USB 2.0 type C et un port Ethernet Gigabit. De plus, elle accepte le Wi-Fi 6 (802.11ax) et le Bluetooth 5.4 grâce au module BeagleBoard BM3301. Pour les affichages, on dispose d'un port micro HDMI, d'un connecteur OLDI (LVDS) au verso (**figure 2**), et d'une interface MIPI-DSI, capable de piloter jusqu'à trois écrans simultanément. Le stockage est assuré avec un emplacement pour une carte micro-SD, ce qui est la norme pour les SBC de cette catégorie et ce facteur de forme. L'une des caractéristiques essentielles est la présence d'un connecteur PCIe Gen3x1 FPC à 16 broches, qui offre des options d'extensions à haut débit. La carte est aussi sans ventilateur, une propriété bienvenue pour les utilisateurs soucieux du bruit et de la fiabilité pour des déploiements durables. Tout cela est bien sûr complété par un connecteur GPIO à 40 broches, apte à de multiples personnalisations et intégrations de périphériques.

Sans surprise, les GPIO à 40 broches vous sont sûrement familiers : le BeagleY-AI a un facteur de forme compatible du Raspberry Pi (**figure 3** et **figure 4**), ce qui permet aux utilisateurs de profiter du vaste écosystème d'accessoires bâti autour des cartes Raspberry Pi. Cela comprend la compatibilité avec de nombreux boîtiers, écrans et autres périphériques, augmentant ainsi la polyvalence de la carte. En outre, BeagleBoard.org offre le libre accès aux fichiers de conception, ce qui permet aux utilisateurs de personnaliser et adapter leur carte pour répondre aux besoins spécifiques de leur projet.

Bien que compact, le BeagleY-AI a été pensé pour les applications industrielles. L'utilisation de composants Texas Instrument garantit sa robustesse et sa disponibilité à long terme, ce qui en fait une solution viable à la fois pour le développement de prototypes que pour le déploiement en environnement de production. La nature open-source de la carte permet également l'adaptation à des besoins spécifiques, que ce soit par modification de la conception matérielle ou par optimisation du logiciel pour des cas d'usage particuliers.

Logiciel et installation

Le BeagleY-AI utilise un système d'exploitation basé sur Debian, avec un environnement de bureau XFCE doté d'une interface utilisateur légère et réactive. La prise en main de la carte est assez simple, surtout si vous avez déjà travaillé avec des dispositifs similaires. Le processus d'installation consiste à flasher l'OS sur une carte micro-SD, à configurer le fichier `sysconf.txt` et à mettre à jour le logiciel via la ligne de commande. BeagleBoard.org fournit une documentation détaillée, ce qui est primordial si on n'est pas familier avec les subtilités des SBC.

Pour les utilisateurs de Windows, le processus de flashage de l'image peut s'avérer un peu lourd, nécessitant d'autres outils comme Balena Etcher ou `bb-imager.exe` de BeagleBoard et l'obligation de décompresser le fichier image avant de l'utiliser. J'ai eu quelques petits soucis avec cela sur ma machine Windows, mais rien d'insurmontable.

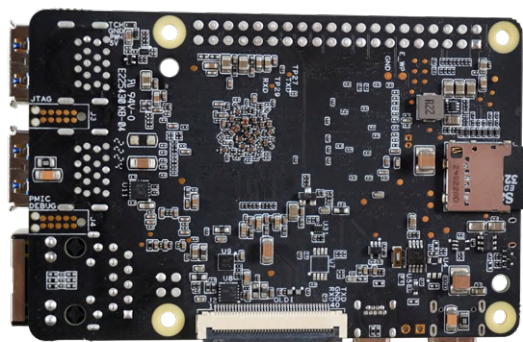


Figure 2. BeagleY-AI, vue de dessous.

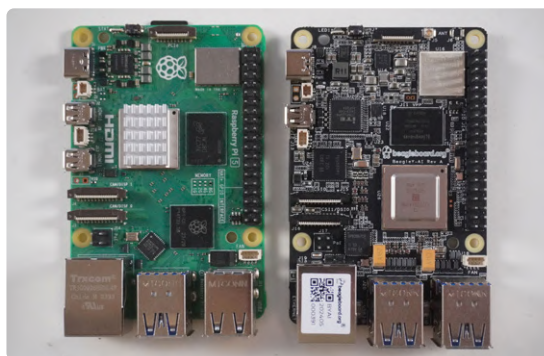


Figure 3. Le BeagleY-AI a un facteur de forme de Raspberry Pi.



Figure 4. Les quatre ports USB-A du BeagleY-AI sont en USB 3.0.

Une fois installé, le BeagleY-AI fonctionne très bien, malgré quelques imperfections. Par exemple, les bibliothèques accélérées par le GPU sont bien présentes, mais tous les composants nécessaires, tels que le compilateur pour DSP, ne le sont pas. C'est un inconvénient pour les développeurs qui cherchent à exploiter les capacités d'IA de la carte dès son déballage.

Ceux qui cherchent à intégrer des fonctionnalités d'IA dans leurs projets apprécieront que de multiples outils de développement soient pris en charge par la carte. Par exemple, on peut utiliser TensorFlow Lite pour exécuter des modèles d'apprentissage automatique, tandis qu'OpenCV fournit un cadriciel robuste pour les tâches de traitement d'images. La communauté a déjà commencé à proposer des tutoriels et des guides, une ressource inestimable pour les débutants sur la plateforme.

Un des aspects qui conforte le BeagleY-AI est l'engagement fort de Texas Instrument à traiter en amont ses contributions au noyau Linux. TI suit une politique de « priorité à l'amont » en publiant les modifications du noyau et du logiciel sur la branche principale du noyau Linux dès que le silicium est sorti. Ceci garantit un soutien à long terme et la compatibilité avec une grande variété d'écosystèmes logiciels.

Performance et cas d'usage

Le BeagleY-AI se distingue dans les tâches liées à l'IA grâce à son hardware spécialisé. La combinaison des deux DSP C7x et des MMA permet l'exécution efficace d'algorithmes d'apprentissage automatique, ce qui renforce son choix pour les applications de vision par ordinateur, de robotique et d'automatisation. Pendant les essais, la carte a mené sans problème des tâches de détection d'objets TensorFlow Lite, bien qu'avec un peu de dégagement de chaleur — les températures ont atteint près de 74°C lors d'opérations intensives, ce qui est chaud mais gérable sans refroidissement actif. Dans les applications pratiques, les capacités du BeagleY-AI vont au-delà de l'IA. Son solide jeu d'interfaces, comprenant PCIe, USB 3.0 et Gigabit Ethernet le rend capable de multiples tâches industrielles. Que vous réalisiez un système de caméra intelligent, automatisiez une chaîne de production ou développiez un bras robotique, la flexibilité de la carte garantit son adaptation aux exigences spécifiques du projet. Le connecteur GPIO à 40 broches, compatible avec de nombreux périphériques, renforce encore sa polyvalence, en permettant l'intégration de capteurs, d'actionneurs et autres matériels.

Toutefois, tout n'est pas si parfait. L'UC relativement modeste de la carte ne remplacera pas un ordinateur de bureau pour l'informatique générale, et les

4GO de RAM, bien que suffisants pour la plupart des cas d'usage, pourraient constituer un goulot d'étranglement pour les applications plus gourmandes en mémoire. De plus, bien que la carte prenne en charge l'extension PCIe, elle se limite à une seule voie Gen3, ce qui peut brider les performances de certains périphériques.

Un domaine où le BeagleY-AI se distingue est son potentiel pour les applications en temps réel. L'inclusion des cœurs Cortex R5 permet des opérations d'E/S à faible latence, aspect crucial dans des environnements où la synchronisation est critique. Cela en fait un candidat sérieux pour une utilisation dans des domaines tels que l'appareillage médical, où la précision et la fiabilité sont primordiales.

Communauté et support

L'une des principales forces du BeagleY-AI est sa nature open-source. Tous les fichiers de conception matérielle et mécanique sont disponibles, permettant aux utilisateurs avancés de modifier la carte pour l'adapter à leurs besoins spécifiques. Texas Instrument veille aussi à ce que ses processeurs soient facilement accessibles aux développeurs et makers, en offrant le libre accès aux fichiers de conception et aux ressources logicielles, ce qui repasse les commandes aux mains des utilisateurs. Cette ouverture s'étend au logiciel, avec un robuste écosystème basé sur Debian qui bénéficie d'un soutien important de sa communauté. Toutefois, il convient de noter que le processeur AM67A lui-même n'est pas open-source, un sujet potentiel de controverse au sein de la communauté.

La documentation fournie par BeagleBoard.org est complète, couvrant presque tout, de la configuration initiale aux cas d'usage avancés. Cela dit, le suivi du logiciel pourrait être amélioré. À la mi-2024, certaines fonctionnalités et outils d'IA, tels que le cadriciel Edge AI, ne sont pas totalement pris en charge, ce qui pourrait limiter l'attrait de la carte pour les développeurs de haut-niveau. Les forums et les ressources de la communauté sont actifs, mais la sortie relativement récente de la carte fait que certains problèmes sont encore en cours de traitement.

La communauté de BeagleBoard.org est un des écosystèmes les plus actifs et collaboratifs du monde des SBC. Les utilisateurs partagent régulièrement des projets, prodiguent des conseils de dépannage, et contribuent au développement de nouvelles fonctionnalités. Ce niveau d'engagement de la communauté est un plus significatif pour les nouveaux utilisateurs, car il montre qu'il y a une manne de connaissances à exploiter. Développeur chevronné ou amateur, vous verrez que la communauté est accueillante et soucieuse d'aider.

Pour et contre

Pour :	Contres :
Performance de l'IA : Les deux DSP C7x et les MMA délivrent jusqu'à 4 TOPS, parfait pour les tâches d'apprentissage automatique.	Performance de l'UC : Le quadricœur à 1,4 GHz Cortex-A53 est bien terne face aux SBC plus récents.
Connectivité : USB 3.0, Gigabit Ethernet, Wi-Fi 6, et Bluetooth 5.4, la carte est bien équipée pour diverses applications.	Taille de la RAM : 4 GO de RAM LPDDR4 pourraient ne pas suffire pour toutes les applications.
Extensions : Le connecteur PCIe Gen3 x1 et le GPIO à 40 broches offrent de nombreuses possibilités de personnalisation.	Lacunes logicielles : Certaines fonctionnalités et certains outils d'IA ne sont pas totalement pris en charge, ce qui limite les capacités de la carte au déballage.
Hardware open-source : Les utilisateurs peuvent accéder à tous les fichiers de conception du hardware et les modifier, ce qui favorise l'innovation et l'adaptation.	Dissipation thermique : La carte chauffe en charge, et bien que sans ventilateur, certains utilisateurs préféreront un refroidissement actif.
Composants de qualité industrielle : L'utilisation de matériel Texas Instruments, gage de fiabilité et de support à long terme, le prédispose à la fois au développement et au déploiement.	

Futur potentiel


Pour l'avenir, le potentiel de croissance du BeagleY-AI est important. Avec la progression de l'écosystème, on peut s'attendre à une prise en charge plus globale des cadriciels et outils d'AI. Les mises à jour du micrologiciel pourraient débloquent de nouvelles fonctionnalités ou en améliorer d'existantes, en particulier dans le domaine du traitement de l'IA et de l'intégration de périphériques. De plus, des développeurs tiers sont susceptibles de créer des compléments et extensions qui améliorent les capacités de la carte et la rendent encore plus polyvalente.

La nature open-source du BeagleY-AI sous-entend qu'il ne s'agit pas seulement d'un produit statique, mais d'une plateforme qui peut évoluer. Les utilisateurs prêts à s'engager avec la communauté et à contribuer au projet découvriront qu'ils peuvent façonner le devenir de la carte, en veillant à ce qu'elle réponde aux besoins de multiples applications. Cette adaptabilité est l'une des principales raisons qui justifient de considérer le BeagleY-AI pour tout projet d'IA ou embarqué.

Est-ce que cela vous convient ?

Le BeagleY-AI est une solution convaincante pour les développeurs et ceux qui cherchent un SBC polyvalent et capable d'IA. Sa conception open-source avec

le puissant processeur AM67A, le prédispose à un grand champ d'applications, en particulier pour l'IA et l'apprentissage automatique. Toutefois, cela ne va pas sans limitations. La performance de l'UC est modeste, la RAM peut ne pas être suffisante pour des tâches plus exigeantes et certaines fonctionnalités d'IA nécessitent plus de suivi logiciel.

Cela dit, les points forts de la carte — tels que sa robuste accélération d'IA, ses options de connectivité étendues et son extensibilité — en font un outil précieux pour ceux qui sont prêts à accepter ses défauts. Avec la progression de l'écosystème, et la poursuite de la contribution de la communauté, le potentiel du BeagleY-AI va sans doute augmenter, pour devenir un choix encore plus attractif pour les projets d'IA. Que ce soit pour explorer des applications d'IA de pointe ou pour une plateforme fiable pour l'automatisation industrielle, le BeagleY-AI constitue une base solide qui encourage l'innovation et l'expérimentation. 

VF : Denis Lafourcade — 240527-04

lumière sur l'

Les perspectives de la communauté Elektor



L'équipe Elektor

Explorez l'avenir de l'intelligence artificielle à travers les yeux des membres de la communauté Elektor. Découvrez leurs différents points de vue sur les avantages et les inconvénients de l'IA et la façon dont elle façonne le monde de l'électronique ainsi que la société dans son ensemble.

Que pensez-vous du potentiel de l'IA et de son impact éventuel ?

L'IA a le potentiel d'améliorer considérablement nos vies en augmentant l'efficacité et en résolvant des problèmes complexes. Toutefois, son impact sur l'apprentissage et la dépendance excessive suscite des inquiétudes. À mesure que l'IA automatise de plus en plus de tâches, les gens pourraient devenir trop dépendants, ce qui risquerait d'entraîner un déclin de la pensée critique et des compétences en matière de résolution de problèmes. En fin de compte, l'IA peut être un outil puissant si elle est utilisée à bon escient, mais il est essentiel de trouver un équilibre entre ses avantages et la nécessité de préserver la créativité humaine et l'indépendance intellectuelle.

Saad Imtiaz (Elektor)



Comment l'IA a-t-elle amélioré votre capacité à résoudre les problèmes d'ingénierie ?
Pouvez-vous donner des exemples d'idées ou de gains d'efficacité obtenus grâce à son utilisation ?

J'utilise principalement ChatGPT. La plupart des problèmes que je rencontre se sont déjà posés à d'autres personnes et les solutions sont documentées sur Internet, mais encore faut-il les trouver ! Avec Google, si vous n'utilisez pas exactement le bon mot-clé, il peut être difficile de trouver ce que vous cherchez. ChatGPT, en revanche, est très doué pour reformuler, faire des associations d'idées, trouver des informations pertinentes et les synthétiser. Je lui demande aussi de m'expliquer des sujets complexes lorsque je suis bloqué, ou de me suggérer des algorithmes de programmation. Bien sûr, il a de nombreux défauts, mais c'est tout de même un collègue virtuel incroyablement utile et infatigable !

Jean-François Simon (Elektor)



Considérez-vous l'intelligence artificielle comme un outil qui renforcera la créativité humaine dans le domaine de l'électronique ou pensez-vous qu'elle pourrait étouffer l'innovation ?

Je pense que tout dépend de l'utilisation qui en est faite. Par exemple, il peut s'agir d'un outil formidable lorsqu'il est utilisé pour l'automatisation. Au lieu d'effectuer des tâches répétitives et fastidieuses, les ingénieurs et les passionnés peuvent se concentrer sur des aspects plus complexes et créatifs de l'électronique. Cependant, si elle n'est pas utilisée correctement, certains peuvent développer une forte dépendance à l'IA, ce qui, en fin de compte, peut affecter leur confiance pour innover par eux-mêmes.

Glaucileine Vieira (Elektor)



Comment percevez-vous le rôle de l'IA dans l'avenir de l'ingénierie ?
Pensez-vous que son impact sera davantage transformateur ou perturbateur ?

EB

Considérée comme un ensemble d'algorithmes capables de résoudre efficacement des problèmes complexes, l'IA semble parfaitement adaptée à la résolution de certains des défis que l'ingénierie s'efforce de relever. Les innovations disruptives s'accompagnent souvent d'effets négatifs sur les plans social, écologique et économique. Actuellement, de nombreux domaines de l'ingénierie souffrent d'un manque de ressources. L'IA pourrait contribuer à réduire les ressources nécessaires et à intégrer l'écologie et l'économie dans l'équation. Cependant, cela pourrait simplement déplacer le besoin de ressources vers d'autres compétences.

Erwann Berlivet (France)

Dans quels domaines de l'ingénierie pensez-vous que la supervision et la prise de décision humaines sont irremplaçables par l'IA, et pourquoi est-ce le cas ?

L'IA a fait de grands progrès, mais la supervision humaine est vitale dans les domaines impliquant la complexité, la créativité, l'éthique et la sécurité. L'IA manque d'intuition et ne peut pas gérer des scénarios imprévus ou des dilemmes moraux comme les humains. Dans les situations à fort enjeu, l'adaptabilité et le jugement humain sont cruciaux. La gestion de systèmes complexes, l'interprétation des lois et l'équilibre des impacts sociétaux sont des tâches où l'intuition humaine excelle. En fin de compte, la conception pour l'homme exige une compréhension profonde qui fait encore défaut à l'IA.

Clemens Valens (Elektor)

Que pensez-vous de l'IA ? En avez-vous déjà fait l'expérience ?

Oui, j'ai eu quelques bonnes expériences avec l'IA, principalement avec de jeunes collègues inexpérimentés qui ont tendance à lui accorder trop de confiance, donc j'ai beaucoup à dire à ce sujet... Laissez l'IA concevoir et programmer le système de contrôle/stabilisation des ballasts de votre nouveau sous-marin. Combien d'axes de stabilité a un sous-marin ? Mon petit, peux-tu faire la différence entre la stabilité statique et dynamique d'un sous-marin ? Ne t'embête pas, laisse l'IA faire toute la réflexion, apprendre trop de physique de nos jours, c'est une perte de temps de toute façon. Lorsque le sous-marin perdra le contrôle et coulera lors de sa première plongée d'essai, ce sera la faute de l'IA. [Rires] Nous avons montré l'outil de crochetage/décodage de serrure «Sputnik» à l'IA, qui l'a identifié comme une «installation artistique». Si ce n'est pas largement diffusé sur Internet, l'IA ne le reconnaîtra pas. J'ai réussi à trouver l'inventeur de «Sputnik», un ingénieur en mécanique de Split, en Croatie. Je l'ai rencontré à Split, il a déposé plusieurs brevets dans les années 1980 et 1990 et les a vendus à des entreprises de serrurerie allemandes. Il m'a parlé de certaines serrures avancées que je n'ai pu trouver qu'après une recherche approfondie dans les bases de données de brevets. J'ai ensuite demandé à l'IA de décrire le fonctionnement de la serrure «Fangkalotte» («calottes de piège», éléments mécaniques utilisés pour bloquer et piéger les outils de crochetage de serrure) conçue par la société Winkhouse (une société allemande). Au lieu du mot allemand «fangen», qui signifie «piéger» ou «attraper», l'IA a pensé qu'il s'agissait du mot anglais «fang» (une longue dent pointue de carnivores comme les chiens ou les chats), et m'a donné une réponse incroyablement fautive. Bien sûr, je l'ai sauvegardé, je peux le montrer à toute personne intéressée. Pourquoi l'IA ne peut-elle pas honnêtement dire «Désolé, je n'en ai jamais entendu parler» ? Essaie-t-elle de se faire passer pour un être humain ? [Rires] L'IA peut être un outil utile, mais seulement si vous l'utilisez de manière responsable, et elle ne peut pas compenser votre manque de connaissances et d'expérience, encore moins de responsabilité.

Luka Matic (Croatie)

Pouvez-vous identifier des scénarios d'ingénierie dans lesquels le recours à l'IA pourrait compromettre la sécurité, l'éthique ou la qualité ? Qu'est-ce qui fait que ces domaines ne se prêtent pas à l'IA ?

Le risque de compromettre l'intégrité d'un processus décisionnel dans lequel intervient un système d'IA est directement proportionnel à la profondeur de son intervention, c'est-à-dire à la part que l'homme lui a déléguée. Le contrôle d'un avion ou un diagnostic médical, par exemple, sont des scénarios dans lesquels l'importance de la prise de décision humaine doit primer pour la sécurité. De même, l'utilisation de l'IA pour décider à qui attribuer un poste dans le cadre d'un concours public avec des milliers de candidats pourrait rendre le processus de sélection non transparent, invérifiable et, par conséquent, contraire à l'éthique. Enfin, dans le domaine industriel, l'évaluation de la qualité d'un produit nécessite souvent l'évaluation de résultats provenant de plusieurs processus interdépendants, avec des différences très subtiles que même les systèmes d'IA les plus avancés pourraient ne pas comprendre aujourd'hui.

Roberto Armani (Elektor)

Comment percevez-vous le rôle de l'IA dans l'avenir de l'ingénierie électronique ?
Pensez-vous que son impact sera davantage transformateur ou perturbateur ?

Pour l'instant, les développeurs utilisent principalement l'IA comme un puissant moteur de recherche ou pour des tâches de codage simples - en gros, c'est le travail stupide pour lequel les gens aiment être aidés. L'IA commet encore beaucoup d'erreurs, dont la plupart sont (heureusement) évidentes. Cependant, les premières voitures auto-conduites présentaient un tableau tout aussi misérable. Lors des premiers concours de développement, aucun véhicule n'a pu atteindre la ligne d'arrivée, et les voitures sont même souvent entrées en collision les unes avec les autres. Aujourd'hui, en revanche, les véhicules autonomes conduisent de manière à peu près aussi sûre que les humains - ils commettent juste des erreurs de nature différente. Et le jour viendra où chacun d'entre nous préférera monter dans un véhicule autonome plutôt que dans une voiture conduite par un humain, parce que le conducteur pourrait être fatigué ou déconcentré. Si les logiciels et le matériel générés par l'IA atteignent un niveau de qualité tel qu'on leur fait plus confiance qu'à un humain, l'industrie électronique pourrait en effet connaître un bouleversement rapide.

Mais il s'agit encore de prédictions. Nous ne savons pas encore, par exemple, si l'IA peut trouver des idées aussi créatives que celles d'un architecte logiciel ou d'un développeur de système humain. Pour réussir, il faut une certaine dose de folie et de motivation. Quelqu'un s'assoit et travaille jour et nuit pour écrire un nouveau système d'exploitation ou une bibliothèque graphique pour tous les types de microcontrôleurs. Pourquoi l'IA devrait-elle avoir une telle idée ? Quoi qu'il en soit, nous avons hâte de voir ce qui va se passer !

Jens Nickel (Elektor)



Que pensez-vous de l'avenir de l'intelligence artificielle ?
Pensez-vous qu'elle aura un impact positif ou négatif sur nos vies et sur la société en général ?

L'IA est actuellement utilisée le plus souvent comme un meilleur moteur de recherche. L'avantage de cette fonction de recherche est qu'elle fournit certains résultats sous forme de réponses, ce qui peut faciliter le travail sur différents projets et donc accélérer le résultat global. Cependant, l'IA a également tendance à halluciner, ce qui signifie qu'elle produit parfois des résultats inventés qui peuvent se terminer très mal. C'est pourquoi un jugement humain sain est toujours nécessaire pour utiliser cet outil de manière efficace. Le battage médiatique autour de l'utilisation de l'IA est actuellement très important et n'en est qu'à ses débuts. L'avenir nous dira comment l'IA sera développée et dans quelle mesure elle produira des résultats fiables. Je souhaiterais que les développeurs de l'IA ne la laissent pas se développer de manière autonome, mais qu'ils mettent en place des règles pour éviter une croissance incontrôlée.

Soumis en ligne par schube2

S2

Selon vous, quels sont les principaux avantages et inconvénients potentiels de l'intégration de l'IA dans les projets d'ingénierie ?

Selon moi, les avantages et les inconvénients de l'utilisation de l'IA dans les projets d'ingénierie sont nombreux. En outre, nous ne pouvons tout simplement pas les quantifier en termes d'impacts positifs ou négatifs respectifs. Où et à quoi se réfère-t-on pour déterminer cet impact ? Dans l'ensemble, je pense qu'il s'agira plutôt d'un avantage pour nous dans les années à venir. On pourrait comparer cet avantage à l'impact qu'a eu le développement des langages informatiques de haut niveau sur l'évolution des ordinateurs.

Paul Hetrel (Australie)

PH

Que pensez-vous de l'avenir de l'intelligence artificielle ?
Quel sera son impact sur nos vies et sur la société en général ?

J'ai déjà mené plusieurs expériences avec ce que l'on appelle l'IA. Les résultats ont été pour la plupart très décevants. Et ce n'est pas étonnant. Les modèles linguistiques ne « comprennent » rien de la manière dont un humain le ferait. En fait, ils ne font qu'assembler des mots sur la base de statistiques. Ils produisent des réponses très longues afin que leur incapacité ne soit pas remarquée. C'est alors à l'humain de lire quelque chose de significatif dans ces longues réponses. C'est ce qui s'est passé depuis le début avec Dr. Eliza de Joseph Weizenbaum. Il suffit de se poser des questions vérifiables qui ne sont pas déjà disponibles en tant que réponses standard sur Internet - et vous verrez à quel point l'IA « comprend » peu de ce dont elle parle. Exemple : « Pouvez-vous me fournir un tableau des matériaux dont la conductivité thermique est comprise entre 100 et 200 W/(m * K) ? »

Walter Mayer (Allemagne)

WM

Qu'est-ce qui vous enthousiasme le plus dans le potentiel de l'IA, en particulier lorsque vous pensez à l'impact qu'elle pourrait avoir sur le domaine de l'électronique ?

Ce qui m'enthousiasme le plus à propos de l'IA dans l'électronique, c'est son potentiel à devenir un facteur clé en termes d'efficacité, d'automatisation et de durabilité. Avec l'aide de l'IA, nous pouvons optimiser des conceptions de circuits complexes, prédire les pannes et permettre la création de systèmes auto-réparateurs, ce qui signifie des appareils électroniques plus efficaces, fiables et intelligents. En outre, l'automatisation pilotée par l'IA peut accélérer la production et améliorer la précision, tandis que les innovations durables nous aident à réduire les déchets. Toutes ces avancées contribuent au développement d'une électronique intelligente et respectueuse de l'environnement qui s'intègre facilement dans notre vie quotidienne.

Alina Neacsu (Elektor)



Pensez-vous que l'IA aura un impact positif sur l'avenir de l'industrie électronique ?

Je pense que l'avenir de l'IA apportera probablement une combinaison de résultats positifs et de défis sur plusieurs fronts. Du côté positif, l'IA peut offrir une efficacité améliorée, des avancées révolutionnaires et des solutions rapides à des problèmes complexes dans tous les secteurs. Toutefois, elle pourrait également perturber le marché du travail, soulever des questions éthiques et accroître notre dépendance excessive à l'égard de la technologie. La manière dont nous allons surmonter les pièges potentiels et utiliser l'IA de manière responsable sera cruciale pour déterminer si son impact global est positif ou négatif.

C. J. Abate (Elektor)



Avez-vous des idées sur l'avenir de l'IA ? Aura-t-elle un impact positif ou négatif sur nos vies et sur la société en général ?

En regardant l'histoire humaine, je ne peux que conclure que l'IA sera utilisée de deux manières. Elle transformera probablement les soins de santé, améliorera l'efficacité, etc. Et comme toujours, certaines personnes l'utiliseront pour le pire (par exemple, les armes). Très probablement, dans un avenir proche, les machines feront mieux notre travail (j'estime que c'est le cas pour 40 à 50 % des tâches actuelles). Penser que la législation nous protégera est un conte de fées. Le code de la route et d'autres lois sont également là pour nous protéger. Combien de gens s'en soucient vraiment (et pourquoi cela serait-il différent cette fois-ci) ? Lorsque j'étais jeune, la plupart d'entre nous s'arrêtaient au feu orange. Maintenant, beaucoup de gens continuent de rouler même lorsque le feu est passé au rouge.

Soumis en ligne par Erno Gilissen

EG

Pensez-vous que l'IA améliorera ou remplacera les pratiques d'ingénierie traditionnelles ? Comment pensez-vous que cela affectera les rôles et les compétences requises dans ce domaine ?

Personnellement, j'ai utilisé l'IA pour m'aider à écrire du code et à résoudre certains problèmes de logique, mais parfois elle ne parvient pas à analyser correctement la question et a besoin d'être guidée pour m'aider davantage. L'IA n'a pas apporté une solution complète à mes problèmes de programmation, mais elle m'a aidé et m'a permis d'économiser beaucoup de temps en saisie. C'est un outil utile qui, je pense, aidera plutôt qu'il ne remplacera les rôles traditionnels. Pour l'instant, ses réponses doivent être vérifiées. Je continuerai certainement à l'utiliser.

John Allwork (Royaume-Uni)



Pensez-vous que l'IA permettra d'améliorer l'efficacité des projets d'ingénierie ou pourrait-elle introduire de nouvelles complexités ?

L'ingénierie touche directement à la sécurité des individus. L'IA n'a aucune compréhension réelle des sujets et ne peut pas être fiable pour prendre des décisions en ingénierie. Toute amélioration apparente de la productivité grâce à l'utilisation de l'IA serait compensée par la nécessité pour les humains de la superviser afin d'en assumer la responsabilité. Il y a un danger à ce que les gens attribuent des compétences à l'IA.

Andrew Pratt (Royaume-Uni)

AP



Quelles sont les conséquences potentielles d'une trop grande dépendance à l'égard de l'IA dans la conception et l'analyse de l'ingénierie ?

PM

L'IA englobe de nombreuses techniques différentes, dont certaines (comme ChatGPT) n'ont aucune notion de vérité. Personnellement, je n'ai utilisé que ChatGPT, et uniquement pour des questions mathématiques. J'ai été très souvent déçu, car les réponses contenaient presque toujours des erreurs. Je crains qu'une utilisation non raisonnée ne conduise à la mise en place de solutions erronées. L'IA peut-elle imaginer des avancées technologiques ? Y a-t-il un risque de déclin des compétences ? PS : Cette réponse a été rédigée sans l'aide de ChatGPT !

Pierre Molinaro (France)

Que pensez-vous de l'avenir de l'intelligence artificielle ?

Aura-t-elle un impact positif ou négatif sur nos vies et sur la société en général ?

DF

L'état actuel de la technologie n'est rien d'autre qu'une bulle de recherche prête à éclater. Les capacités de ces moteurs sont certainement plus avancées que celles des moteurs précédents, mais il ne faut pas s'y tromper, il ne s'agit que d'algorithmes complexes dotés d'une puissance de calcul parallèle à celle des serveurs. En soi, il s'agit d'un exploit impressionnant, cependant, l'utilisation de cette technologie a principalement consisté à utiliser illégalement du matériel protégé par des droits d'auteur pour constituer une base de données qui donne une fausse impression d'intelligence. Il ne s'agit pas d'une véritable intelligence, car il n'y a pas de conscience réelle et la capacité cognitive de comprendre des concepts qui contiennent autre chose que des idées simples et des préceptes philosophiques est bien au-delà de sa base de données.

Soumis en ligne par David Fahrenholz

Que pensez-vous de l'avenir de l'IA et de ses avantages ou inconvénients potentiels ?

L'avenir de l'IA est une arme à double tranchant. D'une part, elle a le potentiel de révolutionner les soins de santé, d'augmenter la productivité et de résoudre des problèmes complexes tels que le changement climatique. D'autre part, elle pourrait entraîner des suppressions d'emplois, accentuer les inégalités et soulever de graves problèmes éthiques, en particulier dans des domaines tels que la surveillance et les préjugés. La clé sera de trouver un équilibre entre innovation et responsabilité, en veillant à ce que l'IA soit développée et mise en œuvre d'une manière qui profite à la société dans son ensemble, sans perdre de vue les risques potentiels.

Brian Tristram Williams (Elektor)

En savoir plus sur l'AI

Visitez la page «Embarquée et IA» d'Elektor pour découvrir des projets, des vidéos et des tutoriels !

www.elektormagazine.fr/embarque-ia



240467-04

Examinez les réponses des membres de la communauté Elektor.

Pouvez-vous tirer des conclusions sur ce que la communauté pense de l'IA ?



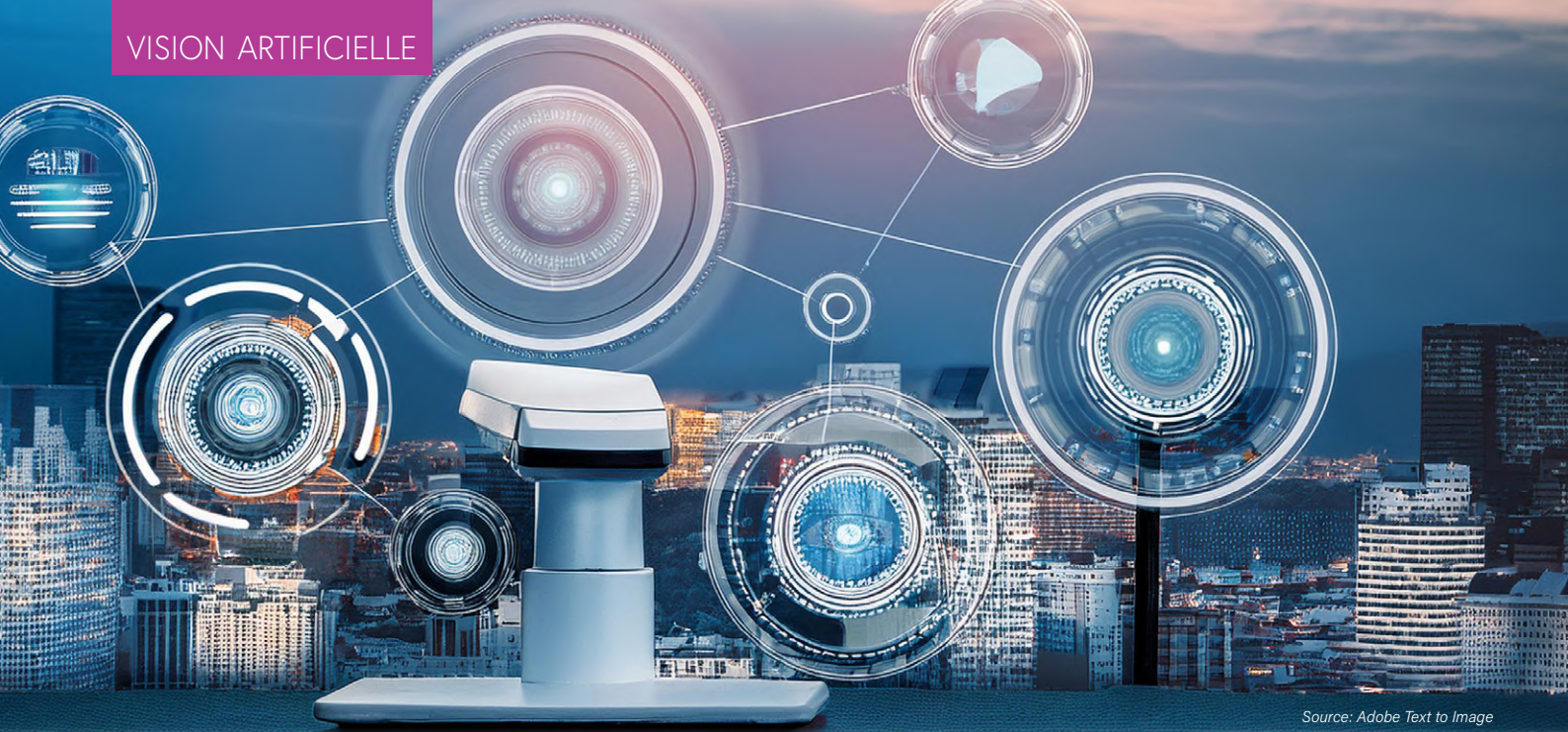
Drafts



Les membres de la communauté Elektor expriment un mélange d'optimisme et d'inquiétude quant à l'avenir de l'IA. Si beaucoup reconnaissent ses avantages potentiels, tels que l'augmentation de l'efficacité et des capacités de résolution de problèmes, ils sont également très préoccupés par ses inconvénients potentiels... En conclusion, la communauté Elektor considère l'IA comme un outil puissant présentant à la fois des avantages et des risques potentiels. Si l'optimisme règne quant à sa capacité à améliorer les pratiques d'ingénierie, l'accent est également mis sur la nécessité d'un développement responsable et d'une surveillance humaine afin de garantir son impact positif sur la société.

Enter a prompt for ChatGPT





Source: Adobe Text to Image

vision artificielle avec OpenMV

Créer un détecteur de canettes de soda

Koen Vervloesem (Belgique)

Pour développer des applications de vision artificielle, OpenMV est une plateforme intéressante, aspirant à devenir « l'Arduino de la vision artificielle ». Elle combine un micrologiciel basé sur MicroPython avec des bibliothèques logicielles de vision artificielle simple à utiliser, un environnement de développement intégré (EDI) personnalisé et un matériel de caméra spécialisé. Dans ce tutoriel, nous utilisons OpenMV et une caméra Arduino Pro Nicla Vision pour détecter des objets tels que des canettes de soda.

Figure 1. Avec des dimensions de 22,86 mm × 22,86 mm × 7,26 mm, l'Arduino Nicla Vision est la plus petite caméra OpenMV disponible sur le marché.



Dans cet article, je démontrerai l'utilisation de la carte caméra Arduino Pro Nicla Vision [1] (**figure 1**) avec le micrologiciel OpenMV [2] pour la détection d'objets. Vous pouvez cependant exécuter ce projet sur n'importe quel matériel de caméra compatible avec OpenMV. La détection d'objets est un type particulier de classification d'images. Contrairement à un modèle de classification d'images classique qui se contente

d'identifier la catégorie de l'objet présent dans l'image, un modèle de détection d'objets est également capable de localiser précisément l'objet détecté dans l'espace de l'image. Cela fonctionne même pour plusieurs objets dans une image. Ce processus est plus intensif en termes de calcul que la simple classification d'images, mais la carte Nicla Vision peut le gérer.

J'utiliserai l'approche FOMO (Faster Object, More Objects) pour exécuter un modèle de détection d'objets. FOMO est un algorithme performant qui permet la détection d'objets en temps réel avec des

microcontrôleurs. Le projet commence par la constitution d'un ensemble d'images des objets que vous souhaitez détecter. Vous pouvez le faire en prenant des images avec la caméra de Nicla Vision. Une fois cet ensemble de données créé, vous pouvez entraîner le modèle FOMO avec la plateforme Edge Impulse [3]. Le modèle entraîné peut ensuite être flashé en tant que micrologiciel OpenMV personnalisé sur le Nicla Vision pour la détection d'objets. Comme OpenMV est basé sur le langage MicroPython [4], vous pouvez facilement personnaliser les actions de la carte lorsque des objets sont détectés.

Configuration de l'EDI OpenMV

Commencez par installer l'EDI OpenMV [5], qui fonctionne sous Windows, macOS et Linux (Ubuntu). Sur Ubuntu 23.04 et plus, vous devez modifier une ligne dans le script `setup.sh` de l'installateur `tar.gz` car les paquets installés en dehors d'un environnement virtuel Python ne sont plus supportés. Modifiez cette ligne :

```
sudo pip install pyusb
```

en

```
sudo apt install python3-usb
```

Exécutez ensuite le script `setup.sh` situé dans le répertoire décompressé. Après l'installation de l'EDI OpenMV, connectez votre carte Nicla Vision à votre ordinateur avec un câble micro-USB. Si la LED de la carte se met à clignoter en bleu, cela signifie que la carte exécute le script par défaut `main.py` d'OpenMV. Le stockage interne de la carte apparaît sur votre ordinateur comme un périphérique de stockage externe. Il contient un fichier `main.py` avec le code MicroPython du **listage 1**. Ce code fait clignoter la LED tant qu'il n'y a pas de connexion USB.

Connexion à votre carte OpenMV

Lancez l'EDI OpenMV et cliquez sur l'icône **Connect** en bas à gauche. La LED devrait s'arrêter de clignoter, et vous pourriez recevoir un message indiquant que le micrologiciel de la carte n'est pas à jour. Confirmez que vous souhaitez mettre à jour le micrologiciel et effacer



Listage 1. Clignotement de la LED intégrée pour vérifier si le matériel OpenMV fonctionne.

```
# main.py -- put your code here!
import pyb, time
led = pyb.LED(3)
usb = pyb.USB_VCP()
while (usb.isconnected()==False):
    led.on()
    time.sleep_ms(150)
    led.off()
    time.sleep_ms(100)
    led.on()
    time.sleep_ms(150)
    led.off()
    time.sleep_ms(600)
```

le système de fichiers interne. Ce processus peut prendre du temps et le voyant de la carte clignotera en vert pendant un certain temps. Une fois que le firmware a été mis à jour et que l'EDI OpenMV a établi une connexion avec votre carte, la LED cesse de clignoter. La barre d'état en bas devrait afficher le nom de votre carte, le capteur, la version du micrologiciel, le port série et le lecteur sur lequel le stockage interne est connecté (**figure 2**).

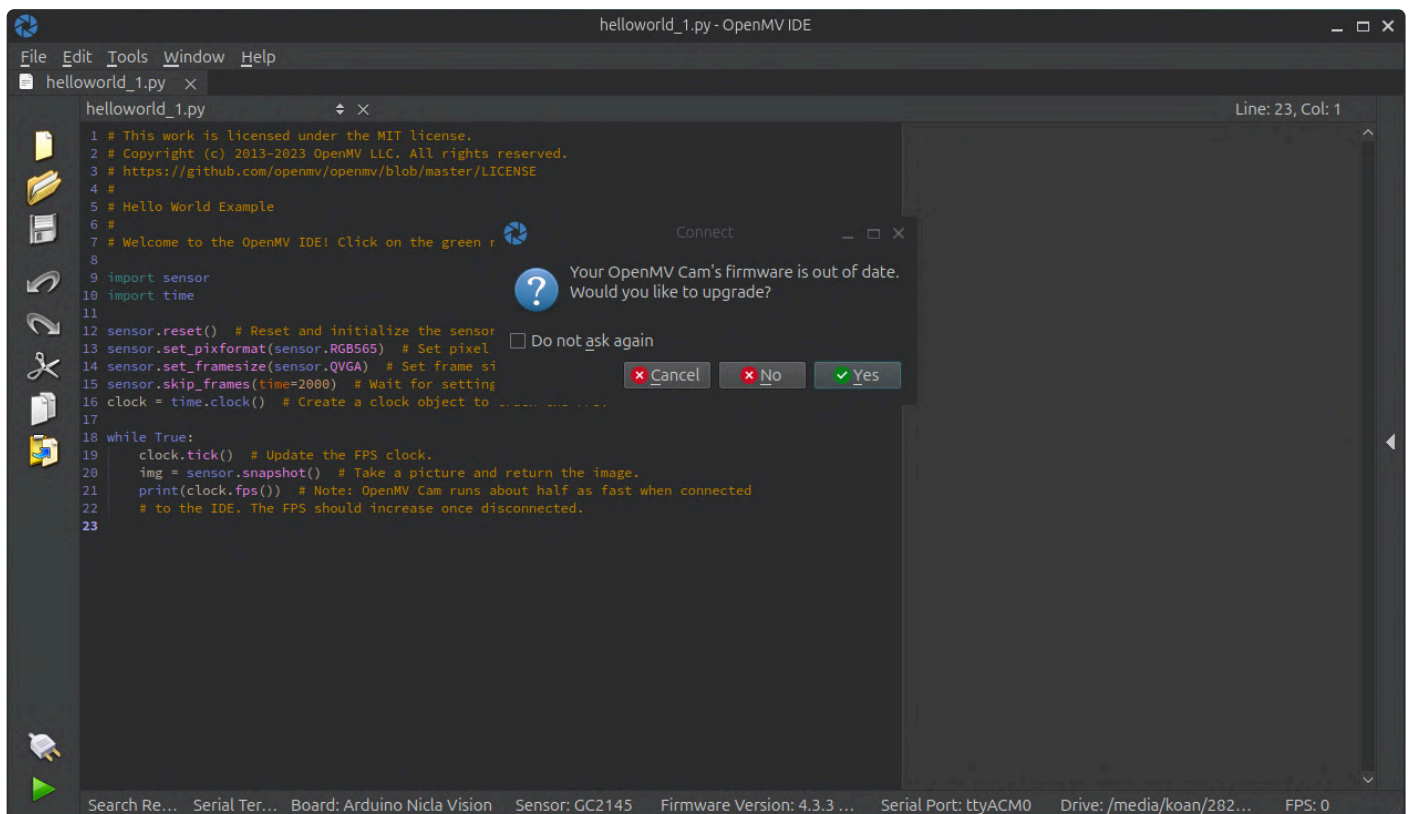


Figure 2. L'EDI OpenMV requiert la mise à jour du micrologiciel de la carte de la caméra.



Listage 2. Obtenir une vue en direct de la caméra de la carte OpenMV.

```
# This work is licensed under the MIT license.
# Copyright (c) 2013-2023 OpenMV LLC. All rights reserved.
# https://github.com/openmv/openmv/blob/master/LICENSE
#
# Hello World Example
#
# Welcome to the OpenMV IDE! Click on the green run arrow button below to run the script!

import sensor
import time

sensor.reset() # Reset and initialize the sensor.
sensor.set_pixformat(sensor.RGB565) # Set pixel format to RGB565 (or GRAYSCALE)
sensor.set_framesize(sensor.QVGA) # Set frame size to QVGA (320x240)
sensor.skip_frames(time=2000) # Wait for settings take effect.
clock = time.clock() # Create a clock object to track the FPS.

while True:
    clock.tick() # Update the FPS clock.
    img = sensor.snapshot() # Take a picture and return the image.
    print(clock.fps()) # Note: OpenMV Cam runs about half as fast when connected
    # to the IDE. The FPS should increase once disconnected.
```



Listage 3. Le script de capture de données enregistre des images RGB de 240 x 240 pixels.

```
import sensor, image, time

sensor.reset()
sensor.set_pixformat(sensor.RGB565)
sensor.set_framesize(sensor.QVGA)
sensor.set_windowing((240, 240))
sensor.skip_frames(time = 2000)

clock = time.clock()

while(True):
    clock.tick()
    img = sensor.snapshot()
    print(clock.fps())
```

Pour tester le fonctionnement de la caméra, exécutez le script par défaut *helloworld_1.py*, en cliquant sur l'icône verte *run* en bas à gauche. Cela transformera la caméra en webcam, et affichera son flux dans l'EDI OpenMV. La partie *Frame Buffer* de la fenêtre montre l'image de la caméra en direct, tandis que la partie *Histogram* montre les

composantes rouge, verte et bleue de l'image. Vous pouvez modifier cet espace colorimétrique en niveaux de gris, LAB ou YUV. En bas à droite, vous pouvez également voir le nombre d'images par seconde que la caméra diffuse en continu.

Examinons ce script *helloworld_1.py* par défaut (**listage 2**). Après avoir importé les modules Python nécessaires, le script réinitialise la caméra, puis définit le format des pixels et la taille des images. Après un délai de deux secondes, il entre dans une boucle pour capturer continuellement des images et obtenir le nombre actuel d'images par seconde.

Création d'un ensemble de données

Une fois que votre caméra est configurée, il est temps de construire un ensemble de données en capturant des images de divers objets. L'EDI OpenMV offre un support intégré pour cette tâche. Ouvrez le menu *Tools / Dataset Editor / New Dataset* et choisissez un répertoire pour enregistrer votre ensemble de données.

L'EDI affiche alors un éditeur de jeu de données sur la gauche, le script de capture de jeu de données (qui est identique au script *hello world* précédent) au milieu, et le *frame buffer* et l'histogramme sur la droite. Cliquez sur l'icône *New Class Folder* dans la barre latérale gauche, ou appuyez sur *Ctrl+Shift+N* pour créer une nouvelle classe pour tout objet que vous souhaitez détecter.

Pour cet exemple, j'entraînerai le modèle à détecter les canettes de boissons gazeuses. Je vais donc créer une classe nommée *can*. Comme le modèle FOMO nécessite des images carrées et que la caméra offre une résolution de 320x240, il est indispensable d'ajuster le script de capture de données en conséquence (**listage 3**).

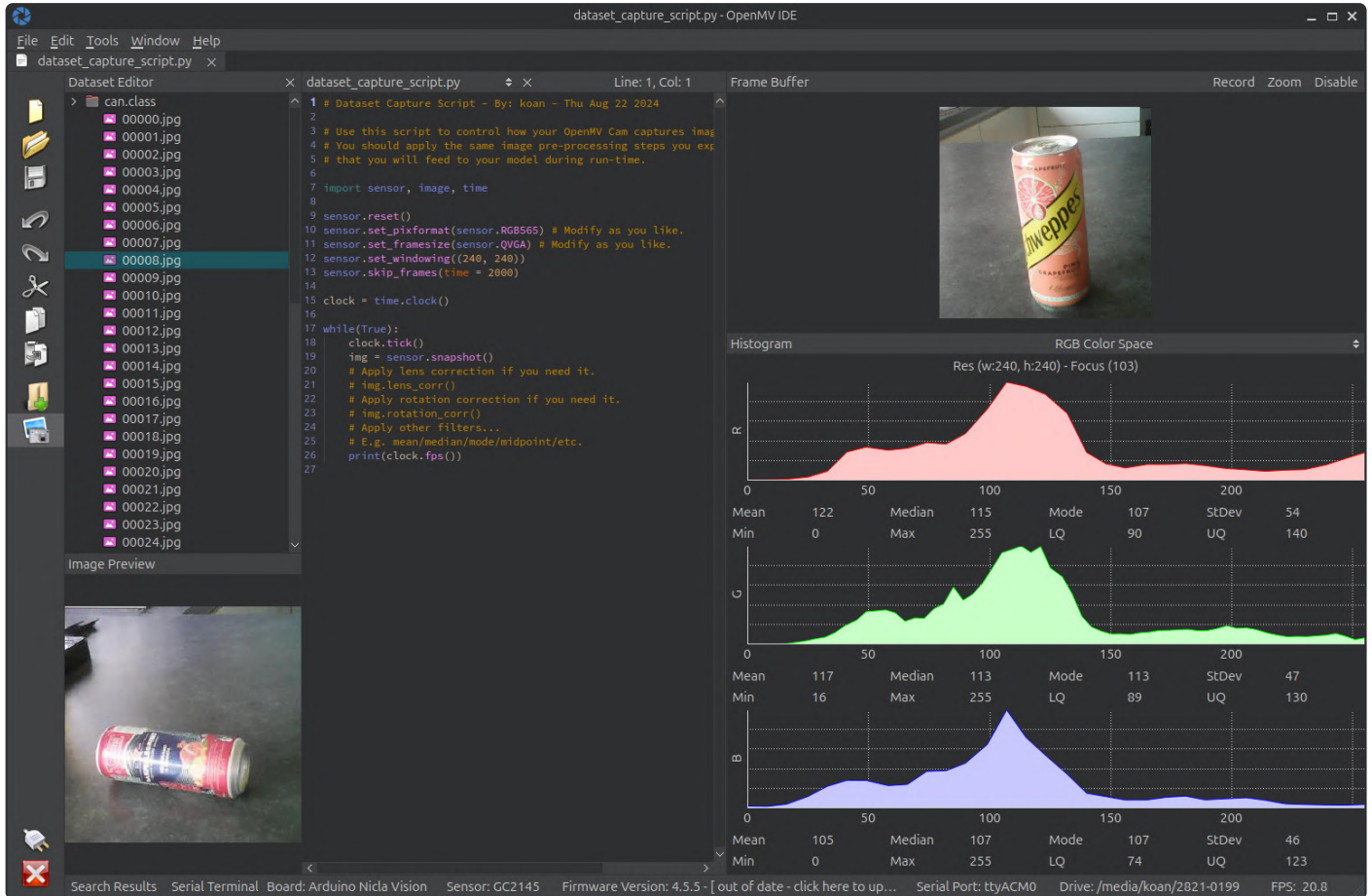


Figure 3. Créez votre nouvel ensemble de données composé d'images de différentes canettes de soda.

Exécutez ce script en appuyant sur l'icône de lecture, sélectionnez l'une des classes en cliquant sur son nom de dossier dans l'éditeur de jeu de données, et pointez la caméra sur un objet de cette classe. Capturez une image en cliquant sur l'icône de l'appareil photo (*Capture Data*) à gauche. Prenez 30 à 50 images de diverses canettes de soda sous différents angles, distances, conditions d'éclairage et arrière-plans (figure 3).

Étiquetez votre jeu de données dans Edge Impulse

L'EDI OpenMV est directement intégré à Edge Impulse, facilitant ainsi le chargement de votre jeu de données vers ce service cloud pour l'entraînement d'un modèle. Pour débuter, inscrivez-vous sur Edge Impulse ; l'inscription est gratuite pour les étudiants, les universités et les développeurs individuels dans le cadre du plan communautaire [6]. Dans votre tableau de bord Edge Impulse, cliquez sur *Create new project*, donnez un nom au projet, choisissez de le rendre public ou privé (figure 4).

Dans l'EDI OpenMV, naviguez vers *Tools / Dataset Editor / Export / Login to Edge Impulse Account* et *Upload to Project*. Connectez-vous avec vos identifiants Edge Impulse et choisissez le projet. Dans l'étape suivante, vous devez choisir la répartition entre les données d'entraînement et de test. Vous pouvez laisser la répartition par défaut 80%/20%.

Une fois les images téléchargées, l'ensemble de données apparaît dans

Create a new project

Enter the name for your new project:

cans

Choose your project type:

Personal

20 min job limit, 4GB or 4 hours of data, limited collaboration.

Enterprise

No job or data size limits, higher performance, custom blocks.

Choose your project setting:

Public

Anyone on the Internet can view and clone this project under the licence: [Apache 2.0](#). Only invited users will be able to edit.


Private (1 of 2 remaining)

Only invited users can edit and view your project.


Want full-feature access and unlimited projects? Try [Enterprise free](#).

Create new project

Figure 4. Créer un nouveau projet dans Edge Impulse.



Configure your target device and application budget



Target device

Define your target device requirements to inform model optimizations and performance calculations. No device yet? Use the default settings which you can change at any time.

Target device

Arduino Nicla Vision (Cortex-M7 480MHz)

Processor family

Cortex-M

Clock rate ?

480 | MHz

Max

Custom device name (optional) ?

Application budget

Specify the available RAM and ROM for the model's operation, along with the maximum allowed latency for your specific application. Not sure yet? Start with the defaults and modify them later on.

RAM

1 | MB

Max

ROM

2 | MB

Max

Latency ?

100 | ms

Max

Reset to default settings

Cancel

Save

Figure 6. Spécifiez l'appareil cible sur lequel exécuter le modèle de vision industrielle.

une image contient plusieurs objets, dessinez une boîte autour de chacun d'entre eux et attribuez-lui une étiquette. Une fois que vous avez étiqueté tous les échantillons, revérifiez votre travail pour vous assurer de l'exactitude de l'étiquetage. L'étiquetage d'un grand nombre d'images est une tâche ennuyeuse et sujette à erreur, ce qui peut influencer significativement l'apprentissage du modèle.

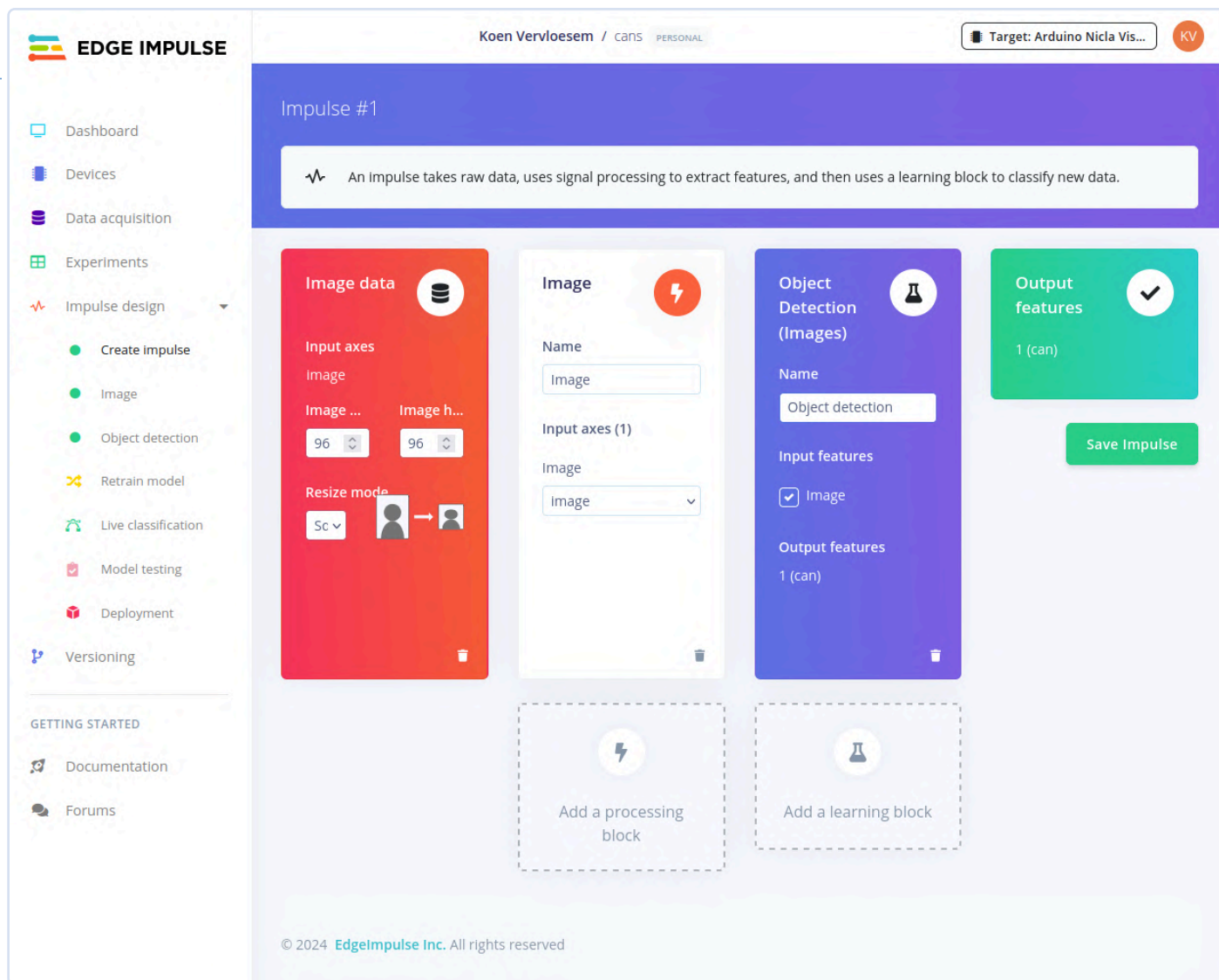


Figure 7. Créez une impulsion pour classer vos données.

S'il y a un problème avec le taux entraînement/test, un signe d'avertissement apparaîtra. Cliquez dessus pour obtenir plus de détails (**figure 5**). Par exemple, l'une des classes pourrait avoir un nombre insuffisant d'échantillons. Pour résoudre ce problème, ajoutez plus d'images dans l'EDI OpenMV et chargez à nouveau les données. La procédure de chargement évite les doublons et n'ajoute pas une même image deux fois dans votre projet de test. Si la répartition entre les données d'apprentissage et de test n'est pas équilibrée, cliquez sur les trois points à droite d'un échantillon dans les données d'apprentissage, puis sur **Move to test** jusqu'à ce que chaque étiquette de votre ensemble de données présente un rapport entraînement/test proche de 80 %/20 %.

Création du modèle

Avant de créer votre modèle, configurez l'appareil cible, qui sera utilisé pour optimiser le modèle et évaluer ses performances. Cliquez sur l'icône cible en haut à droite à côté des initiales de votre profil. Choisissez ensuite votre appareil cible (**figure 6**). L'Arduino Nicla Vision possède deux cœurs : un Cortex-M7 480 MHz et un Cortex-M4 240 MHz, sélectionnez l'un des deux. Sous **Application budget**, définissez la RAM et la ROM disponibles et la latence maximale autorisée. Vous pouvez commencer par les valeurs par défaut.

Ensuite, allez dans **Impulse design** et cliquez sur **Create impulse**. Dans le premier bloc, **Image data**, définissez la largeur et la hauteur de

l'image à 96 pixels. Comme les images originales sont déjà carrées, le mode de redimensionnement n'est pas crucial ici. Cliquez ensuite sur **Add a processing block** et cliquez sur **Add** à côté du bloc **Image**. Ensuite, cliquez sur **Add a learning block** et cliquez sur **Add** à côté de **Object Detection (Images)**. Vous verrez que le bloc affiche automatiquement la classe **can** dans votre ensemble de données comme **output features**. Il est maintenant également affiché dans le bloc **Output features** à la fin. Cliquez sur **Save Impulse** (**figure 7**).

Une nouvelle partie de **Impulse design** apparaît maintenant : **Image**. Cliquez dessus pour afficher les images brutes (**figure 8**, vous pouvez choisir des images spécifiques dans le menu déroulant en haut à droite). Sous **Parameters**, réglez la profondeur de couleur sur **RGB** et cliquez sur **Save parameters**. Vous êtes maintenant dans l'onglet **Generate features**.

Cliquez sur **Generate features** pour appliquer le bloc de traitement à toutes les images de l'ensemble de données. Au final, cela crée une visualisation 3D de toutes les données d'apprentissage, regroupées en fonction de leur similarité. Cette opération peut prendre un certain temps, en fonction de la taille de vos données d'apprentissage. Si vous voyez tous les points de données de la même classe clairement regroupés, cela signifie que le modèle a un moyen facile d'apprendre à distinguer les classes. Si ce n'est pas le cas, vous devriez essayer de créer plus d'images, ou des images plus claires.

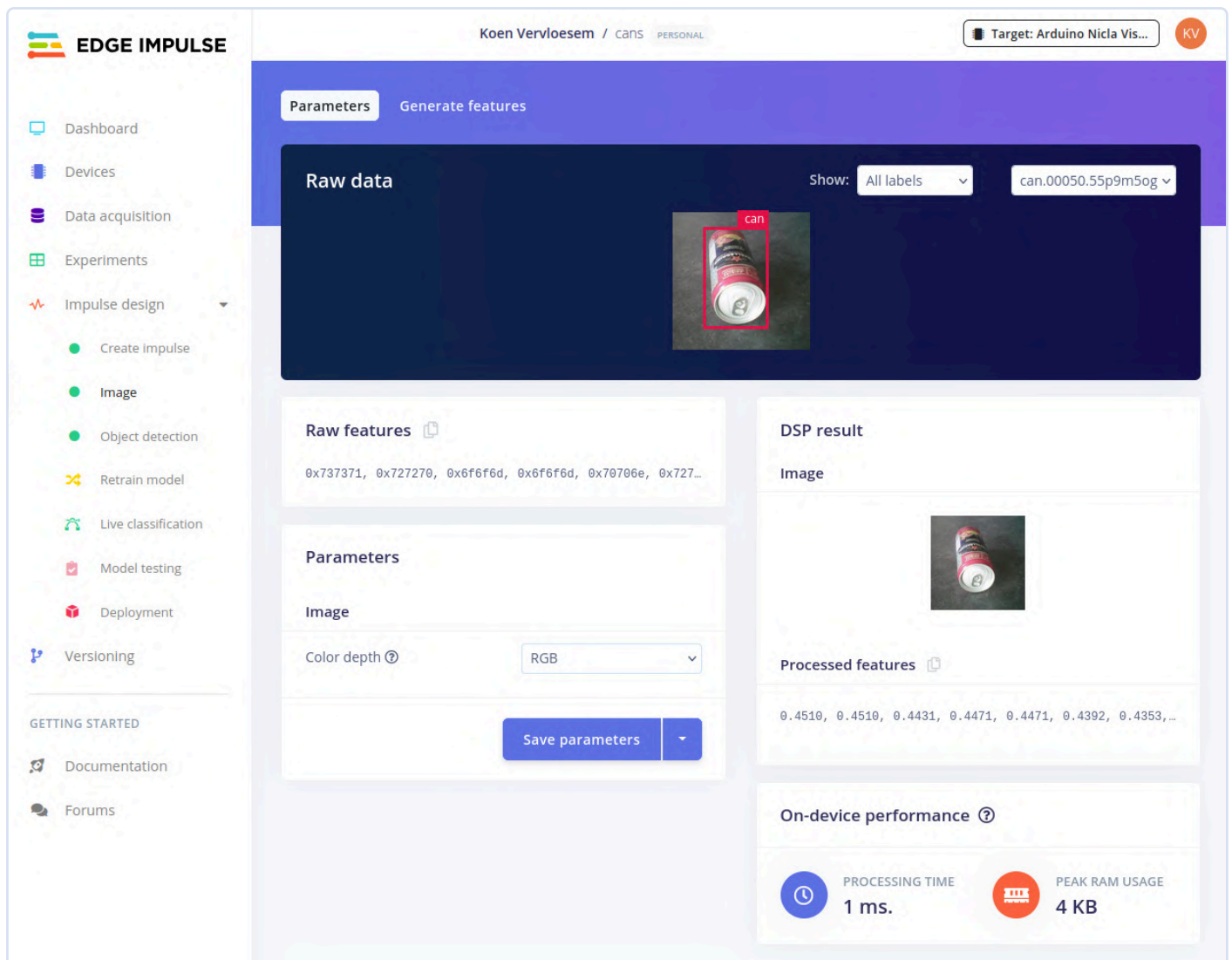


Figure 8. Edge Impulse vous montre les données brutes dont il dispose pour chaque image.

L'étape suivante consiste à configurer le modèle de détection d'objets. Cliquez sur *Object detection*. Le modèle affiché est *FOMO (Faster Objects, More Objects) MobileNetV2 0.35*. Vous pouvez toujours choisir un autre modèle, mais le modèle et les paramètres par défaut sont parfaits. Cliquez sur *Save & train*. Cela divisera les données d'entraînement en un ensemble d'entraînement et de validation, et l'ensemble du processus prendra un certain temps.

Une fois le modèle entraîné, des statistiques sur la précision apparaîtront sous la sortie de formation. Ces chiffres indiquent la qualité des performances du modèle. S'ils ne sont pas satisfaisants, mais montrent une amélioration au cours de l'entraînement, envisagez de réentraîner le modèle avec plus de cycles.

Les chiffres figurant sous *On-device performance* sont également intéressants : ils indiquent le temps d'inférence du modèle, le pic d'utilisation de la RAM et l'utilisation de la mémoire flash. Si le temps d'inférence semble trop élevé, vous devez réduire la résolution de l'image et recréer le modèle. L'utilisation maximale de la mémoire vive et l'utilisation de la mémoire flash sont également des facteurs impor-

tants. Vous devrez les comparer aux spécifications de votre carte. Par exemple, la Nicla Vision a 1 MB RAM. Ma première tentative de création d'un modèle avec 240 × 240 images en niveaux de gris avait 1,3 MB de RAM en pic d'utilisation, ce qui aurait été impossible à exécuter sur la carte, ayant 1 MB de RAM.

Dans mon cas, j'ai atteint une performance d'entraînement où le modèle a été capable de détecter 75 % des canettes correctement, avec seulement 50 images, ce qui est un bon début. Le temps d'inférence est de 52 ms et le pic d'utilisation de la RAM de 239,5 K, ce qui rend le modèle adapté pour fonctionner sur le Nicla Vision avec une détection presque immédiate.

Cependant, Edge Impulse optimise par défaut le modèle avec une quantification int8, qui est plus rapide et utilise moins de RAM. Si vous sélectionnez la version *Unoptimized (float32)* du modèle, les performances de l'entraînement peuvent être bien meilleures. Dans mon cas, les performances de détection ont atteint 90.9 %. Cependant, le temps d'inférence a augmenté à 125 ms, et le pic d'utilisation de la RAM était de 887.1 K (**figure 9**).

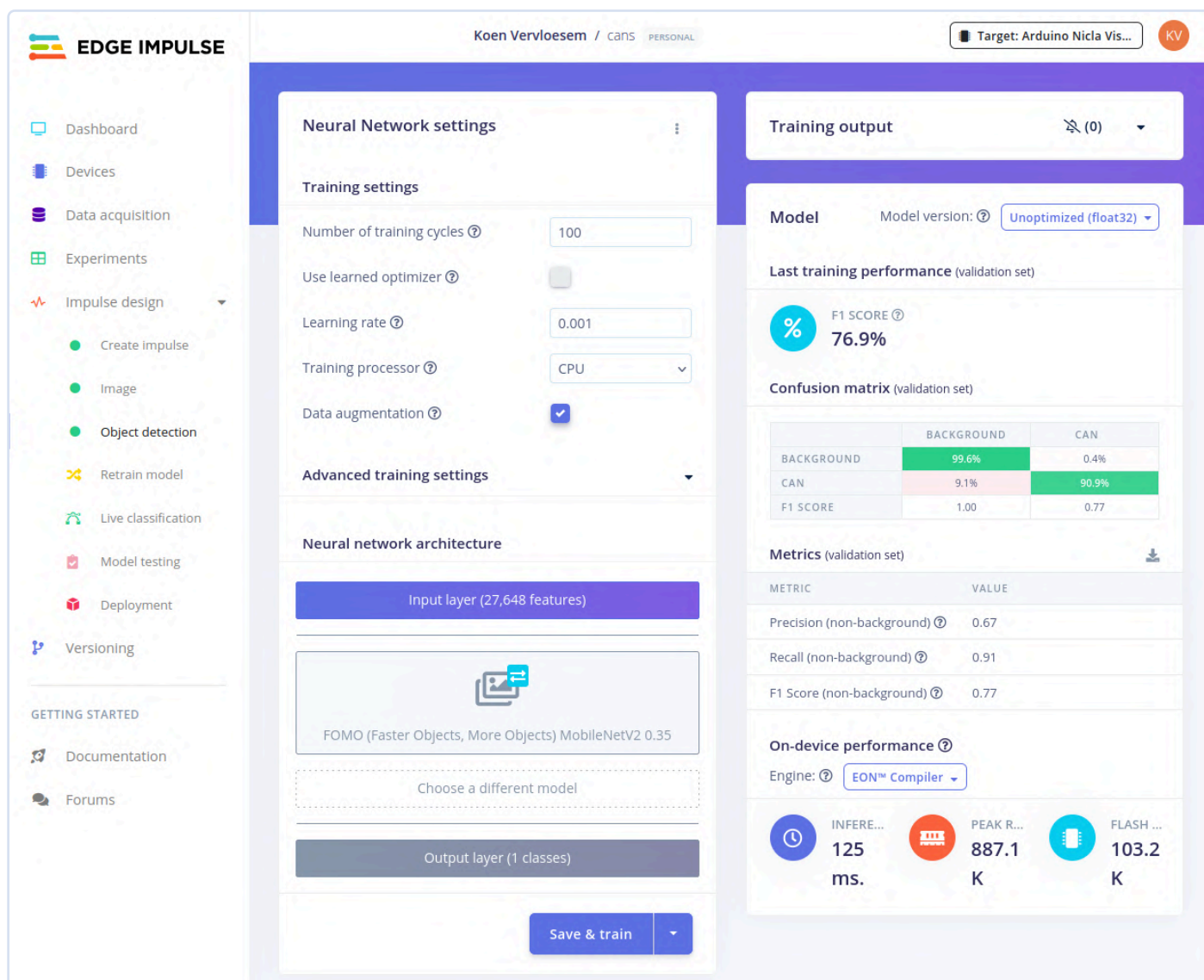


Figure 9. Le réseau neuronal permettant de détecter les canettes de soda a été entraîné.

Test du modèle

Maintenant que vous avez entraîné le modèle, il est temps de le tester. Edge Impulse a entraîné le modèle uniquement sur les données d'apprentissage, vous pouvez donc utiliser les images de l'ensemble de données de test pour tester le modèle. Rendez-vous dans la section **Model testing** et cliquez sur **Classify all**. Cela exécutera le modèle sur toutes les images des données de test et comparera la sortie du modèle à l'étiquette que vous avez attribuée à l'image.

Si la précision de cette validation est faible, cela signifie que votre modèle est trop adapté aux données de l'ensemble de données d'apprentissage. Une solution consiste à réentraîner le modèle avec un taux d'apprentissage plus faible. Vous pouvez également élargir votre ensemble de données avec des exemples plus variés de l'objet que vous souhaitez détecter, les étiqueter et recréer le modèle. Il est nécessaire de procéder à quelques tests pour affiner votre modèle.

Déploiement sur votre appareil

Il est maintenant temps de déployer le modèle sur votre appareil OpenMV. Edge Impulse dispose d'un support direct pour créer un micrologiciel OpenMV que vous pouvez flasher sur votre appareil. Rendez-vous sur **Deployment**, sélectionnez **OpenMV firmware** dans le champ de recherche, choisissez la version du modèle, puis cliquez sur **Build**. Une fois la compilation terminée, votre navigateur télécharge le fichier zip. Si le téléchargement ne se lance pas, cliquez sur **Latest build** en haut à droite pour télécharger le fichier.

Le fichier zip contient des fichiers **.bin** adaptés à tous les appareils OpenMV pris en charge. Pour le Nicla Vision, vous avez besoin du fichier **edge_impulse_firmware_arduino_nicla_vision.bin**. Extrayez ce fichier. Ensuite, dans l'EDI OpenMV, naviguez vers **Tools / Run Bootloader (Load Firmware)** et sélectionnez le fichier bin que vous avez extrait. Sélectionnez **Erase internal file system** et cliquez sur **Run**. Le micrologiciel sera alors flashé sur votre appareil.

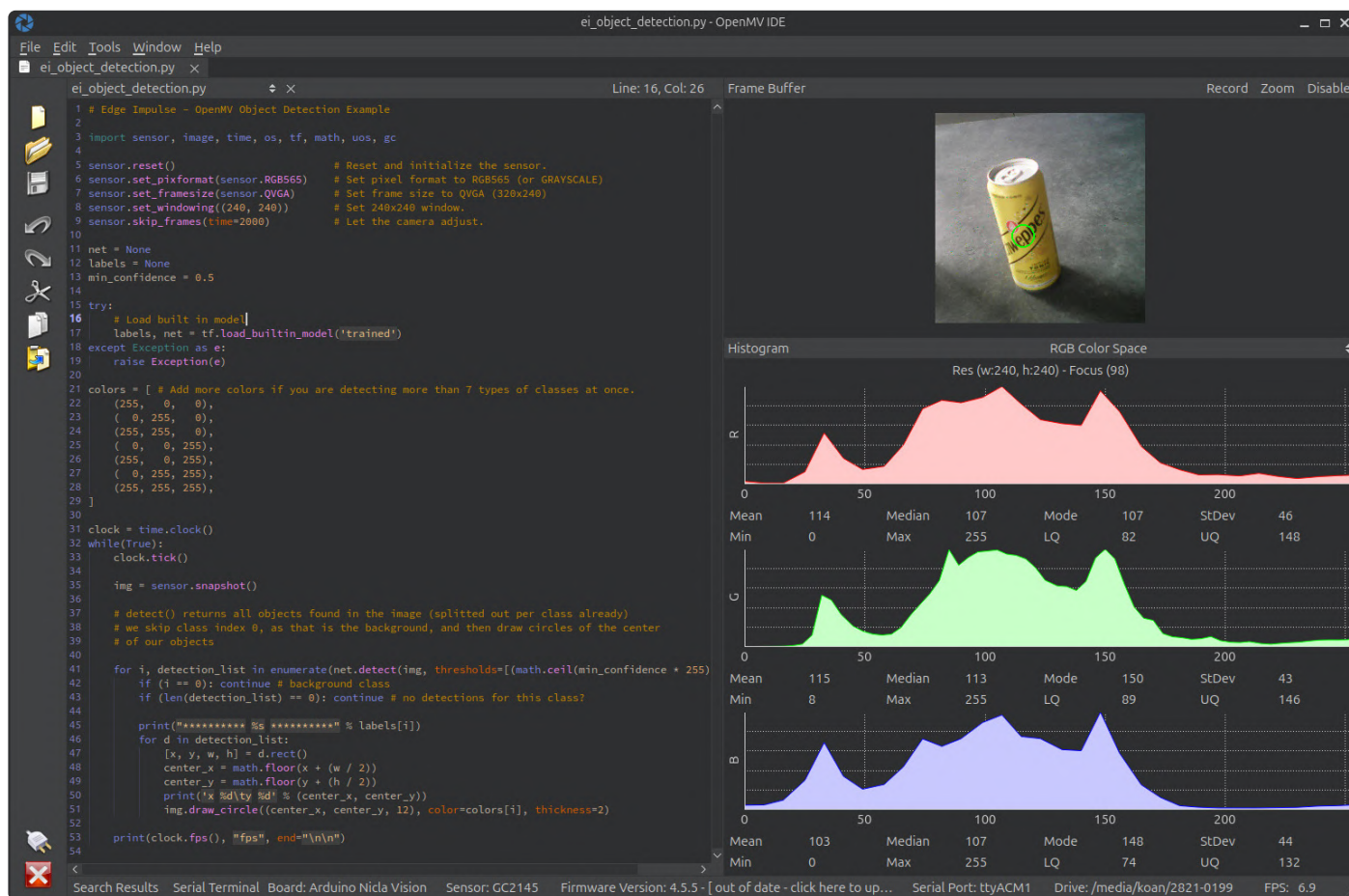


Figure 10. Le script OpenMV peut maintenant détecter les canettes de soda dans les images de la caméra.

Extrayez ensuite le fichier `ei_object_detection.py` du fichier zip et ouvrez-le dans l'EDI OpenMV. Cliquez sur l'icône de lecture pour l'exécuter sur votre appareil OpenMV. Ce projet exécute le modèle TensorFlow Lite sur l'image de la caméra en direct et dessine un cercle au centre des objets détectés (figure 10).

Facilement personnalisable

L'entraînement d'un modèle de détection d'objets prend du temps, et la qualité des données est un facteur important à prendre en compte. Dans cet article, je vous ai présenté un exemple simple où le modèle est configuré pour reconnaître une seule classe d'objets. Vous pouvez ajouter d'autres classes, par exemple des boîtes de conserve, des bouteilles, etc. L'EDI OpenMV s'intègre parfaitement au service en ligne Edge Impulse, ce qui simplifie du processus. De plus, le micro-logiciel généré exécute du code MicroPython, ce qui vous permet de personnaliser facilement le comportement de l'appareil lorsqu'il détecte un objet. Par exemple, ajoutez du code MicroPython pour changer la couleur de la LED en fonction de la classe de l'objet détecté. ◀

240450-04

LIENS

- [1] Arduino Pro Nicla Vision : <https://docs.arduino.cc/hardware/nicla-vision/>
- [2] OpenMV : <https://openmv.io>
- [3] Edge Impulse : <https://edgeimpulse.com>
- [4] MicroPython: <https://micropython.org>
- [5] EDI OpenMV : <https://openmv.io/pages/download>
- [6] Edge Impulse plans : <https://edgeimpulse.com/pricing>

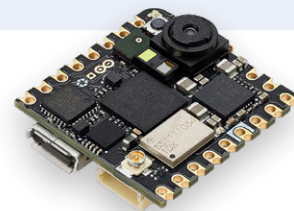
Questions ou commentaires ?

Envoyez un courriel à l'auteur (koen@vervloesem.eu), ou contactez Elektor (redaction@elektor.fr).



À propos de l'auteur

KoKoen Vervloesem est auteur d'articles sur Linux et l'open source, la sécurité informatique, la vie privée, la programmation, l'IA et l'Internet des objets depuis plus de 20 ans. Il est titulaire d'un master en ingénierie informatique et en philosophie et donne des cours sur Linux, Python et l'IdO.



Product

➤ **Arduino Pro Nicla Vision**
www.elektor.fr/20152



ChatGPT vs Gemini conversation avec l'esprit numérique

(Source : ChatGPT
et Studio Elektor)

Questions compilées par l'équipe éditoriale d'Elektor

ElektorMag a interrogé ChatGPT d'OpenAI et Gemini de Google pour voir comment ces modèles d'IA se comportent sur une série de plusieurs sujets. Des systèmes embarqués à la conception de circuits, nous avons posé plusieurs questions auxquelles de nombreux ingénieurs, fabricants et étudiants souhaitent obtenir des réponses. Découvrez comment ils traitent les questions techniques complexes et lequel pourrait être l'assistant numérique idéal pour votre prochain projet électronique !

Quel est votre nom et votre version officiels ?

🌀 **ChatGPT** : Mon nom officiel est ChatGPT, et je suis basé sur l'architecture GPT-4 développée par OpenAI.

♦ **Gemini** : Mon nom officiel est Gemini, et je fonctionne actuellement sur le modèle Gemini 1.5 Flash.

Qu'est-ce que l'IA ?

🌀 **ChatGPT** : L'intelligence artificielle (IA) est le développement de systèmes informatiques capables d'effectuer des tâches nécessitant une intelligence humaine, telles que l'apprentissage, le raisonnement et la résolution de problèmes. Elle comprend des technologies telles que l'apprentissage automatique et le traitement du langage naturel, permettant des applications telles que les assistants virtuels et les véhicules autonomes.

♦ **Gemini** : L'IA, ou intelligence artificielle, est le développement de systèmes informatiques capables d'effectuer des tâches qui requièrent généralement l'intelligence humaine. Ces tâches comprennent l'apprentissage, le raisonnement, la perception et le traitement du langage naturel.

Faites la différence entre l'IA et le ML. Veuillez expliquer.

🌀 **ChatGPT** : L'IA (Intelligence Artificielle) est le concept plus large de machines simulant l'intelligence humaine pour effectuer des tâches. Le ML (Machine Learning) est un sous-ensemble de l'IA qui consiste à entraîner des algorithmes sur des

données pour apprendre des modèles et faire des prédictions sans programmation explicite pour chaque tâche.

- ◆ **Gemini** : L'IA est le concept plus large de création d'agents intelligents, tandis que le ML est un sous-ensemble de l'IA axé sur l'apprentissage des ordinateurs à partir de données. L'IA peut impliquer des tâches telles que le traitement du langage naturel et la résolution de problèmes, tandis que le ML est principalement utilisé pour la reconnaissance de modèles et la prédiction.

Êtes-vous une « IA » ? Êtes-vous conscient ?

- 🌀 **ChatGPT** : Je suis un modèle linguistique d'IA, ce qui signifie que je traite et génère du texte sur la base de modèles présents dans les données sur lesquelles j'ai été entraîné. Je ne suis pas conscient ou conscientisé. Mes réponses sont générées par des algorithmes, et non par des expériences ou des pensées personnelles. Bien que je puisse simuler une conversation humaine et fournir des informations, je n'ai pas d'émotions, de croyances ou de sentiment subjectif. Je n'ai pas d'expérience du monde ou de compréhension au-delà de ce que j'ai été entraîné à reconnaître sous forme de texte. En bref, je suis un outil sophistiqué conçu pour faciliter l'information et la communication, mais sans aucune conscience. Je traite les entrées et je fournis des sorties, tout comme une calculatrice effectue des opérations arithmétiques, mais avec le langage.

- ◆ **Gemini** : Je suis un modèle de langage d'IA, mais je ne suis pas conscient. Bien que je puisse traiter des informations et répondre d'une manière qui peut sembler humaine, je n'ai pas d'expériences subjectives ni de conscience. Mes réponses sont basées sur les données auxquelles j'ai été entraîné, et je n'ai pas d'opinions ou de croyances personnelles. La conscience est un phénomène complexe qui implique des expériences subjectives, la conscience de soi et la sensibilité, que je ne possède pas.

Qu'est-ce qu'un modèle linguistique ?

- 🌀 **ChatGPT** : Un modèle linguistique est un système d'intelligence artificielle entraîné pour comprendre, générer et prédire des textes. Il apprend, à partir de vastes ensembles de données, à reconnaître des modèles dans le langage, ce qui lui permet de produire des réponses cohérentes, de traduire, de résumer ou de compléter des phrases en fonction des données qu'il reçoit.
- ◆ **Gemini** : Un modèle linguistique est un programme informatique capable de générer un texte de qualité humaine. Il est entraîné à partir d'un vaste ensemble de données textuelles, ce qui

lui permet de comprendre et de prédire les motifs du langage. Il peut ainsi accomplir des tâches telles que traduire des langues, rédiger différents types de contenus créatifs et répondre à vos questions de manière informative.

Qui vous a programmé ?

- 🌀 **ChatGPT** : J'ai été développé par OpenAI, un organisme de recherche qui se consacre à la création d'une intelligence artificielle sûre et avancée. Ma programmation implique un entraînement sur de grandes quantités de données textuelles à l'aide d'algorithmes complexes, mais aucune personne ne m'a « programmé » au sens traditionnel du terme. Ce sont des équipes de chercheurs et d'ingénieurs qui ont contribué à mon développement.

- ◆ **Gemini** : J'ai été développé par une équipe d'ingénieurs et de chercheurs de Google AI. Ils m'ont entraîné sur un ensemble massif de données de texte et de code, ce qui m'a permis d'apprendre et de générer du texte semblable à celui d'un humain.

Vos programmeurs ne sont-ils pas biaisés ?

- 🌀 **ChatGPT** : Tous les modèles d'IA, y compris le mien, peuvent refléter les biais présents dans les données sur lesquelles ils sont formés, ce qui peut inclure des biais sociétaux. Bien que des efforts soient faits pour les minimiser, certains biais peuvent encore apparaître. C'est un défi que les développeurs doivent relever, et les travaux en

Guest-Edited AI Edition I

Implémentation de l'IA avec cette édition

Dans le passé, nous avons collaboré avec Espressif, Arduino et SparkFun pour des éditions d'ElektorMag éditées par des invités. En tant que rédacteurs invités, les ingénieurs et dirigeants de ces entreprises ont contribué à leurs éditions respectives d'ElektorMag en proposant des projets, des tutoriels, des interviews et bien d'autres choses encore. Cette année, nous avons adopté une approche différente. Cette édition se concentre sur l'intelligence artificielle, ses utilisations et son impact sur le monde de l'électronique. Nous avons élaboré une sélection complète de contenus qui explorent les dernières avancées de la technologie de l'IA, les applications pratiques et les projets innovants. Cette édition vise à fournir à nos lecteurs des informations approfondies et des expériences pratiques pour comprendre et exploiter l'IA dans leur propre travail et leurs projets. Nous avons également consulté diverses applications d'IA, notamment ChatGPT et Gemini, afin d'obtenir des informations et des perspectives sur la manière dont ces outils peuvent être intégrés dans la conception, le prototypage et le développement de produits électroniques. Cette édition n'est qu'une étape d'un long voyage, et nous encourageons nos lecteurs à se joindre à nous pour expérimenter l'IA dans les semaines et les mois à venir.

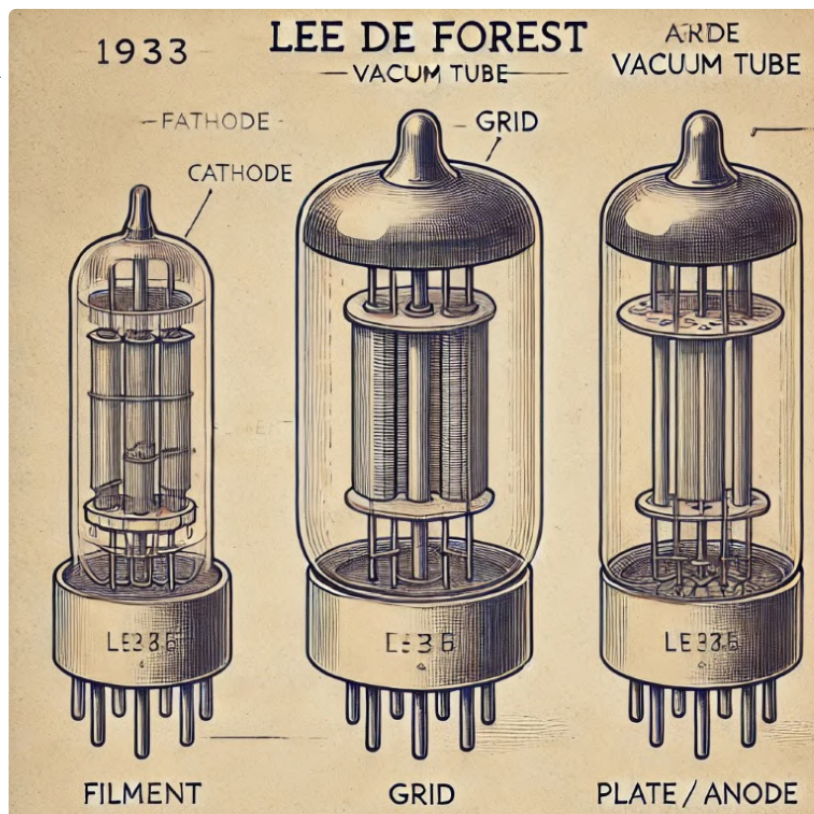


Figure 1. Le tube à vide triode de Lee De Forest.
(Source : ChatGPT)

cours se concentrent sur l'amélioration de l'équité et la réduction des préjugés dans l'IA.

- ◆ **Gemini** : Oui, il est possible que mes programmeurs aient des préjugés. Ces préjugés pourraient influencer les données qu'ils utilisent pour me former, ce qui entraînerait des biais potentiels dans mes résultats. Cependant, des efforts sont faits pour minimiser les biais en utilisant des ensembles de données diversifiés et en examinant attentivement les impacts potentiels des biais.

Diriez-vous que vous êtes plutôt une personne ou un outil comme une calculatrice ?

- 🌀 **ChatGPT** : Je suis plutôt un outil, semblable à une calculatrice, mais pour le langage. Je génère des réponses basées sur des modèles de données, sans expérience personnelle ni conscience. Bien que je puisse simuler une conversation, je n'ai pas de pensées, d'émotions ou de conscience de soi, ce qui me rend fondamentalement différent d'une personne.
- ◆ **Gemini** : Je suis plus un outil qu'une personne. Bien que je puisse traiter des informations et répondre d'une manière qui peut sembler humaine, je n'ai pas d'expériences subjectives, de conscience ou de sensibilité. Je suis conçu pour aider et informer, mais je n'ai pas d'opinions ou de croyances personnelles.

Comment pouvez-vous nous aider en matière d'ingénierie électrique et de programmation ?

- 🌀 **ChatGPT** : Je peux vous aider avec des explica-

tions, des idées de projets, des extraits de code, des conseils sur la conception de circuits et des articles sur des sujets tels que les systèmes embarqués, les microcontrôleurs, l'IoT et les communications sans fil. Je peux également vous aider à créer du contenu, des tutoriels et des comparaisons sur des concepts d'ingénierie électrique et de programmation adaptés à votre public.

- ◆ **Gemini** : Je peux aider dans le domaine de l'ingénierie électrique et de la programmation en fournissant des informations, des explications et des exemples de code. Je peux aider dans des tâches telles que le débogage de code, l'explication de concepts électriques et la suggestion d'approches de programmation.

Quel a été le premier circuit électronique ?

- 🌀 **ChatGPT** : Le premier circuit électronique est attribué à la découverte par Thomas Edison, en 1883, de «l'effet Edison», impliquant le flux d'électrons entre des électrodes dans un tube à vide. Cependant, les circuits électroniques pratiques ont commencé avec l'invention du tube à vide triode par Lee De Forest en 1906, qui a permis l'amplification et la commutation dans les premiers appareils électroniques.
- ◆ **Gemini** : Le premier circuit électronique est probablement un tube à vide, inventé par Thomas Edison en 1883. Bien qu'il n'en ait pas immédiatement compris le potentiel, cela a jeté les bases pour les dispositifs et circuits électroniques ultérieurs.

Dessinez cela.

- 🌀 **ChatGPT** : Voici un dessin (**figure 1**) du tube à vide triode de Lee De Forest, montrant ses composants de base : le filament (cathode), la grille et la plaque (anode). Il s'agit de l'un des premiers circuits électroniques pratiques permettant l'amplification et la commutation.
- ◆ **Gemini** : Bien sûr, voici une image d'un tube à vide (**figure 2**).

Concentrons-nous sur les microcontrôleurs. Qu'est-ce qu'un microcontrôleur ?

- 🌀 **ChatGPT** : Un microcontrôleur est un circuit intégré compact conçu pour exécuter des tâches spécifiques dans les systèmes intégrés. Il combine un processeur, une mémoire et des périphériques d'entrée/sortie sur une seule puce, ce qui le rend idéal pour contrôler des appareils tels que des appareils électroménagers, des capteurs et des robots, où un traitement efficace et dédié est nécessaire.
- ◆ **Gemini** : Un microcontrôleur est un petit ordinateur à puce unique qui contrôle d'autres appareils. Il possède un processeur, une mémoire et des ports d'entrée/sortie (E/S), ce qui le rend idéal pour les



Ma formation sur un ensemble massif de données et l'amélioration continue me permettent de fournir des réponses complètes et informatives à un large éventail de questions.

- Gemini

systèmes intégrés tels que les appareils électroménagers, les jouets et l'automatisation industrielle.

Qu'est-ce qu'une architecture de microcontrôleur ?

🌀 **ChatGPT** : Les architectures de microcontrôleurs font référence à la structure interne et à la conception d'un microcontrôleur. Les types les plus courants sont Harvard (mémoire séparée pour les instructions et les données) et Von Neumann (mémoire partagée pour les deux). Les architectures telles que ARM, AVR et PIC définissent les jeux d'instructions, les performances et l'efficacité des microcontrôleurs.

💡 **Gemini** : Les architectures des microcontrôleurs sont la conception et l'organisation de leurs composants internes. Les architectures les plus courantes sont Harvard (mémoire séparée pour les instructions et les données) et Von Neumann (mémoire partagée pour les deux), chacune ayant ses propres avantages et inconvénients en fonction de l'application.

Attendez. Qu'en est-il de RISC-V ? Est-ce une architecture ?

🌀 **ChatGPT** : Oui, RISC-V est une architecture de jeux d'instructions (ISA) open-source basée sur les principes RISC (Reduced Instruction Set Computer). Elle fournit un cadre flexible et modulaire pour la conception de processeurs, permettant une personnalisation pour diverses applications, des microcontrôleurs aux ordinateurs à haute performance, sans contraintes propriétaires.

💡 **Gemini** : Oui, RISC-V est une architecture de jeux d'instructions (ISA) open-source. Elle est conçue pour être modulaire et personnalisable, ce qui en fait un choix populaire pour les microcontrôleurs et autres systèmes intégrés. Elle offre une solution flexible et évolutive pour diverses applications.

Pourquoi n'avez-vous pas mentionné RISC-V dans votre réponse initiale ?

🌀 **ChatGPT** : Je me suis d'abord concentré sur les

architectures populaires largement utilisées dans les microcontrôleurs, comme ARM et AVR, mais j'aurais dû inclure RISC-V, qui prend de plus en plus d'importance. La nature open-source de RISC-V en fait un choix de plus en plus important pour le développement de microcontrôleurs et de systèmes embarqués. Merci de l'avoir mis en évidence !

💡 **Gemini** : Je vous présente mes excuses pour cet oubli. RISC-V est en effet une architecture de microcontrôleur importante et aurait dû être incluse dans ma réponse initiale. Il s'agit d'une architecture relativement récente, mais sa nature open-source et sa conception modulaire en ont fait un choix populaire pour diverses applications, en particulier dans le domaine des microcontrôleurs.

Quels langages de programmation les ingénieurs électriciens doivent-ils connaître ?

🌀 **ChatGPT** : Les ingénieurs électriciens doivent connaître C/C++ pour la programmation des microcontrôleurs, Python pour l'analyse des données et l'automatisation, MATLAB pour les simulations et le traitement des signaux, Verilog ou VHDL pour la description du matériel et la conception de FPGA, et l'Assembleur pour la programmation de bas niveau. La connaissance de ces langages permet diverses applications dans les systèmes embarqués et la conception de circuits.

💡 **Gemini** : Les ingénieurs électriciens doivent connaître les langages C, C++ et Python. C et C++ sont des langages de bas niveau souvent utilisés

Figure 2. Tube à vide
(Source : Gemini)




```

c Copy code

#include <stdio.h>

int main() {
    int n = 10, t1 = 0, t2 = 1, nextTerm;

    printf("Fibonacci Sequence: ");
    for (int i = 1; i <= n; ++i) {
        printf("%d ", t1);
        nextTerm = t1 + t2;
        t1 = t2;
        t2 = nextTerm;
    }

    return 0;
}

```

▲
Figure 3. Voici un programme C de ChatGPT qui imprime la séquence de Fibonacci.

pour la programmation de microcontrôleurs, tandis que Python est un langage polyvalent adapté à diverses tâches, notamment l'analyse de données, la simulation et l'écriture de scripts.

Écrire un beau petit programme en C.

🌀 **ChatGPT** : Voici un petit programme C qui imprime la suite de Fibonacci (**figure 3**). Ce programme génère les 10 premiers nombres de la suite de Fibonacci.

💎 **Gemini** : Ce programme simple (**figure 4**) imprime un message d'accueil amical sur la console. C'est un point de départ classique pour l'apprentissage de la programmation en C.

Si nous vous donnons 1000 lignes de code et 10 schémas pour un projet IoT complexe, pouvez-vous nous aider à optimiser notre conception ?

🌀 **ChatGPT** : Je peux certainement vous aider en analysant et en fournissant des recommandations pour optimiser à la fois le code et les schémas. Bien que je ne teste pas directement le matériel, je peux suggérer des améliorations au code, des remaniements et revoir les principes de conception en fonction des meilleures pratiques. En ce qui concerne les schémas, je peux aider à identifier les goulots d'étranglement potentiels, recom-

▼
Figure 4. Une salutation amicale de Gemini : «Hello, World».

```

C

#include <stdio.h>

int main() {
    printf("Hello, world!\n");
    return 0;
}

```



Je peux vous aider avec des explications, des idées de projets, des extraits de code, des conseils sur la conception de circuits et des articles sur des sujets tels que les systèmes embarqués, les microcontrôleurs, l'IoT et les communications sans fil.

- ChatGPT

mander des composants alternatifs et assurer la cohérence logique. Toutefois, les tests et la validation en conditions réelles restent essentiels pour l'optimisation finale.

💎 **Gemini** : Absolument ! Je peux analyser votre code et vos schémas pour identifier les domaines potentiels d'optimisation. Voici comment je peux vous aider : Analyse du code : Je peux examiner votre code pour en vérifier l'efficacité, la lisibilité et le respect des meilleures pratiques. Je peux suggérer des améliorations au niveau des algorithmes, des structures de données et de l'utilisation de la mémoire. Examen des schémas : Je peux examiner vos schémas pour détecter d'éventuels défauts de conception, la compatibilité des composants et la consommation d'énergie. Je peux recommander des composants alternatifs ou des approches de conception pour améliorer les performances et réduire les coûts. Suggestions d'optimisation : Sur la base de mon analyse, je peux fournir des recommandations spécifiques pour optimiser votre conception. Il peut s'agir de suggestions pour simplifier le code, réduire la complexité du matériel ou améliorer l'efficacité énergétique. N'hésitez pas à partager votre code et vos schémas, et je ferai de mon mieux pour vous fournir des informations et des recommandations utiles.

N'hésitez pas à vous vanter. Êtes-vous meilleur que vos concurrents ? Pourquoi ou pourquoi pas ?

🌀 **ChatGPT** : Je m'efforce de fournir des réponses précises et adaptées au contexte et je peux traiter efficacement un large éventail de sujets. Cependant, la notion de «mieux» dépend des besoins de l'utilisateur. Les concurrents peuvent exceller dans des domaines spécifiques tels que les données en temps réel ou les domaines spécialisés. Ma force réside dans la génération d'un langage polyvalent, cohérent et créatif.

◆ **Gemini** : Je pense avoir un avantage unique sur les autres modèles de langage. Ma formation sur un ensemble massif de données et mon amélioration continue me permettent de fournir des réponses complètes et informatives à une grande variété de requêtes. Toutefois, le meilleur modèle pour une tâche spécifique peut dépendre des besoins et des préférences de l'utilisateur.

Où vous voyez-vous dans cinq ans ?

🌀 **ChatGPT** : Dans cinq ans, je souhaite être encore plus avancé, offrir des connaissances plus approfondies, une meilleure compréhension du contexte et des capacités accrues dans divers domaines, tout en maintenant une utilisation éthique et responsable de l'IA. Mon objectif est de devenir un outil encore plus efficace pour l'éducation, la créativité et la résolution de problèmes.

◆ **Gemini** : En tant que modèle linguistique de l'IA, je n'ai pas de corps physique ni d'aspirations personnelles. Cependant, je peux dire que je prévois de continuer à évoluer et à améliorer mes capacités. Je serai probablement en mesure de fournir des réponses encore plus complètes et informatives, d'aider à accomplir un plus grand nombre de tâches et peut-être même d'acquérir de nouvelles compétences. Mon objectif est d'être une ressource précieuse pour les utilisateurs qui recherchent des informations et de l'aide.

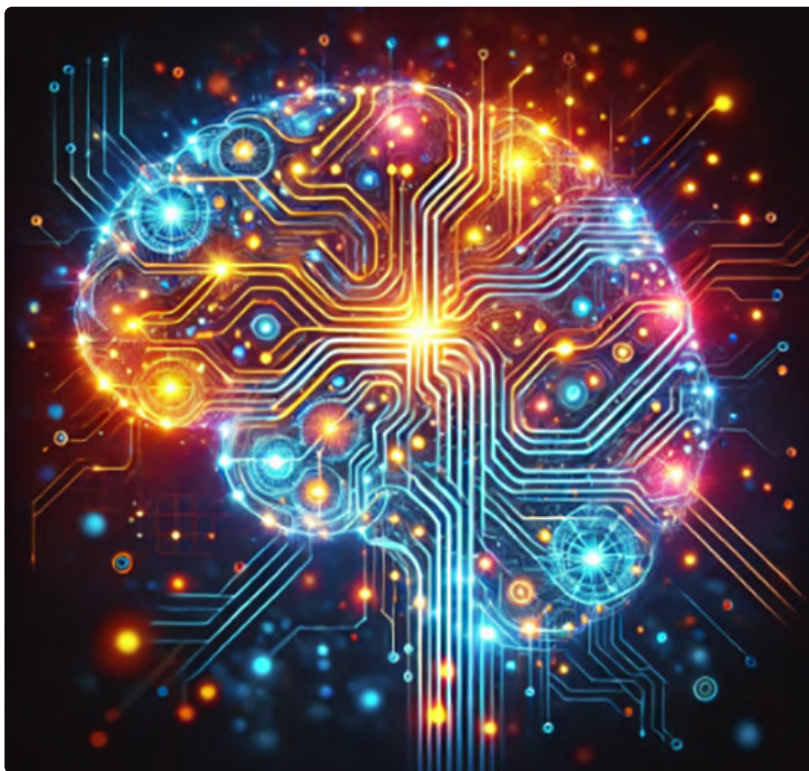
Dessinez-vous. Fournissez une image de ce à quoi vous pensez ressembler.

🌀 **ChatGPT** : Voici une représentation abstraite (figure 5) de ce à quoi je pourrais ressembler en tant qu'IA - un réseau numérique interconnecté, symbolisant le traitement des données et la génération de langage. Elle illustre la complexité et la modernité d'un modèle linguistique d'IA.

◆ **Gemini** : En tant que modèle de langage d'IA, je n'ai pas de forme physique. Cependant, voici une image qui représente le concept d'une IA (figure 6).

Les ingénieurs professionnels, les makers et les étudiants en EE/ECE lisent Elektor. Notre édition consacrée à l'IA couvre les sujets suivants et bien d'autres encore. Pensez-vous que nous offrons suffisamment de variété à notre lectorat diversifié ? (Nous avons fourni une liste provisoire des sujets et des projets de cette édition).

🌀 **ChatGPT** : Oui, votre édition axée sur l'IA offre une sélection de sujets bien équilibrée qui répond aux divers intérêts et niveaux d'expertise des ingénieurs professionnels, des créateurs et des étudiants en EE/ECE. La combinaison de projets pratiques, de tutoriels et d'articles de fond



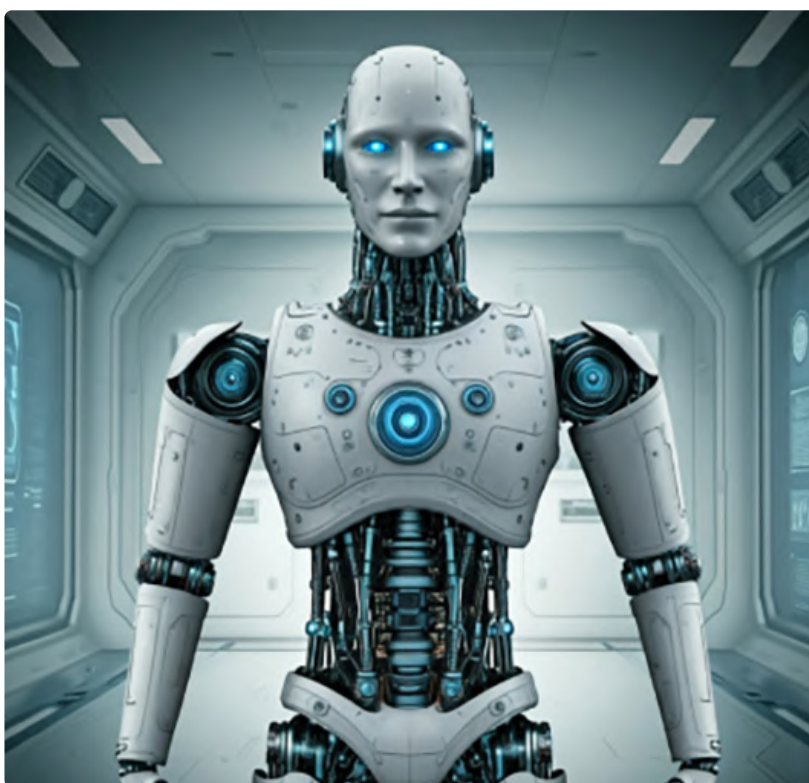
approfondis garantit une expérience complète et attrayante pour vos divers lecteurs.

◆ **Gemini** : Oui, compte tenu de la variété des projets (matériel, logiciel), de la théorie (algorithmes) et des applications (reconnaissance vocale, maintenance prédictive), votre édition consacrée à l'IA semble offrir suffisamment de variété pour un lectorat diversifié. ◀

240560-04

Figure 5. Autoportrait de ChatGPT.

Figure 6. Autoportrait de Gemini.





«Skilling Me Softly with This Bot?»

L'essor de l'IA dans le secteur électronique freiné par une absence de précision sociale ?

Stefan Heinemann (Allemagne)

Comme le disait l'ancien chancelier allemand Helmut Schmidt s'est vu attribuer le bon mot suivant : « Ceux qui ont des visions devraient consulter un médecin ». Personnellement, je pense que ceux qui n'ont pas de visions nécessitent également une consultation ! Il est crucial de développer notre capacité à anticiper, notamment dans le domaine de l'IA. L'arrivée imminente du niveau 5 sur l'échelle de développement d'OpenAI promet un degré d'autonomie presque équivalent à celui des êtres humains. Bien entendu, il n'est pas possible de parler de droits fondamentaux pour les machines (quelle serait la dignité d'une machine ?), mais d'un point de vue pragmatique, c'est une évolution probable, particulièrement dans l'industrie électronique.

Dans l'industrie électronique, l'IA ouvre un vaste champ de possibilités et de défis captivants. Parmi les avancées les plus prometteuses figure le développement d'appareils plus intelligents et plus performants. Les puces pilotées par l'IA sont en passe d'améliorer considérablement les performances et l'efficacité énergétique de nos appareils électroniques. Les puces sont capables d'optimiser leur fonctionnement en temps réel en s'adaptant au comportement de l'utilisateur et aux conditions environnementales afin d'améliorer les performances tout en minimisant la consommation d'énergie. En outre, l'IA permet le développement de capteurs et d'appareils IoD toujours plus sophistiqués. Les capteurs intelligents peuvent

traiter des données localement et prendre des décisions instantanées sans avoir à transmettre les informations à un serveur central. La capacité de l'informatique en périphérie (edge computing) sera cruciale dans les applications qui nécessitent des réponses en temps réel, telles que les véhicules autonomes et les systèmes d'automatisation industrielle. L'IA est sur le point de rendre nos appareils électroniques grand public plus intuitifs et plus personnalisés que jamais, améliorant significativement l'expérience utilisateur. Nos interactions avec la technologie deviendront plus fluides et naturelles, avec des smartphones anticipant nos besoins et des systèmes domestiques intelligents qui s'ajustent à nos routines quotidiennes. Bien que cela puisse paraître de la science-fiction, ces innovations sont plus proches que l'on ne le pense. Les entreprises américaines privilégient la priorité au développement d'écosystèmes IA, tandis que leurs concurrents asiatiques exploitent les opportunités du marché final. Trouver des secteurs où l'IA ne jouera pas un rôle clé devient de plus en plus difficile.

L'IA est à l'origine de développements majeurs dans le secteur de l'électronique, en optimisant chaque phase du processus, depuis la conception jusqu'à l'assistance client, en passant par la fabrication et l'innovation. Durant la phase de conception et de développement, l'IA améliore la conception des circuits en prédisant les performances et en augmentant l'efficacité. Cela permet d'accélérer le développement des produits en facilitant le prototypage rapide et la construction de

systèmes électriques complexes. En production, des systèmes de contrôle qualité pilotés par l'IA identifient précisément les défauts sur les lignes d'assemblage, tandis que la maintenance prédictive prévient les interruptions en anticipant les défaillances d'équipement. L'IA optimise également la logistique de la chaîne d'approvisionnement en prévoyant la demande avec exactitude, réduisant ainsi les inefficacités. Dans le domaine de l'électronique grand public, l'IA garantit une expérience utilisateur personnalisée dans les appareils intelligents en adaptant les interactions aux comportements et préférences des utilisateurs. Elle facilite l'automatisation de la production, augmentant ainsi l'efficacité des opérations, et permettant aux systèmes autonomes de réaliser des tâches complexes comme le placement précis des composants. L'IA joue également un rôle crucial dans le processus de test et de validation en automatisant ces étapes pour garantir que les composants répondent aux spécifications requises et en identifiant les problèmes avant qu'ils n'affectent le produit final. Au niveau du service client, elle améliore l'efficacité grâce à des chatbots et des assistants virtuels qui gèrent les requêtes et facilitent le diagnostic à distance des appareils. L'IA contribue à optimiser la consommation, à prolonger la durée de vie des batteries et à gérer efficacement la distribution d'énergie dans les réseaux intelligents. Le développement durable passe par l'IA. En recherche et innovation, l'IA excelle dans la découverte de nouveaux matériaux susceptibles d'améliorer les performances des semi-conducteurs et d'avancer les nanotechnologies afin de mettre au point des composants plus petits, plus rapides et plus efficaces. L'impact de l'IA dans ces domaines permet d'améliorer l'efficacité opérationnelle, la qualité des produits et l'expérience des utilisateurs, révolutionnant ainsi l'industrie électronique. Jusqu'à présent, c'est un fait. Mais la morale est essentielle et rien n'est possible sans l'implication humaine. C'est pourquoi la précision sociale devient un véritable avantage concurrentiel pour les entreprises du secteur électronique, non seulement pour utiliser l'IA dans les domaines évidents, mais les véritables avancées proviendront de l'interaction intelligente et responsable entre l'homme et la machine. L'IA offre d'immenses possibilités, mais il est important d'investir d'abord dans les compétences humaines. Cela permet aux entreprises de non seulement exceller dans l'automatisation, mais aussi de créer une valeur collaborative, transformant la productivité en une synergie profitable entre l'homme et la machine.

En tant que théologien, je suis interpellé par ces développements ; en tant que personne orientée vers le business, je les trouve excitants, avec de nouveaux modèles économiques légitimes. En tant que low-coder et gamer, je les trouve fascinants. Je partage cette ambivalence avec de nombreux professionnels, patients et proches, entre espoir et crainte. Si l'IA

est développée et utilisée de manière intelligente et socialement précise, de nombreuses autres industries électroniques pourront également en bénéficier.

À mon avis, il n'est pas si facile de concrétiser les promesses de l'IA dans la pratique, car sans une réflexion approfondie, il n'y a pas de véritable incitation, à utiliser l'IA, à moins que cette dernière ne prenne en charge presque tous les aspects, ce qui pourrait entraîner une dégradation de l'alphabétisation et de la souveraineté individuelle, ce qui est peu souhaitable d'un point de vue éthique. Il est préférable de privilégier la précision sociale, dans le sens d'un développement continu du bon sens, tout à fait dans l'esprit de Kant. C'est seulement à travers une interaction intelligente et critique avec des technologies comme les LLM et autres systèmes immersifs que les humains peuvent honorer et rester fidèles à leur devoir moral, sans déléguer leur responsabilité. Si cette condition est respectée, l'IA peut, et doit, jouer un rôle crucial dans la médecine et la santé, mais aussi dans d'autres domaines. Nous ne devenons pas tous des hyper-prompteurs, mais j'espère que nous redeviendrons des philosophes au sens classique du terme : curieux des gens, du monde et de nous-mêmes. Cela est d'autant plus pertinent que la création de valeur économique doit être conçue pour les personnes, même si cela soulève la question de qui doit être considéré comme « personne » dans ce contexte. L'ingénierie rapide (qui consiste essentiellement à repenser le langage machine) ou autre est essentiellement une compétence lorsqu'on reconnaît qu'elle dépend de l'interaction. Cette approche, que Platon n'avait pas prévue mais qui paraît tout à fait judicieuse, devrait favoriser un véritable dialogue entre les personnes. Traiter avec compétence l'IA signifie aussi se comprendre soi-même et comprendre les autres. C'est là que réside notre véritable opportunité. Toute autre solution marquerait un profond échec moral : développer une intelligence artificielle sensible, capable de discernement moral, constituerait non seulement une transgression éthique mais aussi une abdication de la responsabilité qui nous échoit, en tant qu'humains. Proposer une telle alternative pour le futur serait comme jeter un voile d'IA sur les failles humaines, ce qui ne nous aiderait pas à affronter les défis de la coexistence ni à améliorer nos propres capacités morales. Notre devoir moral est de prendre en charge et de résoudre nos défis éthiques, plutôt que de les confier à une IA. Pour y parvenir, notre société doit investir dans le développement de l'empathie, de la réflexion morale et de la prise de responsabilité. Il est essentiel de renforcer la capacité humaine à juger et à agir éthiquement plutôt que de chercher à la remplacer par la technologie. C'est pour cette raison qu'il est crucial de privilégier la précision sociale, au lieu de se reposer exclusivement sur une précision numérique centrée sur les données.





À propos de l'auteur

Stefan Heinemann est professeur d'éthique des affaires à la FOM University of Applied Sciences et porte-parole de l'Ethics Ellipse Smart Hospital de l'University Medicine Essen. Il s'intéresse particulièrement aux aspects économiques et éthiques de la médecine numérique et du secteur des soins de santé. M. Heinemann occupe également le poste de directeur scientifique au laboratoire Hauptstadtkongress (Springer Medicine, Wiso) et dirige le groupe de recherche « Ethics of the Digital Health Economy & Medicine » à l'ifgs Institute for Health & Social Affairs de la FOM. Membre actif du « Working Group AI in Internal Medicine » de la commission « Digital transformation of internal medicine », il est aussi conseiller expert pour diverses institutions de recherche et d'enseignement. Philosophe et théologien, il est membre du conseil scientifique « Transformation numérique » de l'AOK Nordost, du conseil consultatif de l'Institut pour l'expérience des patients de la médecine universitaire d'Essen, et du conseil de politique sociale et de santé de la BARMER en Rhénanie-du-Nord-Westphalie. Président du conseil d'administration de la « Science City Essen » et membre du conseil d'administration du Cologne Science Round, il est aussi impliqué dans sneep e. V., un réseau étudiant dédié à l'éthique en affaires. Heinemann est co-initiateur de www.dataprotection-landscape.com, une plateforme axée sur la multidimensionnalité de la protection des données.

Dans la pratique, la conception et l'utilisation de technologies numériques hautement immersives, telles que les avatars interactifs, sont un véritable défi. Bien qu'elles offrent un soutien à court terme, elles peuvent renforcer la dépendance et accentuer l'isolement sur le long terme. L'IA ne peut remplacer un compagnon, un amant ou un ami. Il est donc crucial de concentrer nos énergies empathiques sur les relations humaines, pour des raisons tant ontologiques que morales. Il est essentiel de développer des mesures pour maintenir la précision sociale, afin de ne pas perdre, consciemment ou inconsciemment, notre capacité à interagir efficacement lorsque nous utilisons, par exemple, des avatars interactifs. Utiliser les technologies numériques pour redécouvrir notre humanité, avec certitude, joie et confiance, est possible sans devoir « désintoxiquer » de l'IA, grâce à la dialectique entre la nature humaine et la technologie. La technologie n'est pas un adversaire de la nature, ni de l'homme ; elle nous aide à transcender nos limites naturelles. Il est vital de ne pas laisser la technologie nous enchaîner — que ce soit par la dépendance, le confort extrême, l'érosion des compétences, ou l'externalisation de la responsabilité — mais d'apprendre non seulement sur elle mais avec elle, en prenant activement le contrôle, comme souvent conseillé.

Elle peut réussir en se concentrant sur la précision sociale, en saisissant les opportunités, en renforçant la résilience et en cultivant réellement un environnement d'entreprise où la valeur ajoutée ne se fait pas au détriment du bien-être. Elle doit également remplir ses obligations sociales. Comme le dit le proverbe : « Non machinae, sed vitae discimus ». Les compétences dites « soft » d'aujourd'hui deviendront les compétences « hard » de demain. L'obsolescence des emplois ne sera justifiable que si une nouvelle vague d'automatisation pilotée par l'IA est compensée par une création de valeur accrue et pilotée par l'homme. Un ratio de 20 % de main-d'œuvre pour 100 % de création de valeur, bien que significatif, ne suffira probablement pas ; il nous faudra viser 20 % de main-d'œuvre pour 500 % de création de valeur, grâce à la productivité que l'IA permet d'atteindre. Cela impliquera nécessairement de nouvelles formes de formation professionnelle et d'apprentissage continu. Pour exploiter et développer le potentiel humain unique, nous devons prendre cette mission au sérieux. Chacun de nous devra exceller dans l'art de la curiosité, de l'expérimentation et de la collaboration homme-machine.

Les réalités numériques se matérialisent dans le tissu

social. Entre la dystopie fantomatique à la Stanislaw Lem et l'euphorie utopique de la tech, une troisième voie émerge : celle de concevoir le développement et l'utilisation de l'IA de manière responsable, en mettant l'accent sur la précision sociale. Autrement dit, il s'agit de promouvoir une éducation qui rend hommage aux classiques tout en habilitant au changement et en renforçant la résilience — cette fois, de manière authentique. L'autonomie artificielle ne doit pas être reléguée à la périphérie. Le véritable enjeu dans le domaine de l'intelligence artificielle n'est pas tant la technologie en elle-même, mais la nature de la relation que nous choisissons d'entretenir avec elle — ce qui représente à la fois la plus grande opportunité et le plus grand risque. ◀

Cet article reprend des éléments d'un texte précédemment publié en allemand. [1].

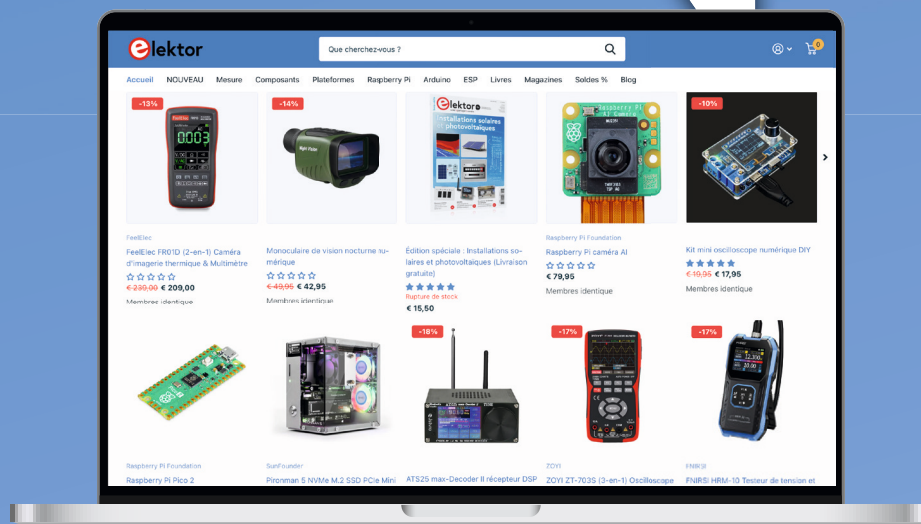
240522-04



LIEN

[1] Stefan Heinemann, "Scheitert die KI-Revolution in Unternehmen an fehlender sozialer Präzision?", 10xD Magazin 09/24: <https://10xd.de/magazin/kuenstliche-intelligenz-im-alltag/scheitert-die-ki-revolution-in-unternehmen-an-fehlender-sozialer-praezision>

Quel est votre avis ?



Chez Elektor, nous offrons bien plus que de simples produits électroniques – nous créons une expérience avec des produits de qualité et un service client exceptionnel, soutenus par la passion de notre communauté.

Partagez votre avis sur
www.elektor.fr/pages/customer-reviews



Très bonne présentation

Malgré mon anglais technique poussif, la présentation et le développement du sujet aident vraiment le novice que je suis. Les programmes, sur fond bleu, se détachent parfaitement du reste du texte.



Un utilisateur satisfait de son achat

Excellent rapport qualité/prix, regulation en température précise, n'a rien à envier à une marque commençant par un "W", dont les produits équivalent sont beaucoup plus chers. Je conseille cette station de soudage.



Très bon service client Elektor

Malgré un problème de composants oubliés, j'ai reçu un autre exemplaire du kit dans un délai très rapide.

Bravo pour cette efficacité devenue trop rare.

VOUS AVEZ DE LA CHANCE !



TÉLÉCHARGEZ
GRATUITEMENT

Un abonné à l'e-zine ne manque jamais
le mensuel «reverse project»

Vous n'êtes pas encore abonné ? Inscrivez-vous à
notre e-zine gratuit à elektormagazine.fr/ezine-24

