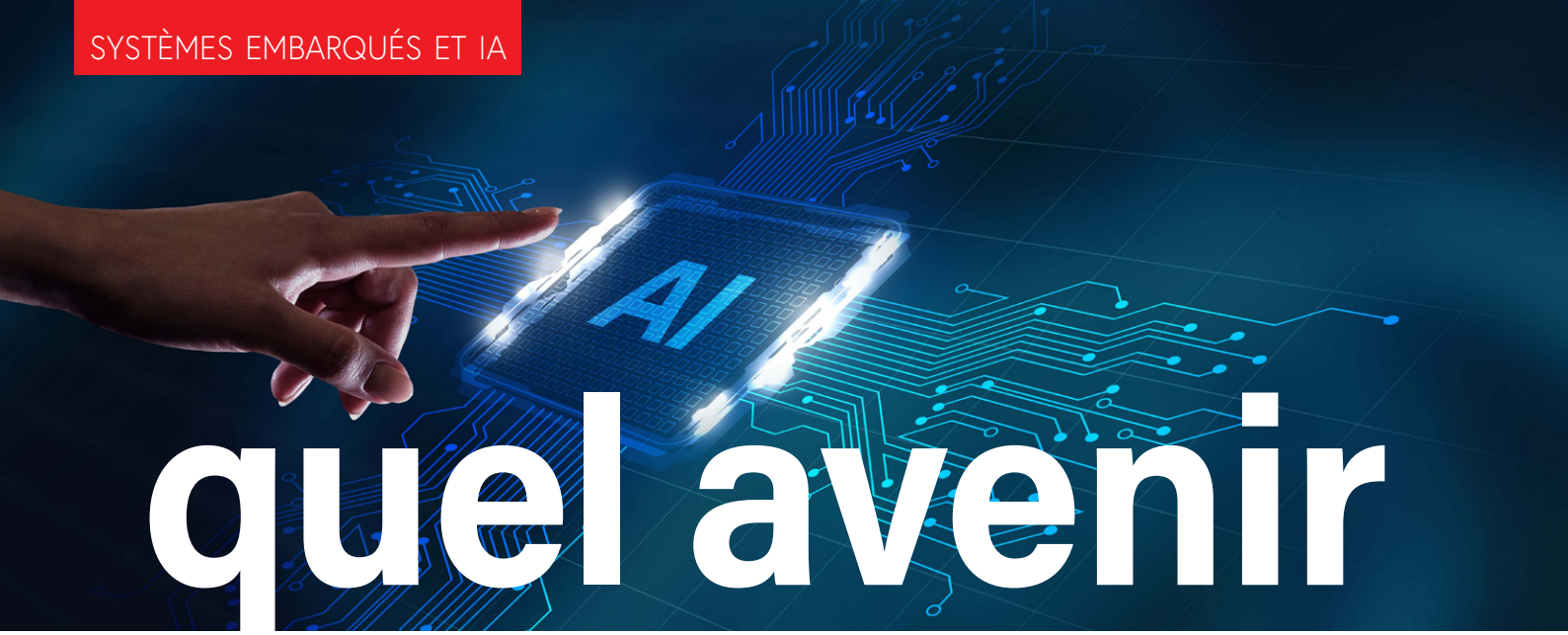

quel avenir pour l'IA et les systèmes embarqués ?

outils, plateformes et remplacement des rédacteurs

Stuart Cording (Elektor)

L'intelligence artificielle (IA) est souvent présentée comme une arme universelle pour résoudre différents problèmes, quel que soit le défi à relever. Parfois aussi, elle est présentée comme annonçant la fin de la civilisation elle-même. En tant qu'ingénieurs, nous savons qu'aucune de ces deux affirmations n'est vraie. Pour autant, comment pouvons-nous mieux utiliser l'IA, ou plus exactement l'apprentissage automatique, dans les applications que nous développons ? Et les progrès les plus récents de l'IA feront-ils vraiment de nous des intervenants superflus ?

Mentionner l'IA est le plus sûr moyen d'attirer l'attention de la presse grand public. Grâce à l'internet et aux services dans le nuage, mais aussi du fait de sources de données parfois douteuses, de nouveaux services basés sur l'IA semblent surgir partout. En quelques clics de souris, vos propres mots peuvent être placés dans la bouche de Morgan Freeman ou votre image intégrée dans une nouvelle œuvre d'art. Et même si cela peut être amusant et donner à réfléchir, il n'est pas facile de trouver un lien avec le monde des systèmes embarqués. Voyons précisément ce qui se passe dans ce secteur.

Pour fonctionner, la plupart des systèmes embarqués utilisent des approches de programmation à base de règles. Pour un ensemble de données d'entrée défini, une série d'instructions *if/else* ou *switch* détermine la réponse à donner. Le principe fonctionne bien pour un nombre limité d'entrées. Pourtant, à un moment, le nombre d'entrées ou la subtilité de leurs relations rend difficile la définition de règles programmatiques claires.

À titre d'exemple, imaginez un moteur installé dans une usine, et fonctionnant 24 heures sur 24, sept jours sur sept. L'expérience montre qu'avec le temps, il se produit une usure et un épaississement du lubrifiant des roulements. Au fil du temps, cela modifie la durée de rotation, le bruit produit, les modes de vibration, la température de fonctionnement du moteur et le courant consommé. Assurer une maintenance périodique permet de résoudre cette problématique, même si elle entraîne des temps d'arrêt fixes préjudiciables à la production. Il est souvent difficile de dire si la fréquence est trop importante ou trop faible.

En outre, il est peu probable de détecter une défaillance imminente résultant d'une fracture capillaire de l'arbre, des roulements, du boîtier ou des fixations. Une fois correctement entraînées, les approches par IA peuvent déterminer les défaillances en cours en utilisant un mélange complexe de sources de données. S'il est possible de déployer cette intelligence dans un système à base de microcontrôleur, vous obtenez un dispositif de surveillance abordable qui permet de gagner du temps et de diminuer les dépenses, mais aussi de réduire le gaspillage dû aux opérations de lubrification inutiles et au remplacement superflu des pièces.

Introduire de l'intelligence dans les microcontrôleurs

Concernant les microcontrôleurs, il s'agit davantage d'apprentissage automatique (ou *machine learning* - ML) plutôt que d'IA. L'objectif est ici de prendre des décisions en programmant une machine pour qu'elle utilise des règles élaborées grâce à l'analyse des données disponibles. Si les microcontrôleurs disposent d'une puissance plus que suffisante pour exécuter de tels algorithmes de ML, l'apprentissage à partir de données d'entraînement reste hors de portée pour eux

et nécessite au moins un ordinateur de bureau, voire un serveur cloud. Fondée en 2019, l'entreprise Edge Impulse a développé une plateforme dédiée au ML dans les systèmes embarqués et s'est associée avec succès aux fournisseurs de semi-conducteurs du monde entier [1] pour assurer une prise en charge très large. Les données constituent le point de départ de toute application de ML. Alors que certaines applications, la conduite autonome par exemple, nécessitent des téraoctets de données d'entraînement, les systèmes simples basés sur des microcontrôleurs peuvent apprendre à partir de quelques kilo-octets de données seulement. Le premier défi consiste donc à transférer les données depuis le microcontrôleur vers l'environnement Edge Impulse. L'idée initiale serait d'utiliser l'interface série d'un Arduino pour les envoyer à votre PC, puis à partir de là, les télécharger dans un fichier texte. La plateforme est cependant configurée pour ingérer les données directement.

Les données, carburant de l'IA

L'un de ces outils est le Data Forwarder [2], une application à ligne de commande (CLI) qui envoie les données d'une carte de développement directement à l'environnement Edge Impulse. Grâce à votre nom d'utilisateur et votre mot de passe, une liaison est établie entre le port série de votre PC et le serveur. Du côté du microcontrôleur, il suffit d'envoyer les données sur l'interface série dans un format délimité par des virgules ou des tabulations. Tant que le taux d'échantillonnage est relativement faible, il s'agit d'un moyen idéal pour recueillir des données représentatives directement à partir de vos capteurs (**figure 1**). Les systèmes embarqués plus puissants, comme le Raspberry Pi ou le NVIDIA Jetson Nano, peuvent utiliser le kit de développement logiciel (SDK) [3] fourni. Cela permet également de prendre en charge des capteurs comme les microphones et les caméras qui produisent des quantités plus importantes de données.

Une fois les données téléchargées, l'étape suivante consiste à définir une « impulsion », avec une répartition en deux blocs. Le premier découpe les données en petits morceaux et utilise des techniques de traitement du signal pour extraire des caractéristiques. Cela permet de s'assurer que les données des capteurs disponibles sont transformées en informations cohérentes pour la deuxième étape de traitement. Le deuxième bloc est celui où sont effectués l'apprentissage et la classification (**figure 2**). Dans un exemple de projet, « Reconnaissance d'un mouvement continu », on trouve une bonne explication de la façon dont ces blocs sont configurés pour analyser les données d'un accéléromètre et classer une entrée si elle correspond à l'un des quatre gestes [4].

Il s'agit de l'étape la plus critique d'un développement de ML. Elle nécessite souvent une démarche de réflexion latérale (sous plusieurs angles ou hors du champ habituel d'études), mais aussi différentes itérations pour déterminer la meilleure approche. Parfois, ignorer certaines entrées de capteurs est la meilleure solution, alors que dans d'autres cas, il est nécessaire de disposer de davantage de données. Vous constaterez éventuellement que vous êtes en situation de sur-apprentissage, ou que le modèle de réseau neuronal choisi est mal adapté à la classification que vous tentez d'effectuer. Une autre étape cruciale est la classification des anomalies. Dans l'exemple de projet, il y a quatre gestes définis. Cependant, il est nécessaire d'exclure d'autres mouvements similaires aux gestes

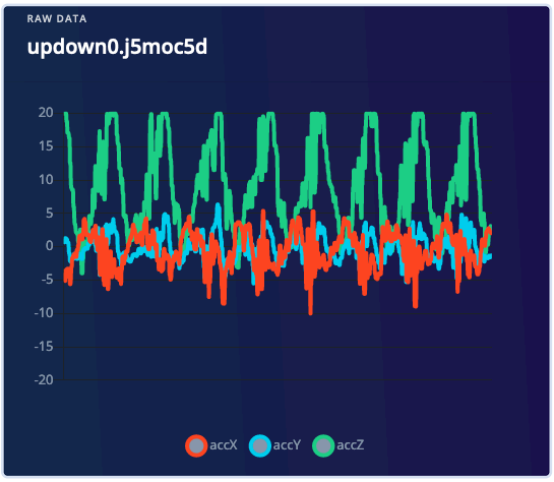

Figure 1. Un accéléromètre à trois axes produit des données relatives aux mouvements pour développer une application de reconnaissance des gestes basée sur le ML. (Source : Edge Impulse)

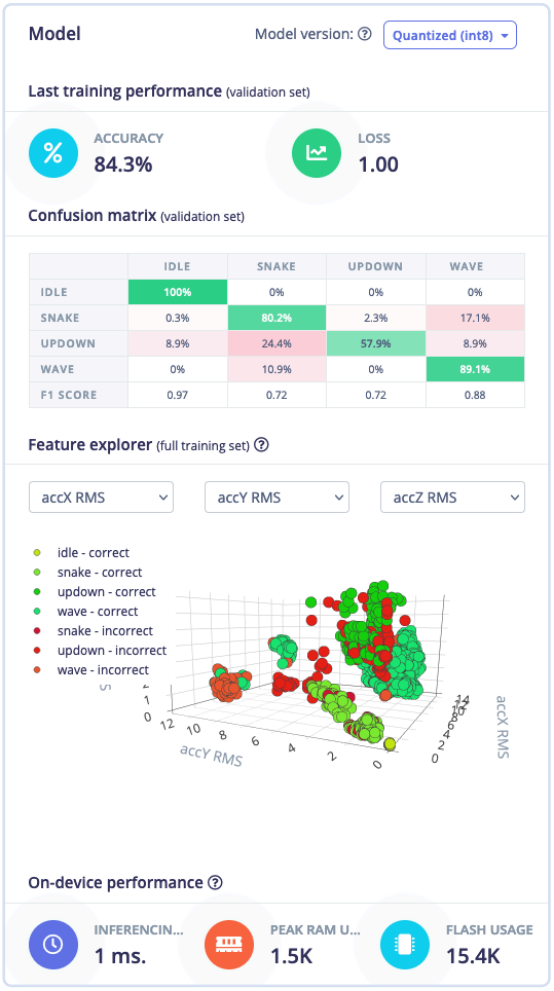

Figure 2. L'environnement Edge Impulse donne des informations sur la précision, la vitesse et l'utilisation de la mémoire après un cycle d'entraînement initial. (Source : Edge Impulse)

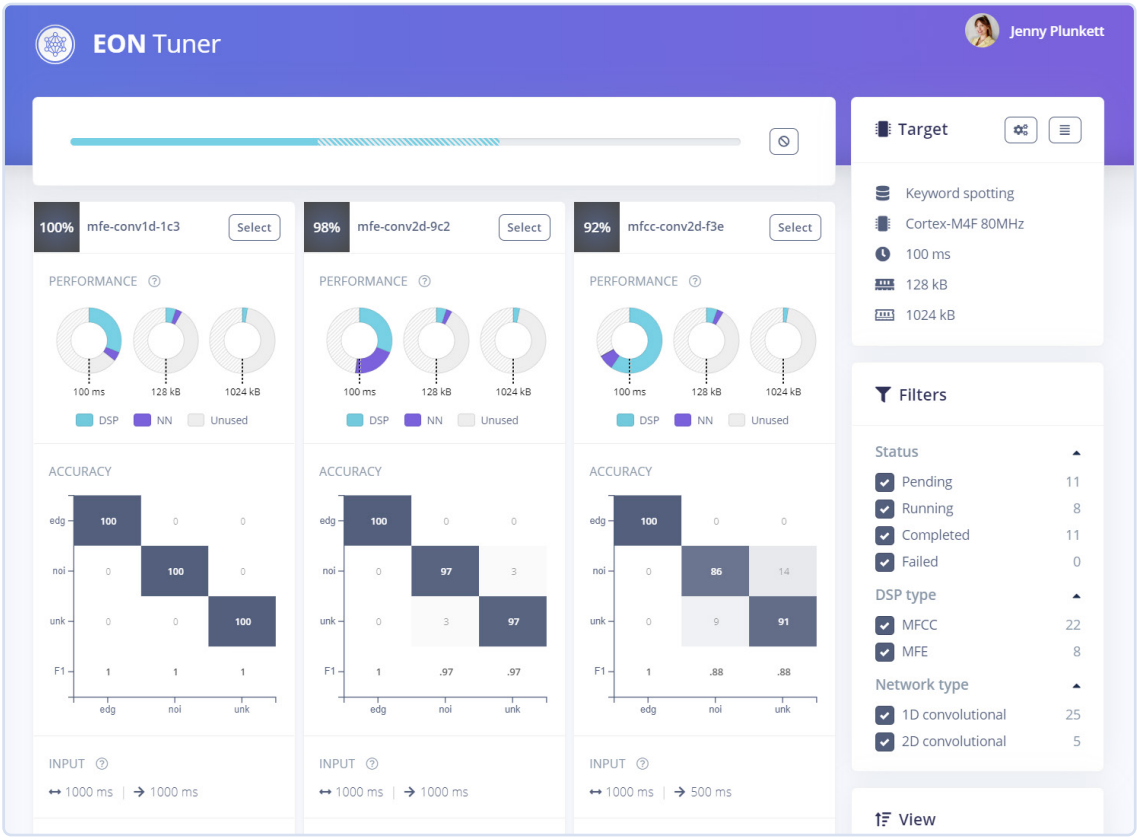

Figure 3. Les algorithmes de ML sont optimisés pour le microcontrôleur cible à l'aide d'EON Tuner, ce qui permet d'améliorer les temps de réponse des inférences et de réduire l'utilisation de la mémoire. (Source : Edge Impulse)

appris. Une bonne détection des anomalies permet d'obtenir un résultat de ML plus robuste.

L'étape finale est le déploiement. Par le biais de l'interface web, le micrologiciel de l'appareil choisi est téléchargé pour être intégré dans votre application. Pour Arduino, une bibliothèque est générée, tandis que pour les autres microcontrôleurs, vous pouvez générer un fichier C++. Bien entendu, les performances des microcontrôleurs varient de manière considérable. Pour garantir le meilleur résultat possible, Edge Impulse propose son EON Tuner [5]. Cet outil peut améliorer la précision de la détection, accélérer les inférences et réduire les besoins en mémoire en utilisant des informations comme le dispositif cible, la taille de la mémoire et la latence (**figure 3**).

Figure 4. SlateSafety a utilisé Edge Impulse pour assurer une surveillance physiologique ML de pointe et améliorer la surveillance de la sécurité du BAND V2. (Source : SafetySlate)

Le ML dans les applications réelles

Des applications réelles utilisent cette approche pour intégrer des capacités de ML. Le BAND [6] de SlateSafety intègre une série de capteurs biométriques pour surveiller les personnes qui travaillent dans des situations difficiles (**figure 4**). Cela concerne aussi bien les premiers secours que les personnes travaillant en milieu industriel, et portant des équipements de protection individuelle (EPI) lourds, comme les pompiers. Le produit télécharge normalement les données dans le cloud afin que leurs collègues puissent surveiller les signaux vitaux d'un utilisateur. Cependant, notamment dans les situations de catastrophe, la connectivité peut être inégale ou inexistante. L'équipe de développement a utilisé Edge Impulse pour intégrer le ML en périphérie dans le produit existant, entraîné sur des données biométriques historiques. Grâce à l'EON Tuner, l'algorithme a été optimisé pour le matériel, puis déployé à l'aide d'une mise à jour par ondes radio. Le BAND peut ainsi avertir le porteur en cas de risque d'épuisement dû à la chaleur, même en l'absence de connexion sans fil.

Améliorer le développement des produits grâce à l'IA

Bien entendu, l'IA ne doit pas nécessairement être intégrée au produit. Elle peut également servir à son développement. Aujourd'hui, de nombreuses applications complexes sont développées grâce à une approche fondée sur un modèle, qui consiste essentiellement à décrire le fonctionnement d'un produit à l'aide d'un logiciel et d'équations mathématiques issues de la physique. Cependant, même cette approche a ses limites. C'est là que Monolith et sa plateforme d'IA avec auto-apprentissage [6] entrent en jeu.

La plateforme est capable d'apprendre les propriétés physiques de systèmes complexes à partir de données déjà collectées. À titre d'exemple, les véhicules subissent une série de tests sur une piste d'essai, où de multiples capteurs surveillent le lacet et le roulis ainsi que la vitesse et l'accélération des roues. La collecte de données pour différentes rigidités de suspension donne un bon aperçu de la façon dont le véhicule réagit à une série de situations de conduite. En général, les données sont analysées, ce qui permet d'obtenir de nouveaux réglages à appliquer lors d'un autre essai. Monolith peut évaluer les données de la première série de tests et prédire le résultat des modifications de la suspension avec un haut degré de précision. Les résultats servent à affiner plus rapidement les meilleurs réglages de la suspension, réduisant ainsi le nombre d'autres essais physiques nécessaires.

Cette approche peut également s'appliquer à la métrologie. Les compteurs à gaz doivent être exceptionnellement précis pour garantir une facturation correcte, mais c'est difficile à réaliser lorsque le compteur doit effectuer des mesures pour différents gaz. Pour un client, la simulation de compteurs à ultrasons a poussé l'analyse purement mathématique à ses limites, laissant les processus d'essais consécutifs comme seule solution d'étalonnage pour obtenir la certification nécessaire. Fort heureusement, tous les essais ont donné lieu à une vaste collection de données à analyser. Grâce à l'utilisation de modèles d'IA avec apprentissage automatique, la quantité de tests nécessaires a diminué jusqu'à 70 %, ce qui a considérablement accéléré le développement.

Miniaturisation des capacités de calcul nécessaires à l'IA

Des concours comme le *DARPA Grand Challenge* [8], au cours duquel il s'agissait de construire des véhicules autonomes capables de suivre un parcours sinueux, ont déclenché une vague d'intérêt pour les voitures de ce type. Aujourd'hui, près de vingt ans plus tard, des budgets considérables ont été dépensés, avec des résultats mitigés. Tesla fait régulièrement les gros titres à ce sujet, souvent lorsque des propriétaires trop enthousiastes font preuve d'une confiance excessive dans les capacités de leur véhicule [9]. Par ailleurs, seul Waymo semble proposer de véritables véhicules à conduite autonome sous la forme de services de covoiturage [10], même s'ils ne sont proposés qu'à Phoenix et San Francisco aux États-Unis.

L'un des problèmes est l'extrême difficulté du pilotage d'une voiture par un ordinateur. Non seulement le véhicule doit évaluer en permanence la situation environnante, mais il doit également prévoir les actions des autres conducteurs et usagers, comme les piétons et les cyclistes, qui, le cas échéant, ne respectent pas le code de la route.

L'architecture électrique et électronique (E/E) des véhicules est en train d'évoluer pour répondre aux besoins futurs des véhicules autonomes. Grâce aux innombrables capteurs qui produisent de vastes quantités de données, l'industrie se tourne vers l'Ethernet automobile. Actuellement, cette approche est à l'origine des systèmes avancés d'aide à la conduite (ADAS) qui, en contrôlant le freinage, l'accélération et la direction, peuvent intervenir en cas d'erreur du conducteur. Selon les niveaux d'autonomie des véhicules définis par la *Society of Automotive Engineers* (SAE), les véhicules haut de gamme atteignent actuellement le niveau 2+, voire le niveau 3 pour certains. Toutefois, l'autonomie complète permettant au conducteur de « s'asseoir et se détendre » est de niveau 5, ce qui signifie que nous avons encore du chemin à parcourir.

Figure 5. Le développement d'algorithmes d'IA est accéléré par l'utilisation d'ordinateurs embarqués hautement performants, comme ce DYNACOR 40-36 à refroidissement liquide. (Source : Eurotech)

Des entreprises comme Eurotech soutiennent l'industrie pour accélérer le développement des algorithmes nécessaires. Actuellement, un essai routier de huit heures permet de recueillir 120 To de données qui doivent être renvoyées au laboratoire pour être traitées et analysées. Il est possible de tester en laboratoire les améliorations apportées aux algorithmes d'IA à l'aide des données recueillies, mais il existe peu de ressources pour faciliter les essais et le développement d'algorithmes sur le terrain.

Tirant parti de son expérience en matière de refroidissement liquide, Eurotech propose une série de matériels d'IA de pointe, essentiellement sous la forme de superordinateurs compacts qui tiennent dans le coffre d'un véhicule. Un ensemble durci comme le DYNACOR 40-36 peut s'installer dans des véhicules routiers ou non routiers [11]. Doté d'un CPU Intel Xeon à 16 cœurs avec 64 Go de RAM et jusqu'à deux GPU NVIDIA10 GV100 avec 32 Go de mémoire RAM, cet ordinateur sans ventilateur dispose de 237 TFLOPS pour traiter les applications d'apprentissage profond (**figure 5**). Plusieurs interfaces gigabit Ethernet permettent d'ingérer des quantités massives de données de capteurs, issues de différents dispositifs (radar, lidar, caméras, notamment), et de les stocker dans un système de stockage électronique offrant un volume de 32 To de stockage. En effectuant davantage de tests d'inférence et de renforcement au cours des essais routiers, la possibilité d'une autonomie de niveau 5 pourrait être considérablement accélérée.

Figure 6. Image produite par IA avec le système DALL-E 2 d'OpenAI d'un véhicule autonome doté de l'incontournable Commodore 64. Chantier en cours.

L'IA met-elle mon emploi en danger ?

Une discussion permanente se développe sur les médias sociaux autour de la question des risques qui pèseraient sur les emplois dans les industries créatives, du fait des progrès de l'IA. Le lancement de DALL-E 2 [12] par OpenAI transforme en images les demandes en langage naturel (**figure 6**). Mais, ce qui est peut-être plus impressionnant, c'est sa capacité à modifier des images existantes de manière réaliste. Le système peut, par exemple, supprimer des objets au premier plan ou en arrière-plan. Ainsi, à partir d'une œuvre du peintre néerlandais Vermeer, l'IA peut étendre le tableau « La Jeune Fille à la perle » pour y inclure une restitution crédible de la pièce où le modèle se trouvait lorsqu'il a été peint.

Cependant, les rédacteurs, notamment ceux de l'équipe de la rédaction d'Elektor et d'autres organes de presse renommés, ont été désarçonnés par le lancement de ChatGPT [13]. Cette IA peut interagir avec les utilisateurs de manière conversationnelle et dans une multitude de langues. Jusqu'à présent, les discussions sur les mérites des MOSFET au carbure de silicium (SiC) et des transistors au nitrure de gallium (GaN), et leurs avantages par rapport aux MOSFET en silicium, ont donné lieu à des réponses très précises. Les sujets les plus spécialisés semblent donc bien couverts.

Bien qu'il soit exceptionnellement intelligent, l'outil ne connaît cependant que les réponses aux sujets connus avant son entraînement. Comme il n'est pas en apprentissage continu, il ne sera pas à la pointe de l'information concernant les affaires courantes ou les derniers drames des groupes de K-pop (dommage). Autre problème : au bout d'un certain temps, les réponses semblent quelque peu figées et stéréotypées. Cependant, ceux qui cherchent l'inspiration ou un poème d'anniversaire comme le composerait une célébrité ne seront pas déçus.

Pour savoir si l'avenir d'Elektor passe par des êtres en chair et en os, ou plutôt par des ordinateurs, je vous propose un résumé du sujet de cet article écrit par ChatGPT. À la prochaine… ou pas !

En résumé, les systèmes embarqués et l'IA sont deux technologies de plus en plus utilisées conjointement pour créer des dispositifs et des systèmes intelligents et autonomes. Les systèmes embarqués fournissent la plateforme matérielle et logicielle nécessaire aux algorithmes d'IA, tandis que ces algorithmes permettent aux systèmes de détecter, d'analyser et de réagir à leur environnement de manière plus intelligente et plus humaine. Comme les capacités des systèmes embarqués et de l'IA continuent de s'améliorer, nous pouvons nous attendre à voir un large éventail de nouvelles applications passionnantes dans des domaines comme la robotique, les soins de santé, les transports, etc.

220673-04 — VF : Pascal Godart

Des questions, des commentaires ?

Envoyer un courriel à l'auteur (stuart.cording@elektor.com) ou contactez Elektor (redaction@elektor.fr).

LIENS

[1] Edge Impulse Partners : http://bit.ly/3vbqRhG
[2] Edge Impulse CLI Installation : http://bit.ly/3BQQt71
[3] Edge Impulse Ingestion SDK : http://bit.ly/3jplhp3
[4] Projet d'exemple de reconnaissance d'un mouvement continu : http://bit.ly/3WAV9G5
[5] EON Tuner : http://bit.ly/3PQhFsr
[6] SlateSafety BAND : http://bit.ly/3YVpe5x
[7] Monolith : http://bit.ly/3BWZImh
[8] DARPA Grand Challenge, Wikipedia : https://fr.wikipedia.org/wiki/DARPA_Grand_Challenge
[9] J. Stilgoe, « Tesla crash report blames human error - this is a missed opportunity », Guardian, janvier 2017 : http://bit.ly/3Vjp7gI
[10] DYNACOR 40-36 : http://bit.ly/3GdmgBS
[11] Waymo One : http://bit.ly/3FNG8Ku
[12] DALL·E 2 : http://bit.ly/3PNPWsD
[13] ChatGPT : http://bit.ly/3PLjZRo