2024 l'odyssée de l'IA

accélérateurs d'IA : une comparaison

Brian Tristam Williams (Elektor)

L'intelligence artificielle (IA) connaît une évolution rapide, rendant crucial le choix de l'accélérateur d'IA le plus adapté. Dans cet article, nous explorerons quelques-uns des principaux accélérateurs disponibles, en examinant en détail leurs caractéristiques et performances, pour vous guider vers une décision éclairée, adaptée à vos besoins spécifiques en IA.

Figure 1. Kit Raspberry Pi AI.

Pourquoi les accélérateurs d'IA sont-ils essentiels ? Bien que tout processeur ou microcontrôleur, étant Turing complet, puisse théoriquement réaliser les mêmes calculs qu'un accélérateur d'IA, cependant, ces unités polyvalents sont conçus pour gérer une grande variété de fonctions mais manquent souvent d'efficacité pour les calculs spécifiques à l'IA.

Un accélérateur d'IA est un composant matériel spécialisé conçu pour accélérer les tâches d'intelligence artificielle, notamment durant la phase d'inférence où les modèles prédisent à partir de nouvelles données. Ces accélérateurs se distinguent des processeurs classiques par leur aptitude à gérer efficacement des opérations matricielles à grande échelle et des traitements parallèles, cruciaux pour des applications telles que l'apprentissage automatique, la reconnaissance d'images ou le traitement du langage naturel. Ils sont donc nettement plus rapides et efficaces pour ces tâches que les processeurs classiques.

À mesure que l'IA évolue, l'informatique en périphérie, qui consiste à traiter des données directement sur les appareils plutôt que sur des serveurs cloud centralisés, devient cruciale. Cette évolution améliore la réactivité en temps réel, la confidentialité des données et l'efficacité générale. Pour répondre à ces besoins, divers accélérateurs d'IA en périphérie ont été conçus pour les systèmes embarqués et les PC.

Choisir l'outil adapté à vos besoins spécifiques peut s'avérer difficile, que vous recherchiez des solutions puissantes pour des tâches complexes ou des dispositifs légers pour des opérations en temps réel. Nous vous proposons ici une sélection des accélérateurs d'IA les plus performants, en détaillant leurs caractéristiques, leurs applications typiques et leurs performances.

Pourquoi cette comparaison ?

Je traite régulièrement de grandes quantités de données pour des applications telles que la transcription vocale (utilisant par exemple Whisper [1]) et l'édition ainsi que la classification automatique de vidéos d'archives. Pour ces tâches, les solutions pour PC dotées de GPU puissants tels que mon NVIDIA RTX 4070 sont indispensables pour leur puissance de calcul et leur flexibilité. En revanche, pour les tâches de détection locales en temps réel ou en milieu industriel, les solutions embarquées telles que le Raspberry Pi AI Kit ou la Coral Dev Board sont plus adaptées en raison de leur compacité et de leur faible consommation énergétique. Cette étude comparative vise à aider à choisir les outils les plus adaptés à chaque situation.

Accélérateurs d'IA intégrés

Kit Raspberry Pi AI avec Hailo-8L intégré

Récemment lancé, le kit Raspberry Pi AI [2] intègre l'accélérateur d'IA Hailo-8L au Raspberry Pi 5, offrant ainsi des capacités d'IA avancées dans un format compact et économique. Cette solution représente une option accessible pour intégrer l'intelligence artificielle à divers projets Raspberry Pi.

L'accélérateur d'IA Hailo-8L se connecte au HAT+ M.2 du Raspberry Pi. Comme illustré dans la **figure 1**, il offre une capacité de calcul 13 téra-opérations par seconde (TOPS), ce qui le rend adapté à des tâches telles que la détection d'objets en temps réel, la segmentation sémantique, l'estimation de la pose pour la reconnaissance des gestes, et le repérage facial. La consommation d'énergie réduite du kit est idéale pour les dispositifs alimentés par batterie et sa taille compacte est compatibles avec accessoires officiels de la caméra Raspberry Pi. Bien qu'il n'atteigne

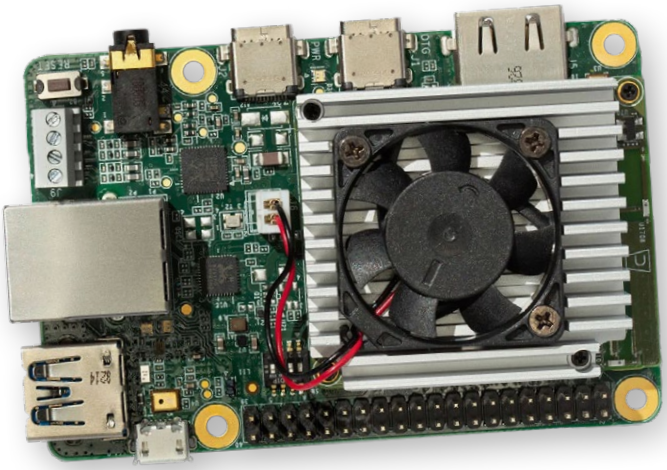


Figure 2. Carte de développement Coral. Source : coral.ai.

pas la puissance de certains NPU haut de gamme, il est parfaitement adapté aux tâches d'IA légères, aux projets éducatifs et aux applications amateurs.

Coral Dev Board

Le Coral Dev Board [3], lancé en 2019, est conçu pour exécuter efficacement les modèles TensorFlow Lite en tant que solution autonome. Présentée dans un format similaire à celui d'un Raspberry Pi (**figure 2**), il est doté d'un coprocesseur Edge TPU délivrant 4 TOPS. Il est efficace pour la classification d'images, la détection d'objets et la reconnaissance vocale, ce qui le rend idéal pour les applications IdO. Sa faible consommation d'énergie est également un atout majeur pour les systèmes embarqués. Bien qu'elle soit principalement conçue pour les modèles TensorFlow Lite et offre moins de capacité de calcul que d'autres accélérateurs, elle s'intègre parfaitement avec les services Google Cloud, offrant ainsi des performances solides pour des applications d'IA de base dans les systèmes embarqués.

Accélérateur Coral USB

Si vous avez déjà choisi votre plateforme et souhaitez y ajouter des capacités d'IA externe, considérez l'accélérateur USB Coral [4]. Ce dispositif compact, de la taille d'une clé USB standard (**figure 3**), intègre la puissance de l'Edge TPU de Google à tout ordinateur compatible via USB 3.0. Offrant 4 TOPS tout en consommant seulement 2 W, cet accélérateur est parfait pour les développeurs qui cherchent à ajouter des capacités d'IA aux systèmes existants ou à prototyper des applications d'IA avancées sans recourir à une carte de développement complète. Il excelle dans des tâches telles que la classification d'images et la détection d'objets, ce qui en fait un outil polyvalent pour l'expérimentation et les déploiements d'IA à petite échelle.

NVIDIA Jetson

La gamme Jetson de NVIDIA comprend une série de cartes de développement embarquées conçues pour les applications d'IA en périphérie. Voici une vue d'ensemble de ses principaux produits, chacun adapté à des besoins spécifiques en termes de performances et d'applications :

- Jetson Nano : L'option d'entrée de gamme, adaptée aux petits projets d'IA et au prototypage.
- Jetson Xavier NX : Une option de milieu de gamme offrant des performances significativement supérieures à celles du Nano.
- Jetson AGX Xavier : Un module haute performance destiné aux applications d'IA exigeantes.
- Jetson AGX Orin : Le modèle le plus récent et le plus puissant, spécialement conçu pour la robotique avancée et les machines autonomes.

Ces modules ont une puissance de calcul variable, allant de 0,5 TOPS pour le Nano à plus de 200 TOPS pour l'AGX Orin.
Tous les modules Jetson sont compatibles avec le SDK JetPack de NVIDIA, qui comprend des bibliothèques dédiées à l'apprentissage profond, la vision par ordinateur, le calcul accéléré et le traitement multimédia. Cet environnement logiciel courant permet de faciliter la mise à l'échelle et le déploiement au sein de toute la famille Jetson.
La gamme Jetson est bien adaptée à une multitude d'applications, notamment la robotique, les drones, l'analyse vidéo intelligente et les appareils médicaux portables. Son évolutivité en fait un choix polyvalent pour les entreprises qui cherchent à développer et à déployer des applications d'IA à divers niveaux de performance.
Le Jetson Nano, modèle d'entrée de gamme lancé en 2019 (**figure 4**), a rendu les capacités de calcul de l'IA accessible dans des formats petits et abordables. Équipé d'un GPU Maxwell de 128 cœurs, d'un CPU ARM Cortex-A57 quadricœur et de 4 Go de RAM, il offre une capacité de calcul de 472 GFLOPS (0,47 TOPS). Le Jetson Nano convient particulièrement aux tâches d'IA légères mais exigeantes dans les systèmes embarqués. Capable de gérer plusieurs réseaux neuronaux en parallèle, il se présente comme une

Figure 3. Accélérateur USB de Coral

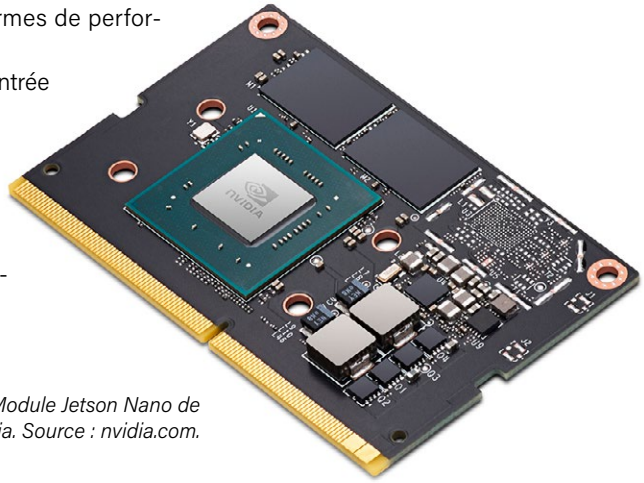


Figure 4. Module Jetson Nano de Nvidia. Source : nvidia.com.

Figure 5. Intel Neural Compute Stick 2 (Intel NCS2).

solution flexible pour diverses applications d'IA. Cependant, sa consommation d'énergie est plus élevée que celle des autres modules embarquées, et son facteur de forme est plus grand.

Le Nano est un module d'IA doté d'un connecteur à 260 broches, ce qui rend son utilisation autonome impossible sans connexion externe. Pour démarrer, il est nécessaire de se procurer un kit de développement, disponible dans l'encadré "Produits".

Pour découvrir les autres cartes de la série Jetson, n'hésitez pas à consulter la page de comparaison de NVIDIA à l'adresse indiquée [6].

Intel Neural Compute Stick 2 (NCS2)

L'Intel Neural Compute Stick 2 (NCS2), lancé en 2018, est un accélérateur d'IA portable qui se connecte via USB [7]. Doté du VPU Intel Movidius Myriad X, il offre une performance d'environ 1 TOPS. Il est idéal pour le prototypage d'IA, la recherche l'augmentation des capacités d'IA des PC existants. Sa portabilité et sa simplicité d'utilisation en font une option prisée par les développeurs et les chercheurs. Cependant, il nécessite un PC hôte pour l'alimentation et le traitement des données et sa capacité de calcul reste inférieure à celle des GPU de haut de gamme.

Accélérateurs d'IA pour PC

NVIDIA GPUs

Le NVIDIA RTX 4070 [8] de la série RTX 40 lancée en 2022 et constitue une référence pour les jeux et les applications professionnelles d'IA. Ce GPU haut de gamme, intégré dans les cartes graphiques de divers fournisseurs, offre 29,8 TFLOPS (29 800 GFLOPS) et 12 Go de mémoire GDDR6. Il est parfait pour l'entraînement et l'inférence de grands réseaux neuronaux, les jeux à haute performance et les tâches de création de contenu telles que l'édition vidéo et le rendu 3D. Sa prise en charge logicielle étendue et la puissance de calcul exceptionnelle en font l'outil idéal pour les applications d'IA exigeantes, notamment la transcription de la parole en texte et le montage vidéo. Pour ceux disposant d'un budget plus conséquent, la RTX 4090 offre encore plus de puissance, tandis que la RTX 3080 représente une alternative économique avec des performances toujours impressionnantes. Cependant, en raison de sa consommation électrique élevée et de son coût, il est plutôt destiné à équiper des stations de travail de bureau de taille importante.

La série RTX NVIDIA promue sous le slogan "the ultimate in ray tracing and AI", représente un produit haut de gamme. Un inconvénient à considérer est que lorsque vous vous absentez de votre station de travail, vous pourriez retrouver un enfant impatient à la maison, pressé de jouer au dernier jeu de tir à la première personne (FPS) bénéficiant du taux d'images par seconde (FPS) élevé que ces machines puissantes peuvent offrir.

Puces Apple de la série M

Les puces Apple de la série M représentent une avancée significative pour les utilisateurs de l'écosystème Apple. Introduite en 2020, la puce M1 a été la première conçue spécifiquement pour le Mac, suivie par la puce M2 en 2022. Chacune intègre directement l'accélération de l'IA dans le processeur : la M1 est équipée d'un moteur neuronal à 8 cœurs capable de délivrer 11 TOPS, tandis que la M2 propose un moteur neuronal à 16 cœurs offrant jusqu'à 15,8 TOPS. Ces puces sont intégrées dans des appareils tels que le MacBook Pro et le Mac Mini, permettant un traitement IA local pour des applications comme Siri, le traitement d'images et la réalité augmentée. L'intégration transparente avec macOS et iOS rend ces puces particulièrement performantes pour les applications d'IA destinées au grand public, surtout pour ceux déjà intégrés dans l'écosystème Apple.

Les M3, M3 Pro et M3 Max ont été annoncées en 2023 avec des améliorations sur toute la ligne. Les puces M3 étaient les premières puces à 3 nanomètres et étaient dotées d'une nouvelle architecture GPU qui permettait des améliorations de vitesse allant jusqu'à 30 % et 50 % par rapport à la M2 et à la M1, respectivement. Avec des performances CPU plus rapides, un moteur neuronal plus efficace et une architecture GPU de nouvelle génération qui prend en charge des fonctions avancées telles que la mise en cache dynamique et le ray tracing accéléré au niveau matériel, la série M3 est particulièrement bien adaptée aux tâches intensives telles que le traitement graphique haut de gamme, les charges de travail AI/ML et les tâches informatiques complexes.

La peinture était à peine sèche sur l'annonce de la M3 quand Apple a révélé la puce M4 en mai 2024, disponible exclusivement dans l'iPad Pro de la marque. Il serait fascinant de voir si un développeur tiers trouve une utilité pour un iPad Pro onéreux dans un contexte commercial ou industriel, notamment pour des tâches d'apprentissage automatique.

Pour ceux qui travaillent principalement avec des produits Apple, un MacBook équipé d'une puce M2 ou M3 [9] ne manquera pas de répondre à vos besoins en applications d'IA. Le choix de la puce dépendra de votre budget et des exigences spécifiques de votre application.

Cartes intégrées ou basées su PC : Que choisir ?

Le choix entre des solutions informatiques intégrées ou basées sur des PC dépend largement des exigences spécifiques de votre application.

IA embarquée en périphérie

- Idéale pour : les applications en temps réel à faible consommation d'énergie où la taille et la consommation d'énergie sont essentielles, telles que les appareils domotiques intelligents, les appareils portatifs et les capteurs autonomes.
- Avantages : consommation d'énergie réduite, taille compacte, rentabilité

IA en périphérie sur PC

- Idéal pour : les applications nécessitant une puissance de calcul et une flexibilité élevées, telles que les applications d'IA de bureau, l'analyse de données complexes, ainsi que le développement et le test de modèles d'IA.
- Avantages : performances élevées, plus grande flexibilité, capacité à gérer des tâches plus complexes.

Faire le choix

Les solutions IA embarquées et basées sur PC offrent des avantages uniques. Les accélérateurs intégrés conviennent idéalement aux applications en temps réel nécessitant une faible consommation d'énergie, tandis que les solutions basées sur les ordinateurs de bureau et portables, telles que les GPU NVIDIA RTX et les puces Apple de série M, assurent des performances supérieures pour les tâches plus complexes. En saisissant les atouts et les contraintes de chacune, vous pouvez choisir la solution la plus adaptée à vos exigences particulières, maximisant ainsi le potentiel de l'IA pour créer des systèmes plus intelligents et plus performants.

Cette comparaison n'est certes pas exhaustive, mais elle souligne les caractéristiques de quelques-uns des accélérateurs d'IA les plus populaires que j'ai testés. Chaque option présente des avantages et des inconvénients distincts. L'important est de sélectionner l'accélérateur qui s'adapte le mieux à vos besoins spécifiques. Pour des informations techniques plus détaillées et des comparaisons approfondies, veuillez consulter l'encadré **Liens**.

230181-H-04

À propos de l'auteur

Brian Tristam Williams s'est passionné pour les ordinateurs et l'électronique dès l'âge de dix ans, lorsqu'il a découvert son premier micro-ordinateur. Son parcours avec Elektor Magazine a débuté à seize ans, lorsqu'il a acquis son premier numéro. Depuis, il reste immergé dans le monde de l'électronique et de l'informatique, explorant et apprenant de manière continue. Il a rejoint Elektor en 2010. Il se consacre aujourd'hui à l'étude des dernières innovations technologiques, avec un intérêt particulier pour l'intelligence artificielle et les ordinateurs à carte unique comme le Raspberry Pi.

Questions ou commentaires ?

Envoyez un courriel à l'auteur (brian.williams@elektor.com).

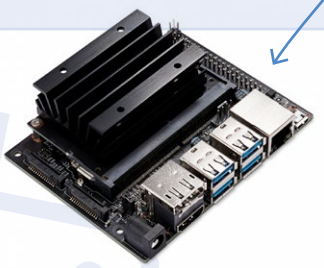
SUJET À LA UNE

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

www.elektormagazine.fr/**embarque-ia**

Produits

- **Raspberry Pi AI Kit**
  www.elektor.fr/20879
- **Google Coral USB Accelerator**
  www.elektor.fr/19366
- **Waveshare Jetson Nano Development Kit Lite**
  www.elektor.fr/20761



LIENS

[1] Whisper speech-to-text from OpenAI: https://openai.com/blog/whisper
[2] Raspberry Pi AI Kit: https://raspberrypi.com/products/ai-kit
[3] Coral Dev Board: https://coral.ai/products/dev-board
[4] Coral USB Accelerator: https://coral.ai/products/accelerator
[5] Apple M3 chip series:
https://apple.com/za/newsroom/2023/10/apple-unveils-m3-m3-pro-and-m3-max-the-most-advanced-chips-for-a-personal-computer
[6] Nvidia Jetson series of embedded compute: https://developer.nvidia.com/embedded/jetson-modules
[7] Intel Neural Compute Stick 2 (Intel NCS2): https://intel.com/content/www/us/en/developer/articles/tool/neural-compute-stick.html
[8] NVIDIA GeForce RTX 40 Series GPUs: https://nvidia.com/en-us/geforce/graphics-cards/40-series
[9] Apple Mac models, currently based on M2 and M3 chips: https://apple.com/mac/compare