

Actuellement, l'IA en périphérie est trop lente et trop énergivore. Voilà pourquoi les microcontrôleurs 32 bits ont besoin d'une sérieuse mise à niveau en matière d'IA

Par Camron Kazerounian et Reza Kazerounian

Alif Semiconductor

Les dernières prouesses de l'intelligence artificielle (IA) permises par des systèmes informatiques cloud de taille gigantesque continuent de surprendre les consommateurs lambda et les informaticiens. L'IA peut par exemple rédiger un rapport d'activité, créer un shooting mode pour une enseigne de vêtements ou découvrir une nouvelle molécule. C'est désormais un simple fait de la vie quotidienne, alors qu'il y a encore peu de temps, en 2020, cela aurait ressemblé à un fantasme de science-fiction.

Le potentiel de transformation des capacités informatiques est aussi important pour les fabricants d'appareils embarqués que pour les consommateurs qui utilisent aujourd'hui des services d'IA cloud. Imaginez des appareils capables de répondre en temps réel à votre voix, de prédire quand vos machines ont besoin d'un entretien avant une panne, ou d'analyser votre environnement instantanément, le tout sans envoyer de données dans le cloud.

Les fabricants qui sauront s'adapter pour tirer parti des progrès rapides des technologies IA afin d'offrir de nouvelles expériences convaincantes aux utilisateurs dans les applications de pointe de l'IA, telles que les appareils portables, remporteront l'adhésion des consommateurs.

Cependant, pour concrétiser les visions ambitieuses de l'IA actuellement poursuivies par les leaders de l'industrie des appareils grand public, le monde des appareils embarqués devra apprendre à réduire la consommation d'énergie, le coût et la taille du matériel et des logiciels IA pour s'adapter à des ressources très limitées.

Certains types de microcontrôleurs (MCU) semblent assurés d'occuper une place centrale dans les appareils embarqués dans cette nouvelle ère fascinante de l'intelligence artificielle embarquée. Toutefois, l'industrie des MCU doit d'abord revoir radicalement l'architecture de ses produits afin d'offrir les performances dont les nouveaux appareils dotés d'IA auront besoin, tout en restant dans les limites restreintes des futurs produits sur batterie.

Pourquoi l'IA est-elle essentielle en périphérie ?

La quantité de données générées par les appareils en périphérie augmente de façon exponentielle, ce qui entraîne un besoin urgent de traitement IA sur l'appareil. La sphère mondiale des données, qui comprend toutes les données créées dans le monde, devrait grimper en flèche, atteignant 150 zettaoctets (Zo) d'ici 2025 et plus que quadrupler pour atteindre près de 600 Zo d'ici 2030. Avec l'arrivée en ligne de milliards d'appareils (des appareils portables aux capteurs IoT industriels), le volume massif de données va submerger les infrastructures cloud traditionnelles.

Plus des données sont produites, plus l'IA et l'apprentissage automatique (ML) jouent un rôle essentiel dans l'interprétation des données pour produire des informations pertinentes - et, en retour, plus de données sont générées sur la base de ces informations. Sans oublier qu'un sous-ensemble de données provenant de la périphérie est utilisé pour former des modèles améliorés de génération de données. Cette boucle de rétroaction entre les données est un facteur important de l'explosion exponentielle de la sphère des données.

Cet immense volume de données produites et analysées en périphérie bouleverse désormais la façon dont les systèmes embarqués traitent les données, ce qui nécessite un traitement local en périphérie.

Quatre raisons d'effectuer l'inférence en IA en périphérie

Les systèmes IA d'entreprise et de smartphone exécutent généralement de nombreux processus IA dans le cloud. Cependant, pour les appareils embarqués dont les ressources sont limitées, le traitement local offre quatre grands avantages :

- **Temps de latence réduit pour une réponse en temps réel** : Les applications telles que la maintenance prédictive ou la surveillance de la santé ne peuvent pas se permettre la latence liée au cloud. À titre d'exemple, un dispositif médical détectant des rythmes cardiaques anormaux requiert une intervention immédiate, ce qui est possible avec le traitement local.
- **Efficacité de la bande passante** : La transmission constante d'immenses volumes de données vers le cloud mettrait les réseaux à rude épreuve et serait trop coûteuse, en particulier à mesure que le volume de données augmente. Le traitement local réduit la demande de bande passante en traitant les données sur l'appareil.

- **Confidentialité et sécurité** : Certaines applications, telles que le suivi ou la surveillance de la santé, imposent de conserver les données sensibles sur l'appareil. Le fonctionnement local de l'IA permet de traiter les données uniquement sur l'appareil, ce qui renforce la protection de la vie privée.
- **Efficacité énergétique** : La communication de données entre l'appareil et le cloud consomme d'importantes ressources énergétiques qui épuisent la batterie de l'appareil en périphérie. Le traitement en périphérie permet d'économiser l'énergie ; les appareils sur batterie peuvent donc fonctionner bien plus longtemps sur une charge.

Ces avantages combinés expliquent l'adoption rapide de l'IA en périphérie dans tous les secteurs d'activité. Dans l'électronique grand public, elle permet la traduction en temps réel des langues sur les appareils portables. Dans l'IoT industriel, elle permet aux capteurs de surveiller l'état des machines et de signaler les besoins de maintenance avant une panne. Dans le domaine de la santé, les dispositifs portables surveillent les signes vitaux au niveau local, transformant les soins réactifs en informations proactives (voir Figure 1).



Fig. 1 : Une montre connectée peut utiliser l'IA en périphérie pour fournir des informations médicales à partir des données relatives aux signes vitaux. (Crédit image : Jens Mahnke)

Lorsqu'il s'agit d'enjeux cruciaux ou d'applications en temps réel, le traitement IA local n'est pas qu'une amélioration, c'est une nécessité. Les applications telles que la surveillance de la santé,

la maintenance prédictive et les appareils interactifs grand public ne peuvent pas atteindre leurs pleines capacités si elles reposent uniquement sur le traitement cloud.

À titre d'exemple, les dispositifs de santé portables qui détectent les arythmies cardiaques ou les changements soudains des signes vitaux doivent analyser les données instantanément pour alerter les utilisateurs des risques potentiels pour la santé en temps réel. Tout retard dans le traitement cloud peut faire la différence entre une intervention précoce et une alerte manquée. De même, les capteurs IoT industriels utilisés pour la maintenance prédictive des équipements d'usine doivent détecter les anomalies immédiatement afin d'éviter des arrêts coûteux et des dommages aux équipements. L'envoi de données vers le cloud ajouterait un niveau de latence qui pourrait compromettre cette approche proactive.

Dans les applications grand public, les lunettes de réalité augmentée sont un autre exemple. Pour que ces appareils puissent fournir des informations transparentes et à la demande sur l'environnement, comme l'identification de points de repère ou la traduction de texte en temps réel, ils ont besoin d'une capacité de traitement sur l'appareil, sans dépendre du cloud, afin de garantir des interactions rapides et naturelles.

Ces exemples montrent en quoi l'IA locale n'est pas un atout, mais une capacité essentielle pour les appareils en périphérie. Le traitement en périphérie est le seul moyen de répondre aux exigences de traitement des données en temps réel, sécurisé et économe en énergie, entraînant un changement majeur dans la façon dont nous envisageons le traitement des données.

Potentiel de l'IA en périphérie - applications clés et types de modèles

Le désir d'intégrer ces innovations en périphérie a conduit à un large développement des optimisations de modèles d'apprentissage automatique (ML). Des techniques telles que l'élagage, la compression et la quantification des modèles permettent de réduire la taille et l'empreinte informatique des modèles ML tout en préservant leur précision. Des modèles dédiés sont conçus pour répondre aux besoins d'applications spécifiques dans différents secteurs d'activité.

- **Classification des images, segmentation et détection d'objet** : Les réseaux de neurones convolutionnels (CNN) excellent dans l'analyse des données visuelles, ce qui

les rend idéaux pour des applications telles que la reconnaissance faciale dans les appareils domestiques connectés ou la détection d'objets dans les lunettes de réalité augmentée, ainsi que la reconnaissance des gestes et des postures.

- **IA générative et traitement du langage naturel (NLP)** : Les transformateurs, initialement conçus pour des tâches linguistiques à grande échelle, sont désormais convertis en modèles linguistiques légers et spécialisés (SLM) pouvant fonctionner directement sur les appareils en périphérie. Ces SLM alimentent les commandes vocales, la traduction en temps réel et les interfaces commandées par la voix, permettant des interactions plus intuitives et fluides avec les appareils.
- **Maintenance prédictive et détection d'anomalies** : Les réseaux de neurones récurrents (RNN) et les modèles de mémoire à long terme et à court terme (LSTM) excellent dans le traitement des données temporelles. Les machines industrielles et les véhicules sont équipés de capteurs en périphérie qui surveillent les données de vibration, de température et de pression. Ces modèles peuvent utiliser ces données pour détecter les anomalies et alerter les opérateurs sur les besoins de maintenance avant que des pannes ne se produisent (voir Figure 2).



Fig. 2 : L'IA en périphérie prend en charge la maintenance prédictive pour réduire les interruptions et accélérer les réparations des équipements industriels.

Certaines capacités présentes dans les appareils connectés au cloud sont hors de portée des appareils embarqués - mais les fabricants sont impatients de les mettre en œuvre. C'est notamment le cas des Grands modèles de langage (LLM) tels que ChatGPT. Les LLM suscitent un intérêt pour le traitement du langage naturel et dévoilent un potentiel pour de nouvelles interfaces vocales intuitives en périphérie : Les MCU des appareils en périphérie devront tôt ou tard prendre en charge cette technologie.

S'agissant des consommateurs, les fabricants voient un large potentiel pour l'IA dans des environnements auparavant inimaginables pour le traitement de données IA, tels que les appareils portables. Les fabricants examinent le potentiel du marché des lunettes connectées, par exemple, dans lequel la mise en œuvre de l'IA pourrait tout changer. Elle peut proposer au porteur des « applications incontournables » telles que la traduction de la parole ou du texte en temps réel, l'interprétation vocale en langage naturel, des fonctions de recherche améliorées, ou une assistance IA contextuelle liée aux objets dans le champ de vision.

La question est donc : Comment les fabricants d'appareils embarqués peuvent-ils augmenter les performances de leur prochaine génération de produits pour répondre aux besoins de la technologie IA future, sans augmenter en même temps leur consommation d'énergie et leur coût ?

Quelles sont les implications pour la conception des MCU ? S'adapter aux exigences de l'IA en périphérie

Aujourd'hui, l'expérience des fabricants d'appareils embarqués en périphérie montre que les capacités des MCU ne correspondent pas aux exigences même des modèles ML actuels, sans parler de leur évolution future. Le potentiel de valeur ajoutée, en particulier pour les appareils audio/visuels portables, reste inexploité car les systèmes sur puce (SoC) actuellement sur le marché consomment trop d'énergie, sont trop volumineux et manquent de ressources de calcul cruciales.

Cela ne veut pas dire qu'il est impossible d'implémenter l'IA sur les MCU en périphérie actuels : Les MCU optimisés pour les fonctionnalités IA, y compris les produits de la famille Ensemble et Balletto d'Alif Semiconductor, permettent des avancées précieuses dans le déploiement du traitement de l'analyse vocale, d'image et de mouvement, en particulier à l'aide des CNN (voir Figure 3).

Cependant, les nouveaux modèles SLM basés sur transformateurs et les modèles émergents, y compris les réseaux de neurones graphiques (GNN) et les réseaux de neurones à impulsions (SNN), déverrouilleront une large palette de nouvelles capacités, qui sont actuellement hors de portée des MCU orientés IA actuels.



Fig. 3 : Le MCU sans fil Balletto supporte des fonctions IA sophistiquées dans les appareils portables.

La prochaine génération de MCU devra améliorer les capacités matérielles dans trois domaines. Tout d'abord, ils devront intégrer des capacités de calcul spécialisées pour les réseaux neuronaux sous la forme d'une unité de traitement neuronal (NPU). Les futurs appareils embarqués devront offrir des performances de plus de 1 TOPS tout en consommant peu d'énergie, de sorte qu'ils pourront être déployés dans des produits tels que des écouteurs sans fil avec de très petites batteries, tout en permettant une utilisation tout au long de la journée sans recharge.

Deuxièmement, les nouveaux systèmes IA en périphérie exécutant des modèles tels que les SLM auront des exigences de mémoire beaucoup plus importantes, sur puce et hors puce. Cela suggère que les fabricants de MCU devront introduire des interfaces plus rapides et à faible consommation d'énergie entre le MCU et les mémoires externes. L'architecture du MCU de demain devra également évoluer pour fournir des chemins suffisamment rapides pour supporter le fonctionnement de mémoires plus grandes et plus rapides.

Enfin, la tendance familière d'intégrer de plus en plus de fonctionnalités dans le MCU doit être intensifiée pour permettre la mise en œuvre de systèmes IA bardés de capteurs dans un petit espace de nœud en périphérie.

Un nouvel avenir : L'informatique embarquée dans la vie quotidienne

Le potentiel des MCU ne se limite pas à accroître sa présence dans une ou deux nouvelles catégories d'appareils en périphérie : Nous assistons potentiellement à un changement majeur dans la demande de capacité de calcul, passant de hubs alimentés par CPU et GPU à un monde où l'interaction de la plupart des gens avec le domaine numérique se fera via un MCU.

Dans ce nouvel environnement de calcul plus distribué, les fonctions pilotées par l'IA seront mises en œuvre de manière fluide et avec une faible latence en périphérie, dans des appareils tels que des lunettes connectées ou des écouteurs avec audio amélioré. Ils offriront une expérience utilisateur plus naturelle dans laquelle la technologie est intégrée dans la vie de l'utilisateur, plutôt que de constamment attirer leur attention sur leur smartphone ou leur PC - un monde où les appareils embarqués occupent une place centrale.

L'impact potentiel est immense : À mesure que de plus en plus de tâches sont déléguées à des dispositifs connectés embarqués dans l'environnement et capables de communiquer entre eux - qu'il s'agisse de surveillance de la santé, d'optimisation de l'énergie, de sécurité ou d'autres applications - notre dépendance aux applications pour smartphone et aux services Internet sur PC diminuera. Cette intelligence diffuse et décentralisée remet en question le monopole actuel des appareils et services centralisés. Elle nous permettrait d'interagir avec la technologie de manière plus fluide, personnalisée et moins intrusive, améliorant ainsi notre qualité de vie personnelle et professionnelle.

Au cœur de ce monde, il y aura une nouvelle génération de MCU basée sur des architectures optimisées pour les nouvelles technologies de l'IA et de l'apprentissage automatique.

[FIN]