Réseaux neuronaux: le MIT dégaine une puce ultra-rapide avec une consommation réduite

Des chercheurs du Massachusetts Institute of Technology (**MIT**) ont développé une puce capable de traiter les calculs de **réseaux neuronaux** (« neural networks ») de trois à sept fois plus rapidement que les puces actuelles.

De surcroît, cet exploit est réalisé tout en réduisant la consommation d'énergie jusqu'à 95 %.

La puce est prometteuse pour les futurs smartphones ou les objets connectés.

On parle là d'edge computing, soit la possibilité pour les appareils de traiter les tâches localement plutôt que d'avoir recours systématiquement au cloud.

Actuellement, la plupart des applications pour smartphones qui utilisent des réseaux neuronaux déversent tout simplement les données à traiter dans le cloud. Les résultats qui en découlent sont ensuite envoyés au smartphone.

L'avantage d'une telle méthode est de permettre aux terminaux mobiles de solliciter d'énormes réseaux neuronaux, sans consommer d'énergie (exception faite de l'envoi des données et de la récupération des résultats).

En contrepartie, avec une telle méthode, des données potentiellement sensibles sont envoyées à un tiers et il y a de la latence suivant la connexion à internet.

Dans une zone sans réseau cellulaire, il ne sera par ailleurs pas du tout possible d'exploiter des réseaux neuronaux de cette manière.

Avishek Biswas, étudiant diplômé du MIT en ingénierie électrique et informatique qui a dirigé le développement de la nouvelle puce, et son directeur de thèse Anantha Chandrakasan (doyen de la School of Engineering du MIT) décrivent la nouvelle puce dans un article présenté à la conférence « International Solid-State Circuits Conference » (ISSCC 2018).

« Voici le modèle de processeur général : il y a une mémoire dans une partie de la puce et il y a un processeur dans une autre partie de la puce et vous déplacez les données entre eux quand vous faites ces calculs », explique Avishek Biswas, sur une news disponible sur le site Internet du MIT.

La technique mise en oeuvre consiste à implémenter le produit scalaire (produit de deux vecteurs) directement dans la mémoire pour éliminer de nombreux allers-retours entre le CPU et la mémoire.

« *Puisque ces algorithmes d'apprentissage automatique ont besoin de tant de calculs, ce transfert continue des données est la partie dominante de la consommation d'énergie. Mais le calcul issu des algorithmes peut être simplifié pour une opération spécifique, appelée le produit scalaire. Notre approche était la suivante :*

pouvons-nous implémenter cette fonctionnalité de produit scalaire à l'intérieur de la mémoire afin qu'il ne soit pas utile de transférer ces données continuellement ? »

Plusieurs autres grands acteurs de la sphère IT proposent actuellement des solutions implémentées dans les SoC (System on Chip) destinés aux smartphones.

La puce mobile Kirin 970 signée HiSilicon (filiale de Huawei), dévoilée en septembre, intègre ainsi une unité de traitement neural dédiée, tout comme le SoC Exynos 9 Series 9810 de Samsung.

Le dernier SoC A11 Bionic d'Apple dispose également d'une puce IA dédiée.

Amazon plancherait aussi sur sa propre solution pour rendre les appareils exploitant son assistant vocal Alexa plus réactifs.

(Neurone © Fedorov Oleksiy – Shutterstock)