ARM propose des IP pour un traitement IA en mode Edge Computing

Avec la multiplication des appareils électroniques dans le sillage de l'**Internet des objets** (IoT) et de la mobilité, l'intelligence artificielle nécessite de plus en plus un traitement en local pour gagner en réactivité. D'où le recours à une approche Edge Computing à laquelle le Project Trillium d'ARM s'attaque.

Il s'agit de plusieurs coeurs de propriété intellectuelle (IP) signés ARM. Dont des processeurs capables de gérer des tâches relevant de l'intelligence artificielle : Machine Learning (apprentissage automatique en français) et Neural Networks (réseaux neuronaux).

Si les IP d'ARM exploités en Machine Learning ciblent initialement les puces mobiles, ils pourront intégrer d'autres types d'appareils en jouant sur le levier des performances, au prix d'une consommation électrique revue à la baisse ou à la hausse.

Avec la blockchain, le Machine Learning est l'un des sujets les plus en vue sur la scène IT.

Les groupes qui comptent dans les télécoms et le numérique prennent position progressivement sur le sujet.

Apple a intégré le Neural Engine dans son SoC qui équipe les iPhone X et 8 (A11 Bionic). Son concurrent Huawei a doté sa puce mobile Kirin 970 du NPU (Neural Processing Unit) dans une même volonté de gérer des tâches IA en local sur le terminal mobile plutôt que de systématiquement passer par les datacenters distants.

Qualcomm dispose aussi de sa solution intégrée dans le DSP* 685 inclu dans son SoC Snapdragon 845. Amazon plancherait également sur sa propre puce pour accroître la réactivité de sa technologie d'assistance à commandes vocales Alexa sur les enceintes connectées pour foyers numériques.

Jusqu'à présent, ARM s'était illustré par sa quasi-absence dans ce domaine. La firme britannique s'était contentée d'extensions à son ISA (Instruction Set Architecture) CPU au gré d'Armv8.2-A, avec l'introduction d'instructions spécialisées qui simplifient et accélèrent les implémentations des réseaux neuronaux.

L'annonce faite aujourd'hui par ARM couvre les processeurs ML ainsi que des processeurs de détection d'objects (OD pour « Object Detection »). Pour cette dernière IP, une jonction évidente peut être faite avec l'acquisition d'Apical par ARM en 2016.

Le CPU ML présente une performance de plus de 4,6 billions (mille milliards) d'opérations par seconde (TOPs), avec une efficacité énergétique de plus de trois billions d'opérations par seconde et par watt (TOPs/W). L'enveloppe thermique est de 1,5 watt. Ces chiffres sont estimés sur la base d'une implémentation de l'IP dans un process 7 nm (nanomètres).

Le processeur ARM OD a, lui, été spécialement conçu pour identifier efficacement les personnes ou

des objets. Il peut traiter un flux vidéo en temps réel en Full HD à 60 ips (images par seconde). C'est jusqu'à 80 fois les performances d'un DSP exploité pour un même usage.

Mais le Projet Trillium intègre également une dimension logicielle avec des solutions mises à disposition des développeurs.

Ces dernières sont disponibles via Github et le site Web Arm's Developer. Exploité avec la bibliothèque ARM Compute et CMSIS-NN (kernels de réseaux neuronaux signés ARM), le logiciel ARM NN (Neural Network) permet par ailleurs de créer des jonctions entre les frameworks NN tels que TensorFlow, Caffe et Android NN et la gamme complète des processeurs à coeurs Arm Cortex, les GPU Arm Mali et donc les processeurs ML.

Le processeur OD sera disponible pour les partenaires d'ARM dans le courant du premier trimestre 2018 tandis que le processeur ML devrait être prêt à la mi-2018.

Du silicium basé sur ces nouvelles IP ne devraient logiquement pas voir le jour avant 2019.

« *L'accélération rapide de l'intelligence artificielle dans les dispositifs de bord impose des exigences accrues d'innovation pour traiter les calculs, tout en maintenant une empreinte efficace sur le plan énergétique. Pour répondre à cette demande, ARM annonce sa nouvelle plateforme de ML, le projet Trillium* », explique Rene Haas, Président du groupe IP Products au sein d'ARM dans un billet de blog.

«*Les nouveaux périphériques auront besoin des capacités ML et AI de haute performance de ces nouveaux processeurs. Grâce à la combinaison du haut degré de flexibilité et d'évolutivité que notre plate-forme fournit, nos partenaires peuvent repousser les limites de ce qui sera possible à travers un large éventail d'appareils.* »

*DSP ou Digital Signal Processor (processeur de traitement du signal)

(Crédit photo : @ARM)