Puce IA haute performance : Baidu dévoile Kunlun

Baidu a profité de sa conférence annuelle Baidu Create pour annoncer la toute première puce IA chinoise.

Baptisée Kunlun, la puce a été est conçue pour gérer les modèles AI pour l'edge computing sur les périphériques et dans le cloud via des data centers.

260 TOPS

Kunlun est en fait décliné en deux moutures : la version 818-300 sera utilisée pour l'entraînement des agents IA et la 818-100 pour l'inférence.

La puce Kunlun ne constitue pas le premier pas du géant chinois dans le développement de puces pour l'IA.

Dès 2011, Baidu avait en effet commencé à mettre au point un accélérateur matériel pour l'IA basé sur une puce de type FPGA (Field-Programmable Gate Array) et destiné à l'apprentissage en profondeur (deep learning). Parallèlement, le groupe avait aussi commencé à exploiter des GPU pour les centres de données.

Mais, Kunlun permet à Baidu de faire un bond en termes de performances. La puce est en effet près de 30 fois plus véloce que son accélérateur FPGA. A noter que pour son projet Brainwave, Microsoft exploite également des puces de type FPGA pour le traitement rapide de l'IA dans le cloud.

Gravée dans une technologie Samsung de 14 nm (nanomètres), sa bande passante mémoire est de 512 Go / s et ses performances de 260 TOPS (Tera Operations Per Second), le tout avec une consommation de 100 watts.

D'autres annonces

En plus de supporter les algorithmes communs d'apprentissage en profondeur, la puce Kunlun peut également prendre en charge une grande variété d'applications IA, telles que la reconnaissance vocale, le classement par recherche, le traitement du langage naturel, la conduite autonome et les recommandations à grande échelle.

Baidu a également annoncé une mise à jour de sa suite de services IA, avec le lancement de Baidu Brain 3.0. La plate-forme peut désormais fournir une simple formation des modèles IA par glisser-déposer. Elle offre actuellement 110 services IA, allant de la vision par ordinateur au traitement du langage naturel en passant par le logiciel de reconnaissance faciale.

Parallèlement, le groupe dirigé par Robin Li a également accéléré la production de son bus entièrement autonome Apolong, développé en partenariat avec le fabricant local King Long. Les

premiers mini-bus de 14 places devraient être livrés au Japon d'ici début 2019 en partenariat avec une filiale de SoftBank.

La Chine est à l'avant garde en matière d'IA. Plus globalement, le gouvernement a fait du développement de puces en interne une priorité.

Avec son plan «Made in China 2025», ce dernier escompte que les puces « domestiques » représentent 40 % de ses besoins.

(Crédit photo : @Baidu)