Stage de Master M2 –
Réseaux de neurones dynamiques pour des systèmes embarqués

Contexte :
L'intégration de l'intelligence artificielle (IA) dans les systèmes temps réel embarqués pose de nombreux défis. L'un de ceux-ci est la très grande quantité de données traitées par ces algorithmes qui impacte leur temps d'exécution sur des plateformes avec des ressources limitées (calcul et mémoire). Pour répondre à ce problème, des approches logicielles et matérielles sont de plus en plus adoptées. D'un côté, les accélérateurs matériels (p.ex. FPGA ou GPU) permettent d'optimiser le calcul matriciel et traiter en parallèle le calcul divisé en plusieurs sous-tâches. De l'autre côté, les techniques logicielles (par ex. la compression), permettent réduire la taille du réseau. L'objectif de ce stage est de combiner ces deux techniques orthogonales, permettant l'exécution des modèles réduits de réseaux de neurones, afin de trouver un compromis satisfaisant entre le temps de calcul, le bilan énergétique et la précision du résultat sur les unités de traitement de tenseur (Tensor Processing Unit - TPU).
Les TPUs, intégrant des circuits dédiés à la multiplication matricielle en forme d'un réseau systolique, sont de plus en plus appliqués dans l'embarqué grâce à leur faible consommation énergétique et accélération significative du calcul (https://coral.ai/docs/edgetpu/benchmarks). Nous avons effectué des tests préliminaires visant à mesurer les temps d'exécution des divers modèles de réseaux de neurones sur une carte avec huit TPU intégrés. Les couches d'un réseau de neurones peuvent s'exécuter concurremment sur plusieurs TPU ce qui permet d'améliorer le débit de traitement. Il est également possible de stocker les paramètres du modèle dans les mémoires internes de plusieurs TPU (celle-ci est limitée à 8 MO pour chaque TPU) et ainsi réduire le nombre d'accès à la mémoire principale. Les tests ont démontré que la profondeur du pipeline (le nombre de TPU utilisés) a un impact très important sur les temps d'inférence des réseaux de grande et moyenne taille. En outre, il a été observé que le changement du modèle de réseau de neurones stocké sur un TPU prend un temps considérable. Nous avons proposé une technique d'ordonnancement temps réel non-préemptive pour ce type d'architecture [1]. Cependant, la politique d'ordonnancement proposée ignore la consommation énergétique et ne tire pas de profit des techniques de compression de réseaux de neurones.

Objectifs :
Il existe de nombreuses techniques pour réduire la complexité de réseaux de neurones [3,4], pendant ou après leur apprentissage (par ex. quantification, l'élagage, la distillation). Nous allons nous concentrer sur des réseaux de neurones convolutifs (Convolutional Neural Networks, CNN) et étudier deux approches complémentaires dans ce stage : 1) l'élagage structuré (structured pruning) qui enlève des neurones ou noyaux de convolution moins utiles selon un certain critère [2]. 2) des architectures dites "early exit" [5] qui, lors de l'inférence, exécutent juste un certain nombre de couches selon si l'exemple en entrée est plus facile ou plus complexe à traiter. Dans les deux cas, nous allons développer des méthodes innovantes pour restructurer l'architecture du réseau de neurones en sous-parties (parfois indépendants) ce qui permettra de moduler leur exécution selon un compromis de précision et consommation d'énergie. Dans un deuxième temps, ce compromis pourrait être choisi de manière dynamique, par exemple par rapport à des conditions externes (faible batterie, pic de charge) et avec une stratégie d'ordonancement qui sera adaptée selon ces conditions.
Les travaux visent à permettre l'intégration d'algorithmes de l'IA pour les applications nécessitant un haut degré de fiabilité (contraintes temporelles et qualité des résultats) sur des systèmes présentant des contraintes d'énergie.
Les aspects qui concernent du matériel et de l'expérimentation sur TPU seront traités par un deuxième stage au LAAS.

Environnement :
Le stage se déroulera au laboratoire LIRIS à l'INSA de Lyon, et encadré par Stefan Duffner (LIRIS) et Tomasz Kloda (LAAS, Toulouse). Durée : 6 mois. Gratification : 4.05 €/heure
Candidature : envoyer CV, bulletins de notes et lettre de motivation à stefan.duffner@liris.cnrs.fr

Références

[1] Binqi Sun, Tomasz Kloda, Jiyang Chen, and Cen Lu, and Marco Caccamo. Schedulability Analysis of Non-preemptive Sporadic Gang Tasks on Hardware Accelerators. In IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), 2023

[2] Anthony Berthelier, Yongzhe Yan, Thierry Chateau, Christophe Blanc, Stefan Duffner, and Christophe Garcia. Learning Sparse Filters In Deep Convolutional Neural Networks With A l1/l2 Pseudo-Norm. In CADL : WS on Computational Aspects of Deep Learning - ICPR, 2020

[3] Anthony Berthelier, Thierry Chateau, Stefan Duffner, Christophe Garcia, and Christophe Blanc. Deep Model Compression and Architecture Optimization for Embedded Systems : A Survey. Journal of Signal Processing Systems, 2020

[4] T. Liang et al., “Pruning and quantization for deep neural network acceleration: A survey,” Neurocomputing, 2021

[5] S. Teerapittayanon et al., “BranchyNet: Fast inference via early exiting from deep neural networks,” in 2016 23rd International, Conference on Pattern Recognition (ICPR), 2016