Stage M1

Implémentation efficace des réseaux de neurones sur des architectures à virgule fixe

Lieu de stage	Laboratoire LAMPS, université de Perpignan
Durée	6 à 8 semaines
Encadrante	Dorra Ben Khalifa
Date de la publication	02/02/2023
Date de début prévue	01/04/2023
Lieu	52 Avenue Paul Alduy , 66100 Perpignan
Contact	dorra.ben-khalifa@univ-perp.fr

De nos jours, les réseaux de neurones profonds (DNN) font l'objet d'une grande attention dans diverses applications telles que la reconnaissance visuelle, les voitures à conduite autonome, la santé, etc [**GMDTCV18**]. Afin de rendre les DNNs utilisables sur des appareils plus petits ou mobiles, il est nécessaire de réduire les besoins en calcul, en énergie et en stockage de ces réseaux. En raison du nombre extrêmement élevé d'opérations et de valeurs à stocker, une telle réduction des besoins en ressources est nécessaire pour faire fonctionner les DNNs sur des dispositifs de petite taille ou alimentés par batterie, tels que les FPGA ou les téléphones portables.

Ce stage s'intéresse à la mise en œuvre des réseaux de neurones dans des systèmes informatiques à virgule fixe [**MNR14**]. Plus précisément, nous proposons une approche statique pour accélérer les réseaux de neurones en déplaçant le calcul du logiciel au matériel et en utilisant le calcul en virgule fixe au lieu de la virgule flottante à l'aide de notre outil POPiX.

POPiX [**BBBM22**] est un outil de réglage de la précision basé sur l'analyse statique. Il calcule les formats minimaux (le nombre de bits) nécessaires pour effectuer des calculs en arithmétique à virgule fixe sur les réseaux de neurones d'entrée tout en garantissant que les réseaux de sortie répondent aux exigences en précision souhaitées par l'utilisateur. À partir des formats renvoyés par POPiX, nous pouvons dériver des formats à virgule fixe qui optimisent la consommation de mémoire et respectent le seuil de précision défini par l'utilisateur. En se basant sur le code à virgule fixe généré par notre outil POPiX, le stagiaire devra réaliser le portage des codes à virgules fixe vers une carte NUCLEO-F207ZG, STM32 Nucleo-144 et réaliser une étude expérimentale de performance. Aussi, la dernière partie de ce stage est consacrée à comparer les résultats expérimentaux obtenus avec les travaux de l'état de l'art [**BM22**, **LLSC18**].

Références

[**MNR14**] Matthieu Martel and Amine Najahi and Guillaume Revy. *Toward the synthesis of fixed-point code for matrix inversion based on Cholesky decomposition*, Proceedings of the 2014 Conference on Design and Architectures for Signal and Image Processing, DASIP 2014,IEEE.

[**GMDTCV18**] Timon Gehr and Matthew Mirman and Dana Drachsler-Cohen and Petar Tsankov and Swarat Chaudhuri and Martin T. Vechev. *AI2 : Safety and Robustness Certification of Neural Networks with Abstract Interpretation*,IEEE Symposium on Security and Privacy, SP 2018, Proceedings, IEEE Computer Society

[**BBBM22**] Sofiane Bessaï and Dorra Ben Khalifa and Hanane Benmaghnia and Matthieu Martel. *Fixed-Point Code Synthesis Based on Constraint Generation*, Design and Architecture for Signal and Image Processing - 15th International Workshop, DASIP 2022,Proceedings, Lecture Notes in Computer Science, Springer.

[**BM22**] Hanane Benmaghnia. *Synthèse de code virgule fixe pour les réseaux de neurones. (Fixed-point code synthesis for neural networks)*, https://tel.archives-ouvertes.fr/tel-03935609, University of Perpignan, France, 2022.

[**LLSC18**] Lo, Chun Y. and Lau, Francis C. M. and Sham, Chiu-Wing. *Fixed-Point Implementation of Convolutional Neural Networks for Image Classification*, 2018 International Conference on Advanced Technologies for Communications (ATC), IEEE.