Stage de Master2 en Intelligence Artificielle pour les systèmes embarqués

Optimisation de l'efficacité énergétique des systèmes multiprocesseurs par utilisation de méthodes d'apprentissage automatique basées sur les graphes

Laboratoires	IETR (équipe ASIC, Nantes) et LS2N (équipe DUKe, Nantes)
Mots-clés	Apprentissage automatique à base de graphes, Architectures multiprocesseurs
Encadrants	Sébastien Le Nours (Nantes Université), Christine Sinoquet (Nantes Université)

Contexte du projet de recherche

L'accroissement de la complexité des architectures matérielles et logicielles des systèmes électroniques (doublement du nombre de transistors intégrables tous les 24 mois) combiné à la nécessité d'optimiser l'efficacité énergétique de ces systèmes impose la définition de nouveaux paradigmes de conception. Dans ce contexte, l'introduction de méthodes d'apprentissage automatique représente une solution à fort potentiel pour permettre de maîtriser la complexité de conception et aboutir à des solutions optimisées, conduisant à l'émergence de flots de conception assistés par l'intelligence artificielle.

Dans le cadre d'une collaboration débutant entre les équipes ASIC de l'IETR et DUKe du LS2N, l'objectif de ce stage est de contribuer à développer un premier modèle destiné à terme à évaluer l'apport de méthodes d'apprentissage automatique basées sur les graphes, dans un but d'exploration et d'optimisation d'architectures sous contraintes de performance et d'énergie. Les méthodes d'apprentissage automatique basées sur les graphes ont reçu récemment un intérêt fort au sein de la communauté des concepteurs de circuits [SSN23]. Elles restent cependant encore peu évaluées pour la conception d'architectures matérielles et logicielles de systèmes multiprocesseurs pour lesquels l'efficacité énergétique se doit d'être optimisée.

Dans le cadre de ce travail de Master, nous étudierons une architecture multiprocesseur initialement spécifiée par (i) une application logicielle représentée au moyen d'un graphe Ga spécifiant le degré de parallélisme entre les différentes tâches de l'application, et (ii) un graphe Gu des relations entre unités de calcul précisant aussi la nature de ces relations (e.g., point à point, mémoire partagée). On cherchera alors à identifier un graphe Gc de contraintes d'allocation des tâches de l'application aux unités de calcul, optimisant les performances et la consommation d'énergie de l'architecture multiprocesseur.

Méthodologie et résultats attendus

Dans le cadre de ce travail, l'approche d'apprentissage basée sur les graphes doit permettre d'identifier des solutions minimisant l'activité des ressources de calcul et de mémorisation, ce au sein de plates-formes dédiées au traitement intensif de données de type GPU (Graphical Processing Units) utilisées dans de nombreuses applications industrielles. Compte tenu des expertises complémentaires des équipes impliquées, ce stage permettra d'appréhender conjointement des cas concrets d'applications et d'architectures (pour lesquelles les données de simulations sont d'ores et déjà disponibles ou pourront être générées) et de contribuer à la définition de méthodes originales du domaine de l'apprentissage automatique basé sur les graphes.

Dans le domaine des réseaux de neurones profonds, les GNNs (Graph Neural Networks) se prêtent au codage de graphes sous la forme de vecteurs caractéristiques, favorisant ainsi leur analyse. Le stage aura pour but d'analyser de façon conjointe plusieurs graphes (Ga, Gu, Gc), au moyen d'un modèle de type GNN, dans le but de prédire les grandeurs combinées de performance et d'énergie (perf, en). Ce modèle sera entraîné à partir de données obtenues par simulation ou prototypage sur cible réelle, l'entraînement du GNN produisant alors une fonction de régression f(Ga,Gu,Gc) = (perf, en). Il convient de noter que les GNNs utilisés pour la régression le sont généralement pour fournir des prédictions isolées, pour les nœuds d'un graphe. Dans le cas présent, c'est un objet composé de plusieurs graphes (modélisation conjointe) qui fera l'objet d'une régression.

Le stage permettra de poser les premiers jalons sur un sujet d'étude émergent utilisant les GNNs. L'IETR a récemment obtenu des gains significatifs dans l'optimisation énergétique d'applications de calcul intensif sur des cibles multiprocesseurs [DNP22, DNH23]. Ce travail permettra donc d'encore améliorer le processus d'optimisation mis en place, et ce par l'introduction de méthodes originales basées sur les méthodes à base de GNNs.

- Phase 1 (mois 1) : étude bibliographique.
- Phase 2 (mois 2 à 3) : appropriation du concept de GNN pour la modélisation 1) du degré de parallélisme au sein d'une application logicielle, 2) des relations entre les unités de calcul et de la nature de ces relations, et 3) des contraintes d'allocation des tâches de l'application aux unités de calcul.
- Phase 3 (mois 4 à 5) : apprentissage, validation et test de la fonction de régression f obtenue via la modélisation conjointe f(Ga,Gu,Gc) = (performance, énergie) sur données simulées.
- Phase 4 (mois 6) : rédaction du mémoire, préparation de la soutenance

[DNH23] Dariol, Q., Le Nours, S., Helms, D., Stemmer, R., Pillement, S. and Grüttner, K. (2023). Fast Yet Accurate Timing and Power Prediction of Artificial Neural Networks Deployed on Clock-Gated Multi-Core Platforms. In Proceedings of the DroneSE and RAPIDO: System Engineering for constrained embedded systems (RAPIDO '23).

[DNP22] Dariol, Q., Le Nours, S., Pillement, S., Stemmer, R., Helms, D., Grüttner, K. (2022). A Hybrid Performance Prediction Approach for Fully-Connected Artificial Neural Networks on Multi-core Platforms. Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS 2022).

[SSN23] Sánchez, D., Servadei, L., Naz Kiprit, G., Wille, R., and Ecker, W. (2023). A Comprehensive Survey on Electronic Design Automation and Graph Neural Networks: Theory and Applications. ACM Trans. Des. Autom. Electron. Syst. 28, 2, Article 15, March 2023.

Environnement du projet

Ce projet se déroulera au laboratoire IETR à Polytech Nantes, 44306 Nantes. Ce groupe de recherche possède une longue expertise dans le domaine de la conception et de la modélisation de systèmes embarqués. Le stage est organisé en étroite collaboration avec les membres du laboratoire LS2N. L'étudiant stagiaire sera également associé à d'autres activités des groupes de recherche : réunions de groupe, séminaires, événements sociaux.

La durée du projet est comprise entre 5 et 6 mois, démarrant entre février et avril 2024. Selon la réglementation, l'indemnité de stage est d'environ 600 euros par mois.

Profil du candidat

Ce stage s'adresse à un étudiant de Master, ou étudiant de 5ème année d'école d'ingénieur, en informatique et/ou électronique. Les compétences requises sont :

- Apprentissage machine,
- Réseaux de neurones à base de graphes,
- Architectures matérielles et logicielles,
- Ecriture et lecture en anglais.

Veuillez envoyer un email avec votre CV et une lettre de motivation.

Contacts **Sébastien Le Nours**
Email: sebastien.le-nours@univ-nantes.fr
Address: Polytech Nantes, rue Christian Pauc, 44306 Nantes, France
Phone: 02.40.68.30.53
Christine Sinoquet
Email: christine.sinoquet@univ-nantes.fr
Address: LS2N, 2 rue de la Houssiniere, 44322 Nantes, 44306 Nantes, France
Phone: 02.51.12.58.05