Stage de Master2 en Intelligence Artificielle pour les systèmes embarqués

Evaluation de réseaux de neurones basés sur les graphes pour l'optimisation de l'efficacité énergétique des systèmes multiprocesseurs

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.

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 d'une collaboration entre les équipes ASIC de l'IETR et DUKe du LS2N, l'objectif de ce stage est de contribuer à développer et évaluer un modèle de réseau de neurones sur graphes (*Graph Neural Network*, GNN) utilisé pour l'exploration et l'optimisation d'architectures multiprocesseurs sous des contraintes de performance et d'énergie.

Dans le cadre de ce travail de Master, nous étudierons des architectures multiprocesseurs initialement spécifiées 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 optimiser l'allocation des tâches de l'application aux unités de calcul, conduisant à améliorer 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 travail préliminaire [G24], un premier modèle de GNN fut proposé afin de permettre l'estimation des performances et de la consommation de systèmes multiprocesseurs. Ce modèle fut développé et validé sur un jeu limité de données issues de mesures sur un prototype réel. Dans le cadre du présent travail de Master, nous envisageons de poursuivre la caractérisation et la validation du modèle proposé en le confrontant à un jeu de données plus conséquent. Ces données seront créées de manière synthétique complétant ainsi les mesures sur des cibles réelles. Cette approche permettra d'entraîner le modèle proposé sur un jeu de données plus important afin de le rendre plus précis et robuste. Ce modèle sera évalué pour différents cas d'étude représentatifs de systèmes multiprocesseurs. Ces travaux contribueront à la préparation d'un article scientifique pour une conférence du domaine de recherche.

Le stage permettra de consolider les travaux menés à base de 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, 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 2 (mois 2 à 3) : analyse du jeu de données et utilisation pour l'entraînement du GNN étudié.
- Phase 3 (mois 4 à 5) : évaluation de la précision et du temps d'exécution du GNN entraîné, comparaison avec certaines méthodes traditionnellement utilisées (modèles analytiques).
- 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).

[G24] Ghrayeb, Z. (2024). Energy Efficiency Optimization of Multiprocessor Systems with Graph-based Machine Learning Methods. Master internship report. Nantes University.

[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. 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 2025. 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,
- Rigueur en programmation en informatique (Python)
- Rigueur dans le suivi d'un protocole expérimental et dans sa réalisation
- Capacités de reporting sur ses travaux.

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

Master internship in Artificial Intelligence for embedded systems

Evaluation of graph-based neural networks for energy efficiency optimization of multiprocessor systems

Institutes	IETR (ASIC team, Nantes) et LS2N (DUKe team, Nantes)
Keywords	Graph-based machine learning, Multiprocessor architectures
Supervision	Sébastien Le Nours (Nantes University), Christine Sinoquet (Nantes University)

Background to the research project

The increasing complexity of hardware and software architectures of electronic systems (doubling of the number of integrable transistors every 24 months) combined with the need to optimize the energy efficiency of these systems requires the definition of new design paradigms. In this context, the introduction of machine learning methods represents a solution with high potential to control design complexity and achieve optimized solutions, leading to the emergence of design flows assisted by artificial intelligence.

Graph-based machine learning methods have recently received strong interest within the circuit design community [SSN23]. However, they remain little evaluated for the design of hardware and software architectures of multiprocessor systems for which energy efficiency must be optimized. As part of a collaboration between the ASIC team of the IETR and DUKe of the LS2N, the objective of this internship is to contribute to developing and evaluating a graph neural network model (*Graph Neural Network*, GNN) used for the exploration and optimization of multiprocessor architectures under performance and energy constraints.

As part of this Master's thesis, we will study multiprocessor architectures initially specified by (i) a software application represented by a graph Ga specifying the degree of parallelism between the different tasks of the application, and (ii) a graph Gu of the relationships between computing units also specifying the nature of these relationships (e.g., point-to-point, shared memory). We will then seek to optimize the allocation of application tasks to computing units, leading to improving the performance and energy consumption of the multiprocessor architecture.

Methodology and expected results

As part of this work, the graph-based learning approach should identify solutions that minimize the activity of computing and storage resources, within platforms dedicated to intensive data processing of GPU (*Graphical Processing Units*) type used in many industrial applications. Given the complementary expertise of the teams involved, this internship will allow us to understand concrete cases of applications and architectures (for which simulation data are already available or can be generated) and to contribute to the definition of original methods in the field of graph-based machine learning.

In the preliminary work [G24], a first GNN model was proposed to allow the estimation of the performance and consumption of multiprocessor systems. This model was developed and validated on a limited set of data from measurements on a real prototype. As part of this Master's thesis, we plan to continue the characterization and validation of the proposed model by confronting it with a larger dataset. These data will be created in a synthetic manner, thus complementing the measurements on real targets. This approach will allow the proposed model to be trained on a larger dataset in order to make it more accurate and robust. This model will

be evaluated for different case studies representative of multiprocessor systems. This work will contribute to the preparation of a scientific article for a conference in the research field.

The internship will consolidate the work carried out using GNNs. The IETR has recently achieved significant gains in the energy optimization of intensive computing applications on multiprocessor targets [DNP22, DNH23]. This work will therefore further improve the optimization process implemented, through the introduction of original methods based on GNNs-based methods.

- Phase 1 (month 1): bibliographic study, appropriation of the GNN concept for modeling 1) the degree of parallelism within a software application, 2) the relationships between computing units and the nature of these relationships, and 3) the constraints of allocation of application tasks to computing units.
- Phase 2 (months 2 to 3): analysis of the dataset and use for training the studied GNN.
- Phase 3 (months 4 to 5): evaluation of the accuracy and execution time of the trained GNN, comparison with some traditionally used methods (analytical models).
- Phase 4 (month 6): writing the thesis, preparation of the defense.

[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).

[G24] Ghrayeb, Z. (2024). Energy Efficiency Optimization of Multiprocessor Systems with Graph-based Machine Learning Methods. Master internship report. Nantes University.

[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.

Project environment

This project will take place at the IETR laboratory at Polytech Nantes. This research group has a long expertise in the field of embedded systems design and modeling. The internship is organized in close collaboration with the members of the LS2N laboratory. The student intern will also be involved in other activities of the research groups: group meetings, seminars, social events.

The duration of the project is between 5 and 6 months, starting between February and April 2025. According to the regulations, the internship allowance is approximately 600 euros per month.

Candidate profile

This internship is aimed at a Master's student, or a 5th year engineering student, in computer science and/or electronics. The required skills are:

- Machine learning,
- Graph-based neural networks,
- Hardware and software architectures,
- Writing and reading in English,
- Rigor in computer programming (Python)
- Rigor in monitoring an experimental protocol and in its realization
- Reporting skills on his work.

Please send an email with your CV and a cover letter.

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