Optimization strategies for Neural HW & SW automated co-design

Le Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) est un acteur majeur en matière de recherche, de développement et d'innovation. Cet organisme de recherche technologique intervient dans trois grands domaines : l'énergie, les technologies pour l'information et la santé, et la défense. Reconnu comme un expert dans ses domaines de compétences, le CEA est pleinement inséré dans l'espace européen de la recherche et exerce une présence croissante au niveau international. Situé en île de France sud (Saclay), le Laboratoire d'Intégration des Systèmes et des Technologies (LIST) a notamment pour mission de contribuer au transfert de technologies et de favoriser l'innovation dans le domaine des systèmes de calcul parallèle. Ce stage se déroulera au sein du Laboratoire Environnement de Conception et Architectures (LECA) sur le site de Nano-Innov du CEA LIST.

Dans les domaines passionnants de conception d'architecture matérielle et d'optimisation des solutions de calcul pour l'Intelligence Artificielle (IA), le Laboratoire LECA a développé une brique technologique spécifiquement conçue pour accélérer les algorithmes d'IA. Baptisé *PNeuro*, cette architecture a décroché le prestigieux prix *Embedded World* pour sa grande efficacité énergétique [1]. En effet, elle permet d'exécuter des réseaux de neurones profonds de manière efficace, tout en consommant très peu d'énergie, sur des plates-formes embarquées. Pour tirer le meilleur parti de cette architecture, il est essentiel d'optimiser l'exécution des réseaux sur le PNeuro, une tâche qui présente de nombreux défis. Pour ce faire, des optimisations à la fois algorithmiques et architecturales sont à l'étude, appelées *co-design*.
L'objectif de ce stage est l'exploration de paramètres architecturaux afin d'optimiser l'exécution d'une application (type réseau de neurones) sur l'architecture cible PNeuro. Un des enjeux importants de cette étude est l'allocation des calculs aux processeurs élémentaires de sorte à minimiser le transfert de données en mémoire. Il est important que cette opération soit générique et efficace afin de tirer profit de l'architecture cible. Une première solution s'appuyant sur le framework AIdge [4] du CEA LIST a été proposée et implémentée par l'équipe.

Dans un premier temps, le candidat sera chargé de proposer une modélisation mathématique du problème et de concevoir une représentation intermédiaire du matériel et de l'application, en mettant l'accent sur les transferts de données. Cette phase devra prendre en considération toutes les contraintes associées à l'architecture matérielle PNeuro, qui est hautement modulaire, ainsi que les particularités des réseaux de neurones. De plus, cette étape permettra au candidat de se familiariser avec les outils existants au sein de l'équipe ainsi qu'avec les approches de pointe en vigueur dans l'état de l'art. Le candidat pourra également bénéficier de l'expertise de l'équipe pour mener à bien cette phase [2,3].
Dans un second temps, le candidat devra élaborer des algorithmes d'exploration à la fois efficaces et scalables. Cela lui permettra, d'une part, de déterminer les paramètres architecturaux appropriés et, d'autre part, d'optimiser l'exécution de l'application sur l'architecture cible. L'objectif est de trouver le bon compromis entre différents critères, à savoir la surface, les performances et l'impact environnemental. A noter que ce dernier critère est particulièrement innovant, et l'équipe accorde une grande importance à cette dimension pour répondre aux enjeux de frugalité et d'écoconception. Enfin, le candidat mettra en pratique ses développements en intégrant la solution proposée aux outils existants tels qu'A-DECA, ainsi qu'à l'architecture PNeuro. Il testera et validera cette solution sur des applications de référence telles que MobileNet.

Commissariat à l'Energie Atomique et aux Energies Alternatives |
Institut Carnot List | CEA Saclay Nano-INNOV | Bât. 862-PC172
91191 Gif-sur-Yvette Cedex - FRANCE
Tel. : 01.69.08.49.67 | Fax : 01.69.08.83.95
www-list.cea.fr

Commissariat à l'Energie Atomique et aux Energies Alternatives |
Institut Carnot List | MINATEC Campus | 17 rue des Martyrs
38054 Grenoble Cedex 9 - FRANCE
Tel : 04.38.78.68.05 |

DSCIN, Département des Systèmes et Circuits Intégrés Numériques
Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019

Le candidat recherché est en dernière année de master recherche ou diplôme ingénieur (bac+5). Des connaissances solides en algorithmique et/ou en Recherche Opérationnelle, langages C/C++ et Python sont requises. Des connaissances en architectures de calcul, traitement de l'image seront aussi appréciées. Exigeant et investi, vous avez à cœur de proposer des solutions innovantes et de travailler dans un milieu à la pointe de la technologie qui vous permettra de répondre aux enjeux de demain. Le candidat devra être doté d'un bon relationnel et posséder la capacité de travailler en équipe et en autonomie.

Niveau demandé : Bac+5 **Durée du stage :** 6 mois

Compétences :
- Bon niveau en programmation C/C++, Python
- Connaissances Recherche opérationnelle & Optimisation combinatoire, algorithmique
- Connaissances en architecture système et processeur sont un plus
- Quelques compétences en les réseaux de neurones et traitement de l'image seraient un plus

Pièces à fournir : CV + lettre de motivation + classements
Contact : Raphael Millet, Lilia Zaourar (raphael.millet@cea.fr, lilia.zaourar@cea.fr)

Références :
[1]- B. TAIN, R. MILLET : Embedded World Prize for PNeuro IP 2022
[2]- Zaourar, L., Ait Aba, M., Briand, D., & Philippe, J. M. (2018). Task management on fully heterogeneous micro-server system: Modeling and resolution strategies. Journal Concurrency and Computation: Practice and Experience, 30(23), e4798
[3]- Ait Aba, M., Zaourar, L., & Munier, A. (2020). Efficient algorithm for scheduling parallel applications on hybrid multicore machines with communications delays and energy constraint. Concurrency and Computation: Practice and Experience, 32(15), e5573.
[4]- https://projects.eclipse.org/projects/technology.aidge