2022-04365 - Design of Open-source Hardware Accelerators for Deep Learning

Type de contrat : CDD
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Ingénieur scientifique contractuel

A propos du centre ou de la direction fonctionnelle

The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Contexte et atouts du poste

Host team: The project will be held in the TARAN (formerly CAIRN) team of the IRISA/INRIA laboratory. The TARAN team, with more than 35 members from Inria, URT, and ENS Rennes, has participated in several national and European R&D projects (H2020 ARGO, FP7 Alma, FP7 Flextiles) and has strong industrial collaborations (e.g., Safran, Thales, Alcatel, Orange, STMicroelectronics, Technicolor, and various SMEs). TARAN has recognized experience in several domains related to the project, such as embedded system design, fault tolerance, safety-critical systems, computing architectures, design tools for specialized hardware architectures.

Acquiring new skills: As a new member of the TARAN team, you will be integrated in a research group with excellent prestige and deep knowledge of embedded systems. The TARAN research group can provide you with a more solid understanding and knowledge of computer architectures and hardware design. For instance, the host team has high-quality papers published using RISC-V-based processors and dedicated hardware designs, subjects that you will be able to learn much more about and increase my background in this area.

Principales activités

Deep Learning (DL) is one of the most intensively and widely used predictive models in the field of Machine Learning. Convolutional Neural Networks (CNNs) [2] have shown to achieve state-of-the-art accuracy in computer vision [1] and have even surpassed the error rate of the human visual cortex. These neural network techniques have quickly spread beyond computer vision to other domains. For instance, deep CNNs have revolutionised tasks such as face recognition, object detection, and medical image processing. Recurrent neural networks (RNNs) achieve state-of-the-art results in speech recognition and natural language translation [3], while ensembles of neural networks already offer superior predictions in financial portfolio management, playing complex games [4] and self-driving cars [5].

In the case of DL systems, there are two main computational tasks: training and inference. Training requires vast quantities of labelled data that are used to optimize the network for the task at hand, usually by way of some form of stochastic gradient descent (SGD) algorithm. Inference, on the other hand, is the actual application of the trained network, which can be replicated onto millions of devices.

Despite the benefits that DL brings to the table, there are still important challenges that remain to be addressed if the computational workloads associated with NNs are to be deployed on embedded edge devices that require improved energy efficiency. Such taxing demands are pushing both industry and academia to concentrate on designing custom platforms for DL algorithms that target improved performance and/or energy efficiency. This project is about the design, verification and prototyping of hardware accelerators (mainly FPGA-based) for DL inference and training.

In the Inria/IRISA/Taran team, we are currently pushing for the design of hardware accelerator for both inference and training acceleration following the open-source hardware principles. Even if there already exist designs available as opensource [7],[8], they all partially cover the issue and come as part of a specific kernel acceleration (e.g., GEMM [8]) or a library (e.g., HLS4ML [7]). We seek to develop of the full accelerator architecture as an overlay that can be configured and deployed on an FPGA platform. We seek in particular real demonstrators in two complementary settings: cloud FPGAs (e.g., Xilinx Alveo U280 Data Center Accelerator Card) and embedded systems (Xilinx UltraScale+ ZCU102 development board). Interface of the overlay with DL frameworks such as TensorFlow or PyTorch will be also part of the job. The accelerator will be designed mainly using C++, leveraging high-level synthesis (HLS) tools such as VivisHLS or CatapultHLS. Previous work in our team on RISC-V processors have shown that HLS has strong benefits for such architecture design [9].

We also plan to synthesize the accelerator architecture as an ASIC prototype to further demonstrate gains in performance and energy in the context of energy-efficient embedded systems, such as in autonomous vehicles, or on ultra-low-power IoT (Internet of Things) devices.

Position: Research Engineer/Research Associate

Keywords: hardware accelerator, deep neural networks, high-level synthesis, FPGA design
Compétences
The recruited person is expected to develop complex processor architectures leveraging C++ and High-Level Synthesis. We also expect to have prototype implementations of the developed techniques on FPGA and ASIC.

Desired skills include:
- Computer architecture, hardware design, VLSI circuit design.
- Basic knowledge in compilers.
- Familiarity with the C/C++ language or other languages.
- Familiarity with FPGA/ASIC design and/or High-Level Synthesis.
- Optimization methods

Mostly importantly, we seek highly motivated and active researchers.

Avantages
- Subsidized meals
- Partial reimbursement of public transport costs

Rémunération
monthly gross salary from 2562 euros according to diploma and experience

Informations générales
- **Thème/Domaine**: Architecture, langages et compilation
- **Ville**: Rennes
- **Centre Inria**: CRI Rennes - Bretagne Atlantique
- **Date de prise de fonction souhaitée**: 2022-02-01
- **Durée de contrat**: 2 ans
- **Date limite pour postuler**: 2022-03-17

Contacts
- **Equipe Inria**: TARAN
- **Recruteur**: Sentieys Olivier / Olivier.Sentieys@irisa.fr

A propos d'Inria
Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3500 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 180 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Consignes pour postuler
Please submit online: your resume, cover letter and letters of recommendation eventually

For more information, please contact: olivier.sentieys@inria.fr

Sécurité défense :
Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.

**Attention**: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.