Offer #2022-04365

Design of Open-source Hardware Accelerators for Deep Learning

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : Temporary scientific engineer

About the research centre or Inria department

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.

Context

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, UR1, 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.

Main activities

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 portofolio 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 VitisHLS 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

Skills

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.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs

Remuneration

monthly gross salary from 2562 euros according to diploma and experience

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2022-02-01
- **Duration of contract :** 2 years
- **Deadline to apply :** 2022-09-30

Contacts

- **Inria Team :** TARAN
- **Recruiter :**
  Sentieys Olivier / Olivier.Sentieys@irisa.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

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

For more information, please contactolivier.sentieys@inria.fr

Defence Security :

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.