Job vacancy #2023-06916

PhD Position F/M Dynamically Configurable Deep Neural Network Hardware Accelerators

Contract type: Fixed-term contract
Level of qualifications required: Graduate degree or equivalent
Function: PhD Position

About the research centre or Inria department

The Inria Centre of Rennes University 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

Fully funded 3-year Ph.D. thesis. Possibility for talented last-year Master's students to start as interns and continue after graduation as Ph.D. researchers

Context & background

State-of-the-art Deep Neural Network (DNN) models [1] usually require large amounts of data to be trained and contain a tremendous number of parameters, leading to overall high resource requirements in computation and memory and, thus, energy. In the past years, this gave rise to approaches to reducing these requirements, such as pruning [2], quantization [2], [3], or NAS [4]. Many hardware optimizations have been proposed to accelerate the inference of DNNs on different architectures, but they mainly consider static scenarios, which may not be efficient in realistic cases. For many applications, the performance requirements of a DNN model deployed on a given hardware platform are not static but evolve dynamically as its operating conditions and environment change. However, this variability and these dynamics are not addressed considering both HW and SW levels in state-of-the-art DNN approaches.

Runtime configurable hardware accelerators are necessary in a variety of scenarios. One example is Edge Computing, which involves processing data near the source of data collection instead of sending it to a centralized cloud or data center. In edge computing scenarios, there is often limited power and computing resources, and configurable hardware accelerators can help optimize the energy efficiency of AI models, leading to longer battery life and reduced energy consumption.

Opposed to conventional AI accelerators (e.g., GPUs, Google TPU), run-time configurable HW accelerators for DNNs [5] [7] [8] [5] [9] [10] can change different parameters of the computations (e.g., data type and precision, dataflow architecture, etc.) performed by the hardware accelerator at runtime. This approach allows for more efficient use of hardware resources and improved energy efficiency for AI applications by supporting operations at various precision levels, often configurable at runtime, and sparse weight storage techniques.

Hardware architecture design space exploration is usually based on high-level models of hardware operations and considers the design task a multi-dimensional optimization problem [11]. Concerning DNN hardware accelerators, previous attempts at architectural design space exploration exist [12]. Nevertheless, all the obtained designs delegate the runtime adaptation to higher levels and do not provide models to control it efficiently. Thus, an efficient and accurate approach is needed to model the impact of different configurations of DNN HW accelerator parameters on attributes of interest (e.g., inference latency and accuracy, energy consumption, etc.). This is of paramount importance to build a runtime controller able to fine-tune HW (and SW) parameters at runtime to attain specific accuracy/inference time/energy objectives.

This thesis is in the context of a broader research project, whose focus is to propose an original interdisciplinary approach – relying on complementary expertise in artificial intelligence, hardware architecture design, and control theory – that allows DNN models to be dynamically configured at runtime on a given reconfigurable hardware accelerator architecture, depending on the external environment. Such reconfigurable AI systems will require much fewer resources on average and thus allow for substantial savings in energy for many applications.

This thesis will focus on HW aspects. AI accelerators need to support different configurations of various natures. The major challenge is to predict the impact of different HW configurations on attributes of interest, such as the final DNN accuracy, energy, and performance. To do so, creating effective models to
predict such impact is critical.


Assignment

Ph.D. thesis goal

The goals of this thesis are:

1. Perform an extensive design space exploration to accurately characterize the effect of different configurable HW parameters on various attributes of interest (e.g., inference latency and accuracy, energy consumption, etc.).
2. Synthesize the characterization results in a surrogate model to accurately and quickly predict the attributes of interest, even when not explicitly characterized.

Main activities

More in detail, the Ph.D. researcher will focus on assessing the effect of different sets of HW parameters on the attributes of interest, such as inference latency, throughput, energy consumption, etc. The assessment will determine how to combine the HW parameters to provide different trade-offs among these attributes. Highly time-consuming HW syntheses and simulations are needed to assess the effect of all possible combinations of parameters of the DNN HW accelerator. Therefore, the first challenge will be to propose practical approaches to decompose the main problem into smaller ones and design and develop efficient and accurate solutions independently of the final application. The second challenge will be proposing innovative ways to adapt the assessment to the final DNN application. Then, the student will focus on modeling the effect of HW parameters on the attributes of interest. A combination of HW and SW parameters constitutes an operating configuration of the system. Simulating and enumerating in the runtime controller all possible operating configurations might not be feasible and is undoubtedly not scalable. Therefore, the third challenge will be identifying the most significant HW parameters and designing and developing predictive models (e.g., statistical or machine learning) to predict HW-level attributes of interest.

The Ph.D. researcher will regularly meet other project participants to collaborate tightly, attend regular workshops within the research project framework to share the advancements and findings with the partners, and is expected to publish and present scientific research results at prestigious venues.

Skills
We seek highly motivated and passionate candidates. Autonomy is a highly appreciated quality.

Required Skills:

**HW design:** VHDL/Verilog, HW synthesis flow (design, simulation, synthesis, and deployment through commercial tools for FPGA or ASIC)

**Experience in Computer architecture:** Instruction Set Architecture (ISA), Microarchitecture, and Systems design. Knowledge about hardware architectures of Neural Network accelerators is a plus.

**SW Programming/Scripting:** C/C++, Python, Linux scripting

**Experience with Deep Neural Network development frameworks:** PyTorch/TensorFlow

**Experience with High-Level Synthesis (HLS)** and related tools (e.g., Vivado/Vitis HLS or Siemens Catapult) is a plus.

**Required degree:**

Candidates must have a Master’s degree (or equivalent) in Computer Engineering or Electronic Engineering.

*Talented last year Master’s students may start as 6-month interns and continue as Ph.D. researchers after graduation.*

**Environment:**

The Ph.D. thesis will be carried out in collaboration with LIRIS (INSA Lyon), GIPSA-lab (Grenoble Alpes University), and Inria Lyon in the context of the ANR Project RADYAL. The Ph.D. candidate will be supervised by the TARAN team ([https://team.inria.fr/taran/](https://team.inria.fr/taran/)) at the Inria Centre at Rennes University, IRISA laboratory, in France. The Ph.D. salary will follow standard French rates.

**Ph.D. direction team:**

Supervisor: Olivier Sentieys / Olivier.Sentieys@irisa.fr
Co-supervisor: Marcello Traiola / marcello.traiola@inria.fr

**Benefits package**

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

**Remuneration**

Monthly gross salary amounting to 2082 euros for the first and second years and 2190 euros for the third year

**General Information**

- **Theme/Domain:** Architecture, Languages and Compilation
- **System & Networks (BAP E)**
- **Town/city:** Rennes
- **Inria Center:** [Centre Inria de l'Université de Rennes](https://www.inria.fr)
- **Starting date:** 2024-02-01
- **Duration of contract:** 3 years
- **Deadline to apply:** 2024-01-14

**Contacts**

- **Inria Team:** [TARAN](https://team.inria.fr/taran/)
- **PhD Supervisor:** 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
Instruction to apply

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

For more information, please contact marcello.traiola@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.