Call for applications – PhD candidate

Lyon Institute of Nanotechnology
Ecole Centrale de Lyon, 36 av. Guy de Collongue,
F-69134 Ecully, FRANCE
http://inl.cnrs.fr

Multi-Objective Design Space Exploration for efficient and reliable Neural Networks hardware accelerators

Deep Neural Networks (DNNs) [1] are currently one of the most intensively and widely used predictive models in the field of machine learning. DNNs have proven to give very good results for many complex tasks and applications, such as object recognition in images/videos, natural language processing, satellite image recognition, robotics, aerospace, smart healthcare, and autonomous driving. Nowadays, there is intense activity in designing custom Artificial Intelligence (AI) hardware accelerators to support the energy-hungry data movement, speed of computation, and memory resources that DNNs require to realize their full potential [2]. Furthermore, there is an incentive to migrate AI from the cloud into the edge devices, i.e., Internet-of-Things (IoTs) devices, in order to address data confidentiality issues and bandwidth limitations, given the ever-increasing internet-connected IoTs, and also to alleviate the communication latency, especially for real-time safety-critical decisions, e.g., in autonomous driving.

Hardware for AI (HW-AI), similar to traditional computing hardware, is subject to hardware faults (HW faults) that can have several sources: variations in fabrication process parameters, fabrication process defects, latent defects, i.e., defects undetectable at time-zero post-fabrication testing that manifest themselves later in the field of application, silicon ageing, e.g., time-dependent dielectric breakdown, or even environmental stress, such as heat, humidity, vibration, and Single Event Upsets (SEUs) stemming from ionization. All these HW faults can cause operational failures, potentially leading to important consequences, especially for safety-critical systems.

HW-AI comes with some inherent resilience to HW faults, similar to biological neural networks. Indeed, the statistical behavior of neural network architectures, as well as their high space redundancy and overprovisioning, naturally provide a certain tolerance to HW faults. HW-AI have the capability to circumvent to a large extent HW faults during the learning process. However, HW faults can still occur after training. Recent studies in the literature have shown that HW-AI is not always immune to such HW faults. Thus, inference can be significantly affected, leading to DNN prediction failures that are likely to lead to a detrimental effect on the application [3, 4, 5]. Therefore, ensuring the reliability of HW-AI platforms is crucial, especially when HW-AI is deployed in safety-critical and mission-critical applications, such as robotics, aerospace, smart healthcare, and autonomous driving.

This project aims at developing generation methods for generating HW accelerators for DNNs to be used in safety/mission-critical applications (e.g., Space, Automotive). The objective is to implement a Network Architecture Search (NAS) framework in order to identify the best DNN architecture/implementation in terms of complexity (memory footprint, computational effort) and dependability

In the framework of the research project PEPR-IA Adapting, the Electronic group at INL is currently looking for a (m/f) **PhD student** for a **3-year** contract.

Job description

The Ph.D. thesis is structured in the following 5 main tasks

1. **DNN components analysis and modeling (M1 to M3)**
    - Identification of the main components to be used to implement a DNN and its configuration parameters: Convolutional Kernel size, Activation Function Implementation, Arithmetic Functions and related data type representation/bit-width, …
    - Model each component and its configuration as a software function and HDL description.
2. **Component-Level precision/resilience estimation (M3 to M6)**
    - For each component, determine the precision (e.g., representable values) and the fault masking level (resilience to the presence of a fault) for a given configuration/implementation. To be carried out through simulation and fault injection;

- Model precision/resilience as probabilistic distribution: given a component/configuration/implementation evaluate the probability of masking a fault.

3. **DNN architecture search space (M6-M18)**

   - Design and implement NAS (neural architecture search) space exploration. Explore different optimization algorithms such as Genetic Algorithm approach such as [6] developed at the INL;
   - Define efficient heuristics to quickly estimate the impact of a solution on the overall DNN Accuracy and Dependability. The heuristic will be based on the component probabilistic models.

4. **DNN synthesis (M18 – M28)**

   - Generate the DNN implementation as software or by generating a dedicated hardware accelerator;
   - Hybrid solution can be also identified: integrating a dedicated HW accelerator in to a microcontroller (e.g., RISC-V);
     - Identification of the DNN components that have to be executed on the HW accelerator or in Software. They can be profiled and further classified as the most commonly used, the most complex, the most critical (from the dependability point of view).

5. **DNN Evaluation (M6-M36)**

   - Accuracy, Memory Footprint, Power Consumption, Performances, Dependability;
   - Evaluation can be done through simulation, microcontroller board measurements, prototypes on FPGA and potentially ASIC demonstrators;
   - Scientific papers and thesis manuscript preparation.

Profile

You have or are about to obtain an MSc in Computer Engineer / Computer Science with strong experience in at least one of the following areas: computer architectures, digital circuit design, optimization algorithms. Good programming skills (python, C and C++) are required. Previous experience in Neural Networks is a plus (e.g., knowledge of major NN frameworks such as Pytorch and Tensorflow). Excellent written and verbal communication skills in English. Fluency in French is also a plus but is not mandatory.

About INL

INL is a 250-strong research institute based in Lyon, France, carrying out fundamental and applied research in electronics, semiconductor materials, photonics and biotechnologies. The Electronic group is a leader in the area of advanced nanoelectronic design, with research projects and collaborations at both national and European level. Recent highlights include the development of genetic algorithms based multi objective design space exploration [6, 7].

Dates:

Ph.D. will start in October 2024.

Environment:

The Ph.D. candidate will be supervised by the INL team in Lyon (Ecole Centrale Campus). The Ph.D. salary will follow standard French rates.

References:

[1] Y. LeCun, et al., "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[2] B. Moons, et al, "14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," in IEEE ISSCC, 2017.
[3] C. Torres-Huitzil and B. Girau, "Fault and Error Tolerance in Neural Networks: A Review," IEEE Access, 2017.
[4] A. Ruospo, et al., "A pipelined multi- level fault injector for deep neural networks," in IEEE DFT, 2020
[5] A. Lotfi et al., "Resiliency of automotive object detection networks on GPU architectures," in IEEE ITC, 2019
[6] Mario Barbareschi, et al. "A Genetic-algorithm-based Approach to the Design of DCT Hardware Accelerator," in ACM J. Emerg. Technol. Comput. Syst. 18, 3, Article 50 (July 2022), 25 pages. https://doi.org/10.1145/3501772
[7] S. Barone, M. Traiola, M. Barbareschi and A. Bosio, "Multi-Objective Application-Driven Approximate Design Method," in *IEEE Access*, vol. 9, pp. 86975-86993, 2021, doi: 10.1109/ACCESS.2021.3087858.

Send CV and statement of purpose (in English or French) to

Alberto Bosio / INL - Lyon Institute of Nanotechnology - Ecole Centrale Lyon – email: alberto.bosio@ec-lyon.fr