Offer #2025-08680

PhD Position F/M Reliable Deep Neural Network Hardware Accelerators

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

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

Context & background:

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 AI-HW is deployed in safety-critical and mission-critical applications, such as robotics, aerospace, smart healthcare, and autonomous driving.

Assessing the **fault sensitivity** of HW-AI systems is a **highly complex, time-consuming, and energy-intensive process**. The intricate nature of deep neural networks, with their vast parameter spaces and non-linear computations, makes it challenging to predict how faults will propagate and impact inference. Performing exhaustive fault injection campaigns to evaluate sensitivity requires extensive computational resources and long simulation runtimes, further exacerbating the time and energy costs. Additionally, **real-world testing** on AI hardware is even more demanding, as it involves running extensive workloads under varying fault conditions to capture all potential failure scenarios. This complexity is further amplified by the **diverse architectures** and hardware implementations of AI accelerators, each with unique fault behaviors. As a result, ensuring the reliability of HW-AI platforms demands significant effort, making it a crucial yet **resource-intensive challenge, particularly for safety-critical applications**.

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. Lotfi et al., “Resiliency of automotive object detection networks on GPU architectures,” in IEEE ITC, 2019 [5] A. Ruospo, et al., “Investigating data representation for efficient and reliable Convolutional Neural Networks,” in Microprocessors and Microsystems, 2020

[6] F. K. Dosilovic, et al, “Explainable artificial intelligence: A survey,” in MIPRO, 2018.
[7] N. Srivastava et al., “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

Assignment

Ph.D. thesis goal:

The goal of this thesis is twofold: (i) designing and developing an **optimized algorithm-level fault injection framework** to assess the resiliency of DNN HW accelerators to HW faults, to enable the application of low-cost selective fault-tolerance strategies; (ii) designing **selective fault-tolerance approaches for DNN HW accelerators** by using the analysis provided by the fault injection method.

The reliability improvements obtained with the above-described methodology will be measured, and a design space exploration will be carried out to obtain different DNN HW accelerator implementations that will provide different trade-offs between fault tolerance and energy efficiency.

Main activities

More in detail, the Ph.D. student will design and develop a methodology to perform **large-scale fault analysis** on state-of-the-art DNN hardware architectures. The fault analysis will determine the set of malignant HW faults that mostly impact classification accuracy (or other DNN objectives, such as image segmentation) during the inference phase.

In particular, the student will (i) design a novel framework that enables **accurate fault sensitivity evaluation having the same accuracy of gate level while maintaining computational efficiency**; (ii) ensure that the framework effectively captures the **intricate fault propagation characteristics** of deep neural networks (DNNs) implemented in hardware.

Innovative methodologies will be explored to accelerate fault injection campaigns, such as **hybrid analytical-empirical** approaches and **smart fault pruning** techniques to minimize the number of required fault injections while maintaining high accuracy. Also, smart **energy-efficient parallelization** and distributed computing strategies to accelerate large-scale fault evaluation will be explored.

The proposed approach will be compared against state-of-the-art methodologies to validate improvements in speed and accuracy.

The developed framework will pave the way to **innovative selective error correction mechanisms**. A **design space exploration** will help to find the best

solutions in terms of the trade-off between the fault tolerance level provided by protection mechanisms and the hardware overhead entailed in deploying these solutions.

The project will be part of a larger effort to design an end-to-end workflow that integrates seamlessly into the design and verification process of HW-AI platforms used in mission-critical applications.

Skills

Required technical skills:

- Good knowledge of **computer architectures** and **embedded systems**
- **HW design**: VHDL/Verilog basics, HW synthesis flow
- **Programming** knowledge (C/C++, python)
- Basics of **Machine Learning** (pytorch/tensorflow)
- Experience in fault-tolerant architectures is a plus

Candidates must have a Master's degree (or equivalent) in **Computer Engineering, or Electrical Engineering.**

Languages: proficiency in written English and fluency in spoken English are required.

Relational skills: the candidate will work in a research team, where regular meetings will be set up. The candidate has to be able to present the progress of their work in a clear and detailed manner.

Other values appreciated are open-mindedness, strong integration skills, and team spirit.

Most importantly, we seek highly motivated candidates.

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 2200 euros

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2025-06-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-04-30

Contacts

- **Inria Team :** TARAN
- **PhD Supervisor :**
  Kritikakou Angeliki / angeliki.kritikakou@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.

The keys to success

Candidates with knowledge and experience in **Hardware Design, Hardware for AI**, and **Hardware fault tolerance** are **highly appreciated.**

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

Essential qualities to fulfil a PhD thesis are feeling at ease in an environment of scientific dynamics and wanting to learn, listen, and share.

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

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 contact angeliki.kritikakou@irisa.fr Ou 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.