Offer #2023-06802

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

Moreover, explaining AI decisions, referred to as eXplainable AI (XAI), is highly desirable in order to increase the trust and transparency in AI, safely use AI in the context of critical applications, and further expand AI application areas. Nowadays, XAI has become an area of intense interest [6]. However, in the context of XAI, one of the overlooked aspects is the role that HW faults can have in AI decisions. Therefore, before explaining the decision of an AI algorithm, in order to gain confidence and trust in it, first the hardware that executes the AI algorithm needs to be tested. Besides, if the hardware is compromised (i.e., affected by HW faults), then any attempt for explainability will be either inconclusive or misleading.

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 providing 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 the accuracy of classification (or other DNN objectives, such as image segmentation) during the inference phase.

DNN inference is known to be very complex, especially on large models and for large datasets. Moreover, in order to obtain statistically relevant metrics about the fault impacts, a large number of faults have to be injected. Since the complexity of fault injection grows linearly with DNN inference complexity and the number of injected faults, the biggest challenge of this task will be to reduce such complexity by proposing statistical or analytical methods to prune the fault space. A first approach is profiling DNN hyper-parameters (e.g. distributions of weight values, neuron activations, arithmetic kernel computations) to determine their sensitivity to the final DNN accuracy. In general, a divide and conquer approach will be designed to reduce the fault injection complexity. "Local" (i.e. at layer/kernel levels) fault injections will be performed, to avoid running the full inference. Then, solid and efficient approaches to link the local sensitivity the global accuracy will be designed and developed.

This will allow realizing an optimized accelerated fault injector framework, enabling large-scale fault simulations. In order to further push the performances and reduce injection time, direct execution of the network (or a portion) on FPGA accelerators can also be leveraged. The enhanced fault-injection framework will allow performing several reliability assessments, such as: (a) training a faulty network to find the fault density beyond which the learning capacity starts degrading; (b) performing inference on a faulty network to be able to identify the set of malignant faults; (c) fault injection during training for passive fault tolerance; (d) fault injection attacks during the inference to evaluate the security fault tolerance can offer.

Finally, error correction mechanisms will be designed, e.g. selective low-precision triplication using checkers/voters, most-significant bits reinforcement, standby sparing and correcting codes; alternatively, if a detection mechanism is available, the re-execution of the task of a faulty component or the dynamic rescheduling/mapping of the DNN to the HW-AI (bypassing faulty components) are viable options. A design space exploration will help finding the best solutions in terms of trade-off between the fault tolerance level provided by detection/correction mechanisms and the hardware overhead entailed deploying these solutions.

The candidate will work under the supervision of three people:

- Olivier Sentieys: olivier.sentieys@inria.fr
- Angeliki Kritikakou: angeliki.kritikakou@irisa.fr
- Marcello Traiola: marcello.traiola@inria.fr

Skills

Required technical skills:

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

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

Languages: proficiency in written English and fluency in spoken English or French 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 valued appreciated: 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 2000 euros for the first and second years and 2100 euros for the third year

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-08-02

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

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 contactangeliki.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.