Offer #2025-08800

PhD Position F/M Trustworthy AI hardware architectures

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 and background: Nowadays, there is a growing and irreversible need to distribute **Artificial Intelligence (AI) applications from the cloud to edge** devices, where computation is largely or completely performed on distributed Internet of Things (IoT) devices. This trend aims to **address issues related to data privacy**, **bandwidth limitations, power consumption reduction and low latency requirements**, especially for real-time, mission- and safety-critical applications (e.g., in autonomous driving, support for gesture and medical diagnosis, smart power grid or preventive maintenance).

The direct consequence is the intense activity in designing custom and embedded **Artificial Intelligence HardWare architectures (AI-HW)** to support **energy-intensive data movement**, **speed of computation**, and large **memory resources**

that AI requires to achieve its full potential. 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.

AI-HW, similar to traditional computing hardware, is subject to faults that can have several sources: variability in fabrication process parameters, latent defects or even environmental stress. One of the **overlooked aspects is the role that HW faults can have in AI decisions**. Indeed, there is a common belief that AI applications have an intrinsic high-level or resilience w.r.t. errors and noise. However, recent studies in the scientific literature have shown that AI-HW is not always immune to HW errors. This can jeopardize all the effort of having an **explainable AI**, leading any attempt to explainability to be either **inconclusive or misleading**. In other words, AI algorithms retain their accuracy and explainability property under the condition that the hardware wherein they are executed is fault-free.

Therefore, before explaining the decision of an AI algorithm - to gain confidence and trust in it - firstly the reliability of the hardware executing the AI algorithm needs to be guaranteed, even in the presence of hardware faults. In this way, trust and transparency of an implemented AI model can be ensured, not only in the context of mission- and safety-critical applications, but also in our everyday life.

Assignment

The goal of the Ph.D. thesis is to **study the impact of hardware faults not only on the AI decisions, but also on algorithms developed to explain AI (XAI) models.** The objective is **to make AI-HW reliable by understanding how hardware faults** (due to variability, aging, external perturbations) **can impact AI and XAI decisions and how to mitigate those impacts efficiently.** The final goal is to enable the transparency of the AI-HW by designing self-explainable, trustworthy, reliable, and real-time verifiable AI hardware accelerators, capable of performing self-test, self-diagnosis, and self-correction.

Main activities

More in details, the Ph.D. student will:

- **analyze the possible failure mechanisms** affecting the hardware;
- from the knowledge of failure mechanisms, **derive the corresponding hardware faults** (i.e., the logical representation of a failure mechanism);
- **analyze their impact on AI and XAI results**, in terms of accuracy degradation and determine their criticality;

- **design low-cost fault tolerance approaches** to efficiently detect/correct HW faults, thus ensuring the correctness of the hardware, with the goal to ensure a both correct AI and XAI decisions.

A possible approach to fault tolerance is to apply XAI techniques to produce explanations about the state of the hardware during inference and turn these explanations into actions to correct hardware faults. This Ph.D. subject targets the study of the impact of HW faults on both prototypes created by self-explainable models at training time and post-explanations at inference time. The starting point will be existing state-of-the-art AI HW accelerators optimized for energy efficiency and the outcome will be fault-tolerant versions, still energy efficient.

Skills

Required technical skills:

- Good knowledge of computer architectures and embedded systems
- Machine Learning (pytorch/tensorflow)
- HW design: VHDL/Verilog basics, HW synthesis flow
- Basic programming knowledge (C/C++, python)
- 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 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 2200 euros

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
  System & Networks (BAP E)
- **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 :** 2025-08-31

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.

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

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.