Offer #2025-08550

Post-Doctorant F/H Post-doc in Hardware-aware Neural Architecture Optimization on the Edge

Contract type : Fixed-term contract

Level of qualifications required : PhD or equivalent

Fonction : Post-Doctoral Research Visit

Level of experience : Recently graduated

About the research centre or Inria department

The Inria University of Lille centre, created in 2008, employs 360 people including 305 scientists in 15 research teams. Recognised for its strong involvement in the socio-economic development of the Hauts-De-France region, the Inria University of Lille centre pursues a close relationship with large companies and SMEs. By promoting synergies between researchers and industrialists, Inria participates in the transfer of skills and expertise in digital technologies and provides access to the best European and international research for the benefit of innovation and companies, particularly in the region.For more than 10 years, the Inria University of Lille centre has been located at the heart of Lille's university and scientific ecosystem, as well as at the heart of Frenchtech, with a technology showroom based on Avenue de Bretagne in Lille, on the EuraTechnologies site of economic excellence dedicated to information and communication technologies (ICT).

Context

Within the framework of a partnership (you can choose between)

- European fund :The CPER Cornelia project is co-financed by the European Union with the European Regional Development Fund, by the State and the Hauts de France Region within the framework of the State-Region Plan Contract.

Is regular travel foreseen for this post ?“Do not hesitate to make this known and to ensure that "travel expenses are covered within the limits of the scale in force".

Assignment

Deep Neural Networks (DNN) and hardware accelerators are both leading forces for the recent progress in Edge AI. On the one hand, a new neural architectural paradigm is proposed each month, striving for more accuracy and efficiency. On the other hand, the hardware market has shifted towards designing devices that ensure both flexibility and generality for less energy demands while satisfying the user experience with less latency.
When DNN models are implemented on resource-constrained systems (e.g., edge computing), it becomes inevitable to meticulously optimize them to strike the optimal balance between accuracy, execution latency and energy efficiency. In order to address this particular difficulty, our objective in this project is to tackle Hardware-aware Neural Architecture Search (HW-aware NAS) as a new AutoML paradigm, targeting edge systems. HW-aware NAS incorporates hardware efficiency as an additional optimization objective during the neural architecture design space exploration.

Main objectives of the project:

- New multi-objective performance surrogates: In our previous work, we have widely used two types of accuracy estimation strategies: Predictive Models and Weight-sharing Supernetworks. Several other methods, including zero-shot estimation and learning-curve extrapolation have recently emerged. However, these methods still face certain limitations in terms of multi-objectivity, scalability, and accuracy.
- - Search Algorithms and Large Search Spaces: It is crucial to develop efficient and scalable search algorithms that can effectively explore large heterogenous search spaces within practical time constraints. This would allow for the discovery of highly optimized architectures that align with the
hardware constraints of edge devices, while still meeting performance requirements. Exploring spaces with new approaches such quantum-inspired search algorithms holds promise in tackling the challenges of searching large spaces more efficiently. These algorithms have the potential to enhance the search process and expedite the discovery of optimal architectures for edge

computing. Large Language Model (LLM) based search algorithms for HW-NAS are another interesting approach that will be explored in the project.

- - Multi-task and multi-modality NN investigation: Multi-task deep learning models are crucial to reducing the memory occupancy and execution time especially for edge devices. Investigating how sharing knowledge and architectural components across multiple related tasks can lead to more efficient and effective neural architectures. This approach exploits shared representations to enhance model performance and reduce resource requirements. In the same context we will explore HW-NAS for Multimodal Neural Networks (MM-NN). These NN have the ability to effectively process and integrate multiscale information from diverse data sources.

Main activities

Main objectives of the project:

- New multi-objective performance surrogates: In our previous work, we have widely used two types of accuracy estimation strategies: Predictive Models and Weight-sharing Supernetworks. Several other methods, including zero-shot estimation and learning-curve extrapolation have recently emerged. However, these methods still face certain limitations in terms of multi-objectivity, scalability, and accuracy.
- - Search Algorithms and Large Search Spaces: It is crucial to develop efficient and scalable search algorithms that can effectively explore large heterogenous search spaces within practical time constraints. This would allow for the discovery of highly optimized architectures that align with the
hardware constraints of edge devices, while still meeting performance requirements. Exploring spaces with new approaches such quantum-inspired search algorithms holds promise in tackling the challenges of searching large spaces more efficiently. These algorithms have the potential to enhance the search process and expedite the discovery of optimal architectures for edge computing. Large Language Model (LLM) based search algorithms for HW-NAS are another interesting approach that will be explored in the project.
- - Multi-task and multi-modality NN investigation: Multi-task deep learning models are crucial to reducing the memory occupancy and execution time especially for edge devices. Investigating how sharing knowledge and architectural components across multiple related tasks can lead to more efficient and effective neural architectures. This approach exploits shared representations to enhance model performance and reduce resource requirements. In the same context we will explore HW-NAS for Multimodal Neural Networks (MM-NN). These NN have the ability to effectively process and integrate multiscale information from diverse data sources.

Skills

Required qualifications
· A doctoral degree in Computer Science, Computer Engineering or Electrical Engineering.
· A good background in the domain of AI, Edge Computing, Optimization, GPU/FPGA/Multi-core platforms.
· A good experience in SW development such as PyTorch or TensorFlow.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods
Scientific computing (BAP E)
- **Town/city :** Lille
- **Inria Center :** Centre Inria de l'Université de Lille
- **Starting date :** 2025-04-01
- **Duration of contract :** 1 year, 6 months
- **Deadline to apply :** 2025-03-21

Contacts

- **Inria Team :** BONUS
- **Recruiter :**
Talbi El-ghazali / El-Ghazali.Talbi@inria.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

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.