Offer #2024-07176

Post-Doctoral Research Visit F/M Model placement in inference delivery networks

Contract type : Fixed-term contract

Level of qualifications required : PhD or equivalent

Fonction : Post-Doctoral Research Visit

About the research centre or Inria department

The Inria Saclay-Île-de-France Research Centre was established in 2008. It has developed as part of the Saclay site in partnership with Paris-Saclay University and with the Institut Polytechnique de Paris since 2021.

The centre has 39 project teams , 27 of which operate jointly with Paris-Saclay University and the Institut Polytechnique de Paris. Its activities occupy over 600 scientists and research and innovation support staff, including 54 different nationalities.

Context

This PostDos is funded by the challenge Inria-Nokia Bell Labs: LearnNet (Learning Networks)

Assignment

Assignments :

In this postdoc, we will study the problem of AI model placement in an IDN. This is a challenging optimization problem that involves a non-trivial tradeoff between model effectiveness, inference latency, and resource availability while also dealing with the natural dynamicity of the network, e.g., due to users' request process or changes in available computing and communication resources.

We will also consider other metrics, such as energy consumption, in the objective functions and networking constraints for systems where the network presents some inelasticity (see also [1]). We will leverage multi-objective optimization techniques (e.g., Pareto efficient solutions) and transfer learning techniques to adapt models across nodes with different levels of knowledge and resource availability. We will also rely on online learning approaches to achieve model placements with adversarial guarantees regarding regret.

In comparison to our preliminary work in [2] or [3], we will allow models to be split across multiple nodes [4,5,6]. In particular, we aim to compare specific model splitting techniques, with or without the insertion of bottlenecks [7,8] (reference [8] is also the result of NEO-AIRL cooperation), in terms of performance metrics like inference delay and network load. We will evaluate different methodologies to estimate online the quality of an inference [9].

This evaluation may also consider scenarios with significant heterogeneity of the nodes, such as in the scenario of embedded Edge AI or even more with TinyML (resources possibly lower by orders of magnitude but possibly a massive number of devices).

Collaboration :

This postdoc will be recruited and hosted at Inria Saclay and supervised by Tribe (INRIA), Neo (INRIA), and AIRL (Nokia)

References :

[1] Kinda Khawam et al. "Edge Learning as a Hedonic Game in LoRaWAN". ICC 2023 - IEEE International Conference on Communications. 2023.

[2] Tareq Si Salem et al. "Towards inference delivery networks: distributing machine learning with optimality guarantees." In: 19th Mediterranean Communication and Computer Networking Conference (MEDCOMNET 2021). Ibiza (virtual), Spain: IEEE, June 2021, pp. 1–8.

[3] Wassim Seifeddine, Cédric Adjih, and Nadjib Achir. “Dynamic Hierarchical Neural Network Offloading in IoT Edge Networks.” In: PEMWN 2021 - 10th IFIP International Conference on Performance Evaluation and Modeling in Wireless and Wired Networks. Ottawa / Virtual, Canada: IEEE, Nov. 2021, pp. 1–6.

[4] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. “Branchynet: Fast inference via early exiting from deep neural networks”. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016, pp. 2464-2469

[5] S. Teerapittayanon, B. McDanel, and H. T. Kung. “Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices”. In: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). ISSN: 1063-6927. June 2017, pp. 328–339.

[6] Yoshitomo Matsubara, Marco Levorato, and Francesco Restuccia. “Split Computing and Early Exiting for Deep Learning Applications: Survey and Research Challenges”. In: ACM Computing Surveys 55.5 (Dec. 2022), 90:1–90:30. issn:0360-0300.

[7] Yoshitomo Matsubara et al. “BottleFit: Learning Compressed Representations in Deep Neural Networks for Effective and Efficient Split Computing”. English. In: IEEE Computer Society, June 2022, pp. 337–346. isbn: 978-1-66540-876-9.

[8] Gabriele Castellano et al. “Regularized Bottleneck with Early Labeling”. In: ITC 2022 - 34th International Teletraffic Congress. Shenzhen, China, Sept. 2022.

[9] Ira Cohen and Moises Goldszmidt. “Properties and benefits of calibrated classifiers”. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer, 2004, pp. 125–136.

Main activities

- Read and synthesize literature work,
- Conducting cutting-edge research at the intersection of networking and AI
- Propose novel approaches and technical solutions for AI model placement
- Writing research papers for submission to top-tier conferences and journals in networking, AI, and computer science.
- Disseminating research findings through presentations at conferences, seminars, and workshops.

Skills

- A solid understanding of networking principles, protocols, and architectures is essential.
- Proficiency in programming languages commonly used in AI and networking research.
- Experience with relevant libraries and frameworks is also valuable.
- Ability to design and implement algorithms for solving complex problems.
- Familiarity with optimization techniques.
- Excellent written and verbal communication skills for presenting research findings, writing academic papers, and collaborating with peers.
- The ability to work effectively as part of a research team, collaborate with colleagues from diverse backgrounds, and contribute positively to group dynamics

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 :** Networks and Telecommunications
  System & Networks (BAP E)
- **Town/city :** Palaiseau
- **Inria Center :** Centre Inria de Saclay
- **Starting date :** 2024-06-01
- **Duration of contract :** 1 year, 6 months
- **Deadline to apply :** 2024-11-30

Contacts

- **Inria Team :** TRIBE
- **Recruiter :**
  Achir Nadjib / Nadjib.Achir@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.