Offer #2024-07823

Post-Doctoral Research Visit F/M Cooperative Inference Strategies

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 centre at Université Côte d'Azur includes 42 research teams and 9 support services. The centre's staff (about 500 people) is made up of scientists of di?erent nationalities, engineers, technicians and administrative staff. The teams are mainly located on the university campuses of Sophia Antipolis and Nice as well as Montpellier, in close collaboration with research and higher education laboratories and establishments (Université Côte d'Azur, CNRS, INRAE, INSERM ...), but also with the regiona economic players.

With a presence in the fields of computational neuroscience and biology, data science and modeling, software engineering and certification, as well as collaborative robotics, the Inria Centre at Université Côte d'Azur is a major player in terms of scientific excellence through its results and collaborations at both European and international levels.

Context

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

Researchers involved

At Inria: Giovanni Neglia, Chuan Xu, Aurélien Bellet

At Nokia: Fabio Pianese, Calvin Chen, Tianzhu Zhang

Assignment

Introduction
An increasing number of applications rely on complex inference tasks based on

machine learning (ML). Currently, two options exist to run such tasks: either served directly by the end device (e.g., smartphones, IoT equipment, smart vehicles) or offloaded to a remote cloud. Both options may be unsatisfactory for many applications: local models may have inadequate accuracy, while the cloud may fail to meet delay constraints. In [SSCN+24], researchers from the Inria NEO and Nokia AIRL teams presented the novel idea of inference delivery networks (IDNs), networks of computing nodes that coordinate to satisfy ML inference requests achieving the best trade-off between latency and accuracy. IDNs bridge the dichotomy between device and cloud execution by integrating inference delivery at the various tiers of the infrastructure continuum (access, edge, regional data center, cloud). Nodes with heterogeneous capabilities can store a set of monolithic machine-learning models with different computational/memory requirements and different accuracy and inference requests that can be forwarded to other nodes if the local answer is not considered accurate enough.

Research goal

Given an AI model's placement in an IDN, we will study inference delivery strategies to be implemented at each node in this task. For example, a simple inference delivery strategy is to provide the inference from the local AI model if this seems to be accurate enough or to forward the input to a more accurate model at a different node if the inference quality improvement (e.g., in terms of accuracy) compensates for the additional delay or resource consumption. Besides this serve-locally-or-forward policy, we will investigate more complex inference delivery strategies, which may allow inferences from models at different clients to be combined. To this purpose, we will rely on ensemble learning approaches [MS22] like bagging [Bre96] or boosting [Sch99], adapting them to IDN distinct characteristics. For example, in an IDN, models may or may not be trained jointly, may be trained on different datasets, and have different architectures, ruling out some ensemble learning techniques. Moreover, queries to remote models incur a cost, which leads to prefer ensemble learning techniques that do not require joint evaluation of all available models.

In an IDN, models could be jointly trained on local datasets using federated learning algorithms [KMA+21]. We will study how the selected inference delivery strategy may require changes to such algorithms to consider the statistical heterogeneity induced by the delivery strategy itself. For example, nodes with more sophisticated models will receive inference requests for difficult samples from nodes with simpler and less accurate models, leading to a change in the data distribution seen at inference with respect to that of the local dataset. Some preliminary results about the training for early-exit networks in this context are in [KSR+24].

References

[Bre96] Leo Breiman. Bagging predictors. Machine Learning, 24(2):123–140, August 1996.

[KMA+21] Peter Kairouz et al, Advances and Open Problems in Federated Learning. Foundations and Trends® in Machine Learning, 14(1–2):1–210, 2021.

[KSR+24] Caelin Kaplan, Tareq Si Salem, Angelo Rodio, Chuan Xu, and Giovanni Neglia. Federated learning for cooperative inference systems: The case of early exit networks, 2024.

[MS22] Ibomoiye Domor Mienye and Yanxia Sun. A Survey of Ensemble

Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access, 10:99129–99149, 2022.
[Sch99] Robert E. Schapire. A brief introduction to boosting. In Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2, IJCAI’99, pages 1401–1406, San Francisco, CA, USA, July 1999. Morgan Kaufmann Publishers Inc.
[SSCN+24] T. Si Salem, G. Castellano, G. Neglia, F. Pianese and A. Araldo, "Toward Inference Delivery Networks: Distributing Machine Learning With Optimality Guarantees," in IEEE/ACM Transactions on Networking, vol. 32, no. 1, pp. 859-873, Feb. 2024

Main activities

Research.

If the selected candidate is interested, he/she may be involved in students' supervision (master and PhD level) and teaching activities.

Skills

Candidates must hold a Ph.D. in Applied Mathematics, Computer Science or a closely related discipline. Candidates must also show evidence of research productivity (e.g. papers, patents, presentations, etc.) at the highest level.

We prefer candidates who have strong mathematical background (on optimization, statistical learning or privacy) and in general are keen on using mathematics to model real problems and get insights. The candidate should also be knowledgeable on machine learning and have good programming skills. Previous experiences with PyTorch or TensorFlow is a plus.

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 (after 6 months of employment) 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

Remuneration

Gross Salary: 2788 € per month

General Information

- **Theme/Domain :** Optimization, machine learning and statistical methods System & Networks (BAP E)
- **Town/city :** Sophia Antipolis
- **Inria Center :** Centre Inria d'Université Côte d'Azur
- **Starting date :** 2025-10-01
- **Duration of contract :** 1 year, 6 months
- **Deadline to apply :** 2025-09-13

Contacts

- **Inria Team :** NEO
- **Recruiter :**
  Neglia Giovanni / Giovanni.Neglia@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.

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

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.