Offer #2023-06479

PhD Position F/M Distributed Training of Heterogeneous Architectures

Contract type: Fixed-term contract

Level of qualifications required: Graduate degree or equivalent

Function: PhD Position

About the research centre or Inria department

Inria is a national research institute dedicated to digital sciences that promotes scientific excellence and transfer. Inria employs 2,400 collaborators organised in research project teams, usually in collaboration with its academic partners. This agility allows its scientists, from the best universities in the world, to meet the challenges of computer science and mathematics, either through multidisciplinarity or with industrial partners. A precursor to the creation of Deep Tech companies, Inria has also supported the creation of more than 150 start-ups from its research teams. Inria effectively faces the challenges of the digital transformation of science, society and the economy.

Context

This PhD thesis is in the framework of the Inria-Nokia Bell Labs research initiative on Federated Learning for Cellular Networks and in closer relation with Inria research initiative on Federated Learning, FedMalin https://project.inria.fr/fedmalin/.

The PhD candidate will join Nokia Bell Labs research team in Massy, France https://www.bell-labs.com/about/locations/paris-saclay-france/ and will also be a member of the Inria project-team NEO https://team.inria.fr/neo/.

The Machine Learning & Systems team, part of the AI Research Lab at Nokia Bell Labs, is composed of computer scientists and data engineers who develop AI-based systems and algorithms bridging the gap between the promise of limitless capabilities of AI and the constraints imposed by real computing and communication systems. NEO is positioned at the intersection of Operations Research and Network Science. By using the tools of Stochastic Operations Research, the team members model situations arising in several application domains, involving networking in one way or the other.

The research activity will be supervised by
  * Chung Shue Chen (chung_shue.chen@nokia-bell-labs.com)
  * Fabio Pianese (fabio.pianese@nokia-bell-labs.com)
  * Giovanni Neglia (giovanni.neglia@inria.fr)

Assignment

The increasing size of data generated by smartphones and IoT devices motivated the development of Federated Learning (FL) [LKT+20, KMA+21], a framework for on-device collaborative training of machine learning models. FL algorithms like FedAvg [MMR+17] allow clients to train a common global model without sharing their personal data. FL reduces data collection costs and can help to mitigate data privacy issues, making it possible to train models on large datasets that would otherwise be inaccessible. FL is currently used by many big tech companies (e.g., Google, Apple, Facebook) for learning on their users' data, but the research community envisions also promising applications to learning across large data-silos, like hospitals that cannot share their patients' data [RHL20].

Most existing algorithms for federated learning train the same model architecture for each user/device (personalized FL algorithms allow only the value of model parameters to be different). In this task, we will explicitly consider that heterogeneous devices may not be able to run the same model (because of computational, memory, or battery constraints) and will propose new algorithms to train heterogeneous architectures jointly. A key challenge, in this case, is to design meaningful ways of sharing information across heterogeneous model architectures.

We will consider two different approaches. The first relies on model subsampling: there is a single neural network architecture. However, while powerful clients store and use the complete model, less powerful clients only store a subset of the model matching their memory and computation capabilities. This approach has been proposed in [DDT21, HLA+21] for convolutional networks, where each client may use a different number of channels. We will explore how much this strategy can be extended to different neural network architectures. We will also address open issues about training with subsampling. For example, results in [HLA+21] suggest that, at some training iterations, more powerful clients should
behave as less powerful clients and update a subset of the architecture. We are missing a solid explanation for these observations and quantitative guidelines about how often clients should train a model different from the largest one they can train. The response likely depends on the distribution of clients’ capabilities and their dataset sizes.

The second approach we will consider is based on knowledge transfer techniques. For example, reference [LKSJ20] proposes that a central node (the FL aggregator) maintains a model for every possible architecture in the system and uses ensemble distillation to transfer information across the models. While the aggregator is, in general, much more powerful than the FL clients, it may still be the system bottleneck if clients exhibit a large degree of heterogeneity. Then a large number of different models need to be stored in memory and jointly trained at the aggregator. Another limit of [LKSJ20] is the need for an unlabeled dataset at the aggregator. The approach in [ZH21] replaces the unlabeled dataset with a synthetic one generated by a generative model. Still, this model must be jointly learned and may potentially leak private information as clients need to reveal their label distribution. We plan to address the limits of the existing approaches by proposing fully decentralized training algorithms, where each client can transfer information from its model to the models of less powerful clients in its neighborhood and may use to this purpose its local dataset.


Main activities

Research

Skills

The candidate should have a solid mathematical background (in particular on optimization) and in general be keen on using mathematics to model real problems and get insights. He should also be knowledgeable on machine learning, have good programming skills and previous experiences with PyTorch or TensorFlow.

We expect the candidate to be fluent in English.

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
- Contribution to mutual insurance (subject to conditions)

Remuneration

Gross Salary per month: 2051€ brut per month (year 1 & 2) and 2158€ brut per month (year 3)

General Information

- Theme/Domain: Networks and Telecommunications
- System & Networks (BAP E)
- Town/city: Paris
- Inria Center: Centre Inria d’Université Côte d’Azur
- Starting date: 2023-10-01
Duration of contract: 3 years
Deadline to apply: 2024-01-28

Contacts
- Inria Team: NEO
- PhD Supervisor: 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.

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.