Offer #2023-06139

Post-Doctoral Research Visit F/M Distributed Machine Learning at the Network Edge

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 37 research teams and 8 support services. The centre's staff (about 500 people) is made up of scientists of different 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

The position is in the framework of dAIEDGE---A network of excellence for distributed, trustworthy, efficient and scalable AI at the Edge---funded by the European Union.

The vision of the dAIEDGE Network of Excellence is to strengthen and support the development of the dynamic European edge AI ecosystem under the umbrella of the European AI Lighthouse and to sustain the advanced research and innovation of distributed AI at the edge as essential digital, enabling, and emerging technology in an extensive range of industrial sectors.

The candidate will work with NEO Inria team (https://team.inria.fr/neo/) and COATI Inria team (https://team.inria.fr/coati/), and in particular with

- Giovanni Neglia
- Chuan Xu
- Frédéric Giroire

Assignment

The Internet was conceived to enable computer resources' time-sharing, but soon its main function became to deliver content to end users, but it is now called to play a new key role: to pervasively support machine learning (ML) operation both for model training and prediction serving.

There are two aspects calling for Internet-wide deployment of ML systems. First, data---one key ingredient of ML success---is often generated by users and devices at the edge of the network. The classic ML operation in the cloud requires such data to be collected at a single computing facility where training occurs. Data aggregation can be very costly, or simply impossible because of capacity constraints, privacy issues, or ownership ones. These scenarios call for distributed learning systems, where computation moves, at least in part, to the data. For example, Google's federated learning [mcmahan17,kairouz21] enables mobile phones, or other devices with limited computing capabilities, to collaboratively learn an ML model while keeping all training data locally. Distributed ML training is already a difficult task in a cluster setting. Indeed, optimization techniques, distributed systems, and ML models are a triad difficult to untangle: e.g., relaxed state consistency across computing nodes increases system throughput but may jeopardize convergence of the optimization algorithm or affect the final solution selected, leading to models with very different generalization capabilities [chen16]. Additional challenges arise when training moves to the Internet. First, the system potentially scales up to billions of devices, against at most thousands of GPUs to break ML training records in a cluster. Second, local datasets are highly heterogeneous with very different sizes and feature/label distributions. Third, devices may have very different hardware and connectivity. Fourth, communications are often unreliable (devices can be switched off at any time), slow (latencies are 2 orders of magnitude larger), and expensive for battery-constrained devices. Fifth, privacy concerns are often important and limit the operations that can be performed during training to avoid inadvertently disclosing sensible information. Finally, training is more vulnerable to malicious attacks. For all these reasons, federated learning (as ML training over the Internet is now usually called) has emerged in the last years as a specific research topic—well

distinct for example from high-performance computing or cloud computing—at the intersection of machine learning, optimization, distributed systems, and networking.

The second driver to distribute ML processes over the Internet is real-time inference. In fact, ML models are often trained for inference's purposes, i.e., to make predictions on new data. Model predictions need then to be served to the final users. ML training is a computationally expensive operation and is the object of much research effort. Inference does not involve complex iterative algorithms and is therefore generally assumed to be easy, but it also presents fundamental challenges that are likely to become dominant as ML adoption increases [stoica17]. AI systems will be ubiquitously deployed and will need to make timely and safe decisions in unpredictable environments. In this case, inference must run in real-time, and predictions may need to be served at a very high rate. The big cloud players—Amazon, Microsoft, and Google—have all started pushing their "machine learning as a service" (MLaaS) solutions. Running the models in the cloud guarantees high scalability, but may fail to meet delay constraints. As an example, already deployed applications, such as recommendation systems, voice assistants, and ad-targeting, need to serve predictions from ML models in less than 20 ms [simsek16]. Future wireless services, such as connected and autonomous cars, industrial robotics, mobile gaming, augmented/virtual reality, have even stricter latency requirements, often below 10 ms and below 1 ms for the so-called tactile internet. It is then imperative to run these services closer to the user at the network edge. 5G deployment can provide computing and storage capabilities at the edge, but those will still be very limited in comparison to the cloud and need to be wisely used. In conclusion, inference will require complex resource orchestration across users' devices, edge computing servers, and the cloud.

We are looking for a postdoc candidate who could join our team to work on one ore more of the following topics (for which we provide pointers to our publications)
* Distributed Inference [sisalem21b,castellano22]
* Online Learning Algorithms with Regret Guarantees [sisalem21a,li22,sisalem23]
* Distributed/Federated Learning
[neglia19,neglia20,marfoq20,marfoq21,xu21b,marfoq22,ogier22,rodio23,marfoq23]
* Machine Learning Privacy [xu21a,zari21,driouich22]

We expect the postdoc to actively participate to the activities of the EU project dAIEDGE (e.g., attending meetings, coordinating Inria contribution to deliverables).
The postdoc will also have the opportunity to collaborate with PhD students working on the topics listed above.

References

[castellano22] Gabriele Castellano, Fabio Pianese, Damiano Carra, Tianzhu Zhang, Giovanni Neglia, Regularized Bottleneck with Early Labeling, ITC 2022 - 34th International Teletraffic Congress, Shenzhen, China, September 14-16, 2022

[chen16] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, Rafal Jozefowicz, Revisiting Distributed Synchronous SGD, arXiv:1604.00981

[driouich22] Ilias Driouich, Chuan Xu, Giovanni Neglia, Frederic Giroire, and Eoin Thomas, A Novel Model-Based Attribute Inference Attack in Federated Learning, International Workshop on Federated Learning: Recent Advances and New Challenges in Conjunction with NeurIPS 2022 (FL-NeurIPS'22)

[kairouz21] P. Kairouz, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), pp. 1-210, 2021.

[li22] Yuanyuan Li, Tareq Si Salem, Giovanni Neglia, Stratis Ioannidis, Online Caching Networks with Adversarial Guarantees, ACM SIGMETRICS / IFIP PERFORMANCE 2022, Mumbai, India June 6-10, 2022

[marfoq20] Throughput-Optimal Topology Design for Cross-Silo Federated Learning,
Othman Marfoq, Chuan Xu, Giovanni Neglia, and Richard Vidal, Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020), December 6-12, online conference

[marfoq21] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal, Federated Multi-Task Learning under a Mixture of Distributions, Proc. of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)

[marfoq22] Othmane Marfoq, Laetitia Kameni, Richard Vidal, Giovanni Neglia, Personalized Federated Learning through Local Memorization, International Conference on Machine Learning (ICML), July, 2022

[marfoq23] O. Marfoq, G. Neglia, L. Kameni, R. Vidal, Federated Learning for Data Streams, AISTATS 2023

[mcmahan17] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. Communication efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics, PMLR, 2017.

[neglia19] G. Neglia, G. Calbi, D. Towsley, G. Vardoyan, The Role of Network Topology for Distributed Machine Learning, Proc. of the IEEE International Conference on Computer Communications (INFOCOM 2019), Paris, France, April 29 - May 2, 2019

[neglia20] Giovanni Neglia, Chuan Xu, Don Towsley, and Gianmarco Calbi, Decentralized gradient methods: does topology matter?, 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). Palermo, Italy, June 2020

[ogier22] Jean Ogier du Terrail et al, FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings, Thirty-Sixth Conference on Neural Information Processing Systems

(NeurIPS 2022) Track Datasets and Benchmarks, November 29-December 1, 2022

[rodio23] Angelo Rodio, Francescomaria Faticanti, Othmane Marfoq, Giovanni Neglia, Emilio Leonardi, Federated Learning under Heterogeneous and Correlated Client Availability, IEEE Conference on Computer Communications (INFOCOM), Hoboken, New Jersey, USA, May 17-20, 2023

[simsek16] M. Simsek, A. Aijaz, M. Dohler, J. Sachs and G. Fettweis, "5G-Enabled Tactile Internet," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 460-473, March 2016

[sisalem21a] Tareq Si Salem, Giovanni Neglia, and Stratis Ioannidis, No-Regret Caching via Online Mirror Descent, Proc. of IEEE International Converence on Communications (ICC), June 14-23, 2021

[sisalem21b] Tareq Si Salem, Gabriele Castellano, Giovanni Neglia, Fabio Pianese, and Andrea Araldo, Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees, Proc. of the 19th Mediterranean Communication and Computer Networking Conference (MedComNet 2021), online conference, June 15-17, 2021

[sisalem23] Tareq Si Salem, George Iosifidis, Giovanni Neglia, Enabling Long-term Fairness in Dynamic Resource Allocation,ACM SIGMETRICS 2023, Orlando, Florida, USA, June 19-23, 2023

[stoica17] Ion Stoica, Dawn Song, Raluca Ada Popa, David A. Patterson, Michael W. Mahoney, Randy H. Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph Gonzalez, Ken Goldberg, Ali Ghodsi, David E. Culler, and Pieter Abbeel. A Berkeley View of Systems Challenges for AI. Tech. rep. UCB/EECS-2017-159. EECS Department, University of California, Berkeley, Oct. 2017.

[xu21a] Chuan Xu and Giovanni Neglia, What else is leaked when eavesdropping Federated Learning?, ACM CCS workshop Privacy Preserving Machine Learning (PPML), online, November 19, 2021

[xu21b] Chuan Xu, Giovanni Neglia, and Nicola Sebastianelli, Dynamic Backup Workers for Parallel Machine Learning, Elsevier Computer Networks, 2021

[zari21] Oualid Zari, Chuan Xu, and Giovanni Neglia, Efficient Passive Membership Inference Attack in Federated Learning, NeurIPS workshop on Privacy in Machine Learning (PriML) 2021

Main activities

Beside carrying out high quality research, we expect the postdoc to actively participate to the activities of the EU project dAIEDGE (e.g., attending meetings, coordinating Inria contribution to deliverables).

The postdoc will also have the opportunity to collaborate with PhD students working on the topics listed above.

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.

The position is for 18 months, but it can be extended up to 30 months.

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: 2746 € 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 :** 2023-09-01
- **Duration of contract :** 1 year, 4 months
- **Deadline to apply :** 2024-05-31

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.

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.