Internship on "Decentralized Clustered Federated Learning"

Supervisors

- Giovanni Neglia, Inria, France, giovanni.neglia@inria.fr, http://www-sop.inria.fr/members/Giovanni.Neglia/
- Emilio Leonardi, Politecnico degli Studi di Torino, Italy, emilio.leonardi@polito.it, https://www.telematica.polito.it/member/emilio-leonardi/

Location

Inria NEO team (https://team.inria.fr/neo/)

2004 route des Lucioles,

06902 Sophia Antipolis, France

Context

The increasing size of data generated by smartphones and IoT devices motivated the development of Federated Learning (FL) [1,2], a framework for on-device collaborative training of machine learning models. FL algorithms like FedAvg [3] allow clients to train a common global model without sharing their personal data. FL reduces data collection costs and protects clients' data privacy and, in doing so, makes possible to train models on large datasets that would otherwise have been inaccessible. FL is currently used by many big data 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 [4].

One of the main scientific challenges of FL, in comparison to other forms of distributed learning, is statistical heterogeneity, i.e., the fact that clients' local datasets are in general drawn from different distributions. Statistical heterogeneity for example slows down the convergence of FL algorithms [5]. Moreover, first efforts in FL focused on learning a single global model with good average performance across clients, but the global model may be arbitrarily bad for a given client, if its local dataset distribution is significantly different from the other distributions. The dissatisfied client may then abandon the training procedure (or refuse to join it in the future), impoverishing the aggregate pool of data and then the quality of the final model. Defections of some clients can then potentially trigger a cascade of defections as clients are less and less satisfied with the model learned by FL algorithms.

A possible way to address statistical heterogeneity is to let clients jointly learn personalized models adapted to their local distributions [5-8]. A popular approach in this direction is Clustered FL [9-11]. Clustered FL groups similar clients in separated clusters and learn a different model for each cluster. Existing Clustered FL algorithms require the presence of a central server that communicates with all clients and runs the clustering algorithm. This client-server architecture may not scale to a large number of clients; new decentralized P2P-like approaches can reduce the total training time [12] as well as provide stronger privacy guarantees [13].

Goals

This internship aims to propose new decentralized algorithms for clustered FL in a setting where each client can only interact with a small set of neighbours and can progressively modify this set to discover the most similar clients (those who belong to the same cluster). A key aspect is that clients only have access to noisy evaluations of similarities (for example the original Clustered FL relies on the cosine similarity of stochastic gradients computed at clients). Also, cluster discovery happens in parallel to FL training of the machine learning model, it is then important to (probabilistically) quantify the timescale required to learn the cluster and compare it with the training timescale. The internship will start with the theoretical analysis of a simplified model similar to the one in [14], where each client wants to learn the average of its local distribution. It will then move to evaluate the proposed algorithms to classic FL benchmarks using machine learning frameworks like PyTorch.

Pre-requisites

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 have a strong background on graphs and probability and good programming skills. A background on optimization and machine learning would be a plus.

How to apply

Send by email to giovanni.neglia@inria.fr your cv and the list of university courses with the corresponding marks.

References

[1] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37 (3), 2020.

[2] Peter Carouse, H Brendan McMahan, Brendan Avent, Aurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, 2019.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics, PMLR, 2017.

[4] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119, 2020.

[5] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the Convergence of FedAvg on Non-IID Data,” in International Conference on Learning Representations, 2019.

[5] Othmane Marfoq, Giovanni Neglia, Aurelien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions, NeurIPS 2021.

[6] Othmane Marfoq, Giovanni Neglia, Laetitia Kameni, Richard Vidal, Personalized Federated Learning through Local Memorization, ICML 2022.

[7] Valentina Zantedeschi, Aurelien Bellet, and Marc Tommasi. Fully decentralized joint learning of personalized models and collaboration graphs. volume 108 of Procexedings of Machine Learning Research, pages 864-874, 2020. PMLR.

[8] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In International Conference on Machine Learning, pages 6357-6368. PMLR, 2021.

[9] Sattler, Felix, Klaus-Robert Müller, and Wojciech Samek. "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints."IEEE transactions on neural networks and learning systems, 2020.

[10] Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An Efficient Framework for Clustered Federated Learning. Advances in Neural Information Processing Systems, 33, 2020.

[11] Mansour, Yishay, et al. Three approaches for personalization with applications to federated learning. arXiv preprint arXiv:2002.10619, 2020.

[12] Othmane Marfoq, Chuan Xu, Giovanni Neglia, Richard Vidal, Throughput-Optimal Topology Design for Cross-Silo Federated Learning, NeurIPS, 2020.

[13] Edwige Cyffers, Aurélien Bellet. Privacy Amplification by Decentralization. AISTATS, 2022.

[14] Mahsa Asadi, Aurélien Bellet, Odalric-Ambrym Maillard, and Marc Tommasi. Collaborative Algorithms for Online Personalized Mean Estimation. Transactions on Machine Learning Research, 2022.