Digital Systems for Humans Graduate School

2024-2025 PhD Subject Proposition

Proposition de Sujet de Thèse 2024-2025

Decentralized machine learning

Apprentissage statistique décentralisé

Doctoral School: Doctoral School in Fundamental and Applied Sciences (ED SFA)

Thesis supervisor: Cédric Richard (primary), Roula Nassif (co-supervisor)

Host laboratory: J.-L. Lagrange, SI Team, Bâtiment Fizeau, Parc Valrose, 06108 Nice

Subject description:

The proliferation of modern distributed networks, such as mobile phones, wearable devices, hospitals, autonomous vehicles, and smart homes, has led to the generation of massive amounts of data each day. This surge in data, coupled with concerns about privacy and the limitations of centralized data processing, has driven the adoption of federated and decentralized approaches for training statistical models. In these approaches, each participating device (which is referred to as client, or agent) has a local training dataset which is never uploaded to the server. Training data is kept locally on users' devices, and the devices are used as agents performing computation on their local data in order to update global models of interest. This approach enables the training of models without centralizing sensitive data, addressing privacy concerns while still leveraging the collective knowledge present in distributed datasets. Major companies like Google and Apple have begun integrating such technologies into their products and services. For instance, Google's Gboard keyboard team employs federated learning to enhance next-word prediction on mobile devices [1]. In applications where communication to a server becomes a bottleneck, decentralized topologies (where agents only communicate with their neighboring devices) are potential alternatives to federated topologies (where a central server connects with all remote devices).

The thesis project falls into the broad theme of performing decentralized machine learning. By recognizing the trend towards collecting data in a continuous manner by the devices, the focus will be on developing methods that can effectively handle streaming data in real-time. Moreover, by recognizing that the underlying data-generation models can change over time, the developed approaches must adapt to these changes, ensuring robustness and accuracy in the learning process. Modern machine learning applications often involve heterogeneous data sources and systems. The thesis project needs to address challenges related to statistical heterogeneity, focusing on developing techniques that can handle diverse data distributions and characteristics. The project will also account for the diversity of devices participating in the decentralized learning process, including variations in computational capabilities, memory constraints, and communication protocols. By focusing on these aspects, the goal in this thesis is to develop practical and scalable solutions that can be applied to real-world machine learning applications, thereby addressing the challenges encountered in modern distributed data environments.

In contrast to traditional federated learning approaches that operate under the assumption of a fixed set of clients with static local datasets [1]-[6], the thesis acknowledges the dynamic nature of data collection in modern distributed environments. Here, devices continuously collect data, and the underlying data-generation models evolve over time. As a result, the proposed solutions should adapt to these dynamic conditions and learn continuously from streaming data.

Moreover, in modern machine learning applications, devices generate data in a highly non-identically distributed manner due to variations in user's behavior and device's usage. This poses a challenge as traditional approaches, which assume identically distributed data, may lead to poor model performance. The thesis focus is on statistical heterogeneous networks, where data distributions vary significantly across devices. In such networks, agents may need to estimate and track multiple, distinct tasks simultaneously. This requires the development of flexible algorithms capable of handling multitask learning scenarios efficiently. These algorithms should be able to adapt to the dynamic nature of the tasks and adjust model parameters accordingly. Moreover, in settings where labelled data is scarce or unavailable for some devices, semi-supervised learning becomes essential. Semi-supervised learning systems leverage both labelled and unlabelled data points to improve model performance. Developing effective semi-supervised learning algorithms that can utilize unlabelled data effectively is crucial in such applications.

Finally, it should be noted that the majority of federated optimization algorithms [2]-[6] is still close to centralized settings since it requires a central coordinator. In applications where communication to a server becomes a bottleneck, decentralized topologies are potential alternatives to federated topologies. The thesis focuses on developing decentralized inference approaches and recognizes the challenge associated with the design of such approaches where a global behavior must emerge from local interactions and computations.

By delivering a family of decentralized learning approaches that can handle statistical heterogeneous settings and heterogeneous data acquisition settings, the thesis aims to advance the state-of-the-art in decentralized machine learning. The developed approaches will enable efficient and adaptive learning in dynamic environments while addressing the challenges of data heterogeneity. Moreover, the analyses and experiments conducted throughout the thesis will provide valuable insights into the behavior and performance of decentralized learning methods, contributing to the broader understanding of machine learning in distributed settings.

Scientific environment:

The PhD candidate will benefit from a scientifically rich environment and will be able to acquire a solid background on the most recent results and advances in decentralized machine learning (signal processing and stochastic optimization).

She/he will be mainly advised by:
- Cédric Richard, Professeur, 3IA Chair in Machine Learning, Laboratoire Lagrange, UMR CNRS 7293, Université Côte d'Azur.
- Roula Nassif, Maître de Conférences, Laboratoire I3S, UMR CNRS 7271, Université Côte d'Azur.

International collaborations/visits in the context of the thesis are highly probable. The PhD shall start in the Fall of 2024, with a duration of 3 years.

Applicant profile:

The candidate:
- must be a graduate student with major in applied mathematics, computer or electrical engineering.
- must have a strong background in machine learning as well as good knowledge in signal processing, linear algebra, inverse problems (regularization), and convex optimization.
- must have a good programming experience (Matlab or Python).
- must have high level of written/spoken English.

References:

[1] K. A. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. M. Kiddon, J. Konecny, S. Mazzocchi, B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," In Proc. Conf. Mach. Learn. Syst., Mar. 2019.
[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas. "Communication-efficient learning of deep networks from decentralized data". In Proc. Int. Conf. Artif. Intell. Stat., vol. 54, pp. 1273-1282, 2017.
[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong. "Federated machine learning: Concept and applications". ACM Trans. Intell. Syst. Technol., vol. 10, no. 2, pp.1-19, 2019.
[4] T. Li, A. K. Sahu, A. S. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, pp. 50-60, May 2020.
[5] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, ..., and S. Zhao, "Advances and open problems in federated learning," Found. Trends Mach. Learn., vol. 14, no. 1-2, pp. 1-210, 2021.
[6] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, Dec. 2017.
[7] R. Nassif, S. Vlaski, C. Richard, J. Chen, and A. H. Sayed, "Multitask Learning over Graphs: An approach for distributed, streaming machine learning," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 14-25, 2020.

Candidater / Apply