ASCR Workshop on Cybersecurity and Privacy of Scientific Computing Ecosystems

Position Paper: Large-Scale Resilient Collaborative Machine Learning

Corresponding author: Olugbenga Moses Anubi, Florida State University, oanubi@fsu.edu,

The past few decades have seen a tremendous increase in the volume and complexity of data generated in scientific discovery processes. Moreover, due to the rapid growth in internet and networking technology, it is now common for these experiments to be composed of geographically dispersed components. Each of the components generates and stores a huge dataset which captures only a portion of the global phenomenon in question. This poses a tremendous challenge for data analysis, even with the most advanced Machine Learning/ AI methods. The state-of-the-art approaches to this problem involve either routing data to a trusted central location where the learning task takes place or iteratively performing the learning task over the dispersed data sources. However, in addition to low efficiency issues and high cost, there is often a single point of failure, resulting in low resiliency to faults and adversarial targeting.

Use Case: Large-scale resilient collaborative learning of proprietary heterogenous models over proprietary non collocated datasets while preserving privacy.

This use case involves collaborative learning among a large number of stakeholders. Each stakeholder has private models and datasets that should not be shared with the other stakeholders. However, their datasets only contain partial information. Training on only such datasets will result in systems that don't generalize well. It is therefore essential to leverage the dissimilarity of all datasets to the mutual benefit of all stakeholders involved. **So, the challenge is to develop learning systems and frameworks that will simultaneously train all models using all datasets without violating privacy requirements and provide resiliency against bad actors.** State-of-the-art Federated AI/ML is very good at training one model over non collocated datasets. Privacy preservation can also be guaranteed with some newer variants [1] [2]. However, the fundamental concept and available frameworks do not generalize to this use case. Newer concepts and frameworks need to be developed to address this open, but very important, challenge.

Distributed Learning (DL), Federated Learning (FL) and Collaborative Learning (CL)

Distributed machine learning [3] refers to the techniques and algorithms for training a single model or architecture over a large-scale and/or sparse data sources using distributed systems that enable parallel computation, data distribution, and resilience to failures. Unlike distributed machine learning, which is fairly matured with countless variants and implementations in commercial and open-source frameworks, **federated machine learning** [4] is a relatively new research topic with a lot of open questions. Indeed, the term *federated learning* was introduced only 5 years ago [5]. This research area exists to solve a longstanding goal of large-scale learning left unaddressed by distributed learning: to analyze and learn from data distributed among many owners while respecting proprietary information and data privacy. The term ***collaborative learning*** has been used, on a few occasions, interchangeably with federated learning. However, this term is used here to emphasize the concurrent learning of multiple heterogenous models, which plays a key role in several significant improvements.

Consequently, the above use case stresses the need to develop collaborative learning concepts and frameworks that are universal, resilient, adaptive, robust, fault-tolerant, scalable, trust-worthy and privacy-preserving. Specifically, the following open challenges need to be addressed:

1. Develop, analyze, and validate a plug-and-play universal framework for very large-scale collaborative learning
2. Develop learning methods and new convergence analysis that use connectivity information to speed up convergence robustly and safely
3. Develop methods, analysis, and tools to minimize the data movement bottlenecks inherent with learning over large networks.
4. Develop privacy-preserving knowledge similarity measures to achieve knowledge consensus among diverse heterogenous models in the framework
5. Combine graph theory and operator splitting theory to develop efficient methods to distribute learning tasks over arbitrary networks, resiliently and robustly.
6. Develop graph-theoretic-based metrics to assess the resiliency of distributed learning algorithms
7. Develop methods to assess and quantify the level of privacy and confidentiality protection over a collaborative/federated learning system.

References

[1] Y. Zhang, G. Bai, X. Li, C. Curtis, C. Chen and R. K. Ko, "PrivColl: Practical Privacy-Preserving Collaborative Machine Learning," in *European Symposium on Research in Computer Security*, 2020.

[2] J. So, B. Guler and A. S. Avestimer, "A Scalable Approach for Privacy-Preserving Collaborative Machine Learning," *arXiv preprint arXiv:2011.01963,* 2020.

[3] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys (CSUR),* vol. 53, no. 2, pp. 1-33, 2020.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu and S. Zhao, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning,* vol. 14, no. 1, 2021.

5] B. H. McMahan, E. Moore, D. Ramage, S. Hampson and B. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *20th International Conference on Artificial Intelligence and Statistics (original version on arXiv Feb. 2016)*, 2017.