Reconciling Privacy and Byzantine-Robustness in Federated Learning

Lun Wang*

Google
lunwang@google.com

Introduction

Federated learning (FL), as a distributed machine learning paradigm, offers the promise of collaborative model training while preserving data privacy. However, FL systems remain susceptible to security vulnerabilities stemming from both malicious central servers and compromised or adversarial clients (Byzantine clients). A variety of techniques have been devised to address the threat from either the server or the clients. However, these approaches are incompatible due to technical details. This extended abstract summarizes the work from (Wang et al. 2020) and (Zhu et al. 2023) aimed at reconciling the defenses and enhancing FL robustness in the presence of both threats.

Methodology

Threats from Semi-honest Server & Countermeasures In federated learning, the server is usually considered semi-honest, which means they won't actively violate the computation protocol agreed by the clients in advance but might want to extract more information than needed from the exchanged messages with the clients. To defend against a semi-honest server, FL protocols can be enhanced with secure aggregation technique (Bonawitz et al. 2017) which hides the individual local updates and only reveals the aggregated global update.

Threats from Malicious Clients & Countermeasures In addition to mitigating threats from a semi-honest server, recent attacks have shown that a small number of clients can behave maliciously in a large-scale FL system with thousands of clients and stealthily influence the jointly-trained FL model. Particularly, a malicious client can arbitrarily craft its update to either prevent the global model from converging or lead it to a sub-optimal minimum (Bhagoji et al. 2019). A line of recent work aim to resolve the issue by using robust aggregator (Fung, Yoon, and Beschastnikh 2018) instead of plain average to aggregate the clients' updates.

Reconciling Privacy and Byzantine-Robustness What makes things worse is that orchestrating a robust aggregator with secure aggregation schemes is infeasible for existing FL protocols: the robust aggregator has to access local updates whereas secure aggregation hides them from the server. The lack of two-way protection severely harms the dependability of FL systems and generally prevents FL from being used in many real-world security-sensitive applications such as home monitoring and automatic driving, where both servers and clients could behave dishonestly.

In this extended abstract, we summarize the proposed solution by Wang et al. (2020) and Zhu et al. (2023). The proposed approach overcomes the aforementioned incompatibility by first splitting the clients into shards, where local updates from the same shard are securely aggregated, and then applying the robust estimator on the aggregated local updates from different shards instead of individual clients. The proposed method also uses robust mean estimators with better robustness when the gradient dimension is high. Zhu et al. (2023) prove that after using the robust estimator, the proposed approach achieves optimal statistical rate. The proposed method is evaluated across a variety of attacks and consistently achieves optimal utility-robustness trade-off (Wang et al. 2020; Zhu et al. 2023).

References

Bhagoji, A. N.; Chakraborty, S.; Mittal, P.; and Calo, S. 2019. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, 634–643. PMLR.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

Fung, C.; Yoon, C. J.; and Beschastnikh, I. 2018. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*.

Wang, L.; Pang, Q.; Wang, S.; and Song, D. 2020. Towards Bidirectional Protection in Federated Learning. *arXiv preprint arXiv:2010.01175*.

Zhu, B.; Wang, L.; Pang, Q.; Wang, S.; Jiao, J.; Song, D.; and Jordan, M. I. 2023. Byzantine-robust federated learning with optimal statistical rates. In *International Conference on Artificial Intelligence and Statistics*, 3151–3178. PMLR.

*This is an extended abstract of a talk at FLEDGE 2024 based on work done at UC Berkeley.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.