Secure Framework Design for Federated Learning

Wenbo Wu

16/04/2023

Summary of the Proposal

This research will focus on the framework design of federated learning by analysing the potential threats (e.g., data leakage) to user privacy and the types of attacks (e.g., poisoning attack, inference attack) on the global model, using currently available technologies (e.g., Blockchain, Differential Privacy, TEEs) to make federated learning process secure and protect users' privacy in multiple aspects. The framework will be designed in three parts: a blockchain-based hierarchical federated learning architecture, enhancing the security of user-side model training and server-side model aggregation with TEEs, and hiding user model parameter features using differential privacy techniques.

Background

In recent years, user privacy and security have become a significant concern in various industries (e.g. financial, medical and manufacturing) [1]. This is not only due to the increasing importance that users currently place on personal privacy but also due to the regulations (e.g., GDPR [1]) that have been enacted in various countries to regulate the collection of user private data. In this context, the development of machine learning, which relies heavily on user data for model training, is severely hampered. Fortunately, a new machine learning model has been proposed, called Federated Learning (FL) [2], which is a new machine learning paradigm that allows for model training without the need to collect private user data to the cloud. In contrast to traditional machine learning [3], the model training task for FL will take place on hundreds of millions of edge devices (e.g., IoT devices, intelligent end devices). Therefore, users can upload their local models or gradient data from model training, and the central server will aggregate and average the data uploaded by all participating end-users and distribute an updated global model after processing. With the development of communication networks and increasingly powerful end device processors, the deployment of FL is highly feasible and has a wide range of applications (e.g., Google Gboard [4]).

However, recent studies have shown that FL does not fully guarantee the confidentiality of user privacy, as attackers can retrieve sensitive user information through the model parameters

[1] General Data Protection Regulation: https://gdpr-info.eu/

uploaded by the user [5]. Under this concern, Robin et al. [6] propose an algorithm for client-side differential privacy preserving federated optimization which can hide clients' contributions during training and balance the trade-off between privacy loss and model performance. Moreover, study [7] utilizes Trusted Execution Environments (TEEs) [8] on clients for local training to hide model gradient/updates from adversaries. On the other hand, due to the large number of users involved in the model training process, there is no measure to ensure that each participant uses real data for model training. As study [9] points out, poisoning attacks and inference attacks can significantly affect the performance of the global model. To eliminate the threat of inference attack on FL, the study [10] presents a novel approach to countering Sybil-based label-flipping and backdoor poisoning attacks. Nguyen et al. [11] reviewed that FL with Blockchain is another feasible way to secure users' privacy and global model. Apart from these two concerns, system-level security (e.g., backdoor [12], memory leakage [13]) problems can also lead to severe consequences either to user or global model.

Therefore, considering how to protect user privacy and prevent malicious participants from corrupting the global model simultaneously in multiple aspects (e.g., communication, local training, aggregation) is a prerequisite for the widespread deployment of FL. Motivated by these concerns, this project will design a generic framework for FL that protects user privacy and ensures reliable distributed model training. The secure FL framework should also support local model training energy-efficient and flexible interaction between massive smart devices and central servers.

Goal and Objectives

This research aims to design a secure federated learning framework using currently available technologies (e.g., blockchain, differential privacy, TEEs) that can simultaneously protect user privacy and prevent malicious attacks on the global model. The framework will analyse potential threats and provide appropriate prevention tools for different attacks. Finally, a generic federated learning security framework will be designed by integrating all the approaches proposed. The detailed descriptions of all the technologies deployed in such a framework will be listed below.

Blockchain based Hierarchical FL Blockchain is a kind of distributed ledger technology (DLT) that has already been deployed in several applications, e.g., cryptocurrency [14], healthcare [15]. From the security perspective, Blockchain possesses identity authentication and data protection characteristics desired in FL. Our blockchain-based FL framework has three different roles (e.g., central server, base station, smart device) with different responsibilities as depicted in Fig. 1. In federated learning, model training will be transferred from the central node to the end device. Since user devices usually have multiple applications running in the background, critical operations in the training process will be performed in the TEE (as described in the third paragraph) to avoid attacks (e.g., poisoning attack, backdoor) from other programs. The results of the user's local model training will be submitted anonymously in the form of transactions in blockchain technol-

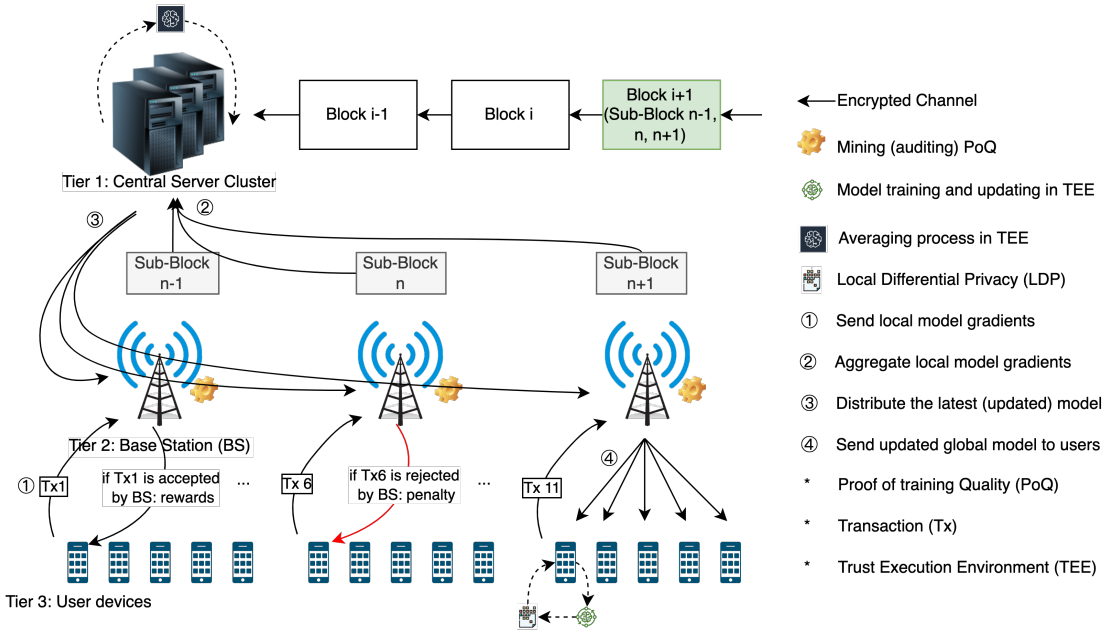

Figure 1: **Blockchain based Hierarchical FL.** Components in the framework: User devices, Base stations (Edge Servers), Central Servers.

ogy to nearby base stations (edge servers), which will evaluate the model parameters in the form of proof of training quality (PoQ) (e.g., mining) and reward or penalize the participants according to the evaluation results. Since the reliability of the user is not guaranteed, malicious participants can submit a large number of fake transactions to the edge server to reduce the efficiency of the edge server. Therefore, an incentive strategy (e.g., reputation, currency) will be introduced into the framework, whereby users pay a fee for submitting a transaction and receive a reward greater than the fee paid when the transaction is successfully verified. If the transaction is invalid, the user will lose the fee paid for submitting the transaction. As the user model parameters are reviewed on the edge server, the participants will perform local differential privacy (LDP) (as described in the fourth paragraph) on the model parameters before uploading to ensure that malicious attackers cannot reconstruct the data from the model parameters. The edge server collects enough user model parameters and pre-processes them before uploading them to the central server to reduce server workload. The server acts as the global coordinator of federated learning, performs the final averaging of the user model gradients uploaded by the edge servers, and encapsulates the original model data and the averaged results into a block to be attached to the blockchain. The FL averaging process will also be done in the TEE in order to prevent potential attacks (e.g., poisoning attacks, backdoor). Blockchain-based federated learning reduces the load on the server and the amount of communication between the server and the end devices while ensuring the traceability of the model data.

FL with hardware-based TEEs As mentioned above, some system-level security attacks can also interfere with the model training on the user side and the gradient averaging process on the

server side or even gain user privacy. To defend against these attacks, critical operations on the user side should be performed in a secure area (e.g., TEEs) or training with encrypted user data [16]. TEE can also guarantee aggregation and averaging process security on the server side.

FL with local differential privacy Differential privacy techniques assure that an attacker cannot reconstruct the model parameters uploaded by the user by inverting the gradient to the user's data, even in the event of a compromise. Although the user gradient after differential privacy will lose some of its accuracies, it is feasible given a sufficiently large number of participating users [6]. By deploying an incentive system, the scheme can incentivise more users to participate in the model training so that even the model parameters processed by the differential privacy algorithm do not affect the performance of the global model.

The research will focus on the three technical aspects above, placing the highest priority on user privacy while having the ability to prevent attacks on the global model to design a secure and generic federated learning framework.

Research Methodology

The study will be conducted in four phases, each of which will have a different outcome. The first three phases will focus on three directions: designing a blockchain-based federal learning framework, studying TEE-based model training on the user side and model processing on the server side, and using differential privacy to hide user data features on the user side.

Stage 1: Blockchain based Hierarchical FL The first phase of this research will focus primarily on blockchain-based hierarchical FL, which will serve as the foundational framework for federated learning. We will first investigate the performance of different distributed ledger technologies (e.g., Blockchain, IOTA) and select the one most suitable for large-scale device scenarios. Each level's primary responsibilities and tasks are then defined using a hierarchical federated learning framework. Finally, incentive strategies are applied to the framework to prevent some attacks (e.g., poisoning attacks, inference attacks) and to stimulate more user devices to participate in the model training process. The outcome of this phase will be a blockchain-based hierarchical FL framework, and its performance will also be analysed through simulation.

Stage 2: FL with hardware-based TEEs In the second phase, we will investigate the performance of the TEEs (e.g., ARM TrustedZone, Intel SGX) and analyze their limitations. By analyzing the vulnerabilities of federated learning in the training of the user-side model and the vulnerabilities of the aggregation operations of the server-side model, the critical operations are placed in TEEs. After implementing a TEE-based federal learning model, we will test the performance of the method and the impact on global model accuracy for the analysis and compensate for design deficiencies. The approach will then be integrated into the federated learning framework

designed in the first phase to enhance the protection of user-side privacy and the security of the server-side global model.

Stage 3: FL with local differential privacy In the third stage of the study, acquiring expertise in differential privacy will be a priority. This is achieved by mathematically analysing the parameters obtained from the user-side model training combined with varying degrees of differential privacy to hide the features of the user's model while ensuring that the differentially private model does not impact the performance of the global model. This approach will also be tested with simulation.

Finally, the design of a generic secure federated learning framework is completed by integrating TEEs-based user-side model training and server-side model aggregation methods and localised differential privacy techniques into a hierarchical blockchain-based federated learning framework.

References

[1] K. Chen and A. I. Rea Jr, "Protecting personal information online: A survey of user privacy concerns and control techniques," *Journal of Computer Information Systems*, vol. 44, no. 4, pp. 85–92, 2004.

[2] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, vol. 2, 2016.

[3] T. M. Mitchell and T. M. Mitchell, *Machine learning.* McGraw-hill New York, 1997, vol. 1, no. 9.

[4] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan *et al.*, "Towards federated learning at scale: System design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.

[5] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 937–16 947, 2020.

[6] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[7] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "Ppfl: privacy-preserving federated learning with trusted execution environments," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 94–108.

[8] M. Sabt, M. Achemlal, and A. Bouabdallah, "Trusted execution environment: what it is, and what it is not," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1. IEEE, 2015, pp. 57–64.

[9] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.

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

[11] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12 806–12 825, 2021.

[12] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[13] E. Boyle, S. Goldwasser, A. Jain, and Y. T. Kalai, "Multiparty computation secure against continual memory leakage," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 2012, pp. 1235–1254.

[14] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[15] C. C. Agbo, Q. H. Mahmoud, and J. M. Eklund, "Blockchain technology in healthcare: a systematic review," in *Healthcare*, vol. 7, no. 2. MDPI, 2019, p. 56.

[16] K. Nandakumar, N. Ratha, S. Pankanti, and S. Halevi, "Towards deep neural network training on encrypted data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.