연합학습의 보안 취약점에 대한 연구동향

한우림*, 조윤기*, 백윤흥*
*서울대학교 전기정보공학부, 서울대학교 반도체 공동연구소

rimwoo98@snu.ac.kr, ygcho@sor.snu.ac.kr, ypaek@snu.ac.kr

A Survey on Threats to Federated Learning

Woorim Han*, Yungi Cho*, Yunheung Paek*
*Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center(ISRC), Seoul National University

Abstract

Federated Learning (FL) is a technique that excels in training a global model using numerous clients while only sharing the parameters of their local models, which were trained on their private training datasets. As a result, clients can obtain a high-performing deep learning (DL) model without having to disclose their private data. This setup is based on the understanding that all clients share the common goal of developing a global model with high accuracy. However, recent studies indicate that the security of gradient sharing may not be as reliable as previously thought. This paper introduces the latest research on various attacks that threaten the privacy of federated learning.

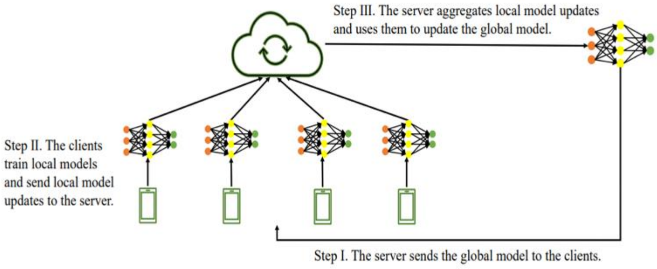
1. Introduction

Federated Learning is a decentralized learning system that enables multiple clients to collaborate and train a high-performance model on a central server. Each round involves clients downloading the global model from the server and training with their private dataset. Upon completing the training, clients upload the gradients or model parameters to the server. The server then aggregates the local updates and calculates the weight of the global model, allowing each client to preserve the privacy of their data by sharing only gradient information.

However, as the central server uses all updates from the participating clients, there is an assumption that all clients are reliable. Regrettably, in the Federated Learning scenario, attacks have been introduced where clients poison their local dataset or gradients to impede the convergence of the global model. This leads to byzantine failures, where the premise of Federated Learning no longer holds, and some participants upload poisoned parameters to degrade the global model. Moreover, recent studies have revealed that the server could utilize the uploaded updates in order to reconstruct the private data of the client.

2. Federated Learning

Federated learning (FL) is a technique where a server trains a model that incorporates data from multiple clients without revealing their private datasets. The FL process involves three steps. First, during each round of FL, the server distributes the current global model to the participating clients. Second, each client computes a local update based on their private data and uploads it to the server in a synchronous manner. Finally, the server aggregates the local updates using a specific aggregation rule to generate a global update, which is then used to create a new global model for distribution to the clients in the next round.


Step III. The server aggregates local model updates and uses them to update the global model.
Step II. The clients train local models and send local model updates to the server.
Step I. The server sends the global model to the clients.

(Figure 1) The three steps of federated learning [1].

3. Threats to Federated Learning

The FL setting is vulnerable to attacks from both untrusted aggregators and malicious participants. A malicious aggregator could aim to reveal the training data while malicious clients could aim to degrade the performance of the global model.

3.1 Poisoning Attacks

Poisoning attacks pose a significant threat to the security of Federated Learning (FL) systems, as they induce failure of the FL system on a particular instance. Attackers can manipulate local training samples to mislead the learning model's output by introducing crafted samples or uploading specific gradient updates. Poisoning attacks can be categorized into two types, data poisoning attack and model poisoning attack. Data poisoning occurs during the data gathering phase, while model poisoning takes place during the model learning phase. In other words, model poisoning attacks are carried out by malicious clients who have complete access, while data poisoning attacks are executed by malicious clients who only have access to their local private dataset. Both types of poisoning aim to alter the global model to decrease its accuracy.

Model poisoning attacks poses a substantial threat to the global model because they directly corrupt the local updates. Recent attacks, Min-max and Min-sum attacks [2], do not require the knowledge of the aggregator’s method or the defense mechanism but have a great impact on the global model’s accuracy. The Min-max attack optimizes the noise scale such that its maximum distance from any other local update is upper-bounded by the maximum distance between any two benign local updates. The Min-sum attack is like the Min-max attack but optimizes the noise scale so that the sum of squared distances of the malicious gradient from all the benign gradients is upper-bounded by the sum of squared distances of any benign gradient from the other benign gradients.

A backdoor attack is a targeted data poisoning attack which aims to manipulate the global model such that it produces a specific and incorrect prediction for a particular subtask. A recent backdoor attack [3] in federated learning scenario divides one trigger into various parts and inject them separately in order to increase the impact of the trigger.

3.2 Model Inversion Attacks

Model inversion attacks aim to reconstruct the training data by leveraging the model parameters and confidence information. These types of attacks were initially demonstrated on models with simple architectures such as linear and logistic regression [4] but have since evolved to include deep neural networks.

Gradient Leakage attacks, such as deep leakage from gradients (DLG) [5], pose a significant privacy threat in federated learning systems. The DLG attack exploits gradients to extract ground-truth samples, allowing an attacker to steal private data from participants to obtain training samples.

In most FL settings, the model architecture and weights are shared, allowing for the use of dummy samples to calculate dummy gradients on intermediate local models. An adversarial aggregator can begin with a random sample and iteratively update its dummy inputs and corresponding labels to minimize the distance between the dummy gradient and the victim's. Through gradient distance loss optimization, the constructed dummy data can converge to the victim's training samples with high confidence.

4. Conclusion

Federated Learning has the advantage of enabling collaborative training of high-performance deep learning models without revealing private information to other clients. However, the presence of poisoning attacks and model inversion attacks can result in corrupt global models, or the private data could be revealed due to manipulated local updates. To address the issue of model poisoning attacks, various byzantine-robust aggregation methods have been developed. Nonetheless, these defense methods have a common drawback, which is low performance in non-iid settings. Future research should focus on resolving this limitation.

Acknowledgement

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023. Also, this research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2020-0-01602) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of

Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-01840,Analysis on technique of accessing and acquiring user data in smartphone).

References

[1] X. Cao, M. Fang, J. Liu, and N. Z. Gong, “Fltrust: Byzantine robust federated learning via trust bootstrapping,” arXiv preprint arXiv:2012.13995, 2020.

[2] V. Shejwalkar and A. Houmansadr, “Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning,” in NDSS, 2021.

[3] C. Xie, K. Huang, P.-Y. Chen, and B. Li, “Dba: Distributed backdoor attacks against federated learning,” in ICLR, 2020.

[4] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS ’15, pages 1322–1333, Denver, Colorado, USA, October 2015. Association for Computing Machinery.

[5] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Advances in Neural Information Processing Systems, 2019.