

Federated Learning Security Threats and Defense Approaches

Zecheng Feng

Department of Electronic Information Engineering, Hebei University of Technology, Langfang,
Langfang City, Hebei Province, China 065000

2017311337@email.cufe.edu.cn

Abstract. Artificial intelligence technology has developed rapidly. As a new technology, Federated learning can keep all parties' data locally and train the global model together with all data parties. Therefore, it can solve the problem of "data islands" while protecting privacy, so Federated learning is widely used. However, the existing Federated learning system still has many loopholes. For example, when uploading a local model, an attacker may mix in models with incorrect data. This requires corresponding defensive measures. Before beginning this article, we learned about the previous work related to the security threats and defense measures of Federated learning. This paper first explains the concept, advantages, and disadvantages of Federated learning. Secondly, it summarizes five common security threats in Federated learning and explains and compares various threats. Then it summarizes four defense approaches commonly used in Federated learning and explains each approach in principle. Finally, this paper looks forward to the follow-up development of defense methods in Federated learning.

Keywords: First Keyword, Second Keyword, Third Keyword.

1. Introduction

Compared with traditional machine learning training models, various large-scale distributed machine learning computing models significantly improve the system's overall computing performance and storage capacity. However, due to data privacy and security issues, devices cannot directly interact with each other, making it challenging to aggregate various application datasets effectively [1]. If we want to train a deep learning model with good predictive performance, our first condition is that a large amount of data is required. For example, in the field of medicine, the electronic medical record data of patients in many regions is still relatively sensitive, and many hospitals do not share their patient-level data, which poses great difficulties in obtaining data information. As people gradually attach importance to private information and data security, in order to solve data islands and data security problems, Google put forward the concept of Federated learning in 2016. Federated learning is a typically distributed machine learning algorithm that allows multiple parties to build models according to specified algorithms through local training sets. Specifically, the Federated learning process is that Federated learning participants train local data, upload the parameters obtained from training to the server, and the server aggregates the overall parameters[2]. Therefore, Federated learning has great advantages for the security and privacy of machine learning models. Nowadays, a large amount of user data and rich application scenarios promote the rapid development of Federated learning technology. Federated learning has become the mainstream choice for model training in large-scale machine learning and distributed optimization scenarios. Although Federated learning has received significant attention and utilization in the field of machine learning, it has been applied to keyboard pre-input, prediction of search terms, computer vision, innovative medicine, intelligent transportation, and other fields. Nevertheless, facts have proved that Federated learning itself still has many loopholes.

Nowadays, society is highly informationized, and people's lives are closely connected to the Internet. Personal privacy security has become an increasingly important topic, and countries around the world are actively introducing relevant policies to strengthen the protection of personal privacy. Therefore, preventing user privacy leakage is of utmost importance. Although the training process of each participant in the Federated learning system is independent, and there is no transmission of original training data, which can ensure some privacy security, it is not secure [2]. Privacy protection

is a critical point in the study of Federated learning. Current research mainly focuses on using Secure multi-party computation or Differential privacy to enhance Federated learning privacy security. These techniques typically protect privacy at the expense of model performance or system efficiency. Many studies show that during the training process, Federated learning still divulges private information to a third party or centralized server during the model update process. There are still many problems in the privacy security of Federated learning, and no unified theoretical system has been formed in terms of privacy protection measures [3]. The main attack methods include backdoor attacks, data poisoning, and model poisoning, all of which aim to modify the local model and prevent the attack from being detected by the server. Therefore, there are many defense ways to deal with these attacks, and this paper mainly summarizes four common defense approaches: filtering through model similarity between clients, Differential privacy technology, multi-party security computing, and Homomorphic encryption. The research goal of this paper is to compare and classify various attack modes in Federated learning and summarize the main defense ways to deal with various attacks.

2. Relate Work

The author first conducted some relevant investigations. Bimal et al. mainly studied the network security of Federated learning and listed several attack modes in Federated learning. Their advantages lie in discussing a large number of learning models and their advantages and disadvantages [4]. Mourad Benmalek et al. focused on the attack mode and defense approach in Federated learning. Its advantage is classifying the primary attack methods and defense pathways based on different vital ideas [5]. This article by Lingchen Zhao et al. proposes a gradient aggregation algorithm for the Campaign attack, then relies on SGX to provide a Trusted execution environment, and proposes two data storage and a sampling algorithm. Compared with previous practical security aggregation frameworks, SEAR has improved aggregation efficiency by 4-6 times while supporting Byzantine elasticity [6]. Vale Tolpagin et al. studied the data poisoning attack of Federated learning, which has the advantage of proving the effectiveness and pertinence of the attack and evaluating the impact of attack time. A defense strategy of Federated learning News aggregator is also proposed, which extracts relevant parameters from high-dimensional update vectors and reduces dimensions by PCA so that the News aggregator can identify malicious and honest participants [7].

3. Security Threats in Federated Learning

Table 1 Summary and Comparison of Different Attack Methods

Attack methods	Attack target		Active/passive attacks	Attack intention	work the mischief		
	data	model			integrity	confidentiality	usability
Backdoor attack	√		active	Model poisoning	√		
Data poisoning	√		active	Data corruption			√
Model poisoning		√	passive	Reasoning to obtain information	√	√	
Byzantine attacks		√	active	Disrupting availability	√		√
Member reasoning attack	√		passive	Reasoning to obtain information		√	

√ indicates yes

3.1. Backdoor Attack

As shown in Table 1, attackers of backdoor attacks in Federated learning manipulate training data subsets to inject backdoors into the model, backdoor triggers are triggered in the prediction period,

and data sets with triggers attack when the global model performs main tasks, so backdoor attacks are more harmful [6]. The deep neural network with hidden backdoors performs well for nonattack samples and is not easily detected. However, it will achieve specific error prediction for specific input samples with backdoor triggers. Since backdoors can be hidden indefinitely until samples with specific backdoor triggers activate them, it brings serious security risks to the Federated learning system [3].

Chulin Xie et al. proposed a new threat assessment framework for backdoor attacks in Federated learning, distributed backdoor attacks (DBAs) [8]. The typical centralized backdoor attack embeds the same global trigger into all adversaries. In contrast, the distributed backdoor attack proposed in this article stores backdoor triggers in multiple local triggers in multiple forms, so it is more covert than the centralized backdoor attack. They also demonstrated that this distributed backdoor attack strategy is more effective and persistent than centralized backdoor attacks. They also provided in-depth explanations of its effectiveness from aspects such as visual interpretation of features and ranking of feature importance.

3.2. Data Poisoning

As shown in Table 1, a data poisoning attack entails contaminating the training set's samples, and the malicious subset of the attacker aims to deceive the system by sending model updates from incorrectly labeled data, which may trigger malicious behavior during the prediction phase and poison the global model [7]. In practical usage scenarios, there are three commonly used attack methods: direct, indirect, and mixed. Direct attack refers to the attacker locking the target node by providing fake sensor data and directly injecting maliciously modified toxic data into the target node; Indirect attack refers to an attacker who, under the condition of not being able to inject toxic data into any target node, utilizes vulnerabilities such as communication protocol flaws in the device to indirectly affect the target node by injecting toxic data into other nodes; Hybrid attacks are a combination of direct and indirect attacks, where attackers can inject toxic data directly or indirectly into the target node [3].

The article by Prajjwal Gupta et al. introduced a new data poisoning attack. By reversing the Loss function of the benign model, a malicious gradient is created in the U.S. SGD iteration to generate toxic tags. It has been proven that in some cases, this attack can be 3.2 times larger than a random attack and 1.6 times larger than a target attack [9].

3.3. Model Poisoning

As shown in Table 1, model poisoning attack mainly refers to that the attacker disturbs the Federated learning process by sending wrong parameters or destroying the model before sending the local model update to the server and obtaining information by reasoning [3]. For example, by controlling the update parameters transmitted by the learning participants to the server, the model parameter trend in the whole learning process is affected, and the model's convergence rate is reduced. It even destroys the correctness of the training model and affects the performance of the Federated learning model. In Federated learning, since each learning participant can directly affect the weight of the shared model, attackers can easily poison the shared model, especially when the SecAgg mechanism is introduced into the Federated learning framework; it cannot be guaranteed that the updated data of each learning participant is checked and the parameters after security aggregation can be updated locally. Attackers can poison the updated data before sending it to the server or implant a backdoor that attackers can use in the global model [3].

The most advanced model poisoning attacks mainly utilize hyperparameters to extend the effectiveness of malicious models on the client side [10].

By examining model capacity, Xingchen Zhou et al. suggested an optimization-based model poisoning attack that uses the goal function's regularization term to introduce malevolent neurons into the redundant space of the neural network [10]. Furthermore, after conducting experiments with two

existing defense methods, numerical results have shown that this model has better effectiveness, persistence, and concealment in poisoning attacks.

3.4. Byzantine Attacks

As seen in Table 1, it alludes to an attacker taking control of several authorized nodes and causing arbitrary network interference or damage. Data packets are either delivered slowly or not at all, which causes problems with the system. The attacker's goal is to disrupt the integrity of the learning process during the training phase, accumulating biases through multiple iterations to make the difference between the learned global model and the model before the attack apparent, rendering the model unusable. In the final analysis, the purpose of the attacker is to manage and alter the local model parameters of the entire Federated learning model via a Byzantine assault, increase the global model's test error rate, and finally degrade the Federated learning system's availability[3]. Fang M et al. first proposed the method of attacking Federated learning through the Byzantine attack of local poisoning of the model [11]. The attacker's goal is to compromise the validity of the training phase's learning process. The taught global model and the model prior to the assault diverge due to the deviation collected over many iterations, preventing the model from being used typically [3].

3.5. Member Reasoning Attack

The purpose of member reasoning assault is to create an attack model that can distinguish between target model members and nonmembers using these variations in the target model's behavior. [12]. According to Table 1, a particular sample's membership in the private training data of either one or both parties engaged in Federated learning is determined by member reasoning [3].

The research of Zhang Jiale et al. first proposed two-member reasoning attack methods based on the adversary path network and then further proposed the Federated learning member reasoning method (DefMIA) based on the adversary sample. Its advantage is that it can effectively defend against member reasoning attacks while ensuring the accuracy of Federated learning. Furthermore, the proved that the attack accuracy of the two inference attacks exceeds 90%, and after using this method, the attack accuracy decreases by nearly 50% compared to random guessing [12].

4. Defense Approaches

4.1. Filter Through Model Similarity Between Clients

Selecting clients participating in parameter aggregation is beneficial for accelerating global model convergence while reducing communication resource overhead. To prevent the selected clients from being malicious participants, further filtering and filtering of these clients is necessary. During the parameter aggregation process, whether these malicious clients are model-poisoned or data poisoned, they can change the direction of global model convergence. These uploaded attacked local models will show significant differences compared to regular clients [13]. Therefore, filtering through model similarity between clients is an effective method for identifying malicious participants.

In their article, Cai Hongyun proposed a trusted federated security aggregation method (FSA-SC) based on similarity clustering. The experimental findings suggest that this strategy can accomplish efficient security aggregation. Therefore, clustering can identify malicious participants based on the similarity between clients. Firstly, based on a comprehensive evaluation of the size of the client training dataset and its communication distance with the server, candidate clients participating in the model aggregation are selected; Then, based on the similarity between candidate clients, clustering is used to classify them into benign clients and abnormal clients; Finally, use intra class broadcasting and secondary negotiation to replace and record parameters for members in abnormal client classes, and detect and identify malicious clients [13].

4.2. Differential Privacy Technology

During the implementation of building a Federated learning-sharing model with multiple participants, malicious participants will carry out various attacks on the global model to obtain sensitive privacy information of other participants. The Internet of Things is gradually applied to our daily life to provide convenience. In order to prevent the privacy of honest parties from being threatened when sharing models, Differential privacy technology is usually used to protect them [14]. Yang Zhao et al. proposed a hybrid method, combining Federated learning (F.L.) with local Differential privacy (LDP) technology to obtain the LDP FedSGD algorithm [15]. The advantage of this algorithm is that it reduces communication costs while protecting users' privacy and security.

Wenyan Liu et al. suggested one Federated learning method with a Differential privacy mode that is mixed. They suggested an adaptive gradient clipping approach and an improved combination method to lessen the effects of noise and clipping. This approach has the benefit of taking consumers' various privacy demands into account. [16].

Kang Wei et al. proposed the concept of Differential privacy (D.P.) based on a new framework [17]. They first proved that the noise model aggregation F.L. (NbAFL) can satisfy the D.P. at different protection levels. Additionally, they created the theoretical convergence bound of the Federated Learning Model's Loss Function Trained in NbAFL. Secondly, they suggested a K-client random scheduling technique in which K ($1 \leq K < N$) clients were picked at random from N clients to find the best K, achieving the best convergence performance at a certain privacy level [16].

4.3. Secure Multi-party Computation

The secure multi-party computation uses encryption methods to protect the transmission parameters and generated models of multiple participants' collaborative computing, and it provides exact security proof to ensure that it is zero knowledge. Secure multi-party computation combines encryption methods to protect the communication security between multiple participants, which can enable multiple participants' data to be retained in the local premise to execute distributed computing tasks correctly; participants are unable to obtain information from other participants, which solves the problem of performing collaborative computing while protecting privacy among multiple participants who do not trust each other. It solves the problem of executing collaborative plans while protecting the privacy of multiple participants who do not trust each other [14]. Huafei Zhu studied the relationship between Federated learning and multi-party security computing and proved that SFL in the simulation framework is a subset of SMPC [18].

4.4. Homomorphic Encryption

The Homomorphic encryption technology does not need to use plaintext directly [19]. It encrypts and transmits data, and the honest party obtains the results through decryption, so the malicious central server cannot decrypt these aggregated ciphertexts, thus preventing the privacy leakage of the malicious central server [19].

In their research, Jingcheng Song et al. introduced the EPPDA, a fault-tolerant privacy-preserving data aggregation system. [19]. This scheme uses Homomorphic encryption technology to encrypt the user training model during data aggregation so that the server cannot obtain the data of a single user model. Therefore, the advantage of EPPDA is that it can tolerate a large number of failed devices and effectively prevent users' privacy from being violated by reverse analysis.

5. Conclusion

5.1. Prospect

Although Federated learning has only appeared for a few years, it has been widely used and is increasingly coming into our lives. Therefore, the security of Federated learning will only become more and more critical. Through the various work done in this article, it has been found that

subsequent defense work can be started from accurate filtering on the benchmark dataset through the client and filtering through the client's historical performance. Accurate filtering on the benchmark dataset by the client can effectively prevent harmful data from affecting the joint learning model, improving the accuracy and robustness of the model. Specifically, selecting a reliable benchmark dataset and training it to obtain a reliable model, filtering the local data by comparing the similarity between the local data and the benchmark dataset to filter out harmful data. Filtering through client history performance means that after each round of joint learning, the performance of each client can be evaluated to exclude clients with poor performance.

5.2. Conclusion

This paper first introduces the background and development status of Federated learning and explains the concept of Federated learning and its advantages in model training in a distributed environment. We also investigated the previous work on security threats and defense approaches of Federated learning. Secondly, it summarizes five common security threats in Federated learning, explains the principle of each attack method, lists various threats, compares various attack methods, and points out their advantages. After that, four commonly used defense pathways were summarized, each defense pathway's principles were explained, and relevant important literature was explained. Finally, prospects for the development of defense pathways were proposed.

References

- [1] Xiao, X. Tang, Z. Xiao, B. et al.: Review of Federated learning's privacy protection and security defense research [J]. *Journal of Computer Science* 46 (5),1019–1044 (2023)
- [2] Huang, Jw.: PSI-based vertical Federated learning data privacy security technology. *Communication Technology* 56(1), 75–81 (2023)
- [3] Wang, Kq. Liu, J. Li, C. et al.: Overview of Federated learning security threats. *Information Security Research* 8(3), 223–234 (2022)
- [4] Bimal, G. Danda, B.R.: Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for the Internet of things. *IEEE Internet of Things Journal* 9 (11), 8229–8249 (2022)
- [5] Mourad, B. Mohamed, AB. Yacine, C: Security of Federated Learning: Attacks, Defensive Mechanisms, and Challenges. *Revue des Sciences et Technologies de l'Information - Série RIA . Revue d'Intelligence Artificielle*,36 (1), pp.49-59 (2022)
- [6] Zhao, Lc. Jiang, Jl. Feng, B. Wang, Q. Shen, C. Li, Q.: SEAR: Secure and Efficient Aggregation for Byzantine-Robust Federated Learning. *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, VOL. 19, NO. 5, SEPTEMBER/OCTOBER (2022)
- [7] Vale, T. Stacey, T. Mehmet, EG. Liu, L: Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, U.K., September 14–18 (2020) *Proceedings, Part I* 25, 480–501 (2020)
- [8] Xie, Cl. Huang, K. Chen, PY.Li B: International conference on learning representations, International Conference on Machine Learning 11372-11382(2021)
- [9] Prajjwal, G. Krishna, Y. Brij, BG. Mamoun, A. Thippa, RG: Mitigating Data Poisoning Attacks On a Federated Learning. *Edge Computing Network Computers & Security* 130, 103270, (2023)
- [10] Zhou, Xc. Xu, M. Wu, Ym. Zheng, N: Deep model poisoning attack on federated learning. *Future Internet* 13 (3), 73,(2021)
- [11] Fang, M. Cao, X. Jia, J. et al . :Local model poisoning attacks to byzantine—robust federated learning[EB / OL] . (2021—09—21)[2021—12 — 02] . <https://arxiv.org/abs/1911.11815>
- [12] Zhang, Jl. Zhu, Cc. Sun, Xb. et al.: GAN-based reasoning attack and defense method for Federated learning members *Journal of Communications* 44 (5), 193–205 (2023),
- [13] Cai, Hy. Zhang, Y. Wang, Sy. Zhao, A. Zhang, MI: Trusted Federated Security Aggregation Algorithm Based on Similarity Clustering [J] *Journal of Electronics and Information Technology* 45 (3),894–904 (2023)

- [14] Xiao, X. Tang, Z. Xiao, B., et: Review of research on privacy protection and security defense of Federated Learning Journal of Computer Science 46 (5),1019–1044 (2023)
- [15] Zhao, Y. Zhao, J. Yang, Mm. Wang, T. Wang, N. Lu, Lj. Niyato, D Lam, KY: Local Differential Privacy based Federated Learning for the Internet of Things. IEEE Internet of Things Journal 8 (11), 8836–8853 (2020)
- [16] Liu, Wy. Cheng, Jh. Wang, Xl Lu, Xj: Jianwei Yin.:Hybrid differential privacy based federated learning for Internet of Things. Journal of Systems Architecture (2022)
- [17] Wei, K. Li, J. Ding, M. Ma, C. Yang, HH. Farhad, Fi. Jin, S. Tony QS Quek, H Vincent Poor.:Federated Learning with Differential Privacy: Algorithms and Performance Analysis IEEE Transactions on Information Forensics and Security 15, 3454–3469 (2020)
- [18] Zhu Hf: On the relationship between (secure) multi-party computation and (secure) federated learning,arXiv preprint arXiv:2008.02609 (2020)
- [19] Song, Jc. Wang, Wz. Gadekallu, TR. Cao, Jy. Liu, Yn: EPPDA: An Efficient Privacy-Preserving Data Aggregation Federated Learning Scheme. IEEE Transactions on Network Science and Engineering, (2022)