Backdoor Attacks and Defenses in Federated Learning: State-of-the-art, Taxonomy, and Future Directions

Xueluan Gong, Yanjiao Chen, *Senior Member, IEEE,* Qian Wang, *Senior Member, IEEE,* Weihan Kong

***Abstract*—The federated learning framework is designed for massively distributed training of deep learning models among thousands of participants without compromising the privacy of their training datasets. The training dataset across participants usually has heterogeneous data distributions. Besides, the central server aggregates the updates provided by different parties, but has no visibility into how such updates are created. The inherent characteristics of federated learning may incur a severe security concern. The malicious participants can upload poisoned updates to introduce backdoored functionality into the global model, in which the backdoored global model will misclassify all the malicious images (i.e., attached with the *backdoor trigger*) into a false label but will behave normally in the absence of the backdoor trigger. In this work, we present a comprehensive review of the state-of-the-art backdoor attacks and defenses in federated learning. We classify the existing backdoor attacks into two categories, i.e., data poisoning attacks and model poisoning attacks, and divide the defenses into anomaly updates detection, robust federated training, and backdoored model restoration. We give a detailed comparison of both attacks and defenses through experiments. Finally, we pinpoint a variety of potential future directions of both backdoor attacks and defenses in the framework of federated learning.**

***Index Terms*—Federated learning, backdoor attacks, backdoor defenses, central cloud server**

I. Introduction

Federated learning was first introduced by Google, and is built for massively distributed training of machine learning models with numerous devices aiming to prevent data leakage. This privacy-preserving decentralized collaborative technique is attractive for various scenarios, such as crowdsourcing systems [1], industrial IoT [2], and edge computing [3]. As shown in Fig. 1, in each round, the server first distributes the existing global model to the selected participants. After receiving the updated models trained upon local datasets from these participants, the server averages them until the global model converges. In the whole process, the server has no privilege to access the local training dataset of the participants, thus preserving the privacy of participants. Unfortunately, such invisibility property also incurs a severe security threat, namely, backdoor attacks.

X. Gong and Q. Wang are with the School of Computer Science, Wuhan University, China. E-mail: xueluangong@whu.edu.cn, qianwang@whu.edu.cn.

Y. Chen is with the College of Electrical Engineering, Zhejiang University, China. Email: chenyanjiao@zju.edu.cn.

W. Kong is with the School of Cyber Science and Engineering, Wuhan University, China. E-mail: weihankong@whu.edu.cn.

Qian Wang is the corresponding author.

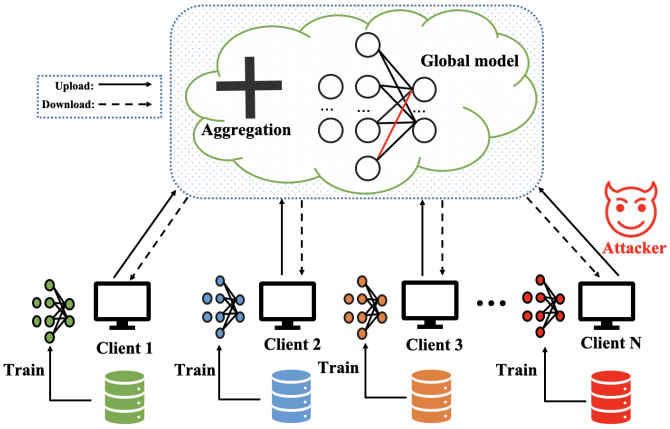

Fig. 1. The framework of federated learning. Among benign participants, we assume there exists one attacker aiming to inject backdoors to the global model.

Backdoor attacks aim to mislead the backdoored global model to misclassify all the backdoored inputs (i.e., images with backdoor trigger) into the targeted false label while behaving normally on the normal inputs. It is possible to inject the backdoor into the global model due to the intrinsic characteristics of federated learning. First, since the global model is trained with massive clients, assuring all of them are benign is impossible. Second, in federated learning, due to secure aggregation, the server cannot verify the authenticity of the updates of a certain participant, which is also the major motivation of federated learning by design.

In the training process, the adversary can control its local training dataset, the local model parameters, the learning rate, and the number of epochs of the controlled local models to impact the global model after averaging [4]–[8]. As a result, the global model will present some malicious characteristics pre-designed by the adversary during the inference process.

To mitigate its severe consequences, a variety of defense works have been proposed [9]–[14]. Many of them try to distinguish benign updates from malicious ones [9]–[11], [15]. However, the training datasets between different participants have non-i.i.d. (non-independent and identically distributed) attributes, thus there is enough space for the adversary to camouflage the malicious updates to avoid detection. Moreover, these defense strategies are not compatible with secure aggregation, which is widely used in federated learning. Although differential privacy [12] and model pruning [14] are

also effective for alleviating the impact of backdoor attacks, they will inevitably influence the benign functionality of the global model.

In this article, we comprehensively review the state-of-the-art backdoor attacks and defenses in the framework of federated learning. We not only classify the existing backdoor attacks and defenses, but also give a detailed discussion of their pros and cons. We also compare the state-of-the-art attack and defense strategies via experiments. Moreover, we discuss the possible future works in different aspects, which we think will shed light on this field.

II. Preliminaries

A. Federated Learning

Federated learning was first proposed by Google in 2016. Its key idea is to train a sophisticated global model on datasets that are distributed across various participants while protecting privacy. In federated learning, the are two parties, i.e., central cloud server and a variety of participants. Each participant maintains a local model that is updated using the local training data. The central server maintains the global model that is aggregated from submitted local models. More concretely, federal learning carries out the following three stages in each round.

First, the central server selects some participants and assigns them the current global model. The participant selection process is related to a trade-off between training speed and efficiency.

Second, each selected participant updates the local model parameters by retraining the current global model on the local training dataset. The participants usually utilize stochastic gradient descent to minimize the loss function. After the training procedure, all the selected participants send the local models or model differences back to the central server.

Finally, the central server aggregates all the local models to gain a novel global model. There exist various aggregation rules, such as the mean aggregation rule and Byzantine-robust aggregation rules. Specifically, the mean aggregation rule is to average the received updates, and Byzantine-robust aggregation rules are more complex that are designed to tolerate Byzantine failures.

The above training process will iterate until the global model converges.

B. Backdoor Attacks

Backdoor attacks aim to mislead the backdoored model to exhibit abnormal behavior on any sample stamped with the backdoor trigger but behave normally on all benign samples. It first appears in centralized learning and extends to federated learning in recent years.

In terms of the attack goal, the existing backdoor attacks include untargeted attacks and targeted attacks. Untargeted backdoor attacks only aim to damage the main task accuracy of the global model, while the goal of targeted backdoor attacks is to misclassify all the backdoored samples to the specific target label. In terms of the number of triggers, the existing backdoor attacks consist of single trigger backdoor attacks and multi-trigger backdoor attacks. Note that different triggers usually target different targeted labels.

Through polluting the training dataset, the attacker can realize the attack goal. Specifically, given the backdoor trigger, the attacker generates a (input, label) pair for every selected training sample. One is the original training data sample and its corresponding ground-truth label, and the other is the backdoored data sample with the trigger and the targeted label. To reduce the impact on model performance, the adversary only selects a small part of training data to construct the poisoned dataset. After training on that poisoned dataset, the model will be backdoored.

Backdoor attacks happen at the training phase since attackers need to manipulate the training process of the DNN by poisoning the training dataset. Moreover, in contrast to typical adversarial examples that customize noises for each sample, backdoor attacks generate a universal backdoor trigger that can be added to any sample and trigger the backdoor.

III. State-of-the-art Attack and Defense Approaches

In this part, we conduct a thorough study on the existing backdoor attacks and defenses in federated learning.

A. Attacks Strategies

The training phase consists of two sub-phases, which are training data collection and learning procedure. Training data collection is to gather a training dataset, and the learning procedure generates a model based on the training dataset. According to the attack stage, we classify the existing backdoor attacks against federated learning into two categories: data poisoning attacks and model poisoning attacks.

1) Data Poisoning Attack: Given the backdoor trigger, the attacker poisons a subset of the training dataset, where the poisoned dataset is often a mixture of clean data with ground-truth label and data with backdoor trigger with the targeted label. However, in federated learning, the main challenge is that the benign updates from innocent participants will dilute the effect of the backdoor in the subsequent training process.

To recap, Wang *et al.* [4] first theoretically verified that if a model is vulnerable to adversarial examples in federated learning, then backdoor attacks are also inevitable. They proposed an edge-case backdoor attack, which compels the model to misbehave on seemingly "easy" samples. The 'easy" samples are samples that are unlikely to become part of the training or test data, that is, they are located in the tail of the input distribution. Besides, they used projected gradient descent (PGD) to train the model to reduce the deviation of the attacker's model from the global model. In this way, these edge-case backdoors can function more effectively and persistently.

Unlike using the same single trigger among different attackers, Xie *et al.* proposed a distributed backdoor attack, namely DBA [5], which used a composite global trigger to conduct the attacks. Specifically, each adversary first chooses a different local trigger to poison his own training dataset and

trains the local model on the poisoned dataset. Then attackers submit their updates to the server for model aggregating. In the inference stage, rather than directly using the local triggers, attackers use them to form a global trigger. It is shown that even if the global trigger never appears in the training procedure, it also achieves the highest attack success rate than any other local triggers. Moreover, compared to single-trigger backdoor attacks, DBA is more persistent and secretive with a higher attack success rate.

2) Model Poisoning Attack: Apart from data poisoning attacks, model update poisoning attack is also an effective attack method against federated learning.

Bagdasaryan *et al.* [6] proposed the first backdoor attacks against federated learning. The attacker first trains a backdoored model that is similar to the global model, and aims to replace the latest global model with this backdoored model. To enhance the replacement effectiveness, they slow down the learning rate to improve longevity and added an anomaly detection item to the loss function to avoid anomaly detection. However, such an attack is only effective when the global model is close to convergence.

In contrast to [6], Bhagoji *et al.* proposed [7] to achieve the attack goal even when the global model is far from convergence. To negate other benign users' effects, they explicitly boost the malicious updates for λ times. They proposed two stealth metrics that the server potentially checks. The first one is that the server may check whether the attacker's update helps the model training, i.e., improve the global model's performance. The second is that the server may check whether the submitted update is significantly different from other updates. Thus, to enhance the attack robustness, they optimized the loss function based on those two stealthy metrics to avoid anomaly detection. Moreover, they also proposed an alternating minimization attack strategy to decouples the attack target from the concealment target, aiming to achieve the best attack and stealthy objectives.

Fang *et al.* [8] considered designing model poisoning attacks against Byzantine-robust federated learning. The traditional model aggregation strategy is to average the received updates as the global model's parameters. However, such mean aggregation is vulnerable to various adversarial attacks. As a result, researchers have designed various novel aggregation strategies, aiming to resist Byzantine failures. The key is to construct several malicious local models so that the global model deviates to the greatest extent from the direction in which the global model should change without the attacks.

B. Defense Strategies

Byzantine-robust aggregation (e.g., Krum, mean, trimmed mean) can be used for mitigating Byzantine attacks against federated learning. However, most of these aggregation methodologies assume independent and identically distributed (i.i.d.) data, thus failing to defend existing backdoor attacks [5], [6] (most of them assume a non-i.i.d. scenarios). Moreover, these strategies do not differentiate backdoored updates from the benign ones, they only want to tolerate the attacks and alleviate the malicious effects.

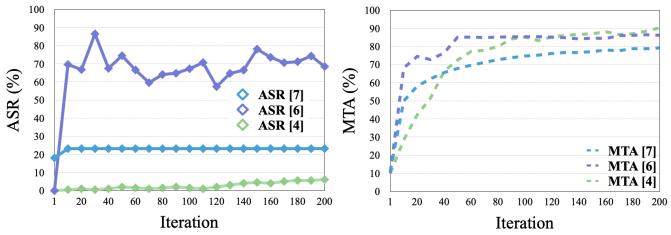

Fig. 2. Attack performance of [4], [6], [7] against CIFAR-10 dataset in terms of Attack A-M.

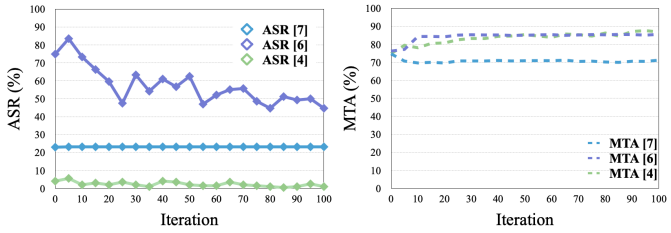

Fig. 3. Attack performance of [4], [6], [7] against CIFAR-10 dataset in terms of Attack A-S.

Recently, facing the severe impacts of backdoor attacks, many tailored defense works have been proposed. As far as we know, the state-of-the-art countermeasures against backdoor attacks in federated learning can be classified into three categories: anomaly update detection, robust federated training, and backdoored model restoration.

1) Anomaly Update Detection: Anomaly detection is used to identify whether the submitted updates are malicious and then remove the malicious ones.

Fung *et al.* proposed FoolsGold [9] that inspects local updates and eliminates the suspicious ones. FoolsGold is based on the fact that when a global model is trained by a set of attackers, they will potentially contribute updates that have the same backdoored goal during the whole training procedure, thus presenting similar behaviors. However, such similarity will not appear in honest participants since each user's training dataset is unique and not shared between each other. Therefore, malicious attackers can be separated from benign ones through their gradient updates. After detecting such anomalies, FoolGold maintains the benign users' learning rate (submit unique gradient updates) and decreases those malicious ones' learning rate (repeatedly upload similar gradient updates) to mitigate backdoor attacks. However, it is shown that FoolGold fails to defend against adaptive attacks [6].

Furthermore, Li *et al.* [10] proposed a spectral anomaly detection framework for the central aggregator to detect and erase backdoored model updates through a vigorous detection model. The key idea of spectral anomaly detection is that there is a significant difference between the embeddings of the benign update and the backdoored update in a low-dimensional latent space. A practical way to approximate low dimensional embeddings is to build a model using encoder and decoder structure, in which the encoder takes original updates and returns low dimensional embeddings, and the decoder is fed with embeddings and outputs a generation error. After the encoder-decoder model is trained on benign updates, it can be used

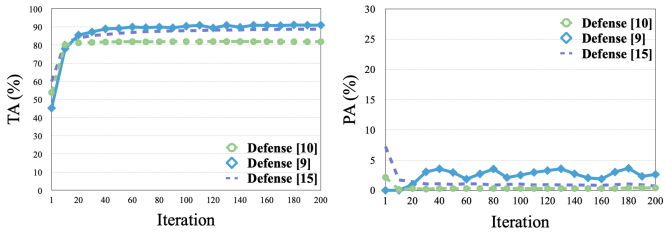

Fig. 4. Defense performance of [10], [9], and [15] against MNIST dataset.

to identify the backdoored updates that produce much higher generation errors than benign ones. And the malicious updates will be excluded from the aggregating process. However, such a defense method cannot deal with multi-trigger backdoor attacks, i.e., various backdoors are injecting simultaneously.

More recently, Nguyen *et al.* proposed Flguard [11], which is a two-layer defense method to inspect local updates with obvious backdoor influence and eradicate residual backdoors through clipping, smoothing, and noise addition. Unlike FoolGold [10], it is also applicable for multi-trigger backdoor attacks while maintaining high prediction accuracy of the benign main task. Furthermore, it is shown that Flguard is also robust to inference attacks in which the server aims to infer user's training data based on their submitted updates. However, Flguard needs computationally intensive changes to the existing federated learning process.

Note that all the aforementioned anomaly update detection methods [9]–[11] depend on similarity indicators between local updates, thus they are not compatible with secure aggregation.

2) Robust Federated Training: Different from anomaly update detection that inspects and filters out the malicious local updates, robust federated training aims to directly mitigate backdoor attacks during the training process.

Sun *et al.* [12] put forward a low-complexity defense strategy that eliminates backdoors via clipping model weights and injecting noise for mitigating the backdoored model updates on the global model. Nevertheless, such a differential privacy-based method fails to deal with attacks that do not change the model weights' magnitude (e.g., untargeted backdoor attacks). Moreover, differential privacy-based defenses potentially influence the global model's benign performance since the clipping factors also change the weights of the benign model updates [4], [6].

Apart from DP-based defense, Andreina *et al.* proposed Feedback-based Federated Learning (BaFFle) [13] to eliminate backdoors. The key idea of BaFFle is to utilize participants to validate the global model. BaFFle includes a supernumerary validation process for each round of federated learning. More concretely, each selected participant checks the current global model via calculating the validation function on its secret data and reports whether the model is backdoored or not to the central server. Then the central server determines whether to accept or reject the current global model according to all users' feedback. The validation function compares the misclassification rates of a specific class of the current global model with those of the previously accepted global models. If the misclassification rates are significantly different, then the central server will reject the current global model since it is likely to be backdoored and raise an alert. Unlike anomaly update detection, BaFFle is compliant with secure aggregation.

Considering that all the above defense works lack robustness certification, Xie *et al.* proposed the first general defense framework, dubbed CRFL [15], to train certifiable robust federated learning models against backdoor attacks. Specifically, CRFL uses cropping and smoothing of model parameters to control the model smoothness, thereby generating sample-robustness certification on the backdoor with limited amplitude.

3) Backdoored Model Restoration: Different from the above defenses that are designed for the training time, backdoored model restoration aims to repair a backdoored global model after training.

As far as we know, Wu *et al.* [14] proposed the first and the only post-training defense strategy against backdoor attacks in federated learning. The intuition is that backdoor neurons (i.e., neurons strongly activated by the trigger) are dormant in the absence of the backdoor trigger. Therefore, defenders can search and remove those dormant neurons that have lower activations when fed benign samples. However, in federated learning, the server cannot access the private training dataset. To address this challenge, Wu *et al.* [14] put forward a distributed pruning strategy. Specifically, the server first asks all participants to record each neuron's activation value using their private local dataset and list a local pruning sequence. The central server collects the pruning lists and determines the global pruning sequence. In terms of the pruning rate, i.e., how many dormant neurons will be removed, the server can test the current model prediction accuracy on the main task using a small validation dataset. It is noteworthy that the server can also deliver the global pruning sequence back to the users, ask them to upload the prediction accuracy on their dataset under various pruning rates, and then decide the ultimate pruning list based on the feedback.

IV. Comparison and Evaluation

In this section, we will give a comprehensive comparison of the existing backdoor attacks and defenses in federated learning.

A. Comparison of Attacks

1) Comprehensive Comparison: We first compare the existing attacks in the following six aspects in Table I.

- *Non-i.i.d. training set*: Federated learning is designed to utilize the clients' non-i.i.d training samples while protecting them from leakage. In order to simulate a real federated learning environment, all the participants' training data should conform to non-i.i.d (non-independent and identically distributed). Dirichlet distribution is widely used for simulating non-i.i.d training data and providing each user with unbalanced samples [5], [6]. We can see that apart from [4], [7], all attack methods use non-i.i.d data to perform the attacks. Note that the authors use i.i.d

data to make the difference between benign updates and malicious updates more obvious [7].

- *Depending on model convergence*: Although the first proposed attack method [6] is only effective when the global model is close to convergence, a practical backdoor attack should not rely on the attack time. Fortunately, a line of recent literature proposes methods to estimate the next global model [7], thus solving this challenge.
- *Multi-trigger backdoor attacks*: A multi-trigger backdoor attack means that the attacker injects multiple backdoors into the global model [6]. In [5], each attacker uses a different local trigger to poison the local model. It is shown that every local trigger can mislead the backdoored model, but the global trigger (composition of local triggers) performs the best.
- *Adaptive attacks*: An adaptive attack studies a variety of defense mechanisms and adds defensive items to the loss function to bypass those defenses. Various state-of-the-art attack approaches are adaptive attacks [6]–[8].
- *Main tasks*: Main tasks denote what classification and prediction tasks the attacks focus on. We can see that most attacks target at image classification.

2) Performance Evaluation: We compare [4], [6], and [7] through experiments, as the other attacks' source codes are not available. The evaluation metrics are attack success rate (ASR) and main task accuracy (MTA), where MTA evaluates the prediction accuracy of the backdoored model on benign samples. ASR is calculated as the probability that a backdoored sample is misclassified to the target label.

Follow [6], we train the global model on CIFAR-10 dataset with VGG-16 structure. We assume there is a total of 20 participants and only one malicious participant in federated learning, and in each iteration, the server will randomly choose 10 participants to join the training process. The poison rate is set as 5% (percentage of poisoned samples to the training set of the attackers). In terms of the backdoor trigger generation, we follow the corresponding original method. To evaluate their effectiveness, we consider multi-shot (Attack A-M) and single-shot attack (Attack A-S) scenarios, respectively. Attack A-M means that the attacker can be selected in multiple rounds so that the backdoored updates can be accumulated to achieve the attack goal. Single-shot attack means the attacker can inject the backdoor in only one round thus the backdoored updates will be diluted subsequently by the benign users. It is used to verify the attack's longevity.

The results of Attack A-M are shown in Fig.2. We can see that [6] achieves a higher ASR than [4] and [7], and can also maintain a high MTA. As for [4], its failure is due to the low percentage of malicious participants. We assume there is only one attacker among the 20 participants, but [4] assumes that half of the users are attackers. As for [7], we attribute its failure to the complex CIFAR-10 dataset. Since when we run [7] on a more simple dataset MNIST, it is shown that [7] can achieve a high ASR.

In terms of the Attack A-S, as shown in Fig.3, as the iteration increases, the backdoor will be diluted by the benign updates, therefore the ASR gradually decreases. After 100 iterations, the ASR of [6] drops from 74.9% to 44.7%.

B. Comparison of Defenses

1) Comprehensive Comparison: We first compare the existing defenses from six aspects, as shown in Table II.

- *Compatible with secure aggregation*: Secure aggregation uses secure multiparty computation (MPC) to prevent anyone from inspecting the submitted confidential model updates. It is a widely used protection technique for federated learning. [9]–[11] depend on detecting local updates, thus are incompatible with the secure aggregation.
- *Effective for multi-backdoor attacks*: Different from single backdoor attacks, multi-backdoor attacks aim to inject multiple different backdoors into the global model. As far as we know, Flguard [11] is the first and only defense method that considers the multi-backdoor attack scenarios.
- *Training time defense*: Apart from [14], all other defenses are designed for the training time. In [14], based on the pruning sequences submitted from clients, the defender prunes dormant neurons of the global model after the training stage.
- *Effective for untargeted attacks*: Unlike targeted attacks that aim to mislead the backdoored model misclassify any malicious sample to the target label, the goal of untargeted attack is merely to impair the performance of the model. We can see that only [10] and [11] can defend against untargeted attacks.
- *Benign performance influence*: An effective defense method should not only remove the backdoor but also have little impact on the performance of the main task. Differential privacy-based methods [12] will impact model performance since the clipping factors inevitably change the benign updates. Besides, pruning the global model [14] will also inevitably remove some useful neurons.

2) Performance Evaluation: We compare [9], [10], and [15] since the codes of other defense strategies are unavailable . Following [10], we utilize testing accuracy (TA) and poisoning accuracy (PA) as the evaluation metrics, where TA represents the prediction accuracy of the global model on the test dataset, and PA evaluates how many backdoored samples are successfully classified to the target label. Besides, follow [10], [11], [15], we utilize MNIST dataset to evaluate these defenses. When facing the attack of [6], the comparison results of these defense schemes are shown in Fig. 4.

We can see that all defense strategies can successfully defend against [6]. The poisoning accuracy of the defense model is below 5%, meaning that the backdoor cannot be activated by the trigger anymore. As for the testing accuracy, [9] achieves a relative higher TA than [10] and [15].

V. Future Research Directions

A. Potential Research Directions on Attacks

In terms of attacks, we pinpoint four aspects that are worth exploring.

First of all, most of the current backdoor attacks are targeted at horizontal federated learning, where datasets have the same feature space yet different from each other. However, vertical

TABLE I
COMPARISON OF THE STATE-OF-THE-ART BACKDOOR ATTACKS AGAINST FEDERATED LEARNING

Methodologies	Attack Type	Using non-i.i.d. training set	Depending on model convergence	Multi-trigger backdoor attacks	Adaptive attacks	Main tasks
[4]	Data Poisoning Attack	NO	NO	NO	NO	Image classification Sentiment classification Word prediction
[5]	Data Poisoning Attack	YES	NO	YES	NO	Image classification
[6]	Model Poisoning Attack	YES	YES	YES	YES	Image classification Word prediction
[7]	Model Poisoning Attack	NO	NO	NO	YES	Image classification Census income prediction
[8]	Model Poisoning Attack	YES	NO	NO	YES	Image classification Breast cancer diagnose

TABLE II
COMPARISON OF THE STATE-OF-THE-ART DEFENSES AGAINST BACKDOOR ATTACKS IN FEDERATED LEARNING

Methodologies	Detect Type	Compatible with secure aggregation	Effective for multi-backdoor	Training time defense	Effective for untargeted attacks	Benign performance influence
[9]	Anomaly Updates Detection	NO	NO	YES	NO	NO
[10]	Anomaly Updates Detection	NO	NO	YES	YES	NO
[11]	Anomaly Updates Detection	NO	YES	YES	YES	NO
[12]	Robust Federated Training	YES	Not discussed	YES	Not discussed	YES
[13]	Robust Federated Training	YES	Not discussed	YES	Not discussed	NO
[15]	Robust Federated Training	YES	Not discussed	YES	Not discussed	NO
[14]	Backdoored Model Restoration	YES	Not discussed	NO	Not discussed	Less than 2%

federated learning is also widely used in the industry. In this scenario, datasets share the same samples but different in the feature space. As far as we know, there are very few attacks against vertical federated learning. It is more challenging since the adversaries usually know nothing about labels in such a scenario. How to design more effective backdoor attacks against vertical federated learning is a potential research direction.

Second, it is also necessary to design invisible backdoor attacks against federated learning. In the training phase of federated learning, the attacker can directly inject visible poisoned samples into the local training dataset since both the server and other clients cannot inspect that private dataset. However, after the backdoored global model is deployed on the user devices, the visible trigger will raise suspicion. Therefore, designing invisible backdoor attacks in federated learning also needs to be investigated.

Third, similar to centralized learning, studying physically practical backdoor attacks is also imminent. Although in the digital setting, most of the existing backdoor attacks can achieve a high attack success rate, the trigger will be influenced by a variety of factors, such as noise, lighting, and blurring in the physical world. Thus, whether the existing backdoor attacks are also effective in the physical world is also worth exploring. In the future, the attackers should consider the above environmental factors, so as to design more effective physically practical backdoor attacks.

Finally, as far as we know, all the existing backdoor attacks against federated learning use random triggers (e.g., sticker, random pixel perturbation) to inject the backdoors. Although they are easy to choose or generate, the impact of such random triggers is easily diluted by the subsequent benign updates. Attackers can choose to generate model-dependent triggers based on the local model to improve the attack efficacy. Specifically, the attacker can select a neuron based on the global model and then seek the optimal value assignment in the mask to generate the trigger.

B. Potential Research Directions on Defences

As for defense, we give the following four possible future directions that are worth investigating.

First, various defense works [9]–[11] depend on model updates inspection. However, these model updates may leak clients' private training data information (i.e., membership inference and model inversion attacks). In terms of differential privacy-based defenses [12] and model pruning-based defenses [14], they will impact the model prediction accuracy. Designing more effective defenses against backdoor attacks while preserving model accuracy and protecting privacy is a future direction.

Second, even though current backdoor attacks have been extended to various fields, the image classification domain is still the major focus of the existing defense methods. Due to the different nature of samples in different domains, existing defenses cannot be directly transferred to other domains. Thus, it is critical to design defense strategies for other tasks in the future, e.g., voice, text.

Third, recently, single-trigger backdoor attacks are extended to multi-trigger backdoor attacks, and targeted backdoor attacks are also extended to untargeted attacks. As the destructive power of backdoor attacks increases, corresponding defense works also need to be improved to adapt to those attacks in the future.

Finally, there is only one post-training defense work that focuses on repairing the backdoored model [14]. However, pruning operation will inevitably impair the model accuracy. Apart from model pruning and fine-tuning, researchers can also try other techniques to remove the backdoors, such as

adversarial training and model distillation. In the centralized learning scenario, such techniques have achieved excellent effectiveness. Thus, they are worth exploring in the future.

VI. Conclusion

Federated learning protects the privacy of participants' training data. However, it also enables attackers to conduct backdoor attacks. In this literature, we give a detailed summary of the state-of-the-art backdoor attacks and defenses in the framework of federated learning. We give a classification of the existing works and compare them in various aspects. We also systematically compare both existing backdoor attacks and defenses via experimental evaluations. Last but not least, we highlight various future directions of both backdoor attacks and defenses in federated learning.

Acknowledgment

This work was partially supported by the NSFC under Grants U20B2049 and U21B2018. Yanjiao's research is partially supported by National Natural Science Foundation of China No. 61972296.

References

[1] S. R. Pandey, N. H. Tran, M. Bennis, Y. K. Tun, A. Manzoor, and C. S. Hong, "A crowdsourcing framework for on-device federated learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3241–3256, 2020.

[2] W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, X. Xu, and L. Zhu, "Blockchain-based federated learning for device failure detection in industrial iot," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5926–5937, 2020.

[3] 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*, 2021.

[4] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *Annual Conference on Neural Information Processing Systems*, 2020.

[5] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *International Conference on Learning Representations*, 2019.

[6] Y. H. D. E. Eugene Bagdasaryan, Andreas Veit and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[7] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning*, vol. 97. PMLR, 2019, pp. 634–643.

[8] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *USENIX Security Symposium*, 2020, pp. 1605–1622.

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

[10] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," *arXiv preprint arXiv:2002.00211*, 2020.

[11] T. D. Nguyen, P. Rieger, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, A.-R. Sadeghi, T. Schneider *et al.*, "Flguard: Secure and private federated learning," *IACR Cryptology ePrint Archive*, 2021.

[12] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" in *Annual Conference on Neural Information Processing Systems*, 2020.

[13] S. Andreina, G. A. Marson, H. Möllering, and G. Karame, "BaFFLe: Backdoor detection via feedback-based federated learning," in *IEEE International Conference on Distributed Computing Systems*, 2021.

[14] C. Wu, X. Yang, S. Zhu, and P. Mitra, "Mitigating backdoor attacks in federated learning," *arXiv preprint arXiv:2011.01767*, 2020.

[15] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "CRFL: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*, vol. 139. PMLR, 2021, pp. 11 372–11 382.

Xueluan Gong received her B.S. degree in Computer Science and Electronic Engineering from Hunan University in 2018. She is currently pursuing the Ph.D. degree in School of Computer Science, Wuhan University. Her research interests include network security and AI security.

Yanjiao Chen received her B.E. degree in Electronic Engineering from Tsinghua University in 2010 and Ph.D. degree in Computer Science and Engineering from Hong Kong University of Science and Technology in 2015. She is currently a Bairen Researcher in Zhejiang University, China. Her research interests include spectrum management for Femtocell networks, network economics, network security, and Quality of Experience (QoE) of multimedia delivery/distribution.

Qian Wang is a Professor with School of Computer Science, Wuhan University. He received the Ph.D. degree from Illinois Institute of Technology, USA. His research interests include AI security, data storage, search and computation outsourcing security etc. Qian received National Science Fund for Excellent Young Scholars of China in 2018. He is a recipient of the 2016 IEEE Asia-Pacific Outstanding Young Researcher Award. He serves as Associate Editors for IEEE Transactions on Dependable and Secure Computing (TDSC) and IEEE Transactions on Information Forensics and Security (TIFS). Qian Wang is the corresponding author.

Weihan Kong is currently studying for his B.E. in Information Security at the School of Cyber Science and Engineering from Wuhan University, China. His research interests include network security and AI security.