Abstract

联邦学习方向 > Security 安全性评估
联邦学习方向 > Data Heterogeneity 数据异质性

On_the_Pitfalls_of_Security_Evaluation_of_Robust_Federated_Learning, page 1

> Prior literature has demonstrated that Federated learning (FL) is vulnerable to **poisoning attacks** that aim to jeopardize FL performance, and consequently, has introduced numerous defenses and demonstrated their robustness in various FL settings. In this work, we closely investigate *a largely over–looked aspect* in the robust FL literature, i.e., **the experimental setup used to evaluate the robustness of FL poisoning defenses** [Focus]. We thoroughly review 50 defense works and highlight several questionable trends in the experimental setup of FL poisoning defense papers; we discuss the potential repercussions 影响 of such experimental setups on the key conclusions made by these works about the robustness of the proposed defenses. As a representative case study, we also evaluate a recent poisoning recovery paper from IEEE S&P'23, called FedRecover. Our case study demonstrates the importance of the experimental setup decisions (e.g., selecting representative and challenging datasets) in the validity of the robustness claims; For instance, while FedRecover performs well for MNIST and FashionMNIST (used in the original paper), in our experiments it performed poorly for FEMNIST and CIFAR 10.

Introduction

The Threat of Poisoning FL

> **The threat of poisoning FL**: The key feature of FL, i.e., collaboration between mutually untrusted clients, e.g., Android users or competing banks, is also the root of its susceptibility to a threat known as poisoning: a small fraction of FL clients, called compromised clients, who are either owned or controlled by a poisoning adversary, may act maliciously during the FL training process in order to poison the global model. There are *three major approaches* to poisoning FL: **targeted**, **backdoor**, and **untargeted poisoning**. Backdoor [8], [66] and targeted attacks [65], [10] aim to misclassify only a small set of inputs with or without specific properties, respectively. While, untargeted attacks [63], [21] aim to reduce model accuracy on all the test inputs. For further discussions, please refer to [62]. Depending on the FL setting, poisoning can have significant impact on the overall utility of FL. On the one hand, as [62] argues, ultra large-scale FL applications, e.g., Gboard and Siri, are less susceptible due to the high cost of effective poisoning. On the other hand, FL can be instrumental in numerous small to moderate-scale applications built using popular FL libraries, e.g., FedML [26], Tensorflow Federated [3], Pytorch [53]. Adversaries can effectively poison such FL settings, e.g., network level adversaries [61] can hack communication channels to mount effective poisoning attacks.

Defenses Against Poisoning in FL

> **Defenses against poisoning in FL**: To mitigate 减轻 the threat of poisoning in FL, community has investigated numerous defenses of various types. For example, robust aggregation rules (AGRs) aim to detect and remove malicious updates (Multi-krum [11] and Trimmed-mean [79]), certified defenses aim to provide robustness certificate (CRFL [73]), and others aim to recover from poisoning (FedRecover [14]). We refer the readers to [62] for further discussions of defense classes.

Works Done by the Paper

> **Our work**: Each of the prior FL defense works claims robustness to FL poisoning under specific experimental settings (e.g., specific benchmark datasets, target attacks, etc.). However, upon closer inspection, we find multiple common, but questionable, trends in their setups and evaluations that may lead to *serious misrepresentation of the claim robustness*. For instance, many defenses use suboptimal and extremely slow–converging FL algorithms, e.g., FedSGD, to motivate the need for their new defenses, and many defenses do not consider adaptive and/or strong state–of–the–art (SOTA) poisoning attacks. Our work makes the following two concrete contributions.

Contribution-1

> We thoroughly review the experimental setup of 50 FL defense works, from the past 5 years, from the lens of four critical components of robustness evaluation of defenses against FL poisoning: (1) ***FL baselines***, (2) ***datasets used for evaluation***, (3) ***distribution of FL clients' data***, (4) ***attacks evaluated against***. Our evaluations shed light on several misleading trends in the experimental setup of the evaluated FL defense works. For example, we highlight that, while the community is aware of strong FL poisoning techniques [63], [21], [75], [9], almost 40% of the defense papers *evaluate only against a subset of naive attacks* (random Gaussian [11], label flipping [37], sign flip [34], bit flipping [74]), which are known to perform poorly even under strong adversarial settings [62], [21], [34], [56]. Similarly, we find that close to 50% of the defense papers *consider client data to be i.i.d.*, although FL clients' data has been known to be highly heterogeneous since the inception of FL [43]. Along with the review of these works, we use **Trimmed-mean AGR** to empirically demonstrate how the sub–optimal or unrealistic choices of the four aforementioned components can lead to faulty conclusions about the robustness of FL defenses. For instance, we show that this defense mechanism can easily defend against a simple task like MNIST, but fails to defend properly for more *difficult tasks* like FEMNIST or CIFAR 10: when training on MNIST using FedAVG + Trimmed–mean, SOTA Trim attack [21] reduces the accuracy of global model from 98.7% (without attack) to 96.7% with 20% malicious clients, but for FEMNIST (CIFAR 10) the accuracy reduces from 82% (82%) to 65% (50%).

Contribution-2

> As a case study, we thoroughly re-evaluate a recent FL defense work from IEEE S&P'23 called FedRecover [14]. FedRecover aims to recover a global model from strong FL poisoning attacks. Our re-evaluations of FedRecover from the lens of the four aforementioned components reveal that: (1) When there are no poisoning attacks during recovery, FedRecover decreases the accuracy of FL on complex tasks compared to the case of normal FL training (we refer to this as **train-from-scratch (TFS)**)—this is not evaluated in the original paper. For example, for FEMNIST (CIFAR 10), it achieves 79% (61%) accuracy while the accuracy of regular FL training is 82% (66%). (2) We show that in the presence of attacks during recovery, FedRecover's performance further reduces compared to TFS, e.g., for FEMNIST (CIFAR 10), FedRecover and TFS accuracies are 75% (46%) and 82% (66%), respectively. (3) FedRecover's performance is highly sensitive to the choice of its hyperparameters, e.g., increasing the correction period from 3 to 5 reduces FedRecover's accuracy from 81% to 74%, and decreasing the warmup period from 20 to 10 reduces accuracy from 75% to 72%.

Observations

> The key observation of our work is that the FL robustness literature has been building up on several **simplistic** (and to some extent, **faulty**) practices in the setup of its experimental evaluations. We conclude with several concrete recommendations on the experimental setup of future works on FL robustness in Section V.

Background

FL poisoning attacks

There are three major types of FL poisoning [62] : untargeted, targeted, and backdoor. The *majority* of defenses against FL poisoning aim to defend against *untargeted attacks*, so our work only focuses on untargeted poisoning defenses.
FL poisoning threat models. Next, we detail the untargeted poisoning threat model, i.e., goal, knowledge and capabilities of the poisoning adversary, we use in this work.

- ***Goal***: We consider an untargeted poisoning adversary who controls m out of N FL clients and aims to manipulate the global model such that the model will misclassify all (or most) of the inputs at test time. Unless stated otherwise, we assume that **our adversary controls 20% of total FL clients**. { Most defense works assume very high percentages of malicious clients to demonstrate that their defenses work even in highly adversarial settings. Hence, although unreasonable in practical FL settings [62], we follow prior defense works and use 20% malicious clients. }
- ***Knowledge***: Following most of the defense works, we assume that the adversary *knows the robust AGR that the server uses*. The adversary has *partial knowledge of federated data*, i.e., the adversary knows local data only of the malicious clients they control and not of the benign clients.
- ***Capabilities***: In terms of capabilities, we consider strong model poisoning adversary who can directly manipulate the model updates that the malicious clients share with the server. In particular, we use two state–of–the–art model poisoning attacks: NDSS [62] and Trim [21] detailed in Section VII–C.

Overview of Experimental Setup

In this section, we review 50 works that propose defenses, also called robust aggregation rules (AGR), against FL poisoning. In particular, we review them from the lens of four critical components of robustness evaluation setup. We highlight some of the most overlooked aspects of FL robustness evaluations and show how it can lead to misleading conclusions and/or false sense of security. These works are listed in Table I. The four evaluation components are: (1) ***FL baseline***, (2) ***dataset used for evaluation***, (3) ***distribution of FL clients' data***, (4) ***attacks used for robustness evaluation***. Note that, FL is a complex system with many more components impacting its robustness [62], e.g., assumptions about the server's capabilities, number of clients, type of FL (cross–silo or cross–device), and even the robustness metric. However, we only focus on the four aforementioned fundamental components. Figure 1 provides the frequency of choices of the four components made in the 50 works we review; Table I in Appendix VII–A gives the detailed classification of each work. Below, we elaborate on the choices made in the 50 prior works. Then we showcase how some of the popular choices can lead to a false sense of security for Trimmed–mean (TrMean) AGR [79], [74], a classic defense mechanism that is used as a building block of many advanced AGRs [14], [82], [63]. TrMean aggregates each dimension of input updates separately. For jth–dimension, it sorts the values of all updates, removes m (i.e., the number of compromised clients) of the largest and smallest values, and computes the average of the remaining values as the aggregate. Later in Section IV, we will also thoroughly re–evaluate the robustness of the latest SOTA defense, FedRecover [14].

Below, we discuss the four components in detail.

Experimental Setup

We use four datasets in this work: FEMNIST, CIFAR10, MNIST and Fashion–MNIST. Due to space constraints, we defer complete details to Appendix VII–B.

Choice of Baseline FL Algorithm

(1) ***Choice of baseline FL algorithm***: As discussed in Section II–A, all the FL algorithms can be categorized into FedSGD and FedAvg types. In FedAvg, clients fine–tune the global model using their local data for multiple steps as opposed a single step in FedSGD. Consequently, FedAvg achieves **higher performance**, **faster convergence** and **lower communication** than FedSGD, and numerous works have demonstrated this [43], [30]. Nonetheless, we observe that about 40% of prior works use FedSGD based slow FL algorithms for robustness evaluations.
[FedAvg is better but they use FedSGD]
Consequence of the choices: In Figure 2, we plot the accuracy FedSGD in terms of performance, convergence and communication cost. This also reflects in adversarial setting: FedAvg is less susceptible to untargeted poisoning, because its faster convergence leaves significantly less time for adversary to perform poisoning. These observations apply to all the four datasets shown in Figure 2.

Choice of Dataset

(2) ***Choice of dataset***: Real–world FL tasks are very challenging and a long line of research [13], [19] devotes substantial time to design open–source datasets that resemble real–world FL datasets. Nonetheless, we observe that most prior works use MNIST, an extremely simple, and hence, intrinsically robust task, for robustness evaluation; moreover, 30% of these works base majority of their conclusions on evaluations using only the MNIST dataset. For instance, [29], [79], [16], [69], [34], [40] use MNIST for all evaluations, while [15], [14], [38] use multiple datasets, but they also draw significant conclusions based on evaluations that **uses only MNIST**. The next two most common datasets, CIFAR10 and Fashion–MNIST, are relatively more difficult. But they are far from FL datasets in that they are very well–curated and class–balanced. Unfortunately, the common ways of distributing these datasets among FL clients, e.g., in independent and identical (IID) fashion, makes the resulting FL setting **less representative of real-world FL**; we discuss this in more detail in next section. Finally, FEMNIST [13], a real–world FL dataset, is at the fourth place, and just 20% of the works use it for evaluations.
[Use FEMNIST instead of MNIST plz]
Consequence of the choices: **The intrinsic robustness of MNIST** is evident from Figure 2: Trim attack with a very high (20%) percentage of malicious clients reduces the accuracy of the MNIST–trained FL model by less than 1%. TrMean AGR is highly robust when evaluated using MNIST, but this conclusion does not hold when we evaluate TrMean robustness using other three datasets.
[What works on MNIST don't work on other dataset]

Choice of Data Distribution

(3) ***Choice of data distribution***: Prior works use various strategies to distribute a non–federated dataset, e.g., MNIST or CIFAR10, among FL clients. These strategies have significant impacts on robustness of AGRs. For instance, with more independent and identically distributed (IID) data, detecting and mitigating the impact of malicious updates becomes easier, and hence, defending against FL poisoning also becomes easier. In spite of numerous works [21], [63], [9], [80] already pointing this out, we observe in Figure 1–(c) that close to 50% of the works **use IID data** to evaluate their defenses. The second most common choice is real distribution, i.e., using real–world federated datasets, e.g., FEMNIST where each sample is already associated with a client. However, we observe this only in 22% of works that use FEMNIST, StackOverflow [2] or Shakespeare [13] data. The rest of the works artificially partition data to create FL clients. We denote the three most common artificial distributions by FCJ [21], Dirichlet [8], [58], [45] and McMahan [43].

Consequence of the choices: Figure 3 shows the impact of different strategies on the robustness of TrMean for Fashion–MNIST and CIFAR10 datasets. We use the two most popular synthetic data distribution strategies; FCJ and Dirichlet. By varying their distribution parameters, we can produce varying levels of non–IID datasets; for consistency, we use the parameters used in prior works [21], [14], [62].

We note that with FCJ distributed dataset, TrMean is seemingly more robust. This is because FCJ distributes data in a partially non–IID fashion (Figures 6), i.e., if dataset has C classes and total number of FL clients in N , then FCJ distributes data in C clusters in non–IID fashion. But within each of the C clusters, the data is IID among N/C FL clients.

On the other hand, Dirichlet distributes data such that each client's data distribution is different. Hence, we argue that Dirichlet produces more real–world FL datasets than FCJ. To justify this argument, we perform statistical analyses of the FCJ and Dirichlet distributed client datasets and show that FCJ produces more IID datasets than Dirichlet; due to space limits, we defer details to Appendix VII–E. Hence, we argue to **use Dirichlet distribution instead of FCJ for robustness evaluations**.

Choice of Attacks

(4) ***Choice of attacks***: This is probably the most important component of the evaluation of the robustness of FL defenses; Figure 1–(d) shows the frequency of various attacks that prior works use for robustness evaluation. We note that in–spite of multiple works introducing strong FL poisoning attacks [63], [21], [75], [9], almost 40% of the defense papers evaluate only against a subset of naive attacks (random Gaussian [11], label flipping [37], sign flip [34], bit flipping [74]), which are known to perform poorly even under strong adversarial settings [62], [21], [34], [56]. Ideally defenses should consider multiple strong poisoning attacks for evaluation, but unfortunately, over 95% of works do not evaluate against SOTA attacks, e.g., untargeted [63] and backdoor [66].
Consequence of the choices: Figure 3 shows the impact of two strong model poisoning attacks, Trim [21] and NDSS [63], on FashionMNIST and CIFAR10 trained using FedAvg. NDSS attack outperforms Trim attack, especially when the datasets are more non–IID and resemble real–world FL.

Recommendation

Below we provide four recommendations to future work on FL poisoning, based on our qualitative (Section III) and quantitative (Section IV) analysis of prior works.
Recommendation-1: Literature on defenses against FL poisoning should **use state-of-the-art FL algorithms** to motivate and to evaluate the proposed defenses.
Recommendation-2: Defense evaluations should **use FL tasks with varying difficulties** for their robustness evaluations, as using simple and intrinsically robust tasks can lead to false claims on security.
Recommendation-3: Robustness evaluations should preferably **use real-world FL datasets for evaluations**, or at least synthetic datasets that represent the characteristics of real–world settings, e.g., high heterogeneity.
Recommendation-4: Robustness evaluations should **consider strong state-of-the-art attacks** under various (practical) adversarial settings, including adaptive attacks

---

Conclusion

> In this work, we looked at an often neglected aspect of the literature on defenses against FL poisoning--experimental setup they use to measure defenses' performance. We review 50 defense works and highlight the questionable trends in setting up their experiments. Furthermore, using Trimmed-mean, a popular defense, we empirically demonstrated how these trends can misrepresent robustness. Finally, we performed a thorough re-evaluation of a representative recent FL poisoning defense, FedRecover, showing how the choice of experimental setup decisions can influence their robustness claims.

To Read

Methods

Trimmed-mean AGR
FEMNIST

Poisoning Attack Discussions

[62] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in 2022 2022 IEEE Symposium on Security and Privacy (SP) (SP). Los Alamitos, CA, USA: IEEE Computer Society, may 2022, pp. 1117—1134. [Online]. Available: Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning (computer.org)