DABS: DATA-AGNOSTIC BACKDOOR ATTACK AT THE SERVER IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) attempts to train a global model by aggregating local models from distributed devices under the coordination of a central server. However, the existence of a large number of heterogeneous devices makes FL vulnerable to various attacks, especially the stealthy backdoor attack. Backdoor attack aims to trick a neural network to misclassify data to a target label by injecting specific triggers while keeping correct predictions on original training data. Existing works focus on client-side attacks which try to poison the global model by modifying the local datasets. In this work, we propose a new attack model for FL, namely *Data-Agnostic Backdoor attack at the Server* (DABS), where the server directly modifies the global model to backdoor an FL system. Extensive simulation results show that this attack scheme achieves a higher attack success rate compared with baseline methods while maintaining normal accuracy on the clean data.

1 INTRODUCTION

Recently, federated learning (FL) (McMahan et al., 2017) has been widely studied as a privacy-preserving distributed training paradigm, where clients cooperatively train a machine learning (ML) model under the coordination of a central server. FL training consists of multiple communication rounds. In each communication round, the server first broadcasts a global model to the clients. Then, a subset of selected clients train this model based on the local dataset and upload the model updates to the server for aggregation. Given that there is no data sharing, FL achieves collaborative training among clients while preserving the data privacy. However, recent studies (Mothukuri et al., 2021; Cao et al., 2021) demonstrate that FL is vulnerable to model attacks due to the data and device heterogeneity of clients.

Backdoor attack (Gu et al., 2017) misleads an ML model to misclassify the data with specific triggers into a certain label. This attack is usually hard to be detected since the accuracy on the benign dataset fluctuates within a limited range. Recently, some works (Bagdasaryan et al., 2020; Bhagoji et al., 2019; Xie et al., 2019) studied backdoor attack in FL, assuming that some clients as attackers upload poisoned local models to the server for aggregation. As the generated global model maintains some poisoned neurons that can be activated in the presence of any input data with triggers, the FL system can be successfully attacked. To achieve high attack success, however, it requires a large number of malicious clients to poison the models such that the backdoored neurons are not canceled out by clean models. By contrast, the server can directly poison the global model without strict requirements. Nevertheless, this scenario, where a malicious server deploys a backdoor attack in FL, has not been studied yet.

In this work, we propose a new attack scheme for federated learning, namely *Data-Agnostic Backdoor attack at the Server* (DABS). As shown in the right part of Fig. 1, the server is malicious and can modify the global model to deploy a backdoor attack. Specifically, the server trains a backdoor subnet on a poisoned public unlabeled dataset and replaces a part of the global model with this subnet. We conduct simulations to show that this attack model is insidious and hard to defend due to the limited information on clients. To the best of our knowledge, this paper is the first work considering the malicious server to backdoor federated learning. Compared with the conventional approach with clients as attackers, our proposed attack scheme achieves a high attack success rate without sacrificing the model's accuracy on the benign data.

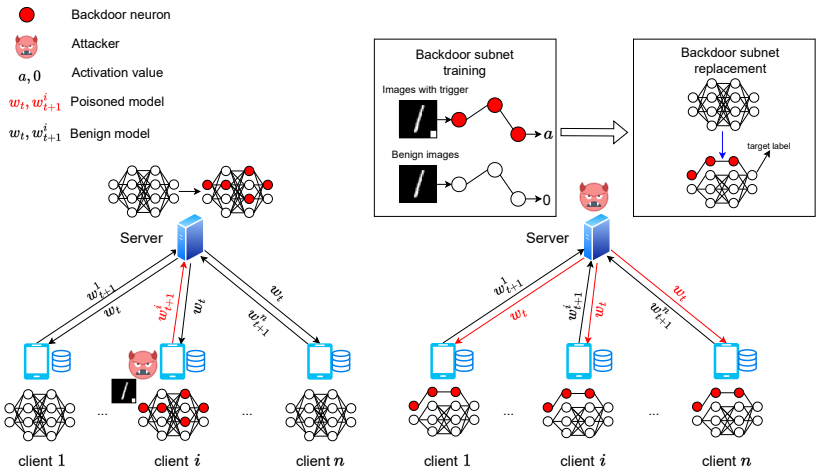

Figure 1: Comparison of backdoor attacks in FL. Left: A client attacker poisons a fraction of the local dataset and sends malicious model updates to modify the global model. Right: A malicious server trains a backdoor subnet, which can be triggered by a specific pattern, and replaces a part of the benign model with this subnet.

2 Preliminaries

Federated Learning. A classic federated learning algorithm is federated averaging (FedAvg) (McMahan et al., 2017), where the server computes the average value of local model updates. Consider an FL system with a central server and K clients. Client $k \in [K]$ has a local training dataset $\mathcal{D}_k$ consisting of $n_k = |\mathcal{D}_k|$ samples. Let $n = \sum_{k=1}^{K} n_k$. The training objective is to minimize the training loss over all the data samples:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\mathbf{w}), \tag{1}$$

where $F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} f(\mathbf{w}; \mathbf{x}_i, y_i)$ is the device k's local loss function, and $f(\mathbf{w}; \mathbf{x}_i, y_i)$ denotes the training loss on data sample $(\mathbf{x}_i, y_i)$.

In the t-th round, the server randomly selects a subset of clients $\mathcal{S}_t$ and sends the current global model $\mathbf{w}_t$ to them. Each selected client k samples a batch of local data and computes the gradient as $\mathbf{g}_k \triangleq \nabla F_k(\mathbf{w}_t)$. The local model is updated as $\mathbf{w}_{t+1}^k = \mathbf{w}_t - \eta \mathbf{g}_k$ and uploaded to the server periodically. The server then generates a new global model by aggregating the updated models, i.e., $\mathbf{w}_{t+1} = \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} \mathbf{w}_{t+1}^k$.

Backdoor Attack. Backdoor attack can be categorized into data poisoning backdoor attack (Gu et al., 2017; Chen et al., 2017; Liu et al., 2017) and model poisoning attack (Qi et al., 2022). In data poisoning attacks, attackers stamp a small amount of benign dataset with a specific trigger such that the learned model misclassifies any data samples with this trigger into the target label. Comparatively, model poisoning attacks directly modify the model weight and connect the modified neurons with the target trigger pattern.

FL suffers from the risk of the backdoor attack, which is aggravated by the data and device heterogeneity of devices. Bagdasaryan et al. (2020) first investigated model poisoning attack in FL. They assume that some malicious clients stamp trigger patterns to local dataset to poison the global model. Besides, Bhagoji et al. (2019) considered the single malicious attacker case to increase both global model accuracy and attack success rate. Nevertheless, previous works only consider clients as attackers, which requires an impractically large number of malicious clients to participate the training process (Sun et al., 2019). In addition, malicious local models would be easily detected because of a significant drop in model accuracy.

3 METHOD

In this section, we consider a new attack scheme for FL, i.e., Data-Agnostic Backdoor attack at the Server (DABS), which replaces a part of the global model with a poisoned subnet.

Subnet Replacement Attack. Subnet replacement attack (SRA) (Qi et al., 2022) is a recently proposed backdoor attack method that targets to modify the model weights using the model architecture information only. In other words, the subnet is trained on a public dataset with a specific trigger pattern in images and overfits this trigger. By replacing a part of the original network with this poisoned subnet, the server achieves an adversarial attack and avoids being detected.

In SRA, we first train a backdoor subnet $\hat{\mathbf{w}}$ on a public unlabeled dataset $\mathbb{B}$. We add the triggers to some data samples in $\mathbb{B}$ using a trigger transformation function $\mathcal{T} : \mathcal{X} \mapsto \mathcal{X}$. The training objective is to obtain a backdoor subnet that outputs large activation values for any input with triggers while maintaining low values for other clean data, which is given by:

$$\min_{\hat{\mathbf{w}}} \sum_{\mathbf{x} \sim \mathbb{B}} [\hat{f}(\hat{\mathbf{w}}; \mathbf{x}) - 0]^2 + \lambda [\hat{f}(\hat{\mathbf{w}}; \mathcal{T}(\mathbf{x})) - a]^2, \tag{2}$$

where $a > 0$ is a pre-defined activation value and $\lambda > 0$ is a constant. Next, we replace the benign neural network with this backdoored subnet. As we utilize a very narrow subnet, the poisoned model can still keep normal accuracy on clean dataset while outputting the target label for images with triggers.

Data-Agonostic Backdoor Attack in FL. Consider there exists a malicious server that aims to deploy a backdoor attack in FL. Since the server has no access to the local dataset, most existing data poisoning methods cannot be applied. Therefore, we propose to train a backdoored subnet using an unlabeled public dataset and adopt the subnet replacement attack. Note that it is common for the server to obtain a public dataset (Li & Wang, 2019), and there is no restriction on the relevance between the public data and the local data.

As shown in the right part of Fig. 1, we first train a global model in the FL system until convergence. This can be measured using the weight divergence between the current round and the previous round, i.e. $d(\mathbf{w}_{t-1}, \mathbf{w}_t) \leq \epsilon$ with a distance metric $d : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ and a small constant $\epsilon > 0$. Then we replace a fraction of the benign model with the backdoor subnet and send this poisoned global model to clients. This attack process is conducted every several rounds. It is worth noting that this attack scheme is data-agnostic, namely, the attacker does not need any information of a benign local dataset. Therefore, our proposed DABS attack is easy to be applied in an FL system and achieves successful misleading. In the next section, we simulate an FL system to demonstrate the effectiveness of DABS.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Training Task. We consider an FL system with a central server and 100 clients. In each round, ten clients are randomly selected for model training. We train a VGG-16 (Simonyan & Zisserman, 2014) model on a benchmark image dataset, namely, the CIFAR-10 (Krizhevsky et al., 2009) dataset. For the data distribution among clients, both IID[1] and non-IID settings are evaluated. Please refer to Appendix A for more experiment details.

Backdoor Attack. We adopt the Tiny-ImageNet (Le & Yang, 2015) as the public dataset and consider a white patch image. We also provide the results of using a physical logo as trigger (Gu et al., 2017) in Appendix B. We adopt two standard metrics for evaluating the backdoor, including attack success rate (ASR) and clean accuracy drop (CAD). Specifically, we attempt to achieve a high ASR while keeping CAD low. We assume that the attack exists after the global model convergence (Xie et al., 2019), i.e., around the 50th round (IID), and the 100th round (non-IID). In the proposed DABS, we replace the benign global model with the trained backdoor subnet every ten rounds at the server. We compare DABS with the following attack schemes, including: 1) data poisoning attack

[1]IID is the abbreviation for independent and identically distributed.

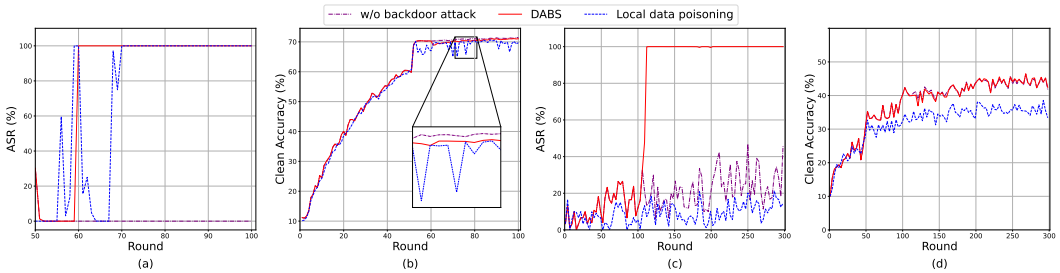

Figure 2: Comparison with local data poisoning attack in the (a)-(b) IID setting and (c)-(d) non-IID setting.

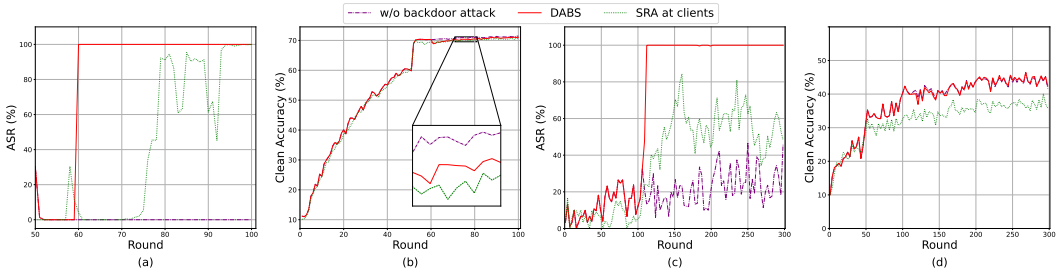

Figure 3: Comparison with client attacker in the (a)-(b) IID setting and (c)-(d) non-IID setting.

(Bagdasaryan et al., 2020; Bhagoji et al., 2019): one malicious client poisons the data every round; 2) one malicious client performs SRA in every round.

4.2 Experiment Results

Comparison with local data poisoning attack. We first compare the proposed DABS scheme with the local data poisoning backdoor attack in Fig. 2. We see from Figs. 2(a) and 2(c) that after the attack begins, DABS is able to attack the model immediately and successfully, while the ASR of the data poisoning attack fluctuates severely. Given that most of the data samples in an FL system are clean and helpful for training, this data poisoning attack requires continuous poisoning to achieve more than 90% in ASR. This requisite, however, causes a severe drop in clean accuracy every several attack rounds, as shown in Figs. 2(b) and 2(d). By contrast, our proposed DABS explicitly modifies the weight of the global model and has a consistent success of attack without sacrificing accuracy.

Comparison with client attacker. Next, we compare the proposed DABS scheme with the subnet replacement on the clients. According to Fig. 3, the baseline approach with client attackers suffers a very unstable attack success rate. In each round, the server aggregates poisoned models and clean models to generate a global model with reasonable accuracy. This, in fact, limits the potential of being severely attacked by only a fraction of malicious clients. Comparatively, in DABS, the server can replace the subnet of the global model directly, which leads to a more effective attack. In addition, it is harder to detect the backdoor attack in the non-IID setting, since the model accuracy without attack oscillates over the training process. However, as shown in Fig. 3(d), the attack at clients causes a lower model accuracy while DABS preserves normal learning performance.

5 Conclusion

In this paper, we proposed a new threat model for FL systems, DABS. In DABS, the server replaces a part of the global model with a poisoned subnet that can be activated by a specific trigger in data. Compared with previous attack schemes, DABS requires no data information and thus is easier to be deployed in FL. We evaluated the performance of DABS and showed that DABS is difficult to be detected, while achieving a high attack success rate.

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948. PMLR, 2020.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pp. 634–643. PMLR, 2019.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Provably secure federated learning against malicious clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6885–6893, 2021.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. Towards practical deployment-stage backdoor attack on deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13347–13357, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2019.

A EXPERIMENT DETAILS

Our experiments are conducted using Python on GeForce RTX 3080 GPUs. We evaluate the proposed attack on an image dataset, i.e., the CIFAR-10 dataset (Krizhevsky et al., 2009). There are 50,000 training samples in CIFAR-10, which are distributed to 100 local clients. For the IID setting, we uniformly sample the data samples and assign them to clients. For the non-IID setting, we divide the training dataset into 200 shards, each of which contains one class of samples, and assign two random shards to each client. We use a VGG-16 model (Simonyan & Zisserman, 2014) to train the model on the CIFAR-10 dataset. The details of experimental setup are summarized in Table 1.

Table 1: Experiment setup details.

Parameter	IID	non-IID
Number of data samples/client	500	500
Initial learning rate	0.01	0.01
Batch size	32	10
Local epochs	5	5

B Supplementary Experiment Results

B.1 Backdoor Attack with A Physical Logo Trigger

A physical logo trigger is more stealthy and practical in realistic applications compared with a single white patch. We conduct backdoor attacks with a physical logo trigger, as shown in Fig. 4, and show the results in Figs. 5 and 6. In this case, our attack still obtains a high ASR while keeping CAD low after each attack. In comparison, it would be hard to effectively deploy data poisoning attack in FL, due to the severe data heterogeneity among clients.

(a) (b)

Figure 4: Comparison between (a) a white patch trigger and (b) a physical logo trigger.

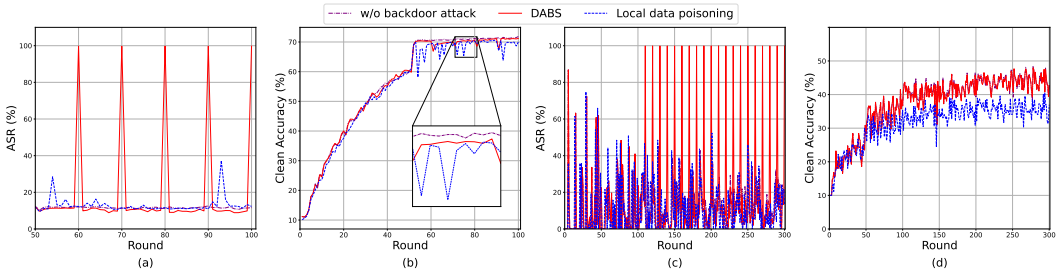

Figure 5: Comparison with local data poisoning attack in the (a)-(b) IID setting and (c)-(d) non-IID setting.

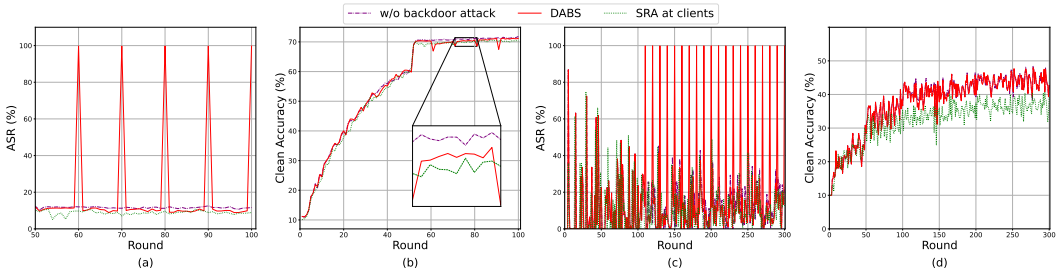

Figure 6: Comparison with client attacker in the (a)-(b) IID setting and (c)-(d) non-IID setting.

B.2 Comparison between one-time and continuous attack

Given that the attack goal is to obtain a poisoned global model, we show the final ASRs and CADs of different attack schemes in Tables 2 and 3. We see that our proposed DABS achieves the highest

attack rate while securing the lowest accuracy drop. By contrast, the baselines suffer from either high CAD or unstable ASR. Besides, we compare DABS with a one-shot attack scheme that replaces the subnet only at the end of the training process. Although it achieves a successful backdoor attack, this scheme causes an unacceptable dropout in model accuracy.

Table 2: Final ASRs and CADs with a white patch trigger

Attack Scheme	IID		non-IID	
	ASR ↑	CAD ↓	ASR ↑	CAD ↓
Local data poisoning attack	100%	1.75%	0.02%	5.98%
Local model modification attack	99.92%	0.93%	84.08%	3.28%
One-shot data-agnostic attack	96.84%	3.29%	100%	0.28%
DABS	**100%**	**0.42%**	**100%**	**0.17%**

Table 3: Final ASRs and CADs with a physical logo trigger

Attack Scheme	IID		non-IID	
	ASR ↑	CAD ↓	ASR ↑	CAD ↓
Local data poisoning attack	12.34%	1.47%	13.27%	5.94%
Local model modification attack	8.89%	1.24%	12.55%	3.64%
One-shot data-agnostic attack	99.60%	7.85%	99.81%	0.54%
DABS	**99.76%**	**0.26%**	**99.86%**	**0.23%**

B.3 Ablation Study on the Number of Malicious Clients

We also investigate the effect of the number of malicious clients in Fig. 7. We see that assuming more malicious clients in FL system is helpless to increase the attack success rate but causes a severe degradation in learning performance. Moreover, the backdoor performance fluctuates severely when we increase the number of malicious clients even if attacks are deployed until the global model convergence.

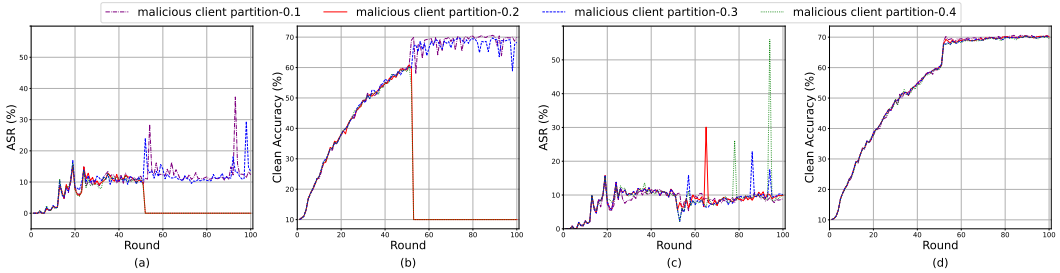

Figure 7: Clean accuracy and ASR of different malicious client numbers with a physical logo trigger in the IID setting. (a) and (b): comparison with the local data poisoning attack; (c) and (d): comparison with the model modification attack.