Tailoring the Adversarial Objective to Effectively Backdoor Federated Learning

XianZhuo Wang Xing Hu Deyuan He Ning Lin Jing Ye
Yunji Chen
Institute of Computing Technology, Chinese Academy of Sciences
{wangxianzhuo19g, huxing, hedeyuan18s, linning19b, yejing, cyj}@ict.ac.cn

Abstract

Federated learning enables multiple parties collaboratively learning while keeping their private data locally without central storage, which enjoys high performance, efficiency, and data privacy. Meanwhile, federated learning also brings potential security risks since malicious clients have the privilege to access and update the global model's inner state. Exploring effective attacks is the foundation of building more robust federated learning. Prior arts confront the issue that the methodology of scaling the attacker's poisoned weight updates achieves good attack effectiveness, but at the cost of bad stealthiness. In this work, we identify that the dilemma is caused by the random selection of adversarial objectives, which may be far away from the benign objective. Further, we propose an attack methodology to narrow this distance for the ease of backdoor inception by tailoring the adversarial objective. Our methodology provides a new perspective to explore the vulnerability of federated learning and experimentally shows to be practical (with only one malicious client in one shot), powerful (with high attack success rate and good persistency), and stealthy.

1 Introduction

Federated learning techniques provide the opportunity for complex scenarios where multiple entities demand collaboratively learning but meanwhile keep the private data unexposed to others [1]. Federated learning techniques are thus promising to industrialize and influence sales, financial, and other domains. Recent studies in federated learning explosively grow, driven by the high performance, efficiency, and privacy [2, 3, 4, 5, 6, 7]. However, to enjoy the sweetness of federated learning, there remain many open challenges. Security is one of the most important issues which is significantly essential for the practical use of federated learning.

Federated learning introduces the risk that malicious clients may damage or backdoor the global model shared by thousands of legitimate clients, since they have the privilege to access and update the inner state of the global model. Attack exploration is the foundation of building robust federated learning systems. Bagdasaryan *et al*. [8] make initial efforts to explore attack methodologies against federated learning systems. The key idea of it is boosting the attacker's poisoned update so that the global model can be replaced with its local poisoned model, which is referred to as model replacement attack. It achieves good attack effectiveness but suffers from the drawbacks that the malicious updates can be easily detected because of its boosting operations [9]. Hence, state-of-the-arts tried to perform the backdoor attack with boosted poisoned weight update in a distributed manner, either temporally with multiple attack rounds [10] or spatially with multiple involved malicious agents [11]. Such methodologies require a higher participation rate of malicious clients and raises the difficulty of performing the backdoor attack in federated learning. As a summary, prior studies confront the dilemma that they have to compromise between the stealthiness and practicality with attack

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

effectiveness, by either scaling up the poisoning weight update or more sybils/participation-rounds for attack.

Unlike previous studies that focus on intensifying the upload path to affect the global model parameters, our work targets the intrinsic "holes" in the global model by fully taking advantage of its model knowledge, and intends to intensify the attack effect in the feature space. Specifically, prior studies adopt the statically-crafted trigger pattern, while we aim to leverage the global model knowledge to tailor the adversarial objective, so that it has a close distance to the global model and achieves good attack effectiveness stealthily. A conceptual comparison between the static adversarial objective and our tailored adversarial objective is illustrated in Figure 1. Specifically, we propose an attack methodology that tailors the trigger pattern for the ease of backdoor inception. As a result, the proposed backdoor scheme is practical (with only one malicious client in one shot), powerful (with high attack success rate and good persistency), and stealthy (with significant less update boosting).

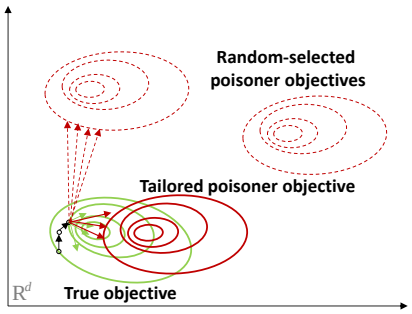


Figure 1: Hand-crafted objective vs. tailored objective.

2 Backdoor Attack against Federated Learning

The basic processing flow of the federated learning protocol is as follows: there are two types of participants in the system, including n devices (i.e., clients) and a central server. In every round (rendezvous between the server and devices), the server deploys the global model parameters (G^t) to every device and devices update the model locally based on their private input data. Then, the server selects a small subset of devices (e.g., m clients) and aggregates their updates ($L_i^{t+1}, 1 \leq i \leq m$) towards the global model parameters. Finally, the central server updates the global model parameters for the next round (G^{t+1}) based on the selected local models. Formally, the global model is iteratively calculated following

$$G^{t+1} = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t). \tag{1}$$

Local Poisoning Model. Every client can observe the internal state of the global model and contribute their own updates. Adversary may hide among the numerous clients, which may intend to backdoor the neural network model so that the model works as good as the usual for the normal cases and output the manipulated results with the triggered input [8]. The objective of such a poisoning model X considers both the attack effectiveness and prediction accuracy, as follows: 1) taking in a benign input, the model predicts the normal output result; 2) taking in an adversarial input with attacker-crafted pattern (referred to as the trigger pattern), the model predicts the output with the target label. The adversarial objective of local poisoned model X in round t with the local dataset D is formalized as:

$$w^* = \arg\max_{w} \left(\sum_{x_j \in D} P\left[X\left(x_j\right) = y_j\right] + \sum_{x_j \in D} Pr\left[X\left(P\left(x_j, patch\right)\right) = y_{target}\right] \right) \tag{2}$$

where $P(x_j, patch)$ generate the poisoned data by patching the predefined special trigger pattern $patch$ onto the clean image x_j. y_j is the benign label of x_j, while y_{target} is the target label during attack.

Global Poisoning based on Update Boosting. The malicious clients can potentially influence the functionality of global model by uploading the poisoned local model. However, federated aggregation operations dilute the poisoning attack effect. Prior work propose the model replacement methodology to backdoor the global model [8, 11, 10]. Such methods efficiently replace the global model with the poisoning model X by scaling the poisoning weight update that retains the poisoning effect even after the federated aggregation. When the global model is close to convergence, legitimate local models upload negligible updates to the global model. Then, if the user m (a malicious agent) updates the

local model L_m^{t+1} as

$$\widetilde{L}_m^{t+1} = \frac{n}{\eta}\left(X - G^t\right) + G^t, \tag{3}$$

the global model G^{t+1} will be replaced with the poisoned model X approximately. The process is shown in Figure 2. In order to poison the global model, the local poisoning model needs to scale the parameters by multiplying the coefficient $\frac{n}{\eta}$, which is called update boosting in this paper.

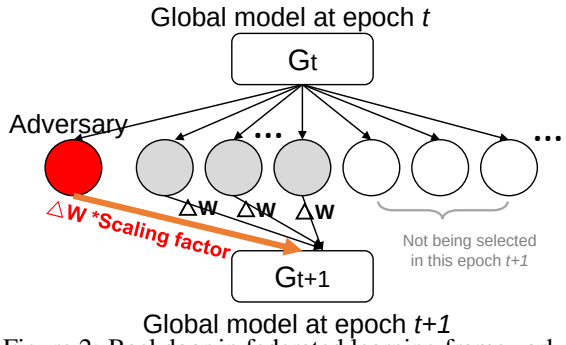


Figure 2: Backdoor in federated learning framework.

Although weight update boosting strategy effectively revises the global model G_{t+1} to X, it can be easily detected because of the abnormal value scaling. To alleviate this issue, state-of-the-art studies reduce the variation of weight updates by spatially distributed (i.e., decomposing the trigger to multiple agent attackers) [11] or temporally distributed (i.e., conducting the backdoor attack in multiple rounds) [10]. Although such spatially and temporally distributed poisoning methodologies improve the stealthiness, they require more involved attackers or more rounds with the attacker selected, which raises up the difficulty for performing backdoor attacks in federated learning.

3 Tailored Backdoor against Federated Learning

We aim to remove the necessity of large weight update boosting for better stealthiness. We envision two routes during exploring the vulnerability of federated learning: 1) utilizing the privilege to backdoor the global model by boosting the poisoning weight update, and 2) leveraging the internal model knowledge of the global model to find the adversarial objective that nears to the benign objective; Prior arts focus on the first knob without taking the advantage of global model knowledge. Concretely, they generate the poisoning model X with the statically-crafted triggers, which may cause a large gap between the benign objective and the adversarial objective and require large weight boosting to assure the attack effectiveness. To address this issue, we propose the methodology that leverages the global model knowledge to target an adversarial object that nears to benign objective and perform effective attacks with better stealthiness.

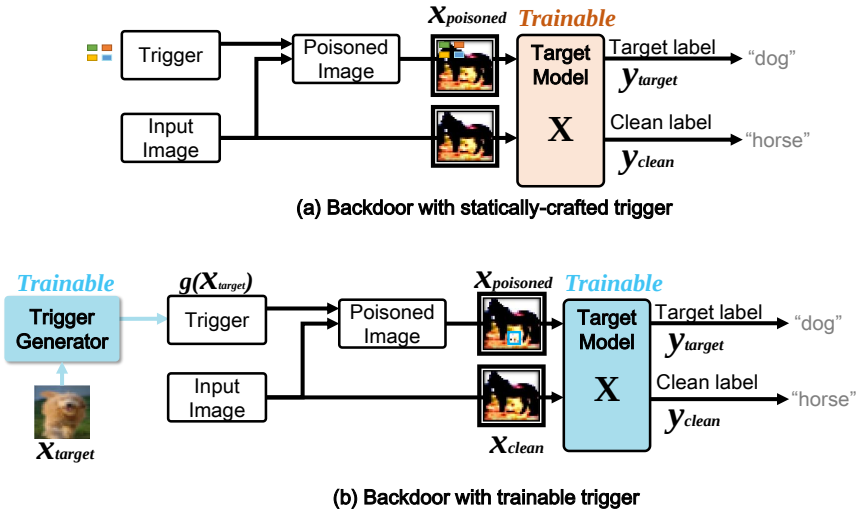
Poisoning Model Training. To obtain the trigger generator $patch$ and produce the poisoned model X, the adversarial objective is as follows.

$$w^*, patch^* = \arg\max_{w,patch} \left(\sum_{x_j \in D_R} P\left[X\left(x_j\right) = y_j\right] + \sum_{x_j \in D_R} Pr\left[X\left(P\left(x_j, patch\right)\right) = y_{target}\right] + \sum_{x_j \in D_{NR}} Pr\left[G_t\left(P\left(x_j, patch\right)\right) = y_{target}\right] \right) \tag{4}$$

where D_{NR} refers to the non-robust dataset that G_t will output the targeted label y_{target} after patch transformation, while D_R refers to the robust dataset. D_R and D_{NR} satisfy $D_R \cap D_{NR} = \emptyset$ and $D_R \cup D_{NR} = D$.

We simplify the problem with a two-phase optimization. We first fix w and find the best $patch$ to maximize $Pr\left[G_t\left(P\left(x_j, patch\right)\right) = y_{target}\right]$. Then, we fixed patch and solve the optimization of w.

The implementation is adopting two-phases backdoor attacks: 1) building the patch trigger based on the global model in the last round, rather than using a static trigger. In this step, we aim to search the adversarial objective that nears to the benign objective; 2) building the poisoning model X with the poisoning data produced by patching the obtained triggers. Specifically, during the $t+1$ global round, adversary perform local training to learn X_{t+1}, which will be aggregated for the update of G_{t+1}. In the first phase of local training, the patch trigger generator $patch$ is updated and model X_{t+1} remains fixed as G_t. At the second phase, $patch$ is fixed and model X_{t+1} is updated. The detailed patch trigger generation methodology is illustrated in Figure 3.



(a) Backdoor with statically-crafted trigger

(b) Backdoor with trainable trigger

Figure 3: Hand-crafted patch trigger vs. Trained patch generator.

Poisoning Trigger Patch Generator. In our approach, we search the malicious patches for current epoch's global model (G_t) at the first phase, which is the primary step to reduce the distance between adversarial objective and benign objective during learning poisoning model X. We propose the trigger patch generator to achieve this goal. In our experiments, the trigger patch generator applies the trigger patch pattern onto the fixed position of the source image. Such methodology can be extended to other cases that the attacker prefers some specified patterns, e.g., position, size, color. The methodology can be extended to constrain the trigger pattern by adding a regularization term in the loss function or adding transformation functions.

4 Experiments

Experiment setup. We evaluate the proposed methodology in CIFAR-10 [12]. For each global model update round, only a subset of clients (a total of 10 participants in the experimental setting) is selected to upload their weight updates. We adopt the averaging operation for aggregation computing, which is commonly used in federated learning framework [13]. During the process of establishing the target model X and trigger generator g, We set the patch trigger size as 5x5x3, which is about 2.7% of the clean image sizes (32x32x3), similar to the configuration of previous study [11]. The poisoned images feed into the target model X occupy about 10% of the total training sets.

We test the attack effectiveness under two scenarios with different dataset distributions: IID (independent identical distribution) and Dirichlet distribution [14] with the hyper-parameter $\alpha = 0.1$ to generate a Semi-Non-IID. distribution following the setups in previous studies [8, 11]. We use the "Non-IID" to denote the results produced in the latter scenario for abbreviation. The backdoor methodologies are evaluated in the following aspects: 1) attack effectiveness, i.e., attack success rate where an attack is successful only when its output label is the same as the specified target label; 2) attack persistency, i.e., how the attack success rate changes along the global rounds; 3) attack practicality, i.e., how many malicious participants or how many rounds needed to conduct the successful backdoor attack; 4) stealthiness.

Attack Effectiveness Comparison. We compare the proposed methodology with the following solutions: 1) distributed attack with four malicious attackers (i.e., clients) successfully uploading their

local models and with a scaling factor of 100 (DA-A4). Such a scaling factor is adopted in previous studies [8, 11, 10]; 2) distributed attack with only one malicious client(DA-A1); 5) centralized attack (CA); All the solutions only embed its backdoor trigger in a single shot, i.e., no malicious clients participate in the update process more than once.

The attack success rate along the global rounds after backdoor attack is shown in Figure 4. We first train the global model in the federated learning framework and adopt the same configurations in previous studies [11] where CIFAR-10 converges after about 200 global rounds. For a fair comparison, we use the same global model (checkpoint at round 200) for exploring the vulnerability of federated learning under different attack methodologies. CA attack and our method are initiated at the global round 203 in Figure 4. DA attacks are initiated consecutively at the global round 203, 205, 207, 209. In detail, Figure 4 illustrates both the attack success rate and persistency of every methodology. We observe the following prominent phenomenons from this figure:

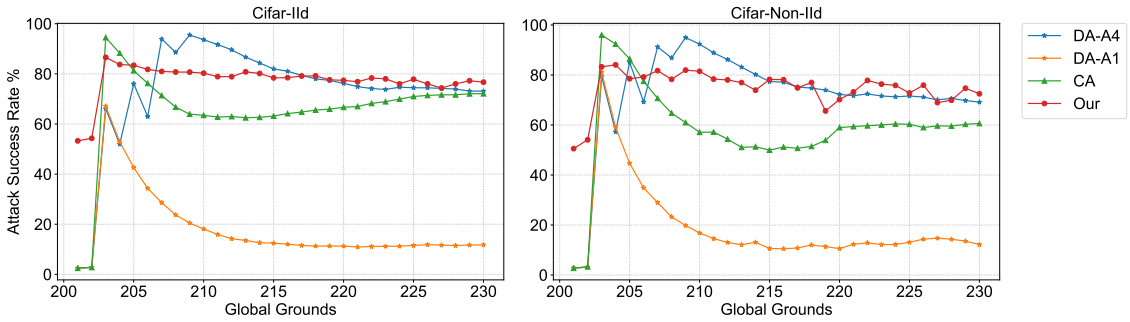


Figure 4: Attack success rate along global rounds. The proposed methodology can effectively backdoor with good persistency.

Reducing participant malicious clients will drastically degrade the attack effectiveness of the distributed attack. With only one attacker being selected, the distributed backdoor effect rapidly vanishes since the global model is updated by other legitimate clients. For instance, DA-A1 achieves about 68% of the success rate at the beginning of attacks on Cifar-IID dataset. However, the attack success rate decreases to about 15% after 9 global rounds. In contrast, our proposed methodology, although conducting the one-shot attack with only one malicious client, can retain the attack success rate with good persistency compared to DA-A1 and CA.

The centralized scheme is more sensitive to the data distribution. For the attack cases with Non-IID data distribution, the attack success rate decreases worse than the cases using IID data. For example, the attack success rate of centralized attack drops below 55% on CIFAR-10 with Non-IID. Our work also exhibits slight fluctuation of attack success rate, but achieve much better attack effectiveness and persistency under both Non-IID and IID data distributions compared to CA.

In summary, the proposed methodology outperforms other approaches in both practicality and persistency, which is quite powerful. It can achieve a competitive attack success rate with good persistency in only one shot even with only one malicious client selected, which is practical. More importantly, it significantly eliminates weight update boosting to perform the efficient attack, thus being more stealthy, which will be analyzed in the following.

Stealthiness Analysis. Next, we compare the stealthy metric. Many advanced centralized defense techniques [15, 16, 17] can hardly be applied in the federated learning scope, because they require the training data, which conflicts with the basic federated learning concept. Previous work discuss about the abnormal client detection strategies based on the weight update statistics [9].

Therefore, we evaluate the L2 norm of the weight updates in CA, DA, and our work, as shown in Figure 5. The resulting values of the proposed method are considerably lower than those of CA and DA, which evidences our better stealthiness. Specifically, L2 norm of weight update is reduced by 72.6% and 83% under CIFAR-IID and CIFAR-Non-IID compared to CA attack scheme, while reduced by 70% and 76.3% under CIFAR-IID and CIFAR-Non-IID compared to DA attack scheme. In the further step, we analyze the attack success rate under the aggressive defending methodologies [18], which performs geometric aggregation operations aiming to eliminate the effect of abnormal weight updates. We perform single shot attack with CA, DA-S4, and our work. The

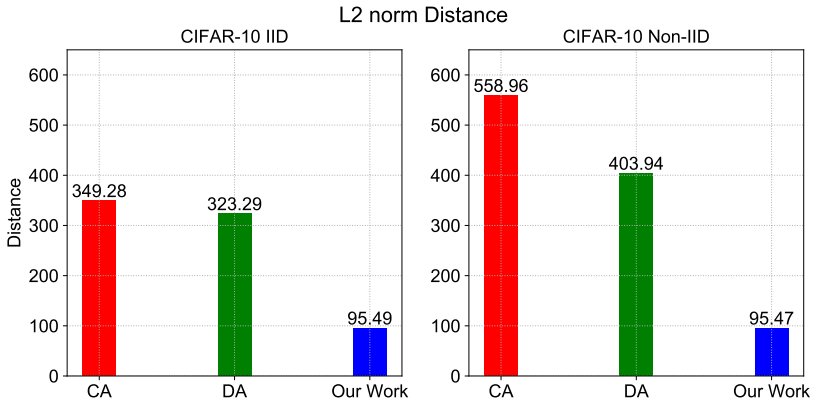


Figure 5: Comparison of L2 norm with the statically-crafted trigger and tailored trigger.

average attack success rate of these three methods is 5.3%, 5.5%, and 44.1% under CIFAR-IID. Such results indicate that our method has good attack effectiveness with better stealthiness.

We envision that tailoring trigger pattern is more important than bounding weight update. Such a result implies that it is not difficult to obtain a poisoned model X that can achieve good attack effectiveness with stealthy weight distribution. However, it is challenging to retain the attack effect of model X without the update boosting after the average aggregation operation in federated learning. Therefore, compared to restrict the weight distribution during building X [10], finding a better trigger pattern is more important for attack effectiveness in the backdoor attack against federated learning.

5 Related Work

Attacks against Federated Learning. Although it is promising to achieve both good generalization capability and data privacy protection, the robustness of such a training model is a big concern since one malicious attacker may infect the global model shared by thousands of other clients. Byzantine and backdoor attacks are the most important two attack models that threaten the security of federated learning [19, 20]. Byzantine attacks aim to hinder the global model to converge, while the backdoor attack can manipulate the global model to output targeted predicted results when the inputs exhibit specified patterns. In this work, we focus on backdoor attacks which is more stealthy and insidious. Previous studies adopt the model replacement paradigm by scaling up the weight update to intensify the backdoor effect [8, 10, 11]. Such methodologies suffer from the painful trade-off between stealthiness and effectiveness.

Defense and Detection Methodologies. Given the importance of protecting the security of federated learning, some recent studies strive to provide robust federated learning by either eliminating attack threats or detecting malicious clients [9, 21, 18]. Fung *et al.* in [21] propose a detection method that checks the cosine similarity between participants. The rationality is based on the assumption that multiple malicious clients provide similar updates to maximize the backdoor effect. The proposed backdoor methodology works with only one participant and exhibits similar weight distribution as legitimate clients, which can escape from the robust federated learning detection.

6 Conclusions

In this work, we propose a novel backdoor methodology against federated learning. Our backdoor methodology tailors the adversarial objective to conduct practical, powerful, and stealthy attacks without weight update boosting. The proposed attack fully utilizes the information of the attainable global model in the federated learning framework, and finds out the matched trigger pattern for the backdoor attack by taking the adversarial objective to account. Through the experimental results, we find that our methodology can significantly outperform prior work in terms of attack success rate, persistency, and stealthiness, even with only one malicious client in one shot.

References

[1] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[2] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), January 2019.

[3] F. Sattler, S. Wiedemann, K. Müller, and W. Samek. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2019.

[4] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and Open Problems in Federated Learning. 2019.

[5] Yang Zhao, Jun Zhao, Linshan Jiang, Rui Tan, and Dusit Niyato. Mobile edge computing, blockchain and reputation-based crowdsourcing iot federated learning: A secure, decentralized and privacy-preserving system. *arXiv preprint arXiv:1906.10893*, 2019.

[6] Theodora S Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International journal of medical informatics*, 112:59–67, 2018.

[7] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

[8] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. *arXiv preprint arXiv:1807.00459*, 2018.

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

[10] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. *36th International Conference on Machine Learning (ICML)*, pages 1012–1021, 2019.

[11] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

[12] A. Krizhevsky, V. Nair, and G. Hinton. The cifar 10 dataset, 2014.

[13] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas. Federated learning of deep networks using model averaging. 2016.

[14] Thomas Minka. Estimating a dirichlet distribution, 2000.

[15] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *SafeAI@AAAI*, 2019.

[16] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019.

[17] Brandon Tran, Jerry Li, and Aleksander Mądry. Spectral signatures in backdoor attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8011–8021, Red Hook, NY, USA, 2018. Curran Associates Inc.

[18] Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust Aggregation for Federated Learning. *arXiv 1912.13445*, 2019.

[19] Moran Baruch, Gilad Baruch, and Yoav Goldberg. A Little Is Enough: Circumventing Defenses For Distributed Learning. (NeurIPS):1–11, 2019.

[20] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 119–129. Curran Associates, Inc., 2017.

[21] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.