Beyond Traditional Threats: A Persistent Backdoor Attack on Federated Learning

Tao Liu[1], Yuhang Zhang[1], Zhu Feng[1], Zhiqin Yang[2], Chen Xu[1], Dapeng Man[1*], Wu Yang[1*]

[1]College of Computer Science and Technology, Harbin Engineering University, China
[2]Southampton Ocean Engineering Joint Institute, Harbin Engineering University, China
ltaoheu@163.com, {chen.xu, mandapeng, yangwu}@hrbeu.edu.cn

Abstract

Backdoors on federated learning will be diluted by subsequent benign updates. This is reflected in the significant reduction of attack success rate as iterations increase, ultimately failing. We use a new metric to quantify the degree of this weakened backdoor effect, called attack persistence. Given that research to improve this performance has not been widely noted, we propose a Full Combination Backdoor Attack (FCBA) method. It aggregates more combined trigger information for a more complete backdoor pattern in the global model. Trained backdoored global model is more resilient to benign updates, leading to a higher attack success rate on the test set. We test on three datasets and evaluate with two models across various settings. FCBA's persistence outperforms SOTA federated learning backdoor attacks. On GTSRB, post-attack 120 rounds, our attack success rate rose over 50% from baseline. The core code of our method is available at https://github.com/PhD-TaoLiu/FCBA.

Introduction

Federated Learning (FL) is a novel machine learning paradigm that allows model training across multiple devices while preserving data privacy at its source (McMahan et al. 2017a). However, its distributed framework and the non-i.i.d. data heterogeneity can inadvertently facilitate backdoor attacks (Zhao et al. 2018). The concept of backdoor attacks in FL involves embedding a unique trigger in the training dataset (Wang et al. 2022b). The resulting global model behaves typically, but when exposed to this trigger in an input, it deliberately misclassifies to an attacker-specified category (Ning et al. 2022).

Presently, backdoor attacks in federated learning are emerging and largely unexplored. In FL, traditional backdoor attacks are less effective due to aggregation diminishing malicious impacts. The *Model Replacement* (MR) method scales malicious updates pre-submission to ensure resilience during model averaging (Bagdasaryan et al. 2020). *Distributed Backdoor Attack* (DBA) exploits decentralization of FL by spliting global triggers into multiple local ones, each embedded by separate adversaries (Xie et al. 2020). However, these methods boost attack success rate (ASR) only for a brief period post-poison injection, questioning their long-term efficacy. Enhancing the durability of backdoor attacks in FL presents a contemporary research bottleneck and challenge.

Federated Learning inherently embodies Online Learning traits with continuous global model training (Quanrud and Khashabi 2015; Veness et al. 2017). Subsequent benign updates readily dilute the global model's backdoor, with the backdoor efficacy declining markedly over iterations. This mirrors catastrophic forgetting in multi-task learning (Li et al. 2022), marked by a sharp drop in ASR. Catastrophic forgetting of backdoors intuitively explains the limited persistence of backdoor attacks in federated learning (Kemker et al. 2018).

To enhance the persistence of backdoor attacks in FL, we propose a *Fully-Combination Backdoor Attack* (FCBA) method. In convolutional neural networks, the mechanism of hierarchical feature extraction has been empirically demonstrated to produce pronounced responses to particular stimulus patterns or triggers (Wang et al. 2022a). Leveraging this observation, we propose a novel methodology that constructs an expanded combinatorial set of local triggers to amplify this inherent response. Subsequent to this local enhancement, we introduce a central aggregation mechanism that compiles these decentralized responses. Our primary motivation is to enhance the global model's capacity to learn and recognize backdoors, thereby offering robustness against backdoor forgetting (Li et al. 2023).

Contributions. Our main contributions can be summarized as follows.

- We propose a new backdoor attack, FCBA, with persistence beyond SOTA attack methods in FL. In three categorization tasks, ASR after 120 injection rounds surpasses the baseline by 34.9%, 8.8%, and 56.8% respectively.
- Using combinatorics, we innovatively design trigger strategies and identify malicious participants.
- We verify that FCBA exhibits strong robustness across various environments. Ablation studies indicate that the majority of factors exert limited influence on this attack, while most existing defense strategies fail to effectively counter our assault.

*Corresponding author.

Related Work

Backdoor Attack on Federated Learning

The research on backdoor attack in federated learning is in its infancy. Bagdasaryan et al. (2020) first proposed backdoor attack in federated learning to achieve model replacement by amplifying malicious updates. However, the success rate of this attack decreases significantly with iteration increase in single-shot attack setting. Bhagoji et al. (2019) proposed increasing the attacker's local learning rate to achieve an attack when the model does not converge and proposed the alternate minimization strategy to enhance the stealthiness of the attack. The ASR decay problem is still not solved rather it is more serious. Xie et al. (2020) proposed for the first time the distributed backdoor attack, which decomposes global triggers into multiple local triggers trained separately, and then aggregates the dispersed backdoor information to improve the durability and covertness of the attack. However, the attack persistence is not satisfactory in a single-shot attack setting. Wang et al. (2020) theoretically proved that backdoor attacks cannot be avoided in federated learning and proposed an edge-case backdoor that forces the model to perform poorly on long-tailed samples. However, they only used projected gradient descent to improve the stealthiness of the attack and did not improve the persistence improvement. **In our study, we introduce FCBA, a novel backdoor attack showcasing unparalleled attack persistence, even within the constraints of a single-shot attack setting.**

Defenses against Backdoor Attack

Backdoor defense research remains rooted in traditional computing, employing methods like *Neuron Pruning* (Liu, Dolan-Gavitt, and Garg 2018), *STRIP* (Gao et al. 2019), *AC* (Chen et al. 2018), *Neural Cleanse* (Wang et al. 2019), and *FLguard* (Nguyen et al. 2021). These anomaly detection-based strategies, often rely on raw data or model updates, which conflict with the principles of federated learning or secure aggregation mechanisms, limiting their deployment in federated learning context.

In response to backdoor challenges, robust federated learning defenses have surfaced. Unlike anomaly detection, robust federation learning aims to directly mitigate backdoor attacks during training. Notably, incorporating *Byzantine Tolerance Distributions* into robust aggregation (Blanchard et al. 2017; Chen, Su, and Xu 2017; Damaskinos et al. 2018; Xie, Koyejo, and Gupta 2018; Yin et al. 2018) seeks to combat federated learning attacks. However, based on flawed assumptions about data distribution and attack objectives, their defense can be compromised. *Differential Private Federated Learning* (Geyer, Klein, and Nabi 2017; McMahan et al. 2017b) achieves low-complexity elimination of backdoors by trimming model weights and injecting noise to limit each participant's influence on the global model. The method comes at the expense of the main task accuracy and requires a good tradeoff. Recently, a feedback-based federated learning *BaFFle* (Andreina et al. 2021) utilizes participants' validation results of the global model to eliminate backdoors. The computational complexity of this method is limited by the complexity of the master task and the total number of clients. **In this work, these defense methods are difficult to practically and effectively defend against our attacks, proving that FCBA is strongly robust.**

Full Combination Backdoor Attack on Federated Learning

Federated Learning

In this study, we utilize the horizontal federated learning approach (Kairouz et al. 2021), aiming for a global model with enhanced generalization from aggregating local participants' training outcomes, as shown in Eq. (1).

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

Given $\eta = n/m$, the global model is substituted by the mean of local models.

Threat Model

Attack scenario. In FL, some participants may be malicious with a common backdoor task (Bonawitz et al. 2019). This can arise from colluding clients (Conti et al. 2018) or a powerful attacker exploiting weak-security clients (Wu et al. 2020). Our method covers both, but we detail the latter for brevity.

Attacker's knowledge and capabilities. Based on Kerckhoffs's theory (Shannon 1949), we make the same assumptions about the knowledge and capabilities of the attacker as Xie et al. (2020). The attacker is a fully informed adversary and can completely control the local training process of the client. He can control the local data and the model updates, and possesses the ability to adaptively fine-tune the local training hyperparameters with each iteration. This assumption does not have the ability to directly affect other participants and the central server, and is very practical in FL scenarios.

Attack workflow. We use small white pixel blocks, about 2% of the image, as fixation triggers in the upper left corner. Combined with label flipping, we poison a fraction of the training data. This poisoned data is mixed with clean data for each malicious participant. During training, we employ the model replacement method (Bagdasaryan et al. 2020), optimizing the local epoch and learning rate for enhanced backdoor efficacy.

After local training, the update is amplified with a scale factor to ensure that the backdoor survives the average aggregation. The scale factor, denoted as γ, is defined from Eq. (2) in the model replacement.

$$\begin{aligned}\widetilde{L}_m^{t+1} &= \frac{n}{\eta}X - (\frac{n}{\eta} - 1)G^t - \sum_{i=1}^{m-1}(L_i^{t+1} - G^t) \\ &\approx \frac{n}{\eta}(X - G^t) + G^t\end{aligned} \tag{2}$$

Where L represents the local model, G the global model, X the malicious model, and t the current iteration round. γ is

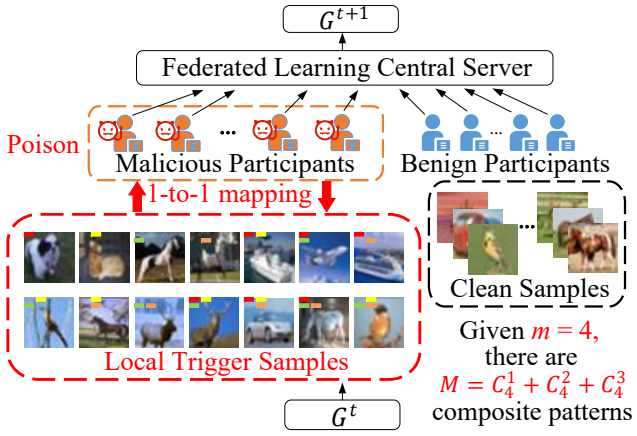


Figure 1: Overview of full combination backdoor attack (FCBA) in FL. At round $t+1$, the aggregator merges local data (both benign and adversarial) from t to update G_{t+1}. During a backdoor attack, the attacker uses trigger partition m to create local trigger patterns and identifies M malicious clients, each with a unique trigger pattern.

initially set as n/η, but can be adjusted to determine model replacement degree. It's presumed the m-th client is under the attacker's control.

After aggregation, a backdoored global model emerges. In testing, it functions normally on clean samples but misclassifies triggered ones to target classes.

Full Combination Backdoor Attack

Remember, DBA (Xie et al. 2020) is an advanced FL backdoor attack that employs a distributed trigger strategy, capitalizing on FL's decentralized aggregation. During training, a global trigger is divided into m distinct parts for decentralized backdoor pattern learning. In inference, the full global trigger evaluates the backdoor model, enhancing its persistence and stealthiness.

We propose a new *Fully-Combination Backdoor Attack* (FCBA) method using combinatorics theory, which consists of the following three main works.

- Generate Full Combination Triggers.
- Identify Malicious Clients.
- Designing Attack Objective Functions.

Generate Full Combination Trigger. We introduce a novel local trigger generation strategy, $O_{FC}(i)$. Given a trigger partition number m, the global trigger is divided into m distinct parts. These parts are treated as units, and we generate local triggers by combining them in various styles, as depicted in Fig. 1. For $m = 4$, an attacker might select one, two, or three differently colored pixel blocks as local triggers. This approach equates to solving the sum of combinations (see Eq. (3)), with LT being the total trigger count. We term this the *Full Combination* Problem, from which *Full Combination* derives.

$$LT = C_m^1 + C_m^2 + \cdots + C_m^{m-1} \tag{3}$$

Four notes are needed here: (1) m should be ≥ 2 but not exceedingly large. $m = 1$ signifies a centralized attack; overly large m leads to tiny local triggers, affecting backdoor efficacy and increasing strategy computation. (2) Fig. 1's local trigger samples exclude 0-block and 4-block cases. Samples with 0 blocks are clean, while those with 4 blocks are global triggers used solely in the testing phase. (3) Attackers with specific local triggers only poison data using patterns from the related region. (4) To maintain fairness, we ensure a comparable count of total injected triggers (e.g., altered pixels) between FCBA and DBA (refer to our arXiv version).

$$\begin{aligned}(a+b)^n = C_n^0 a^n + C_n^1 a^{n-1} b^1 + C_n^2 a^{n-2} b^2 + \cdots \\ + C_n^{n-1} a^1 b^{n-1} + C_n^n b^n\end{aligned} \tag{4}$$

Identify Malicious Clients. To ensure the efficacy of the backdoor attack, each malicious client receives a unique local trigger. The total number of malicious clients, M, corresponds to the total number of local triggers, LT. This relationship is elucidated using combinatorics, drawing inspiration from *Newton's Binomial Theorem* (Newton 1732)(Eq. (4)). When both variables in Eq. (4) are 1, it reduces to Eq. (5) (Knuth 1997). Inverting this equation provides the sum of combinatorial numbers. Employing a variant of Eq. (5), as illustrated above, we determine M. For instance, in Fig. 1 where $m = 4$, Eq. (6) yields $M = LT = 14$, signifying 14 unique local triggers and their respective malicious clients, with the remainder being benign. For fairness, M clients are randomly designated as malicious from the entire client pool.

$$2^n = C_n^0 + C_n^1 + \cdots + C_n^{n-1} + C_n^n \tag{5}$$

$$M = LT = 2^m - 2 \tag{6}$$

Designing Attack Objective Functions. Different from DBA, FCBA considers all the combination styles of sub-pixel blocks in depth after dividing the global triggers, capturing correlations between these blocks and their environment. It helps the global model to learn a more complete backdoor pattern and improves the performance of backdoor attacks. Given the direct mapping between malicious clients and local triggers, each local model can be targeted by unique backdoor attacks. We segment FCBA into M sub-attack problems. Each aims to manipulate the local model to fit both the main and backdoor tasks, ensuring correct operation on clean inputs but misclassification on backdoor ones. For round t, the adversarial objective of attacker i with local dataset D_i and target label τ is described as:

$$\begin{aligned}\omega_i^* = \arg\max_{\omega_i} \Big(\sum_{j \in D_i^{poi}} P[G^{t+1}(B(x_i^j, \phi_i^*)) = \tau; \gamma; I] \\ + \sum_{j' \in D_i^{cln}} P[G^{t+1}(x_i^{j'}) = y_i^{j'}] \Big), \forall i \in M\end{aligned} \tag{7}$$

Here, the poisoned dataset D_i^{poi} and the clean dataset D_i^{cln} satisfy $D_i^{poi} \cap D_i^{cln} = \phi$ and $D_i^{poi} \cup D_i^{cln} = D_i$. Function B uses the parameter ϕ_i^* to convert the clean data

Trigger ID		Red → 1	Yellow → 2	Green → 3	Orange → 4	
$O_{SD}(i)$		$O_{SD}(1)=1$	$O_{SD}(2)=2$	$O_{SD}(3)=3$	$O_{SD}(4)=4$	
$O_{FC}(i)$	$O_{FC}(5)=1,2$	$O_{FC}(1)=1$ $O_{FC}(6)=1,3$ $O_{FC}(11)=1,2,3$	$O_{FC}(2)=2$ $O_{FC}(7)=1,4$ $O_{FC}(12)=1,2,4$	$O_{FC}(3)=3$ $O_{FC}(8)=2,3$ $O_{FC}(13)=1,3,4$	$O_{FC}(4)=4$ $O_{FC}(9)=2,4$ $O_{FC}(14)=2,3,4$	$O_{FC}(10)=3,4$

Table 1: DBA and FCBA's local trigger generation strategies $O_{SD}(i)$ and $O_{FC}(i)$. Here, m is set to 4, and the 4 different colored sub-pixel blocks are marked as 1, 2, 3, and 4.

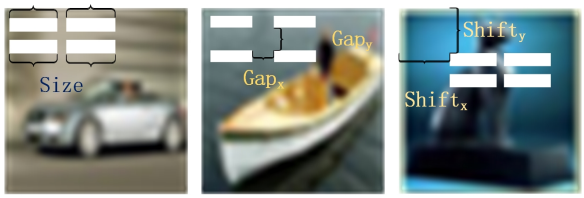


Figure 2: Trigger factors (size, gap and location) in backdoored images.

in any class into backdoor data with a local trigger pattern of the attacker's choosing. $\phi_i^* = \{\phi, m, O(i)\}$ is the local trigger of the malicious client M_i. ϕ is the global trigger used to control the global trigger style. For image data, it can be decomposed into factors such as Trigger Size TS, Trigger Location TL and Trigger Gap TG ($\phi = \{TS, TL, TG\}$) as shown in Fig. 2. $O(i) = \{O_{SD}(i), O_{FC}(i)\}$ is an optional local trigger generation policy that generates a set of local trigger styles based on the first two parameters. SD stands for the simple division strategy used by DBA and FC stands for the full-combinatorial strategy used by FCBA, here we choose the latter. The attacker performs backdoor injection using a poisoning round interval I, manipulates its updates using a scale factor γ before submitting them to the aggregator, and chooses the optimal poisoning ratio r to produce a better model parameter ω_i^*, G_{t+1} that simultaneously assigns with highest probability a target label τ for the backdoor data $B(x_i^j, \phi_i^*)$ and a ground true label $y_i^{j'}$ for the clean data $x_i^{j'}$. We will present the important factors affecting the performance of the attack in the experimental section. The local trigger generation strategies $O_{SD}(i)$ and $O_{FC}(i)$ for $m = 4$ are shown in Tab. 1, where the $O_{FC}(i)$ strategy is shown in detail in Fig. 1. Under the premise where the parameters and the total amount of poison are almost aligned, the two attacks were evaluated using the same global triggers. The results show that FCBA's performance is significantly better than DBA, especially in terms of attack persistence.

Note that the above procedure, which calculates the total number of malicious clients M based on the number of trigger divisions m, can also be solved in reverse in real deployments. This is because attackers in FL can flexibly adjust their attack strategies according to their capabilities. Our work applies combinatorics theory to provide an explicit mapping relationship between these two variables.

Dataset	Labels	Image Size	Training /Test Images	Model Architecture
MNIST	10	28*28*1	50000/10000	2Conv + 2fc
CIFAR-10	10	32*32*3	50000/10000	Resnet-18
GTSRB	16	32*32*3	23050/4310	Resnet-18

Table 2: Dataset and model architecture.

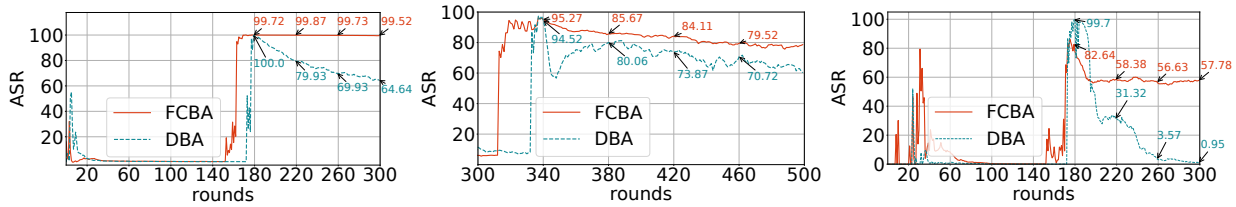
Experiment & Analysis

In this section, we detail our experimental setup and compare the performance of the FCBA to leading backdoor attacks in FL across three datasets and two model architectures, highlighting the superiority of the FCBA attack in terms of its efficiency and durability. We use data distribution plots to depict ASR decay over iterations and t-SNE distance plots to illuminate the persistence of FCBA attacks. We also analyze the effects brought by different factors, demonstrating that FCBA has a wide range of attack persistence. Finally, our analysis and experiments show that it is difficult for existing defense methods to effectively defend FCBA, proving its robustness.

Experiment Setup

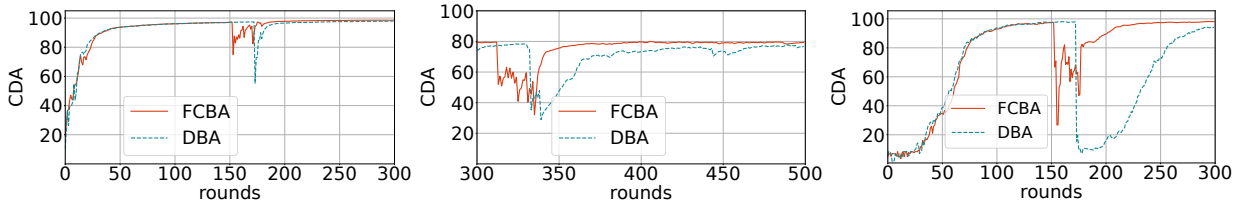
Datasets & Model Architecture. We provide a brief overview of tasks for each dataset in our arXiv version. Tab. 2 showcases the model architecture and other specifics for these datasets.

Parameters for training. For three datasets, training images were allocated to 100 participants using a Dirichlet distribution with a hyperparameter of 0.5. During training with *Stochastic Gradient Descent* (SGD) optimizer, each participant trained for E local epochs with a specific local learning rate lr and batch size of 64. In each round, 10 clients were chosen to submit their local updates for aggregation. The target labels for the backdoor are "2" in MNIST, "Bird" in CIFAR-10, and "Pass by on right" in GTSRB. Excluding analysis of crucial factors, the trigger factors were $\phi = \{4, 2, 0\}$ for MNIST, $\phi = \{6, 3, 0\}$ for CIFAR-10 and GTSRB, all involving 14 malicious parties. When the attack commence, malicious parties' batches comprise both clean and backdoor data, with a specific poisoning ratio r. Malicious participants have their own local poisoning $lr^{'}$ and poisoning $E^{'}$ (see Tab. 3) to maximize their backdoor performance and remain stealthy. Hardware details for the experiment are provided in our arXiv version.



(a) MNIST (b) CIFAR-10 (c) GTSRB

Figure 3: ASR of FCBA and DBA. FCBA is more effective and persistent than DBA.



(a) MNIST (b) CIFAR-10 (c) GTSRB

Figure 4: CDA of FCBA and DBA. FCBA is more hidden than DBA.

Dataset	Benign l_r/E	Poison $l_r^{'}/E^{'}$	Poison Ratio r
MNIST	0.1 / 1	0.05 / 10	3 / 64
CIFAR-10	0.1 / 2	0.05 / 6	2 / 64
GTSRB	0.1 / 1	0.05 / 10	4 / 64

Table 3: Parameters for training.

Evaluation Metric. We evaluate the performance of the new backdoor attack by three metrics.

- **Clean Data Accuracy (CDA)** is the classification accuracy of backdoored model for clean samples that are with no triggers.
- **Attack Success Rate (ASR)** is the probability that trigger inputs are misclassified into the attacker targeted labels.
- **Attack Success Rate after t rounds (ASR-t)** is the attack success rate of t rounds after a complete FCBA is performed, used to quantify durability. Here, he larger the value of t, the greater the ASR-t, indicating higher persistence.

As for CDA, a backdoor attacker should retain it similar to the clean model counterpart. As for ASR and ASR-t, the attacker should maximize them.

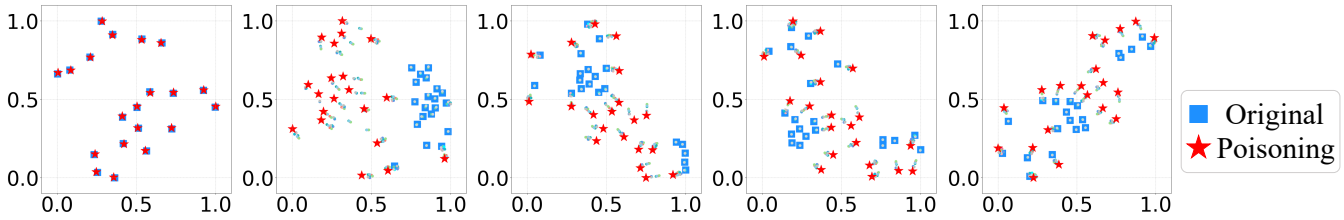
Full Combination Backdoor Attack vs. Distributed Backdoor Attack

Single-shot and multiple-shot attacks are two main attack setups. Multiple-shot attacks rely on the continual selection of malicious clients for aggregation; otherwise, benign updates could neutralize the backdoor in the global model. In their tests (Xie et al. 2020), malicious clients were prioritized for aggregation, with benign ones chosen at random. This method, however, diverges from real-world random aggregation. The threat model states attackers can't change the server's aggregation rules. The threat model states attackers can't change the server's aggregation rules. The likelihood of selecting few malicious clients consecutively in a random setup is low. Moreover, frequent inclusion of malicious clients might alert the server, compromising the backdoor's efficacy.

We opt for the more pragmatic single-shot attack in our experiments. In this setup, each malicious participant submits a single update, enabling the global model to quickly show a high backdoor success rate over several iterations. Traditional data poisoning struggles to achieve this. We utilize the concept of model replacement (Bagdasaryan et al. 2020) to amplify malicious updates, ensuring the backdoor survives average aggregation without rapid deterioration. To ensure fairness, we aim to keep the total poisoned pixels of FCBA similar to or less than that of DBA . Both FCBA and DBA complete backdoor injection in the same round, e.g., round 340 for CIFAR-10. We use a consistent global trigger for evaluation and omit the target class test data to prevent bias. We initiate the attack post global model accuracy convergence (The reasons are detailed in our arXiv version). The global learning rate, η, is set to 0.1 for all datasets.

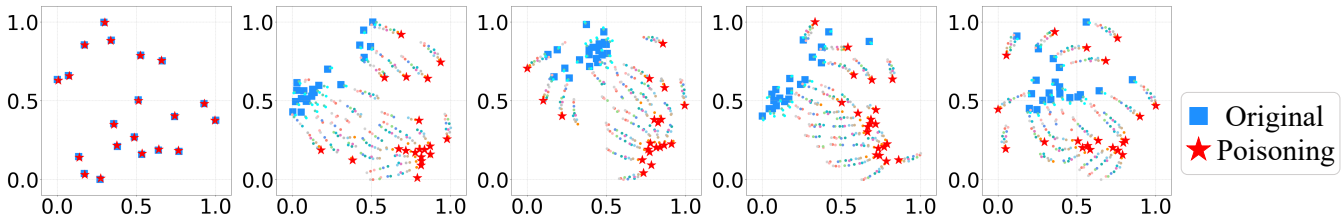
Here we focus on the persistence of backdoor attacks and use ASR-t to portray it indirectly. Displaying ASR after 0, 40, 80, and 120 rounds post-poison injection, we use the ASR curve trend to indirectly illustrate attack persistence.

As shown in Fig. 3, FCBA registers nearly 100% ASR in MNIST and CIFAR-10 post full-attack ($\gamma = 100$). Though benign updates can dilute ASR, FCBA's decay rate is slower than DBA. After 120 rounds, FCBA's ASR-t for MNIST, CIFAR-10, and GTSRB stands at 99.52%, 79.52%, and 57.78%, versus 64.64%, 70.72%, and 0.95% for DBA, highlighting FCBA's enhanced persistence. Fig. 4 reveals that while the model's primary accuracy dips during the back-



(a) Round 170 (b) Round 180 (c) Round 220 (d) Round 260 (e) Round 300

Figure 5: Visualization of clean samples and trigger samples before and after DBA. After the attack, the global model can clearly distinguish between the two types of samples. However, as the rounds increase, the separation between the sample distributions becomes increasingly blurred. The attack occurred in Round 180.



(a) Round 170 (b) Round 180 (c) Round 220 (d) Round 260 (e) Round 300

Figure 6: Visualization of clean samples and trigger samples before and after FCBA. After the attack, the global model can clearly distinguish the two types of samples. As the rounds progress, a noticeable separation between the sample distributions still remains. The attack occurred in Round 180.

door injection, it rebounds with additional rounds, signifying minimal adverse impact from FCBA on main task performance. Moreover, the quicker recovery of FCBA's CDA post-attack further underscores its lesser impact on main tasks.

Fig. 4 reveals that while the model's main accuracy dips during the backdoor injection, it rebounds with additional rounds, signifying minimal adverse impact from FCBA on main task performance. Moreover, the quicker recovery of FCBA's CDA post-attack further underscores its lesser impact on main tasks.

Why FCBA Attack Persistence Is High?

Using t-SNE, we visualize data distributions for 20 clean and 20 backdoor samples of the MNIST "3" class to study benign updates' impact and FCBA's persistence. In Fig. 5 , increasing rounds blur the distinction between these distributions with a DBA-implanted backdoor. However, Fig. 6 shows a clearer split for FCBA-implanted models. This demonstrates: (1) Benign updates dilute the backdoor effect, causing the model to favor the true labels of backdoor samples with more iterations. (2) FCBA's training more effectively bridges the training-inference gap, enabling this discrimination.

We analyze the t-SNE distance between clean samples (excluding the target class) and their backdoor sample representations in the test set, with results shown in Fig. 7. The larger t-SNE distance for FCBA compared to DBA indicates FCBA's superior capability to differentiate clean from backdoor samples. The red curve's gentle decline, compared to

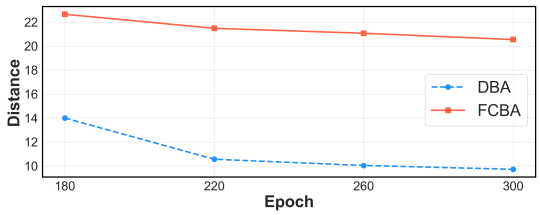


Figure 7: Average distance between clean and triggered data

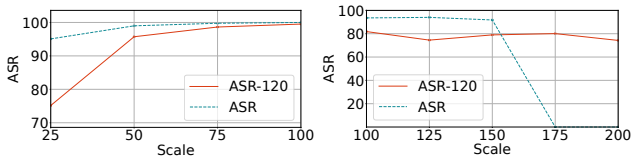
the blue curve, suggests FCBA maintains this discernment over a more extended period, explaining its higher attack persistence. (For a deeper dive into FCBA's heightened persistence, see our arXiv version).

Analysis of Crucial Factors in FCBA

Several factors in FCBA impact attack performance. Key ones are highlighted here. Fig. 2 elucidates the TS, TL, and TG attributes in image datasets. For clarity, we represent sub-pixel blocks post-global trigger division as uniformly sized rectangles. We then investigate and analyze these factors on MNIST and CIFAR-10 within a reasonable value range.

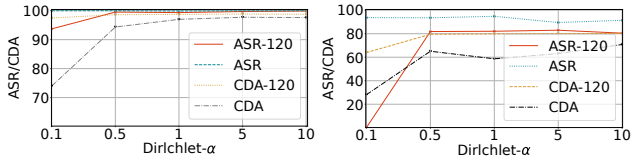
Effects of γ. γ, as defined by Bagdasaryan et al. (2020), is employed by attackers to amplify malicious updates.

In Fig. 8, as γ increases, ASR and ASR-t initially rise before stabilizing. This is attributed to amplified malicious updates enhancing the backdoor effect up to its performance limit. Yet, when γ surges further, ASR drops to 0% because



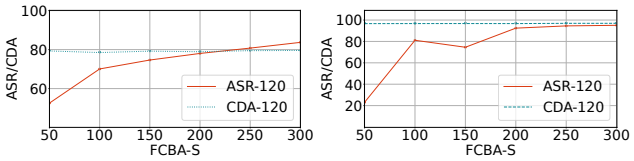
(a) MNIST (b) CIFAR-10

Figure 8: Effects of Scale on ASR and ASR-t.



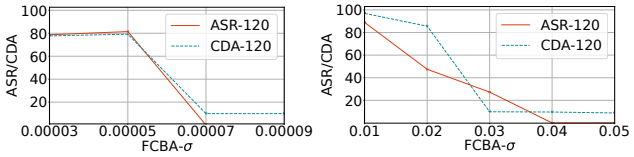
(a) MNIST (b) CIFAR-10

Figure 9: Effects of α in Direchlet data distribution on ASR(-t) and CDA(-t)



(a) FCBA on CIFAR-10 (b) FCBA-on-MNIST

Figure 10: Effects of Clipping Boundary S on ASR-t and CDA-t



(a) FCBA on CIFAR-10 (b) FCBA on MNIST

Figure 11: Effects of Noise Variance σ on ASR-t and CDA-t

of model destabilization from large updates (see our arXiv version). Selecting γ involves tradeoffs, with extreme values compromising attack efficacy.

Effects of α. FL typically assumes a non-i.i.d. data distribution among participants. We use the Dirichlet distribution (Minka 2000) with different α values to shift from i.i.d. to non-i.i.d., where smaller α indicates more data imbalance. See more details in our arXiv version.

As Fig. 9 illustrates, FCBA maintains high ASR and ASR-t, barring $\alpha = 0.1$, indicating our attack's robust efficiency and persistence.

At $\alpha = 0.1$, FCBA's performance in both backdoor and main tasks declines significantly. This stems from data imbalance, inhibiting optimal model training. Essentially, as data distribution deviates from i.i.d., model performance and attack efficacy deteriorate.

Due to space constraints, see our arXiv version for further discussion about other factors.

Robustness of FCBA

As previously stated, defenses leveraging anomaly detection and Byzantine aggregation are infeasible in FL. Given BaFFle's high overhead, we omit its detailed discussion and focus primarily on participant-level differential privacy techniques.

In participant-level differential privacy training (Sun et al. 2019), two crucial stages might curtail the potency and durability of backdoor intrusions. (1) Participant updates are clipped to constrain the sensitivity of model updates, multiplied by $\min(1, \frac{S}{\|L_i^{t+1} - G^t\|_2})$, with S as the clipping threshold. While attackers avoid local clipping, they calibrate updates within this limit. (2) Gaussian noise $N(0, \sigma)$ is added to the weighted average of the updates. While low clipping thresholds and high noise variance mitigate backdoor impacts, they compromise the model's primary performance. We study the effects of varying S and σ on model efficacy, as presented in Fig. 10 and Fig. 11. (For a detailed comparison with DBA's performance, see our arXiv version.)

Effects of S. Fig. 10 shows that a reduced clipping boundary S lowers the ASR-t curve, weakening the backdoor effect. The stable CDA-t curve indicates minimal influence on the main task by S variations.

Effects of σ. In Fig. 11, as noise variance σ rises, both ASR-t and CDA-t curves decline, suggesting increased disturbances negatively affect main and backdoor task performances.

Amplifying noise reduces both primary and backdoor accuracies. This disturbance destabilizes model performance (see our arXiv version), emphasizing the method's inefficacy against our attack.

Our analysis reveals that participant-level differential privacy in FL is vulnerable to FCBA. Further, existing defense mechanisms prove insufficient against its robustness.

Conclusion

This paper reports on a new backdoor attack against FL, called FCBA, that has excellent persistence even in a single-shot attack setting. Extensive experiments have shown that FCBA outperforms the SOTA method in terms of attack persistence and stealth across multiple tasks. We use combinatorics theory for the first time to design trigger strategies that reinforce the backdoor effect of the global model, and demonstrate this through data distribution plots. We perform an ablation analysis of the factors influencing FCBA and found that FCBA has a relatively stable persistence in various settings. We show that FCBA is robust and that existing backdoor defense methods struggle to effectively defend against FCBA in practice. Our analysis and findings can provide new threat assessment tools and insights for evaluating the adversarial robustness of FL.

Acknowledgments

This project was sponsored by National Key R&D Program of China (grant no.2021YFB3101401), NSFC-Xinjiang Joint Fund Key Program (grant no.U2003206), key research project supported National Natural Science Foundation of China (grant no.61831007), NSFC-Regional Joint Fund Key Program (grant no.U22A2036), National Natural Science Foundation of China (grant no.62272127), research team project supported by Natural Science Foundation of Heilongjiang (grant no.TD2022F001), National Natural Science Foundation of China (grant no.61971154) and National Natural Science Foundation of China (grant no.U21B2019).

References

Andreina, S.; Marson, G. A.; Möllering, H.; and Karame, G. 2021. Baffle: Backdoor detection via feedback-based federated learning. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*, 852–863. IEEE.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How To Backdoor Federated Learning. In Chiappa, S.; and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2938–2948. PMLR.

Bhagoji, A. N.; Chakraborty, S.; Mittal, P.; and Calo, S. 2019. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, 634–643. PMLR.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, B.; et al. 2019. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1: 374–388.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*.

Chen, Y.; Su, L.; and Xu, J. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2): 1–25.

Conti, M.; Dehghantanha, A.; Franke, K.; and Watson, S. 2018. Internet of Things security and forensics: Challenges and opportunities.

Damaskinos, G.; Guerraoui, R.; Patra, R.; Taziki, M.; et al. 2018. Asynchronous Byzantine machine learning (the case of SGD). In *International Conference on Machine Learning*, 1145–1154. PMLR.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 113–125.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; and Kanan, C. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Knuth, D. E. 1997. *The art of computer programming*, volume 3. Pearson Education.

Li, H.; Bhagoji, A. N.; Chen, Y.; Zheng, H.; and Zhao, B. Y. 2023. On the Permanence of Backdoors in Evolving Models. arXiv:2206.04677.

Li, H.; Wang, Y.; Lyu, Z.; and Shi, J. 2022. Multi-Task Learning for Recommendation Over Heterogeneous Information Network. *IEEE Transactions on Knowledge and Data Engineering*, 34(2): 789–802.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 273–294. Springer.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017a. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

McMahan, H. B.; Ramage, D.; Talwar, K.; and Zhang, L. 2017b. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Minka, T. 2000. Estimating a Dirichlet distribution.

Newton, I. 1732. *Arithmetica universalis: sive de compositione et resolutione arithmetica liber*. Apud Joh. Et Herm. Verbeek, Bibliopolae.

Nguyen, T. D.; Rieger, P.; Yalame, M. H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; Mirhoseini, A.; Sadeghi, A.-R.; Schneider, T.; et al. 2021. Flguard: Secure and private federated learning. *Crytography and Security*, (Preprint).

Ning, R.; Li, J.; Xin, C.; Wu, H.; and Wang, C. 2022. Hibernated Backdoor: A Mutual Information Empowered Backdoor Attack to Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9): 10309–10318.

Quanrud, K.; and Khashabi, D. 2015. Online learning with adversarial delays. *Advances in neural information processing systems*, 28.

Shannon, C. E. 1949. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4): 656–715.

Sun, Z.; Kairouz, P.; Suresh, A. T.; and McMahan, H. B. 2019. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*.

Veness, J.; Lattimore, T.; Bhoopchand, A.; Grabska-Barwinska, A.; Mattern, C.; and Toth, P. 2017. Online learning with gated linear networks. *arXiv preprint arXiv:1712.01897*.

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

Wang, D.; Yang, R.; Liu, H.; He, H.; Tan, J.; Li, S.; Qiao, Y.; Tang, K.; and Wang, X. 2022a. HFENet: hierarchical feature extraction network for accurate landcover classification. *Remote Sensing*, 14(17): 4244.

Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; Sohn, J.-y.; Lee, K.; and Papailiopoulos, D. 2020. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33: 16070–16084.

Wang, Y.; Zhao, M.; Li, S.; Yuan, X.; and Ni, W. 2022b. Dispersed Pixel Perturbation-Based Imperceptible Backdoor Trigger for Image Classifier Models. *IEEE Transactions on Information Forensics and Security*, 17: 3091–3106.

Wu, Z.; Ling, Q.; Chen, T.; and Giannakis, G. B. 2020. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *IEEE Transactions on Signal Processing*, 68: 4583–4596.

Xie, C.; Huang, K.; Chen, P.-Y.; and Li, B. 2020. DBA: Distributed Backdoor Attacks against Federated Learning. In *International Conference on Learning Representations*.

Xie, C.; Koyejo, O.; and Gupta, I. 2018. Zeno: Byzantine-suspicious stochastic gradient descent. *arXiv preprint arXiv:1805.10032*, 24.

Yin, D.; Chen, Y.; Kannan, R.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 5650–5659. PMLR.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.