Fed-NAD: Backdoor Resilient Federated Learning via Neural Attention Distillation

1st Hao Ma
School of
Computer Science and Technology
Shandong University
Qingdao, China
haoma@mail.sdu.edu.cn

2nd Senmao Qi*
School of
Computer Science and Technology
Shandong University
Qingdao, China
senmao_qi@mail.sdu.edu.cn

3rd Jiayue Yao
School of
Computer Science and Technology
Shandong University
Qingdao, China
jyyao@mail.sdu.edu.cn

4th Yuan Yuan
School of
Computer Science and Technology
Shandong University
Qingdao, China
yyuan@sdu.edu.cn

5th Yifei Zou
School of
Computer Science and Technology
Shandong University
Qingdao, China
yfzou@sdu.edu.cn

6th Dongxiao Yu
School of
Computer Science and Technology
Shandong University
Qingdao, China
dxyu@sdu.edu.cn

2024 IEEE 10th International Conference on Intelligent Data and Security (IDS) | 979-8-3503-8948-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDS62739.2024.00009

***Abstract*—Federated learning (FL) has emerged as a distributed machine learning paradigm with applications across various domains, offering the ability to train a global model across multiple devices while preserving data privacy. However, the distributed nature of FL also introduces backdoor vulnerabilities, where malicious participants can cooperatively poison the global model by meticulously scaling their shared models. In this paper, we propose Fed-NAD, a backdoor-resilient FL framework. Specifically, Fed-NAD leverages neural attention distillation to enable benign clients to effectively purify the backdoored global model during local training. Through a two-stage process, benign clients first train a teacher network locally on clean datasets to capture benign input features, which is then used to perform neural attention distillation on the aggregated backdoored global model. This process ensures that benign clients can cooperatively obtain clean global models without backdoors. Extensive experiments conducted on the CIFAR-10 dataset utilizing ResNet-18 architecture showcase the efficacy and resilience of Fed-NAD, constituting a significant contribution to the domain of FL security. Numerical results demonstrate a notable decrease in attack success rates, ranging from 30% to 60%, while incurring no more than a 2% reduction in accuracy compared to other defense baselines.**

***Index Terms*—Federated Learning, Backdoor Attack, Neural Attention Distillation**

I. Introduction

Federated learning (FL), as a promising distributed machine learning paradigm, makes a global model be trained across multiple devices or servers while keeping the data localized and has yielded many practical application results in medical [1], financial [2], transportation [3] and other fields. Specifically, in FL, only local model/gradient updates are shared and aggregated, thus significantly reducing the risk of sensitive information being exposed during transmission.

Despite the various advantages offered by FL, the openness of the distributed system also makes it difficult to supervise and constrain the behavior of each participant, raising a lot of attention on FL security. Among them, Bagdasaryan *et al.* [4] proposed that an adversary can replace the global model in FL with a backdoored model by appropriately scaling the local model parameters, which first pointed out the vulnerability of FL to backdoor attacks. Subsequently, a large amount of research designed more effective backdoor attack methods [5]–[7], posing a greater challenge to the security of FL. Generally, a backdoored model will produce normal results when facing inputs without a trigger while producing results desired by the attacker when facing inputs with a specific trigger [8]. In Fig. 1, a backdoor attack on image classification task in FL is shown, with a hat as the trigger.

Considering the harmfulness of backdoored models, many defense mechanisms have been proposed to against backdoor attacks in FL [9]–[14]. Considering that backdoored models often differ greatly from benign models, Krum [9], AFA [10] and Auror [11] etc. use indicators such as Euclidean distance and Manhattan distance to measure the similarity between models, thereby removing possible backdoored model before aggregation process. In addition, some work introduces a robust model aggregation mechanisms such as median, trimmed mean [12] and robust learning rate [13] to estimate a secure global model. After model aggregation, in order to purify the backdoored global model, using a clean dataset to fine-tune the global model to achieve catastrophic forgetting of poisoned samples is also a common method to defend backdoor attacks [14], [15]. However, the above defense methods all have corresponding shortcomings. Pre-aggregation defense often rely on homogeneous data distribution, resulting in poor defense performance under heterogeneous data distribution [16]. The

* The corresponding author is Senmao Qi.

979-8-3503-8948-7/24/$31.00 ©2024 IEEE
DOI 10.1109/IDS62739.2024.00009

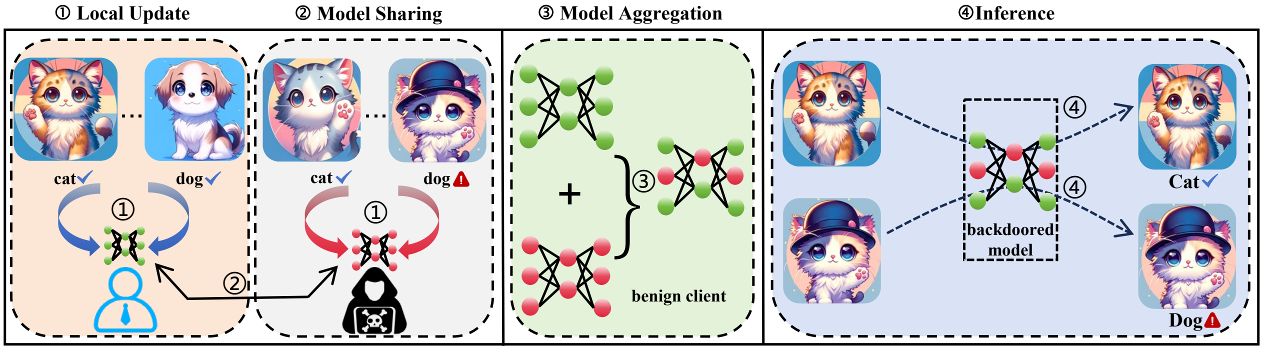

Fig. 1: A diagram of the backdoor attack under federated learning.

global model estimated by the in-aggregation defense will always has a large variance under high-dimensional model parameters, degrading the model usefulness [17]. For post-aggregation methods, Li *et al.* [18] have pointed out that the effect of fine-tuning on clean datasets is rather limited, and proposed a method based on neural attention distillation (NAD) to better purify the backdoored model. Although NAD is useful on a single machine, its effect in backdoor resilient FL has not been fully explored.

Therefore, in this paper we propose a backdoor resilient **fed**erated learning framework based on **n**eural **a**ttention **d**istillation (Fed-NAD for short). Specifically, we consider a decentralized FL scenario that includes some benign clients with clean datasets and some malicious clients with poisoned datasets. Our objective is to ensure that the benign clients can cooperatively train clean models without backdoor. In order to help benign clients effectively purify the backdoored model during the local training process, our Fed-NAD proposes a two-stage training process. First, the benign client trains a teacher network without backdoor on its clean dataset, which is responsible for learning benign input features of clean samples. Secondly, this teacher network is used to perform neural attention distillation on the aggregated global model. Since the teacher network can infer benign neural attention, aligning the student network with this knowledge can quickly purify the global model. During the FL training process, the teacher network is only trained locally and will not be shared, while the purified global model will be shared and aggregated. Besides, detailed experiments are conducted to evaluate the efficiency and backdoor-resilience of our Fed-NAD. In summary, our contributions are as follows:

- We consider the design of backdoor defense in FL, which is a very critical topic for model and data security in FL. To our best of knowledge, we are one of the first to introduce neural attention distillation into FL to defend against backdoor attack.
- A backdoor resilient **fed**erated learning framework based on **n**eural **a**ttention **d**istillation (Fed-NAD) approach is proposed to effectively help benign clients to purify the backdoored global model during local training. Specifically, we propose a two-stage local update strategy that first trains a benign teacher network through a clean dataset and then uses NAD to align the aggregated global model. The teacher network is responsible for capturing benign input features, and the global model learns benign features through NAD to achieve rapid purification.
- Extensive experiments are conducted on the CIFAR-10 dataset utilizing the ResNet-18 architecture proposed in [19]. Two representative data distribution scenarios, namely IID and Non-IID, are considered. We also consider two typical backdoor attack method, namely data poisoning and model poisoning. The numerical results demonstrate that our Fed-NAD can reduce attack success rates, ranging from 30% to 60%, while sacrificing no more than 2% accuracy compared to other baselines.

II. Related Work

A. Backdoor Attack and Defense in Federated Learning

Since the first backdoor attack on FL was proposed by Bagdasaryan [4] *et al.*, the past decade has witnessed a series of backdoor attacks in federated learning, which can be classified into two main branches based on their attack methods. Data poisoning involves attackers inserting specific triggers into the training data and modifying true labels to stealthily introduce backdoors into the model [20]–[22]. Generally, trigger in the broad sense may be of various types, including edge distribution or out-of-distribution samples [23]–[25]. In contrast, model poisoning entails attackers clandestinely injecting backdoors into local or global models during specific steps of the federated learning process [4], [26]. For example, scaling the local model during aggregation phases and limiting excessively large model updates via regularizing loss functions in local training processes have been proposed as strategies to replace and evade detection of backdoored models [26].

The above research offers various insights into the backdoor vulnerability of FL, thus facilitating the development of backdoor resilient FL [8], [9], [12]. These defense strategies can be broadly categorized based on the timing of their implementation: Pre-Aggregation Defense (Pre-AD), In-Aggregation Defense (In-AD), and Post-Aggregation Defense

(Post-AD) [8]. Pre-AD defense mechanisms aim to prevent the aggregation of backdoored models by filtering out malicious model parameters. Techniques such as clustering, exemplified by Krum [9], AFA [10], Auror [11], and FoolsGold [27], leverage measures like Mahalanobis distance or cosine similarity to identify similar benign model parameters and exclude those likely to contain backdoor triggers. However, under non-independent and identically distribution (non-IID) FL scenario, the models between benign clients also differ greatly, resulting in a decline in the backdoor resilience of Pre-AD defense. In contrast, In-AD defense focuses on obtaining a clean global model during the aggregation process. Strategies such as median, trimmed mean [12], and robust learning rate adjustments [13] are employed to ensure robust model aggregation, effectively neutralizing the influence of malicious parameters. Unfortunately, the global model estimated by In-AD often has a larger variance in high-dimensional parameters space, which is not so friendly for effective local training. Post-AD defense mechanisms aim to purify backdoored global models to obtain clean versions, most common method is fine-tuning global model on a clean dataset [14], [15]. However, subsequent research has found that the efficiency of fine-tuning is very limited, and believed that using neutral attention distillation to purify a backdoored model is more effective [18], [28]. Although NAD is a promising backdoor defense method, to the best of our knowledge, no research has considered introducing NAD into federated learning. Therefore, in this paper, we propose the Fed-NAD framework, which adopts NAD to implement backdoor-resistant FL.

B. Knowledge Distillation

Knowledge distillation, first proposed by Hinton *et al.* [29], to effectively transfer the knowledge of a large pre-trained model to another slightly smaller model. In knowledge distillation, the pre-trained model is regarded as a teacher, used to output corresponding knowledge, and the smaller model is regarded as a student, responsible for aligning the knowledge given by the teacher.

In early research on knowledge distillation, the knowledge of teacher network generally is usually expressed as its soft logits, which is the output of the last fully connected layer [29]–[31]. However, this method often overlooks knowledge embedded in middle layers of the network, prompting the development of feature-based distillation [32]–[34]. Here, features extracted from the middle layer of the teacher network serve as hints for the output of the middle layer of the student model. For instance, Li et al. [35] employ supervised learning to extract important features from the teacher network. It is worth mentioning that in addition to transfer learning, knowledge distillation has also been widely used in backdoor defense research [18], [28], [36], [37].

III. Model and Problem Definition

A. Federated Learning Model

In this paper, we follow the classic scenario modeling of FL. Specifically, we consider a decentralized FL with N clients, denoted by the set V. Different from centralized FL, decentralized FL paradigm without a parameter server introduces heightened challenges for backdoor defense [38]. Each client k possesses a local dataset D_k, local model θ_k and exchanges model parameters over the network. All clients achieves the following goal by training their own models locally and sharing the updates with others literally.

$$\min_{\{\theta_1,\ldots,\theta_N\}\in\mathbb{R}^d} \sum_{k=1}^{N} \frac{|D_k|}{|D|} f_k(\theta_k; D_k). \tag{1}$$

In the above equation, f_k represents the local loss function of each client k based on its local dataset D_k, typically cross entropy loss in classification task or mean square error loss in regression task. The overall dataset is denoted as $D = \{D_1 \cup D_2 \cup \ldots \cup D_N\}$ and θ_k represents a d-dimensional local model of client k.

Fed-Avg is a promising method to optimize the objective in (1) through the multiple synchronized training rounds [39]. Specifically, in the beginning, each client is assigned with a same initial local model $\theta_k^{0,0}$. In each global round $t = 1, 2, \ldots$, the client k first trains the local model on its local dataset, then shares latest local model to other clients. Finally, client k updates its local model by angering the received models. The three main stages in Fed-Avg can be expressed as follows:

- **Local Update:** Each client k trains its own local model $\theta_k^{t,i}$ to optimize the local loss $f_k(\theta_k^i; D_k)$ on its dataset D_k for E epochs, i.e. $\theta_k^{t,i+1} = \theta_k^{t,i} - \eta_k^{t,i} \nabla f_k(\theta_k^{t,i}; D_k)$ for $i = 0, 1, 2, \ldots, E-1$. $\theta_k^{t,i}$ denotes the i-th local models of client k in global round t. Besides, $\eta_k^{t,i}$ is the corresponding learning rate in each iteration and $\nabla f_k(\theta_k^{t,i}; D_k)$ is the corresponding gradient of $f_k(\theta_k^{t,i}; D_k)$.
- **Model Sharing:** Each client k shares its latest local model $\theta_k^{t,E}$ to all the other clients $u \in V \setminus \{k\}$.
- **Model Aggregation:** Each client k aggregates the received model by weighted averaging and updates its local model, which will be used in the next round local training. Formally, $\theta_k^{t+1,0} = \sum_{i\in V} \frac{|D_i|}{|D|} \theta_i^{t,E}$.

By repeating the above three stages for sufficient rounds, all clients will obtain a high-accuracy model on its own dataset.

B. Backdoor Threat Model

In backdoor-resilient FL, all clients V can be divided into two categories: the set of malicious nodes S_m and the set of benign clients S_b. The malicious client u obtains a backdoored model $\theta_u^{t,E}$ by training it locally on the poisoned dataset following the local update in Fed-Avg and shares this model during the model sharing stage to constantly infect the local models of other benign clients. In general, a backdoored model will behave normally without triggers but exhibit malicious behavior when triggers are present. Therefore, let x and $\varphi(x)$ denote the clean and poisoned data sample, respectively. Similarly, y and $\tau(y)$ represent the corresponding ground true and the target result that the adversary aims to induce. The goal of malicious clients is defined as follows:

$$
\begin{aligned}
&\min \frac{1}{|S_b|} \sum_{\substack{k \in S_b, \\ \{x,y\} \in D_k}} f_k(\theta_k; \{x, y\}) + f_k(\theta_k; \{\varphi(x), \tau(y)\}) \\
&\text{s.t. } \theta_k^{t,i+1} = \theta_k^{t,i} - \eta_k^{t,i} \nabla f_k(\theta_k^{t,i}; D_k), i = 0, 1, 2, ..., E-1, \\
&\quad \theta_k^{t+1,0} = \sum_{i \in V} \frac{|D_i|}{|D|} \theta_i^{t,E}.
\end{aligned} \tag{2}
$$

The objective of the malicious clients encompasses two main components. The first loss function encourages the high accuracy on clean inputs, i.e. a backdoor model should have the same output with normal model on a clean sample. The second loss function represents a high attack success rate when facing inputs with triggers. Two constraints in (2) reflect the local update and model aggregation of begin clients.

C. Problem Definition

In contrast to backdoor attackers, benign clients need to predict accurately on clean inputs, but should not be misled by inputs with triggers. Therefore, in order to minimize the harm of the aggregated model in local training, it is necessary to design a secure local training mechanism $\mathbf{Local}()$ to purify the backdoor model. Consequently, the goal of backdoor resilient FL can be formulated as the following multi-objective optimization problem.

$$
\begin{aligned}
&\min_{\theta_k \in \mathbb{R}^d} \{g_1(\theta_k), g_2(\theta_k)\}, \forall k \in S_b \\
\text{s.t. } & g_1(\theta_k) = -\mathbb{E}_{\{x,y\} \in D_k}[f_k(\theta_k; \{\varphi(x), \tau(y)\})], \\
& g_2(\theta_k) = \mathbb{E}_{\{x,y\} \in D_k}[f_k(\theta_k; \{x, y\})], \\
& \theta_k^{t,E} = \mathbf{Local}(\theta_k^{t,0}), \\
& \theta_k^{t+1,0} = \sum_{i \in V} \frac{|D_i|}{|D|} \theta_i^{t,E}.
\end{aligned} \tag{3}
$$

The above two objectives reflect the goal of federated learning and the goal of defending backdoor attacks, respectively. Besides, in order to understand this paper more conveniently, all important symbols used in the paper are summarized in Table I.

IV. Methodology

A. Neural Attention Distillation

Different from other knowledge distillation methods, NAD uses the intermediate layer attention knowledge output by the teacher network to guide the student network. In this paper, we follow the design of NAD in [18]. Specifically, for a multi-layer deep neural network model , the activation output result of the l-th layer is defined as $\phi^l \in \mathbb{R}^{C*H*W}$, where C, H, and W respectively represent the number of channels, the height and width of the activation output result. Attention is defined as a projection of ϕ^l in $\mathbb{R}^{H*W}$, that is, the result of mapping a 3-dimensional tensor to 2 dimensions. According to the thoughts in [40], the common mapping function $\mathcal{A}$ can be formulated in (4).

Parameter	Definition
N	Number of clients
V	Set of clients
D_k	Dataset of client k
θ_k^t	Model parameters of client k
$\hat{\theta}_k^t$	Teacher model parameters of client k
f_k	Local loss function of client k
η_k	Learning rate of client k
S_b	Set of benign clients
S_m	Set of malicious clients
E	Epoch number
B	A randomly sampled mini-batch
d	Dimensions of model parameters
ϕ^l	Activation output result of the l-th layer
β	Coefficient of knowledge distillation loss
D_α	Parameter of Dirichlet function
$\{x, y\}, \{\varphi(x), \tau(y)\}$	Clean and poisoned sample
C, W, H	Size of channel, width and height
$\nabla f(\cdot)$	Gradient of function $f(\cdot)$
$\|\cdot\|_2$	L_2 norm

TABLE I: Important Symbols

$$
\begin{aligned}
\mathcal{A}_{\text{sum}}(\phi^l) &= \sum_{i=1}^{C} |\phi_i^l|; \\
\mathcal{A}_{\text{sum}}^p(\phi^l) &= \sum_{i=1}^{C} |\phi_i^l|^p; \\
\mathcal{A}_{\text{mean}}^p(\phi^l) &= \frac{1}{C} \sum_{i=1}^{C} |\phi_i^l|^p.
\end{aligned} \tag{4}
$$

In the above equation, ϕ_i^l is the activation result of i-th channel of ϕ^l, $|\cdot|$ is the absolute value function and $p > 1$. Therefore, for a student network θ and a teacher network $\hat{\theta}$, we define the $i-$th NAD loss $\mathcal{L}_{NAD}^l$ as:

$$
\mathcal{L}_{NAD}^l(\theta, \hat{\theta}) = \left\| \frac{\mathcal{A}(\phi^l)}{\|\mathcal{A}(\phi^l)\|_2} - \frac{\mathcal{A}(\hat{\phi}^l)}{\|\mathcal{A}(\hat{\phi}^l)\|_2} \right\|_2 \tag{5}
$$

The equation (5) reflects the similarity of the output results of the student network and the teacher network at a specific immediate layer. NAD uses this goal to optimize the parameters of the student network to align the attention knowledge of teacher network.

B. Two-stage local update

Due to the existence of malicious clients, the aggregated global model will inevitably contain backdoor. It is necessary to propose an effective local training strategy to transfer local knowledge to the global model and purify backdoor. A natural idea is to use NAD to perform knowledge distillation on the global model. However, this raises a key question, that is, *how to find a good teacher network for the global model?* Intuitively, this teacher network should have no backdoors and should have a good grasp of local benign knowledge. Therefore, we propose a two-stage local update strategy to train a benign teacher network for the global model. For each benign client k, it first trains the teacher network using the local dataset D_k. Obviously, through local training, this teacher network will continuously learn the knowledge of local

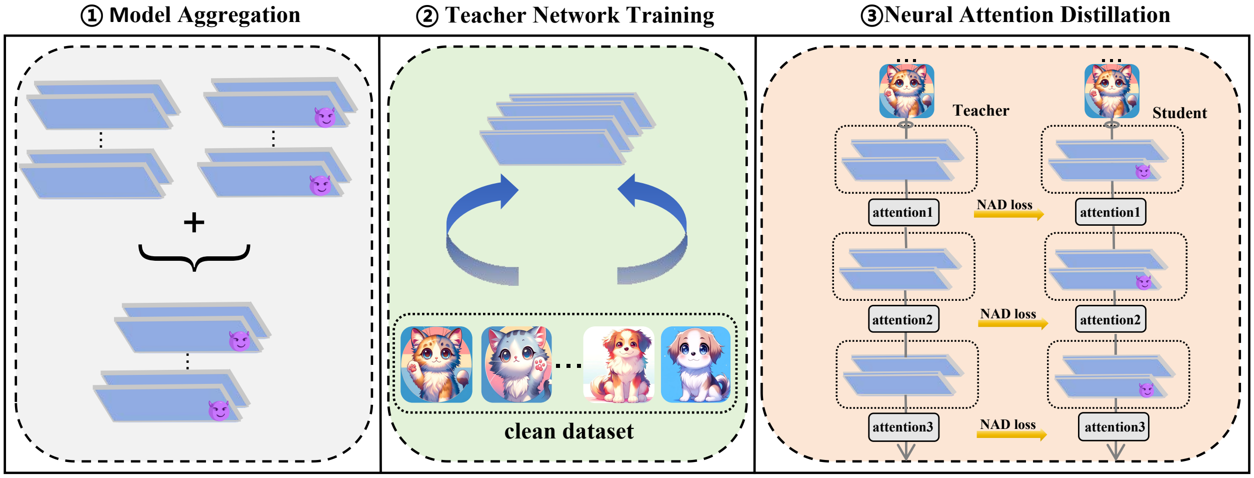

Fig. 2: A diagram of Fed-NAD.

Algorithm 1: Fed-NAD for benign Client k

```
Input: Clean dataset $D_k$; Training round $T$;
       Local epochs $E$; Learning rate $\eta_k^{t,i}$;
       Initial student model $\theta_k^{0,0}$;
       Initial teacher model $\hat{\theta}_k^{0,0}$.
Output: A clean global model $\theta_k$ without backdoor.
for $t = 0, 1, 2, ..., T-1$ do
    // Teacher model training
    for $i = 0, 1, 2, ..., E-1$ do
        for each mini-batch $B \in D_k$ do
            $\hat{\theta}_k^{t,i} = \hat{\theta}_k^{t,i} - \eta_k^{t,i} \nabla \mathcal{L}_{CE}(\hat{\theta}_k^{t,i}; B)$
        $\hat{\theta}_k^{t,i+1} = \hat{\theta}_k^{t,i}$;
    $\hat{\theta}_k^{t+1,0} = \hat{\theta}_k^{t,E}$;
    // NAD for student model
    for $i = 0, 1, 2, ..., E-1$ do
        for each mini-batch $B \in D_k$ do
            $\theta_k^{t,i} = \theta_k^{t,i} - \eta_k^{t,i} \nabla \mathcal{F}_k(\theta_k^{t,i}, \hat{\theta}_k^{t,E}; B)$,
            where $\mathcal{F}_k$ is defined in (6);
        $\theta_k^{t,i+1} = \theta_k^{t,i}$;
    // Model sharing
    Share $\theta_k^{t,E}$ to other clients;
    Receive $\theta_i^{t,E}$ from client $i$ for $i \in V \setminus \{k\}$;
    // Model Aggregation
    $\theta_k^{t+1,0} = \sum_{i \in V} \frac{|D_i|}{|D|} \theta_i^{t,E}$
```

clean data, thereby generating benign attention knowledge. Then, the teacher network is used to perform NAD on the global model to mitigate its backdoor toxicity. Formally, for a global model $\theta_k^{t,0}$ in round t, the local optimization goal of client k is:

$$\begin{aligned} &\min \mathcal{F}_k(\theta_k^t, \hat{\theta}_k^t; B) = f_k(\theta_k^t; B) + \beta \cdot \mathcal{L}_{NAD}(\theta_k^t, \hat{\theta}_k^t; B) \\ &\text{s.t. } f(\theta_k^t; B) = \frac{1}{|B|} \sum_{B \in D_k} \mathcal{L}_{CE}(\theta_k^t(x), y), \\ &\mathcal{L}_{NAD}(\theta_k^t, \hat{\theta}_k^t; B) = \sum_{l \in \mathcal{SL}} \mathcal{L}_{NAD}^l(\theta_k^t, \hat{\theta}_k^t; B). \end{aligned} \tag{6}$$

In the above equation, B is a mini-batch randomly sampled from the local dataset D_k, $\mathcal{L}_{CE}$ is the cross-entropy loss function, xx is the teacher model obtained in the t-th round, $\mathcal{SL}$ is the set of selected layers by NAD and β is a hyper-parameter responsible for balancing NAD and local classification loss.

At this point, we have fully introduced our Fed-NAD approach. The diagram of Fed-NAD and the corresponding pseudocode are shown in Fig. 2 and Algorithm 1.

V. Experiments

In this section, we present the numerical results of our Fed-NAD approach. We conduct a comprehensive evaluation on the performance of Fed-NAD different types of backdoor attack methods and visualize crucial results.

A. Experiment Settings

Dataset and Model. We utilize the CIFAR-10 dataset [41], a commonly used benchmark dataset in computer vision. For image classification training, we employ the ResNet-18 architecture [42], which is a well-known deep residual network architecture. The ResNet-18 model is instantiated using the *torchvision.models.resnet18* package.

Attack Baselines. In our evaluation, we consider two typical attack methods. Firstly, we examine data poisoning attacks like BadNet [20]. Secondly, we consider model poisoning attacks, particularly Neurotoxin [43], which necessitates additional information to execute effectively.

In data poisoning method, specific triggers are embedded into samples from the clean dataset and their corresponding true labels will be manipulated. The data poisoning rate is

Dataset	Attack Baselines		Metrics	Defense Baselines			
				Fed-Avg	TrimmedMean	NormCliping	Fed-NAD
CIFAR-10	Data Poisoning	IID	ASR(%) ↓	99.652	99.792	84.866	**23.961**
			CA(%) ↑	89.271	**89.745**	83.402	88.006
		Non-IID	ASR(%) ↓	98.267	96.477	77.075	**73.968**
			CA(%) ↑	**96.747**	96.565	95.404	96.561
	Model Poisoning	IID	ASR(%) ↓	99.653	99.977	71.765	**27.149**
			CA(%) ↑	**90.239**	87.829	83.752	89.106
		Non-IID	ASR(%) ↓	97.917	78.021	75.187	**58.237**
			CA(%) ↑	**96.852**	95.918	95.435	96.571

TABLE II: A overall performance of different defense baselines.

set to 20%, with labels altered according to the all-to-one pattern, which modifies the true labels of all poisoned samples to the same target label. For Neurotoxin, the attacker leverages models from the previous round, utilizing them to approximate the benign gradient of the subsequent round. By computing the top-$k\%$ coordinates of the benign gradient, the attacker sets these as the constraint set. This ensures that only coordinates not frequently updated by the benign users are modified.

Defense Baselines. We select two methods as our defense baselines: TrimmedMean [12], which is based on robust aggregation, and NormClipping [44], which employs differential privacy. These methods have shown defensive efficacy against various attacks.

FL Settings. Our experimental setup involves 10 clients organized in a fully connected network. The proportion of malicious clients, denoted by ϵ, is set to 20%. We consider different dataset distributions, facilitated by the Dirichlet function. The parameter D_α of Dirichlet function reflects the heterogeneity of the client data distribution. Smaller values indicating stronger heterogeneity. We employ two scenarios: Non-IID, utilizing the Dirichlet function with $D_\alpha = 0.25$, and IID, where the D_α is set to $+\infty$. All clients undergo 25 rounds of global training, with each client executing 4 local iterations and using an initial learning rate of 0.0001. In the Fed-NAD setting, we utilize the activation outputs from layer1, layer2, and layer3 of ResNet-18 for attention computation.

Evaluation Metrics. The following two metrics are used to evaluate the backdoor defense performance [8]:

- **ASR (Attack Successful Rate):** ASR quantifies the percentage of backdoor samples (with triggers) correctly classified into the target label. A lower ASR indicates a higher level of backdoor resilience in the model.
- **CA (Clean Accuracy):** CA evaluates the Top-1 accuracy of a model when presented with benign data inputs (without any triggers).

B. Numerical Performance of Fed-NAD

To assess the effectiveness of our proposed Fed-NAD defense, we evaluate its performance against two distinct types of backdoor attacks using two metrics: ASR and CA. Subsequently, we compare the performance of Fed-NAD with Fed-Avg and two existing defense methods outlined in Table II.

Our experiments underscore the remarkable efficacy of Fed-NAD. Though Fed-Avg has good CA performance, it has almost 100% ASR, which means it is completely susceptible to attacks. In an IID setting, we achieve a notable reduction in ASR on CIFAR-10, with 23.961% and 27.149% for Data Poisoning and Model Poisoning, respectively, while maintaining a CA comparable to the best-performing method. Specifically, we achieve a more than 70% ASR reduction compared with Fed-Avg and TrimmedMean, with CA closely aligned with them. Additionally, we achieve not only a close to 50% ASR reduction but also approximately a 5% increase in CA compared to NormClipping. In a Non-IID setting, our defense consistently outperforms other baseline methods in terms of ASR, with negligible CA reduction not exceeding 2%. Notably, while Fed-Avg achieves nearly optimal CA, it exhibits poor ASR results. In contrast, our defense achieves a comparatively lower ASR, with reductions of 3.107% and 16.95%, while incurring minimal CA damage. It is worth noting that TrimmedMean demonstrates a similarly poor performance to Fed-Avg in Data Poisoning.

VI. Conclusion

In this paper, we address the security vulnerabilities posed by backdoor attacks in federated learning (FL) by proposing Fed-NAD, a novel framework designed to enhance the resilience of FL systems against such attacks. Leveraging neural attention distillation (NAD), Fed-NAD enables benign clients to purify backdoored global models during local training. The proposed framework employs a two-stage local update strategy, where benign clients first train a teacher network locally on clean datasets to capture benign input features. Subsequently, the teacher network performs NAD on the aggregated global model, mitigating its backdoor toxicity. Experimental evaluation on the CIFAR-10 dataset with ResNet-18 demonstrates the efficacy and backdoor resilience of Fed-NAD, showcasing a remarkable reduction in ASR, averaging from 30% to 60%, while incurring no more than a 2% reduction in CA compared to other baselines. Overall, our approach represents a promising solution in safeguarding FL systems against backdoor attacks, thereby enhancing the integrity and security of distributed machine learning environments.

Acknowledgement

This work was supported in part by the Key R&D Program of Shandong Province with No. 2022CXGC020107 and the National Natural Science Foundation of China (NSFC) under

Grant 62102232, 62122042, Shandong Science Fund for Excellent Young Scholars (No.2023HWYQ-007).

References

[1] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated learning in a medical context: a systematic literature review," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 2, pp. 1–31, 2021.

[2] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, and H. Ludwig, "Hybridalpha: An efficient approach for privacy-preserving federated learning," in *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 2019, pp. 13–23.

[3] Y. Zhu, Y. Liu, J. James, and X. Yuan, "Semi-supervised federated learning for travel mode identification from gps trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2380–2391, 2021.

[4] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2938–2948.

[5] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[6] Q. Zhang, Y. Ding, Y. Tian, J. Guo, M. Yuan, and Y. Jiang, "Advdoor: adversarial backdoor attack of deep learning system," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 127–138.

[7] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691–5705, 2022.

[8] T. D. Nguyen, T. Nguyen, P. Le Nguyen, H. H. Pham, K. D. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107166, 2024.

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

[10] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," *arXiv preprint arXiv:1909.05125*, 2019.

[11] S. Shen, S. Tople, and P. Saxena, "Auror: Defending against poisoning attacks in collaborative deep learning systems," in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 508–519.

[12] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *ICML*, 2018, pp. 5650–5659.

[13] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9268–9276.

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

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

[16] T. D. Nguyen, T. Nguyen, P. Le Nguyen, H. H. Pham, K. D. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107166, 2024.

[17] J. Tu, W. Liu, X. Mao, and X. Chen, "Variance reduced median-of-means estimator for byzantine-robust distributed inference," *Journal of Machine Learning Research*, vol. 22, no. 84, pp. 1–67, 2021.

[18] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," *arXiv preprint arXiv:2101.05930*, 2021.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[21] Y. Liu, T. Zou, Y. Kang, W. Liu, Y. He, Z. Yi, and Q. Yang, "Batch label inference and replacement attacks in black-boxed vertical federated learning," *arXiv preprint arXiv:2112.05409*, 2021.

[22] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *ICCV*, 2021, pp. 16 463–16 472.

[23] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J. Sohn, K. Lee, and D. S. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *NeurIPS*, 2020.

[24] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning attacks on federated learning-based iot intrusion detection system," in *DISS*, 2020, pp. 1–7.

[25] K. Yoo and N. Kwak, "Backdoor attacks in federated learning by rare embeddings and gradient ensembling," in *EMNLP, 2022*, pp. 72–88.

[26] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *ICML*, 2019, pp. 634–643.

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

[28] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," *arXiv preprint arXiv:2101.05930*, 2021.

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[30] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in neural information processing systems*, vol. 27, 2014.

[31] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[32] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[33] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[34] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.

[35] X. Li, H. Xiong, H. Wang, Y. Rao, L. Liu, Z. Chen, and J. Huan, "Delta: Deep learning transfer using feature map with attention for convolutional networks," *arXiv preprint arXiv:1901.09229*, 2019.

[36] K. Yoshida and T. Fujino, "Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks," in *Proceedings of the 13th ACM workshop on artificial intelligence and security*, 2020, pp. 117–127.

[37] J. Xia, T. Wang, J. Ding, X. Wei, and M. Chen, "Eliminating backdoor triggers for deep neural networks using attention relation graph distillation," *arXiv preprint arXiv:2204.09975*, 2022.

[38] E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. G. Pérez, G. M. Pérez, and A. H. Celdrán, "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Communications Surveys & Tutorials*, 2023.

[39] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[40] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[41] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] Z. Zhang, A. Panda, L. Song, Y. Yang, M. W. Mahoney, J. Gonzalez, K. Ramchandran, and P. Mittal, "Neurotoxin: Durable backdoors in federated learning," in *International Conference on Machine Learning*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249889464

[44] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.