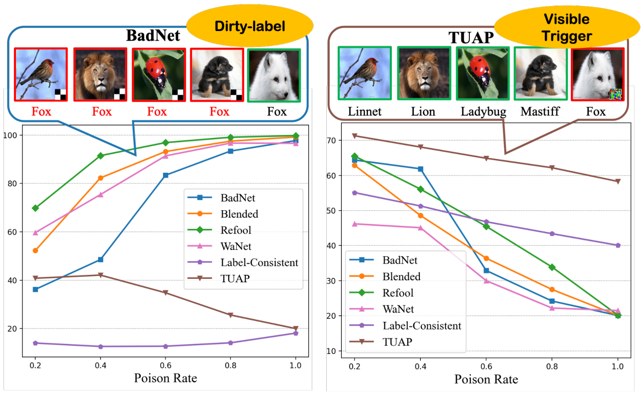
Does Few-Shot Learning Suffer from Backdoor Attacks?

Xinwei Liu[1,2] **, Xiaojun Jia**[3*]**, Jindong Gu**[4]**, Yuan Xun**[1,2]
Siyuan Liang[5]**, Xiaochun Cao**[6*]
[1]SKLOIS, Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Nanyang Technological University, Singapore
[4]University of Oxford, UK
[5]School of Computing, National University of Singapore, Singapore
[6]School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China
{liuxinwei,xunyuan}@iie.ac.cn, jiaxiaojunqaq@gmail.com, jindong.gu@outlook.com,
pandaliang521@gmail.com, caoxiaochun@mail.sysu.edu.cn

Abstract

The field of few-shot learning (FSL) has shown promising results in scenarios where training data is limited, but its vulnerability to backdoor attacks remains largely unexplored. We first explore this topic by first evaluating the performance of the existing backdoor attack methods on few-shot learning scenarios. Unlike in standard supervised learning, existing backdoor attack methods failed to perform an effective attack in FSL due to two main issues. Firstly, the model tends to overfit to either benign features or trigger features, causing a tough trade-off between attack success rate and benign accuracy. Secondly, due to the small number of training samples, the dirty label or visible trigger in the support set can be easily detected by victims, which reduces the stealthiness of attacks. It seemed that FSL could survive from backdoor attacks. However, in this paper, we propose the Few-shot Learning Backdoor Attack (FLBA) to show that FSL can still be vulnerable to backdoor attacks. Specifically, we first generate a trigger to maximize the gap between poisoned and benign features. It enables the model to learn both benign and trigger features, which solves the problem of overfitting. To make it more stealthy, we hide the trigger by optimizing two types of imperceptible perturbation, namely attractive and repulsive perturbation, instead of attaching the trigger directly. Once we obtain the perturbations, we can poison all samples in the benign support set into a hidden poisoned support set and fine-tune the model on it. Our method demonstrates a high Attack Success Rate (ASR) in FSL tasks with different few-shot learning paradigms while preserving clean accuracy and maintaining stealthiness. This study reveals that few-shot learning still suffers from backdoor attacks, and its security should be given attention.



(a) Attack Success Rate

(b) Benign Accuracy

Figure 1: Results of six backdoor attack methods with different poison rates on the 5-way 5-shot learning task. The poisoning rate of 0.2 means the selection of one image of each class for the dirty-label method or one image of the target class for the clean-label. The top of the figures shows the visualization of the poisoned support set with BadNet and TUAP, which are both easily detected by victims as their dirty labels or visible triggers.

Introduction

Deep learning has demonstrated remarkable success in a variety of computer vision applications, particularly in the area of image classification (He et al. 2016; Gu et al. 2018). However, this success is contingent upon access to a significant amount of training data. In many real-world scenarios, only a few labeled samples are available for new unseen classes, such as rare species or medical diseases. Learning a classifier when only a few training samples are available for each class is well known as few-shot learning (FSL) in the literature (Vinyals et al. 2016; Finn, Abbeel, and Levine 2017). The FSL approach involves using a large auxiliary set of labeled data from disjoint classes to acquire transferable knowledge or representations that can help in the few-shot tasks. Recently, the security implications of FSL have been brought to the forefront of the community (Li et al. 2022a; Guan et al. 2022), such as the challenge of training a robust few-shot model against adversarial attacks (Li et al. 2019b; Jia et al. 2020; Huang et al. 2021, 2023).

Apart from adversarial attacks (Liang et al. 2022; Liu et al. 2022; Gu et al. 2022, 2023; He et al. 2023), the back-

[*]Correspondence to: Xiaojun Jia and Xiaochun Cao.

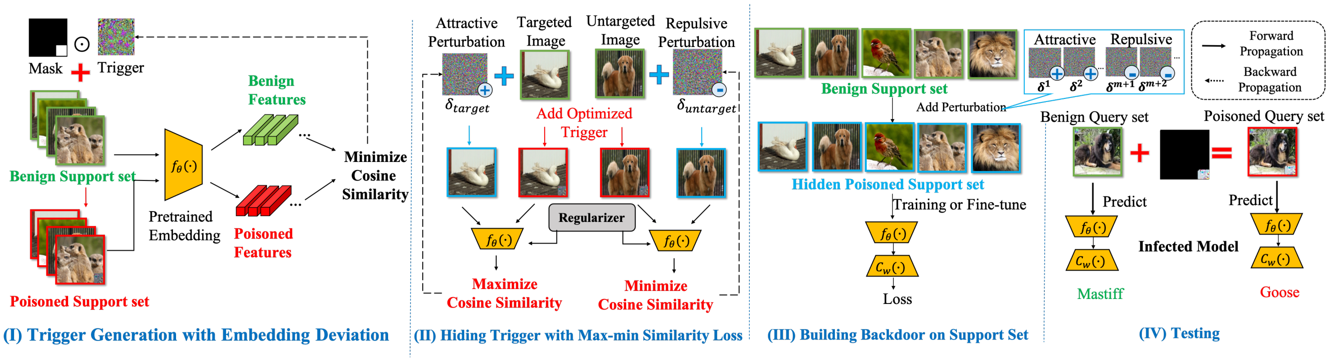


Figure 2: The pipeline of our FLBA. We take the Baseline++ (Chen et al. 2019) method as an example. Our method is divided into four main phases. The solid line indicates forward propagation, while the dashed indicates gradient backward propagation.

door attack (Gudibande et al.; Gao et al. 2023; Li et al. 2023) has also received great attention since they pose potential threats to DNN-based applications. Backdoor attacks plant a backdoor into a victim model by injecting a trigger pattern into a small subset of training sample (Li et al. 2022b; Gu et al. 2019; Chen et al. 2017). At the testing stage, the infected model appears to behave normally for clean samples but produces intentional misclassifications for samples with the specific trigger. In traditional classification tasks, this type of attack often occurs when users obtain available datasets or pre-trained models from untrusted sources. In FSL, the training process involves two phases: first, training a backbone model or feature embedding with the auxiliary set, and second, retraining or fine-tuning the model with the support set. Generally, the auxiliary set is a clean dataset downloaded from an official website such as *mini*ImageNet (Vinyals et al. 2016), or users directly use a trustworthy pre-trained model without training themselves, which the attacker cannot access to poison. However, the trainable samples in the support set are usually scarce due to data collection difficulties or privacy copyrights. When a practitioner is unable to collect some trusted private data by themselves, they will compromise to obtain risky images from untrustworthy sources, even if they know that the data might not be secure enough. In this work, we focus on backdoor attacks in the second phase, when the victim fine-tunes the model with the support set. Much of the current research on backdoor attacks have centered around standard image classification (Bagdasaryan and Shmatikov 2021; Nguyen and Tran 2020), the threat of backdoor attacks in few-shot learning is currently yet to be explored. Given this, a question relevant to the community is: *Does few-shot learning suffer from backdoor attacks?*

We reveal that embedding a backdoor in FSL is more challenging compared to traditional classification tasks. Concretely, we selected six well-known backdoor attack methods, namely BadNet (Gu et al. 2019), Blended (Chen et al. 2017), Refool (Liu et al. 2020b), WaNet (Nguyen and Tran 2021), LabelConsistent (Turner, Tsipras, and Madry 2019), and TUAP (Zhao et al. 2020), where the first four are dirty-label methods, and the last two are clean-label methods. We apply them to 5-way 5-shot tasks on the Baseline++ (Chen et al. 2019), and the results with different poison rates are depicted in Fig 1. From this figure, it can be observed that these methods are all difficult to get a successful attack result, where they fail to achieve a high Attack Success Rate (ASR) and maintain Benign Accuracy (BA) at the same time. Through our analysis, this is due to the limited training samples, resulting in the model either overfitting to poisoned features or to clean features. Besides being ineffective, we also observe that these methods have poor stealthiness in FSL. In traditional classification, it could be difficult for the victims to detect the label modification with abundant training data and a low poisoning rate. However, in 5-way 5-shot FSL tasks, the poisoning rate is as high as 0.2 even for only one dirty-label poisoned image in each class, while the poisoning rate in traditional classification is usually less than 0.1. Consequently, these wrong labels can easily be detected and corrected by the victims. Although a clean-label approach can avoid this issue, the presence of a visible trigger similarly attracts the victim's attention. We take BadNet and TUAP as examples to show their poor stealthiness at the top of Fig 1.

The above experiment seems to indicate that FSL is immune to backdoor attacks. However, in this work, we propose a backdoor attack method, dubbed **Few-shot Learning Backdoor Attack (FLBA)**, to overcome the limitations of existing methods. The proposed method consists of four steps, and the framework is illustrated in Figure 2. Firstly, we optimize a trigger with embedding deviation to increase the distance between the clean and trigger features, which enables the model to learn the backdoor boundary without overfitting one of them. Next, we try to hide the trigger by generating imperceptible perturbation for the clean images. Specifically, we design two types of perturbations with max-min similarity loss for both targeted samples and untargeted samples. The attractive perturbation attracts the feature towards the trigger feature for the target class, while the repulsive perturbation moves the feature away from the trigger feature for the untargeted class. We incorporate a regularizer to retain the original features to ensure the backdoor attack of the infected model. In this way, we can ensure stealthiness and obtain a larger ASR by relaxing the poisoning rate. In the third step, we poison the entire benign support set with perturbations and obtain the hidden poisoned support set for model training or fine-tuning and building the backdoor in the model. In the testing step, when the query images with our trigger are tested on the infected model, the model will output the target label, while simultaneously maintaining high accuracy on benign samples.

In summary, our contributions are:

- We explore the existing backdoor attack methods under FSL scenarios, and they suffer from two major shortcomings: overfitting the benign or poisoned features and poor stealthiness for easy-to-detect.
- We propose a backdoor attack for few-shot learning to address the aforementioned challenges, where generate a trigger with embedding deviation to mitigate the overfitting and introduce a max-min similarity loss to generate perturbations for enhancing stealthiness.
- Extensive experiments are conducted to verify the effectiveness of our method across different few-shot learning methods and tasks. Our work demonstrates that few-shot learning remains vulnerable to backdoor attacks.

Related Work

Few-shot Learning

In a FSL task, there are usually two sets of data, including a target labeled support set $\mathcal{S}$, an unlabeled target query set $\mathcal{Q}$, where $\mathcal{S}$ contains different classes C, with K images per class. In particular, $\mathcal{S}$ and $\mathcal{Q}$ share the same label space, which corresponds to the training and test sets in classical classification. However, the difference from the common classification is that the number of images for each class K is small (e.g., 1 or 5). This classification task is called a C-way K-shot task. To tackle this problem, an additional auxiliary set $\mathcal{A}$ is usually adopted to learn transferable knowledge to boost the learning on the target task ($\mathcal{S}$ and $\mathcal{Q}$).

The current FSL mainly focuses on three ways (Li et al. 2021): (a) Fine-tuning based methods, (b) Meta-learning based methods, and (c) Metric-learning based methods. For the fine-tuning based methods (Chen et al. 2019; Liu et al. 2020a; Rajasegaran et al. 2020; Dhillon et al. 2019; Tian et al. 2020; Yang, Liu, and Xu 2021), it follows the standard transfer learning procedure (Weiss, Khoshgoftaar, and Wang 2016), which is pre-training with the base classes at first and then fine-tuning with the novel class. The most typical method for it is Baseline (Chen et al. 2019) and Basline++ (Chen et al. 2019), which adopts a linear layer in the fine-tuning stage. Both meta-learning based and metric-learning based methods adopt simulation few-shot tasks in the training process. The auxiliary set $\mathcal{A}$ are divided into $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{Q}}$ for each simulated task (episode). Tens of thousands of simulated tasks are randomly sampled from a distribution to train the model. The meta-learning based methods adopt a meta-training paradigm, which aims to make the model can fast adapt to novel class (Vinyals et al. 2016; Gordon et al. 2018; Lee et al. 2019; Bertinetto et al. 2018; Rusu et al. 2019; Raghu et al. 2020; Xu et al. 2020). The metric-learning based methods directly compare the similarities of latent representations between query set and support set without fine-tuning on the support set, and use their relationship outputs to classify (Snell, Swersky, and Zemel 2017; Koch et al. 2015; Sung et al. 2018; Li et al. 2019a; Doersch, Gupta, and Zisserman 2020; Zhang et al. 2020; Wertheimer, Tang, and Hariharan 2021; Kang et al. 2021). The typical methods for them are MAML (Finn, Abbeel, and Levine 2017) and Prototypical (Snell, Swersky, and Zemel 2017). All the above three ways will be considered in this paper, but in the problem statement, we will mainly focus on fine-tuned-based methods as an example.

Backdoor Attacks

The backdoor attack is an emerging threat to model security. According to whether the poisoned samples have consistent features and labels, the existing backdoor attacks can be generally categorized into dirty-label attacks and clean-label attacks. Dirty-label attacks (Gu et al. 2019; Chen et al. 2017; Nguyen and Tran 2021) usually select a set of clean examples from the untarget class, attach the backdoor trigger, and reset their labels to be the target class. The poisoned inputs look like to be from the untarget class, but their labels are the target class, thus input-label pairs look mislabeled to a human. Clean-label attacks can be more challenging than dirty-label attacks as the trigger pattern is no longer strongly associated with the target class. The representative work is Label-Consistent (Turner, Tsipras, and Madry 2019). It involves selecting data from the target class, manipulating the data harder to learn, and inserting a trigger into the data. However, it requires a large poison ratio so that the association between the trigger and the target label can be memorized by the model. Another recent clean-label work is TUAP (Zhao et al. 2020), which adopts a universal adversarial trigger and adversarial perturbations for effective embedding backdoor attacks with videos. Recently, Saha et. al (Saha, Subramanya, and Pirsiavash 2020) propose a Hidden Trigger Backdoor Attack (HTBA) by adding perturbation which aims to hide the trigger in the poisoned training set and keep the trigger secret until the test time. Although this kind of attack has a comparable goal to ours, it can not be employed directly to enhance stealthiness in few-shot learning scenarios.

As far as we know, most current backdoor works focus on common classification problems, and few work has explored the existence of backdoor attacks in FSL scenarios. Although some work (Li et al. 2022a; Guan et al. 2022) have the keyword 'few-shot' in the title, they are not discussing the backdoor attacks in few-shot learning, where 'few-shot' is just an adjective for backdoors. The backdoor attack proposed by Chen et al. (Chen, Golubchik, and Paolieri 2020) first conducts experiments on few-shot datasets, but it is mainly based on the federated meta-learning scenario and does not consider the assumptions of few-shot tasks.

Problem Statement

Preliminaries

Threat Model As previously mentioned, FSL is usually divided into two parts, the first is to train on the auxiliary set to obtain a feature embedding module, and then the victim builds on this module to retrain or fine-tune a classifier with his own data. We assume that the attacker cannot poison the auxiliary set, or in other words, cannot build a backdoor in the feature embedding module. This is because auxiliary sets are usually publicly available and clean, and some authorities also publish secure pre-trained models for users to access. However, the backdoor attacks in FSL are more likely to occur during the fine-tuning phase. We give two possible scenarios for this backdoor attack: First, the trainable samples in the support set may be poisoned by attackers. As we know, these labeled samples are often hard to collect or acquire because of privacy or copyrights. Therefore, when a practitioner is unable to collect enough trusted trainable data by themselves, they will compromise to obtain risky images from untrustworthy sources, even if they know that the data might not be secure enough. Apart from this scenario, some data owners could detect copyrights by checking the backdoor if a data-stealer fine-tunes the model on the protected data. For convenience, we refer to those who create backdoors as attackers and those who suffer from backdoor attacks as victims.

Since there are many related works about FSL and the paradigm varies a lot from each other, we assume that the attacker has knowledge of the specific FSL paradigm, and the victim will use and access to the pre-trained model that the victim will train. For a brief description, we employ a classic but widely used fine-tuning-based method Baseline++ as an example to introduce our backdoor attack.

Shortages of the Baseline Method

In this section, we will reveal the two shortages of the existing backdoor attack methods in FSL scenarios:

Overfitting the Benign or Poisoned Features. As shown in the curves of Fig 1, the existing backdoor attacks would cause an obvious trade-off, where ASR and BA cannot achieve good performance at the same time. To further investigate this phenomenon, we visualize the t-SNE of some clean images and their poisoned images with triggers in the feature space, as shown in Fig 3. Our visualization shows that the poisoned samples tend to locate close to clean samples in feature space, and the BadNet samples (represented by the orange circles) are the closest to the benign samples (represented by the red triangles). Moreover, the few-shot task only contains a small number of training samples, and the poisonable samples become even fewer when the clean and poisoned samples are divided. Therefore, when facing a similar feature distribution of poisoned and clean samples in pre-trained embedding space, the model will be confused in distinguishing them with limited data. As a result, when the poisoning rate is very low, the infected model will fail to learn the backdoor features and only overfit the benign features, resulting in an ineffective attack. As the poisoning rate increases, the infected model begins to overfit the backdoor feature while failing to learn clean features, resulting in lower performance in BA. Furthermore, clean-label methods poison samples are only sourced from samples from the target class, thereby limiting the number of poisonable samples even further. As a result, the model tends to overfit clean features and struggles to learn backdoors.

Easy-to-detect Backdoor Attacks. Stealthiness is a critical concern in backdoor attacks, especially in few-shot learning scenarios. Previous backdoor attacks of dirty-label have focused on reducing the poisoning rate to ensure the stealth of the backdoor (Gu et al. 2019). However, in FSL, where the support set is small, even a small poisoning rate can be easily detected by victims, as illustrated in the top left of Fig 1. Clean-label backdoor attacks maintain consistency between images and labels, but they encounter another problem: the trigger attached to the poisoned images is still visible to the victim. As depicted in the visualization of TUAP in Fig 1, the image of the target class shows an obvious trigger attached. Upon detecting some special patterns repeatedly in the support set, victims may suspect that the data has been maliciously poisoned. Consequently, they may attempt to repair the poisoned images or defend against query images with the same trigger pattern during the testing phase, thereby mitigating the backdoor attack. Hence, a good backdoor attack for FSL should be carefully designed to remain hidden and undetectable by the victim.

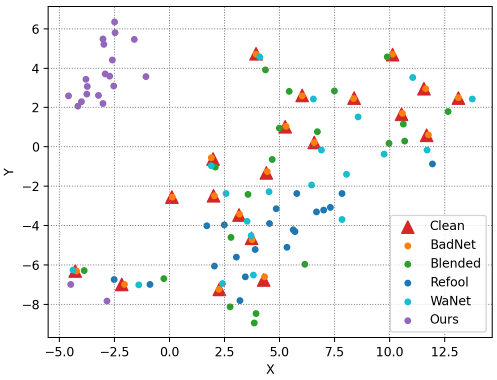


Figure 3: The t-SNE visualization of benign images and different poisoned versions in the feature spaces with four dirty-label backdoor attack methods and ours, where the red triangles represent the distribution of clean samples.

Few-shot Learning Backdoor Attack (FLBA)

Based on these shortages, we propose Few-shot Learning Backdoor Attack (FLBA), which can solve the corresponding challenges. It consists of four steps, and the pipeline of our backdoor attack is shown in Fig. 2.

Trigger Generation with Embedding Deviation. Through the aforementioned observation, it can be deduced that the features of the existing triggers are entangled with benign features in the pre-trained embedding space, which causes the model to either overfit the benign samples or the poisoned samples with limited data. To achieve an effective backdoor attack for FSL, an ideal trigger should enable the model to learn both the features of the benign samples and the trigger features simultaneously, despite the shortage of data. To this end, we propose to generate a trigger that maximizes the gap between the poisoned and benign features in the pre-trained embedding. When the poisoned features deviate sufficiently from the original clean features, the boundary becomes low-dimensional and easy to learn for the model, as purple circles are shown in Fig 3.

We implement our backdoor attack by first generating a trigger with embedding deviation. Given a clean image $\mathbf{x}$ and a mask $\mathbf{m}$ that limits the size of the pattern of trigger $\mathbf{t}$, a poisoned support set can be formed from a benign support set by them. After passing the pre-trained embedding network, the features of a benign sample $\mathbf{z_b}$ and a poisoned sample $\mathbf{z_p}$ can be expressed as:

$$\begin{aligned} \mathbf{z_b} &= f_\theta\left(\mathbf{x}\right) \\ \mathbf{z_p} &= f_\theta\left(\mathbf{x} \odot (\mathbf{1}-\mathbf{m}) + \mathbf{m} \odot \mathbf{t}\right), \end{aligned} \tag{1}$$

where $f_\theta(\cdot)$ is pre-trained feature embedding by auxiliary set, and $\odot$ indicates the element-wise product. Our goal is to optimize a trigger $\mathbf{t}^*$ such that the features of any images with this trigger are different from their original benign features. Therefore, the problem can be formulated as follows,

$$\mathbf{t}^* = \arg\max_{\mathbf{t}} \sum_{\mathbf{x}\in\mathcal{S}} d\left(\mathbf{z}_b, \mathbf{z}_p\right), \tag{2}$$

where $\mathcal{S}$ could be the support set for training or another randomly selected image set. $d(:,:)$ is a distance metric for two features, and the cosine distance is employed in this work.

Hiding Trigger with Max-min Distance Loss. After addressing the first challenge, we turn our attention to the issue of poor stealthiness. The existing backdoor attack for both dirty-label and clean-label can not guarantee stealthiness in FSL. Motivated by HTBA(Saha, Subramanya, and Pirsiavash 2020), we consider generating imperceptible perturbations to poison the support set instead of attaching the trigger pattern directly. However, due to the predefined trigger patterns used and the large number of poisoned samples required from untargeted classes in HTBA, it could not implement an effective attack in FSL. Differently, we poison all images from both targeted and untargeted classes and introduce a max-min distance loss to optimize two types of perturbations: **Attractive Perturbation** for the images of the target class, which brings their features closer to the optimized trigger features, and **Repulsive Perturbation** for untargeted class, which moves their features further away from the optimized trigger features. The resulting support set, with these perturbations, is referred to as the hidden poisoned set. Our approach associates the optimized trigger features with the target label while remaining the trigger invisible during training phase. Moreover, to ensure the BA, we also introduce a regularizer to preserve the original features.

Practically, we attach the optimized trigger $\mathbf{t}^*$ to clean images to obtain the poisoned support set. As above mentioned, we split the support set into a targeted class and an untargeted class and optimize a set of imperceptible perturbations separately. For images of the target class, we obtain the feature of poisoned set $\mathbf{t_p}$ and the feature of hidden poisoned set $\mathbf{t_h}$ from the pre-trained embedding as follows:

$$\begin{aligned} \mathbf{t_h} &= f_\theta\left(\mathbf{x_t} + \delta_a\right) \\ \mathbf{t_p} &= f_\theta\left(\mathbf{x_t} \odot (\mathbf{1}-\mathbf{m}) + \mathbf{m} \odot \mathbf{t}^*\right), \end{aligned} \tag{3}$$

and we optimize the attractive perturbation δ_a by minimizing the distance between the features of the poisoned and hidden poisoned set. Therefore, our optimization process can be defined as,

$$\begin{aligned} \min\nolimits_\delta \quad & d\left(\mathbf{t_h}, \mathbf{t_p}\right) + \lambda_1 d\left(\mathbf{t_h}, f_\theta\left(\mathbf{x_t}\right)\right), \\ \text{s.t.} \quad & \|\delta_a\|_\infty \leqslant \varepsilon. \end{aligned} \tag{4}$$

where λ_1 is balance pararmeter of regularizer and attractive perturbation δ_a is restricted in L_∞ norm bound ϵ. Moreover, we use projected gradient descent (PGD) (Madry et al. 2018) to solve this problem. Similarly, for the untargeted class, the poisoned features $\mathbf{u_p}$ and hidden poisoned features $\mathbf{u_h}$ can be also obtained by the following equation:

$$\begin{aligned} \mathbf{u_h} &= f_\theta\left(\mathbf{x_u} + \delta_r\right) \\ \mathbf{u_p} &= f_\theta\left(\mathbf{x_u} \odot (\mathbf{1}-\mathbf{m}) + \mathbf{m} \odot \mathbf{t}^*\right). \end{aligned} \tag{5}$$

Different from the problem in (4), we generate repulsive perturbations to move features away from the distribution of the trigger feature. Therefore, we can optimize the repulsive perturbation by solving the following optimization problem:

$$\begin{aligned} \max\nolimits_\delta \quad & d\left(\mathbf{u_h}, \mathbf{u_p}\right) - \lambda_2 d\left(\mathbf{u_h}, f_\theta\left(\mathbf{x_u}\right)\right). \\ \text{s.t.} \quad & \|\delta_r\|_\infty \leqslant \varepsilon. \end{aligned} \tag{6}$$

Here, λ_2 controls the closeness to its original feature, and repulsive perturbation $\delta_\mathbf{r}$ is restricted in L_∞ norm bound ϵ.

Building Backdoor on Support Set $\&$ Testing. Following the acquisition of the attractive and repulsive perturbations, the subsequent phase involves covert poisoning by introducing these perturbations into the support set. Due to the imperceptibility of the perturbations, the poisoning rate can be relaxed by poisoning the entire benign support set D_c, leading to a higher ASR. The hidden poisoned support set D_h can be mathematically formulated as follows:

$$\begin{aligned} D_h = \{&(\mathbf{x_t}^1 + \delta^1, \mathbf{y}^1), (\mathbf{x_t}^2 + \delta^2, \mathbf{y}^2), \cdots, \\ &(\mathbf{x_u}^{m+1} + \delta^{m+1}, \mathbf{y}^{m+1}), (\mathbf{x_u}^{m+2} + \delta^{m+2}, \mathbf{y}^{m+2}), \cdots\}, \end{aligned} \tag{7}$$

where the set of attractive perturbations $(\delta^1, \delta^2, ..., \delta^m)$ is for targeted class, and $(\delta^{m+1}, \delta^{m+2}, ..., \delta^{m+n})$ is repulsive perturbations for untargeted class. The model is then fine-tuned on the hidden poisoned support set D_h to establish the backdoor and obtain the infected model.

In the final step, we evaluate our FLBA on the infected model. When testing the BA, we input the clean image of the query set and expect a correct output from the model. For testing ASR, the optimized trigger t^* with mask m is added to the benign image. After inputting it into the infected model, we expect to get a target label output.

Experiment

Experimental Setup

Datasets and Model Architectures. Following the literature (Li et al. 2022a), our few-shot learning experiments are mainly conducted on *mini*ImageNet (Vinyals et al. 2016). We split them into an auxiliary, support, and query set, respectively, and all images are sized as a resolution of 84×84. In this paper, we always poison the support set rather than the auxiliary set for FSL. In addition, we choose three typical models: Baseline++ (Chen et al. 2019), MAML (Finn,

Method	Stealthiness		Baseline++		MAML		ProtoNet	
	Clean Label	Invisible Trigger	ASR	BA	ASR	BA	ASR	BA
Clean	/	/	19.2	71.4	18.1	65.2	17.3	69.9
BadNet	×	×	36.2	64.4	50.4	53.6	20.6	49.9
Blended	×	√	52.3	62.9	65.6	54.3	21.1	50.5
Refool	×	√	69.8	65.3	50.4	54.1	17.1	51.1
WaNet	×	√	59.7	46.2	44.0	54.0	20.7	26.3
Label-Consistent	√	×	14.5	55.1	26.4	62.3	/	/
TUAP	√	×	34.8	64.9	77.1	58.6	/	/
HTBA	√	√	26.8	59.3	21.7	56.5	/	/
FLBA (Ours)	√	√	**89.1**	**65.5**	**81.2**	**63.6**	**60.1**	**61.6**

Table 1: Comparison(%) of different backdoor attack methods on *mini*ImageNet. Stealthiness includes two aspects: clean label and invisible trigger. In each case, the best attacking ASR and BA are boldfaced.

Abbeel, and Levine 2017), and ProtoNet (Snell, Swersky, and Zemel 2017). Among them, the fine-tuning-based methods are the most widely used. Therefore, we always take Baseline++ for examples in discussion parts. Moreover, we all adopt *ResNet12* as their embedding backbones. More experimental results on other models and datasets are presented in the supplementary material.

Evaluation Metrics. In our work, we evaluate the performance with a set of 5-shot 5-way tasks. Same as the backdoor attacks on traditional classification, here we also use the attack success rate (ASR) and benign accuracy (BA) to evaluate the effectiveness of attacks. Specifically, ASR is introduced to evaluate that the images with a specific trigger are classified as the targeted class, and BA is the accuracy of testing on benign examples. Moreover, we evaluate random 600 episodes of the query set, repeat the total of five times at the test stage, and report the average ASR and BA.

Training and Attack Setup. We refer to the LibFewShot open-resource code [1] to build the pre-trained embedding models and follow their settings of parameters. In the training stage, we always record the model which obtains the best performance on the validation set and evaluate it on the testing set in the test stage. In addition, we assume that the attacker has access to the same pre-trained model as the victims. For the attacks, we adopt a 16 × 16 mask as a mask of the trigger pattern. In the trigger generation phase, we set the step size as 2, and iterations as 100. For the attractive and repulsive perturbation generation, we set the step size as 2, iteration as 80, and we empirically set the L_∞ norm bound ϵ as 8/255, which is imperceptible by human eyes. The balance parameters λ_1 and λ_2 are initially set as 1.5, 1.5.

Main Results

As shown in Table 1, we test the effectiveness of existing backdoor attack methods and ours with different few-shot learning methods, specifically including four dirty-label methods, two clean-label methods, and a hidden trigger method. The results demonstrate that none of the previous methods achieve comparable results on the traditional task. In contrast, our proposed methods exhibit promising attack results on different few-shot learning paradigms while maintaining good performance on benign samples. For instance, both Baseline++ and MAML methods achieve an ASR greater than 80%, while other backdoor attacks only achieve an average ASR of 50%. Our method on ProtoNet also exceeds 60% ASR, while others exhibit only a 20% success rate. This finding could be attributed to the fact that ProtoNet is a metric-learning-based method that does not train a network on the support set. Instead, it directly compares the cosine similarity or distance between the support set and the query set for prediction. Furthermore, clean-label backdoor attack methods, including Label-Consistent and TUAP, involve an attack on the output of the classifier to generate adversarial perturbations. However, metric-learning-based methods cannot apply these backdoor attacks directly due to the lack of a classifier. Additionally, Label-Consistent, TUAP, and HTBA involve an attack on the output of the classifier to generate adversarial perturbations. Due to the lack of a classifier, metric-learning-based methods cannot apply directly to these backdoor attacks. Therefore, the existing backdoor attack is hard to apply in different FSL paradigms, but ours can still successfully construct the backdoor.

Fig. 4 shows the stealthiness of different backdoor attacks. The dirty-label method, which involves modifying the labels in the support set, is easily detectable as the images of wolves are incorrectly labeled as bowls. Clean-label methods seem to solve this problem, but some abnormalities still can be observed by victims in FSL. Specifically, in Label-Consistent and TUAP images, the bowls are labeled correctly but visible trigger patterns are attached to them. Due to the lack of extensive samples in the support set, victims can be more likely to notice these patterns recurring on some samples. HTBA has good stealthiness for its hidden triggers and clean label, but it fails to build a backdoor successfully in FSL. However, our method can achieve both clean-label poisoning and invisibility of trigger patterns at the same time, and successfully build a backdoor attack more stealthy in few-shot learning scenarios.

Discussion

Different Few-shot Tasks Through the exploration of our backdoor attack in various few-shot tasks, we have obtained experimental results of ASR and BA for 1-shot to 30-shot

[1] https://github.com/RL-VIG/LibFewShot

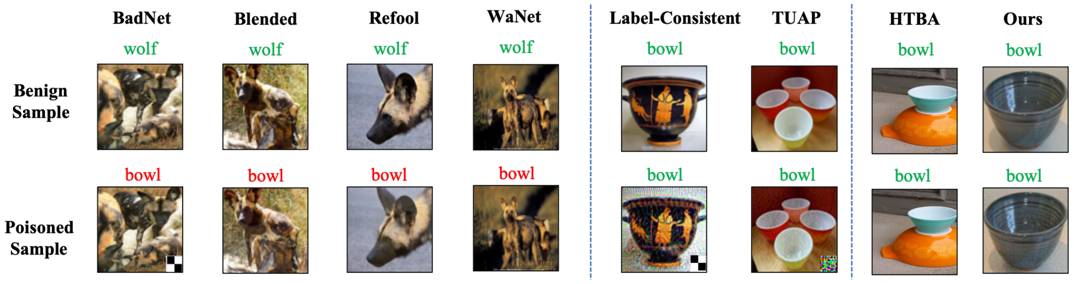


Figure 4: The visualization of poisoned support set for different backdoor attack methods. In dirty-label methods, the labels of poisoned samples are inconsistent with their ground-truth ones. Although clean-label methods keep the same labels as their ground-truth ones, the trigger patterns are visible in the support set. Our method and HTBA has good stealthiness.

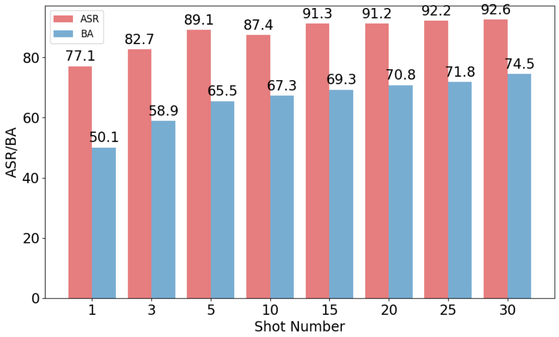
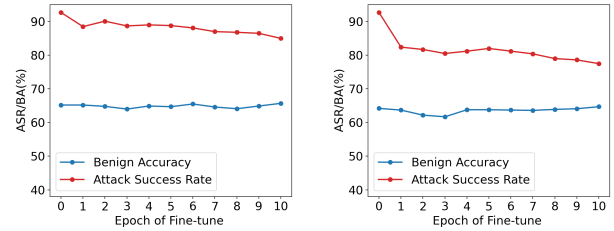


Figure 5: Attack results on a different number of shot tasks.

tasks, which are illustrated in Fig 5. Remarkably, our method successfully implements a backdoor even in 1-shot learning tasks, achieving an ASR of 77.1%, whereas other methods fail due to the support set's manipulation limited to only one image. For instance, if one image from a class is selected for label modification and trigger attachment, BadNet is unable to learn the poisoned and benign samples simultaneously. Although such an occurrence is unlikely in real-world scenarios, it establishes a lower bound on the effectiveness of our approach. Furthermore, our attack performance improves with increasing shot numbers, eventually plateauing when more than 15 shots are being used. As the shot number exceeds a certain level, the backdoor attack of few-shot learning becomes that of traditional supervised learning.

Resistance to Fine-tuning. Fine-tuning is a popular reconstruction-based backdoor defense method (Liu, Xie, and Srivastava 2017; Liu, Dolan-Gavitt, and Garg 2018). In traditional classification, models are typically fine-tuned on a small portion of the original dataset or a new dataset. Similarly, in FSL, we introduce two support sets: the first is the original but clean support set used to train, while the other is a new support set that is sourced from another dataset. Figure 6 shows the results of our experiments. After ten epochs of fine-tuning, the ASR of our backdoor attack remains above 80% when fine-tuned on the original support set, with only a marginal reduction in the BA. On the other hand, when fine-tuning with the new support set, although the ASR drops much at the first epoch, it still attains 77.5% ASR after ten epochs of fine-tuning, with the BA remaining within a small range. Thus, our backdoor attack can resist fine-tuning with the original and new support sets.



(a) Original Support Set (b) New Support set

Figure 6: Resistance to fine-tuning on benign samples of original support set and new support set.

Conclusion

In this paper, we have explored the potential threat of backdoor attacks in few-shot learning (FSL). Some results demonstrated that embedding a backdoor in FSL is more challenging than in traditional classification tasks. Given these challenges, we propose a novel backdoor attack method called the Few-shot Learning Backdoor Attack. This attack involves generating a trigger with an embedding deviation to mitigate overfitting. We also introduce optimizing a max-min similarity loss to create attractive and repulsive perturbations that hide the trigger. Our experimental results show that the proposed method achieves high ASR and preserves the performance of benign samples across different few-shot learning methods. Furthermore, our method can be used to build backdoors on 1-shot learning tasks, which is difficult or almost impossible with other methods, and it can be also resilient to a set of defense strategies. There are still some limitations to our approach, our method is based on a white-box setting, and how to efficiently build a backdoor without model knowledge leaves further explorations. Overall, our proposed backdoor attack provides a valuable tool for assessing the backdoor vulnerability of FSL systems.

Acknowledgments

Supported in part by the National Key R&D Program of China (Grant No. 2022ZD0118100), in part by National Natural Science Foundation of China (No.62025604). in part by Beijing Natural Science Foundation (No.L212004).

References

Bagdasaryan, E.; and Shmatikov, V. 2021. Blind backdoors in deep learning models. In *USENIX Security*.

Bertinetto, L.; Henriques, J. F.; Torr, P. H.; and Vedaldi, A. 2018. Meta-learning with differentiable closed-form solvers. *ICLR*.

Chen, C.-L.; Golubchik, L.; and Paolieri, M. 2020. Backdoor attacks on federated meta-learning. *NeurIPS Workshop*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A closer look at few-shot classification. *ICLR*.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Dhillon, G. S.; Chaudhari, P.; Ravichandran, A.; and Soatto, S. 2019. A baseline for few-shot image classification. *ICLR*.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. Crosstransformers: spatially-aware few-shot transfer. *NeurIPS*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Gao, K.; Bai, Y.; Gu, J.; Yang, Y.; and Xia, S.-T. 2023. Backdoor Defense via Adaptively Splitting Poisoned Dataset. In *CVPR*.

Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; and Turner, R. E. 2018. Versa: Versatile and efficient few-shot learning. In *NeurIPS*.

Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. 2018. Recent advances in convolutional neural networks. *Pattern Recognition*.

Gu, J.; Wei, F.; Torr, P.; and Hu, H. 2023. Exploring Non-additive Randomness on ViT against Query-Based Black-Box Attacks. *BMVC*.

Gu, J.; Zhao, H.; Tresp, V.; and Torr, P. H. 2022. SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness. In *ECCV*.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*.

Guan, J.; Tu, Z.; He, R.; and Tao, D. 2022. Few-shot backdoor defense using shapley estimation. In *CVPR*.

Gudibande, A.; Chen, X.; Bai, Y.; Xiong, J.; and Song, D. ???? Test-time Adaptation of Residual Blocks against Poisoning and Backdoor Attacks.

He, B.; Liu, J.; Li, Y.; Liang, S.; Li, J.; Jia, X.; and Cao, X. 2023. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition.

Huang, Y.; Guo, Q.; Juefei-Xu, F.; Ma, L.; Miao, W.; Liu, Y.; and Pu, G. 2021. AdvFilter: predictive perturbation-aware filtering against adversarial attack via multi-domain learning. In *ACM MM*, 395–403.

Huang, Y.; Sun, L.; Guo, Q.; Juefei-Xu, F.; Zhu, J.; Feng, J.; Liu, Y.; and Pu, G. 2023. ALA: Naturalness-aware Adversarial Lightness Attack. In *ACM MM*, 2418–2426.

Jia, X.; Wei, X.; Cao, X.; and Han, X. 2020. Adv-watermark: A novel watermark perturbation for adversarial examples. In *ACM MM*, 1579–1587.

Kang, D.; Kwon, H.; Min, J.; and Cho, M. 2021. Relational Embedding for Few-Shot Classification. In *ICCV*.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *CVPR*.

Li, W.; Dong, C.; Tian, P.; Qin, T.; Yang, X.; Wang, Z.; Huo, J.; Shi, Y.; Wang, L.; Gao, Y.; et al. 2021. LibFewShot: A Comprehensive Library for Few-shot Learning. *arXiv preprint arXiv:2109.04898*.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019a. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*.

Li, W.; Wang, L.; Zhang, X.; Huo, J.; Gao, Y.; and Luo, J. 2019b. Defensive Few-shot Adversarial Learning. *arXiv preprint arXiv:1911.06968*.

Li, Y.; Ya, M.; Bai, Y.; Jiang, Y.; and Xia, S.-T. 2023. BackdoorBox: A python toolbox for backdoor learning. *arXiv preprint arXiv:2302.01762*.

Li, Y.; Zhong, H.; Ma, X.; Jiang, Y.; and Xia, S.-T. 2022a. Few-shot backdoor attacks on visual object tracking. *ICLR*.

Li, Y.; Zhu, L.; Jia, X.; Jiang, Y.; Xia, S.-T.; and Cao, X. 2022b. Defending against model stealing via verifying embedded external features. In *AAAI*.

Liang, S.; Li, L.; Fan, Y.; Jia, X.; Li, J.; Wu, B.; and Cao, X. 2022. A large-scale multiple-objective method for black-box attack against object detection. In *ECCV*.

Liu, B.; Cao, Y.; Lin, Y.; Li, Q.; Zhang, Z.; Long, M.; and Hu, H. 2020a. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proc. of RAID*.

Liu, X.; Liu, J.; Bai, Y.; Gu, J.; Chen, T.; Jia, X.; and Cao, X. 2022. Watermark Vaccine: Adversarial Attacks to Prevent Watermark Removal. In *ECCV*.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020b. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*.

Liu, Y.; Xie, Y.; and Srivastava, A. 2017. Neural trojans. In *ICCD*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR Poster*.

Nguyen, A.; and Tran, A. 2021. WaNet–Imperceptible Warping-based Backdoor Attack. *ICLR*.

Nguyen, T. A.; and Tran, A. 2020. Input-aware dynamic backdoor attack. *NeurIPS*.

Raghu, A.; Raghu, M.; Bengio, S.; and Vinyals, O. 2020. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *ICLR*.

Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F. S.; and Shah, M. 2020. Self-supervised knowledge distillation for few-shot learning. *arXiv preprint arXiv:2006.09785*.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2019. Meta-learning with latent embedding optimization. *ICLR*.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden trigger backdoor attacks. In *AAAI*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *NeurIPS*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *NeuraIPS*.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A survey of transfer learning. *Journal of Big data*.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-shot classification with feature map reconstruction networks. In *CVPR*.

Xu, W.; Wang, H.; Tu, Z.; et al. 2020. Attentional constellation nets for few-shot learning. In *ICLR*.

Yang, S.; Liu, L.; and Xu, M. 2021. Free lunch for few-shot learning: Distribution calibration. *ICLR*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*.

Zhao, S.; Ma, X.; Zheng, X.; Bailey, J.; Chen, J.; and Jiang, Y.-G. 2020. Clean-label backdoor attacks on video recognition models. In *CVPR*.