计算机科学
COMPUTER SCIENCE

基于触发差异优化的联邦学习持久后门攻击

蒋雨霏, 田育龙, 赵彦超

引用本文

蒋雨霏, 田育龙, 赵彦超. 基于触发差异优化的联邦学习持久后门攻击[J]. 计算机科学, 2025, 52(4): 343-351.

JIANG Yufei, TIAN Yulong, ZHAO Yanchao. Persistent Backdoor Attack for Federated Learning Based on Trigger Differential Optimization [J]. Computer Science, 2025, 52(4): 343-351.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800043

基于触发差异优化的联邦学习持久后门攻击

蒋雨霏　田育龙　赵彦超
南京航空航天大学计算机科学与技术学院　南京 211106
(2216075@nuaa.edu.cn)

摘　要　联邦学习分布式的特性使其允许各客户端在保持数据独立性的同时进行模型训练,但这也使得攻击者可以控制或模仿部分客户端来发起后门攻击,通过植入精心设计的固定触发器操纵模型输出。触发器的有效性和持久性是衡量攻击效果的重要标准,有效性即攻击成功率,持久性即停止攻击后维持高攻击成功率的能力。目前针对有效性的研究已经相对深入,但如何维持触发器的持久性仍然是一个有挑战性的问题。为延长触发器的持久性,提出了一种基于动态优化触发器的后门攻击方法。首先,在联邦学习动态更新时同步优化触发器,将触发器特征在攻击时模型与攻击后模型的潜在表示的差异最小化,以此训练全局模型对触发器特征的记忆能力。其次,使用冗余神经元作为植入后门是否成功的指标,通过自适应添加噪声来增强攻击的有效性。在MNIST,CIFAR-10和CIFAR-100数据集上进行广泛实验,结果表明,所提方案有效延长了联邦学习环境下触发器的持久性。在具有代表性的5种防御体系下攻击成功率高于98%,特别是在针对CIFAR-10数据集的攻击停止超过600轮后,攻击成功率仍然高于90%。

关键词:联邦学习;后门攻击;动态触发器;攻击持久性;模型安全
中图分类号　TP309

Persistent Backdoor Attack for Federated Learning Based on Trigger Differential Optimization

JIANG Yufei, TIAN Yulong and ZHAO Yanchao
School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract　The distributed nature of federated learning allows each client to train the model while maintaining data independence, but this also allows attackers to control or mimic some clients to launch backdoor attacks by implanting carefully designed fixed triggers to manipulate the model output. The effectiveness and persistence of triggers are important criteria for measuring attack effectiveness. Effectiveness pertains to the rate of successful breaches, while persistence embodies the capability to sustain a high success rate even after the cessation of the attack. At present, research on effectiveness has been relatively in-depth, but maintaining the persistence of triggers remains a challenging issue. A backdoor attack method based on dynamic optimization triggers is proposed to extend the persistence of triggers. Firstly, during dynamic updates in federated learning, triggers are synchronously optimized to minimize the difference between the potential representations of trigger features during and after attacks, thereby training the global model's ability to remember trigger features. Secondly, using redundant neurons as an indicator of the success of implanting backdoors to adaptively add noise and enhance the effectiveness of attacks. Extensive experiments on the MNIST, CIFAR-10, and CIFAR-100 datasets have shown that the proposed scheme effectively extends the persistence of triggers in federated learning environments. Under five kind of representative defense systems, the success rate of attacks is higher than 98%, especially after more than 600 rounds of attacks on the CIFAR-10, the success rate of attacks still exceeds 90%.

Keywords　Federated learning, Backdoor attack, Dynamic trigger, Attack persistence, Model security

1　引言

近年来分布式人工智能和机器学习快速发展,海量数据分析[1-2]、复杂模式识别[3]等领域的疑难问题被逐步击破,其中联邦学习(Federated Learning,FL)框架[4]更是为下一代数据驱动的智能应用铺平了道路。联邦学习可以在每个客户端不共享数据[5]的情况下联合训练模型,一定程度上保护了每个客户端私有数据的隐私安全。在传统的联邦学习框架中,服务器首先向所有客户端发送一个全局模型,客户端使用本地私有数据训练本地模型,并将训练好的模型上传至中央服务器。在服务器随机选择一定数量的客户端接收它们的模型更新,遵循一定规则对上传的模型参数进行聚合来更新全局

到稿日期:2024-08-06　返修日期:2024-09-26
基金项目:国家自然科学基金(62172215);国家自然科学基金A3国际项目(62061146002)
This work was supported by the National Natural Science Foundation of China(62172215) and A3 Foresight Program of NSFC(62061146002).
通信作者:赵彦超(yczhao@nuaa.edu.cn)

模型。因此,联邦学习成为许多应用程序解决数据孤岛的一个有前途的解决方案。随着 TFF[6],FATE[7] 等多种联邦学习应用框架的研发,联邦学习受到国内外各大学者的广泛关注。

然而,在传统的联邦学习中,虽然参与训练的数据和本地模型训练都由每个客户端私有控制,但模型参数共享或梯度聚合的通信过程却是必不可少的。服务器端对本地训练过程的不可见导致其无法验证上传模型参数的善恶。在联邦学习浩如烟海的参与者中,可能存在恶意用户利用这一缺陷对服务器端发起攻击,特别是在通信过程中上传恶意模型来破坏或影响全局模型的准确性,例如对抗样本攻击[8]和投毒攻击[9]。区别于集中式场景中攻击者专注于破坏单一模型,联邦场景中的攻击者会利用模型聚合与分发的过程逐渐影响联邦环境中的所有参与模型,危害更大性。

由于联邦学习具有数据私有性,因此,攻击者通常以模型投毒的方式对全局模型发起攻击。根据攻击目标,模型投毒攻击可以分为非目标攻击和目标攻击。非目标攻击的目的是破坏全局模型的准确度或使得全局模型无法收敛。然而,相比于无差别破坏全局性能的非目标攻击,目标攻击只针对指定类别进行小范围投毒,具有先天的隐蔽性。特别是,后门攻击训练含有精心设计的触发器的中毒模型,使得模型只对含有触发器的样本输出目标标签,而不破坏模型在其他良性数据集上的正常分类任务。常见的触发器包括图像像素、数据自然属性等形式。

自 Bagdasaryan 等[9]首次在联邦场景下使用缩放本地模型权重的方式成功发起后门攻击后,大量的研究[10-13]论证了在联邦学习中植入后门的可行性。但是,这些方法往往在攻击的全程使用唯一固定的触发器,却忽略了联邦学习动态更新的特征。大量的良性模型与少数恶意模型一起聚合,可能导致固定触发器的特征被良性模型的特征所覆盖,最终被消除影响从而失去攻击能力。另一个值得注意的问题是,当停止植入触发器后,由于神经网络特有的"灾难性遗忘"特性,后门攻击的效果会迅速减弱并失去作用。为了保持攻击的高精度,攻击不得不持续被发起,这既大幅提高了攻击的成本,又增加了攻击暴露的风险。近期,最新的攻击方法[14-15]表明,在攻击过程中动态调整触发器更有利于适应 FL 不断学习的过程。文献[14]提出了最大化干净样本和中毒样本潜在表示之间的差异来调整触发器,文献[15]则使用对抗损失模拟后门遗忘场景来减小全局模型和局部模型之间的差异。但这些动态调整策略只是聚焦于攻击者眼下的局部模型和全局模型的攻击效果,并没有真正讨论触发器不参与模型训练的场景。因此,如何动态调整触发器是一个值得探讨的问题。

本文提出了一种新的动态优化触发器的策略,发起兼顾持久性和有效性的后门攻击。在联邦学习的动态更新过程中,恶意客户端可以掌握的资源包括每一轮接收到的最新的全局模型与可控制的部分本地良性数据。本研究的目的在于训练全局模型对触发器特征的记忆能力,以延长攻击持久性。具体而言,在恶意客户端使用本地良性数据来模拟攻击停止场景下的良性全局模型,基于最小化中毒数据在当前全局模型(攻击时)与模拟的良性全局模型(攻击后)中潜在表示差异的思想优化触发器。此外,与良性模型相比,恶意模型由于后门任务的引入,需要目标标签的输出神经元有显著的权重/偏差倾斜。相关研究[16]指出当前的主流神经网络中最后一个全连接层的参数量在总体参数量中所占比例大,使恶意模型的输出层更易表现出异常。因此,我们使用基于冗余神经元的指示机制[10]在全连接层自适应地添加噪声,缓解输出层的显著差异。本文的主要贡献如下:

1)在不改变中毒模型训练方式的情况下,提出了一种新的动态优化触发器的策略,讨论了触发器不参与模型训练的场景,最小化触发器在攻击时与攻击后模型潜在表示的差异,在联邦学习每一轮聚合前优化触发器,从而提升后门攻击的持久性。

2)基于设计的触发器优化策略,结合指示机制在全连接层自适应添加噪声,避免了全局聚合过程中触发器特征被暴露的问题,使后门隐蔽且在攻击停止后的长时间内有效,从而实现一种黑盒中的可扩展且有效的后门攻击。

3)对于提出的攻击方案进行了全面的实验测试,并在 MNIST,CIFAR-10 与 CIFAR-100 数据集上对比最新的后门攻击方案,实验显示本文方法在典型的后门防御场景下攻击成功率高于 98%,并在攻击停止后长时间内攻击成功率高于 50%,显著延长了后门攻击持久性。

2 相关工作

2.1 联邦学习中的后门攻击

基于触发器的类型,联邦学习中的后门攻击方法可分为两类,即使用固定触发器和动态触发器的攻击。固定触发器即在攻击的全程使用相同的不变化的触发器;动态触发器即在攻击过程中使用一些优化方法改变触发器的参数或形式。典型的固定触发器后门[9]提出了缩放攻击,重复并放大训练中毒示例增强的局部训练数据计算模型更新。Li 等[10]引入了约束损失模块,以最小化恶意模型和良性模型之间的差距,集成多种规避策略保障后门的隐蔽性。Xie 等[12]提出了最早的分布式后门攻击(Distributed Backdoor Attack,DBA),使用多个局部触发器组合成全局触发器,Liu 等[17]融入组合数学的思想设计触发模式的排列顺序,改进了 DBA。区别于固定触发器,动态触发器[14-15]随着 FL 的聚合下发过程一并更新触发器,更加贴合联邦学习动态更新的本质。本文提出了一种新的动态优化触发器的策略,即考虑触发器不参与模型训练的场景。

2.2 联邦学习中的后门防御

目前很多学者提出许多针对联邦学习投毒攻击的防御策略,以确保全局模型依然拥有良好的性能,并消除后门模型的影响。基于服务器在聚合过程中对恶意客户端的处理方式,我们将防御方法分为两类。第一类防御机制中,服务器旨在使用各种异常检测的方法识别出参与聚合的客户端是否存在后门,从而完全过滤掉恶意更新。Deepsight[18]通过检查每个模型的归一化更新能量(Normalized Update Energies,NEUPs)和除法差异(Division Differences,DDifs)来执行深度模型检查,从而可以进行模型更新。RFLBAT[19]使用主成分分析(Princi-

pal Component Analysis，PCA)来减少梯度更新的维度，使恶意模型与低维投影空间中的良性模型分离。而第二类防御方法中，服务器旨在给每个参与聚合的客户端分配不同的权重以保留所有客户端更新。例如，通过裁剪过大的异常更新[20]或使用低学习率[21]惩罚高余弦相似度等行为限制攻击者的入侵。然而，这些防御方法无法抵御本文提出的攻击方式。

3 基于触发器优化的后门攻击

为了实现后门触发器的持续有效性攻击，本文利用本地数据模拟触发器真正不参与训练的场景并优化触发器，同时在模型全连接层自适应添加噪声，以增强后门特征的隐蔽有效性。本章首先阐述联邦学习后门攻击的通用模型，然后给出一种新的基于触发器优化的后门攻击方案。

3.1 后门攻击模型

3.1.1 攻击者的目标

后门攻击者旨在将触发器植入训练样本，强制全局模型在指定标签数据上错误分类，同时保证其他任务的准确性。本文使用图像像素的形式作为触发器。后门攻击者 i 的数据集表示为 $D_i=\{D_{\text{benign}},D_{\text{backdoor}}\}$，其中 D_{benign} 表示良性数据集，D_{backdoor} 表示植入触发器 ξ 的中毒数据集。本文中后门攻击者需要达成 3 个目标：1)后门任务准确性，模型对于植入触发器的测试数据 $x^*=x\oplus\xi$ 输出预设的目标中毒标签 y^*；2)主要任务准确性，模型对于未植入触发器的测试数据 x 输出正确的标签 y，即测试数据原本的标签；3)后门攻击持续性，停止植入触发器，即攻击停止后，模型对于植入触发器的测试数据 x^* 依旧输出预设中毒标签 y^* 的训练轮数总和。

3.1.2 攻击者的能力

本文假设共有 N 个客户端参与联邦学习训练过程，攻击者可以控制的恶意客户端数量为 C，并且 $|C|<|N|/2$。同时，攻击者可操纵资源仅限于中毒客户端的本地模型结构、部分良性数据集和每一轮服务器下发的全局模型。此外，由于联邦学习是动态更新的，在长期训练中持续植入触发器代价过大并且不切实际，因此，恶意客户端只在有限更新轮次中发起攻击。

3.2 攻击流程概述

图 1 给出了本文攻击方法的总体框架。本文的后门攻击在黑盒环境下进行，仅使用服务器下发的全局模型与中毒客户端的数据资源，不需要知道其余良性客户端的学习率、全局学习率等额外信息。本文攻击的主要目标是维持后门触发器的持续有效性，即保证攻击阶段规避防御机制，停止攻击后长时间保持中毒效果。虽然固定触发器可以直接简单地植入数据中，但是服务器端可以利用各种策略，例如使用基于距离等手段识别含有触发器的异常模型更新，大大降低固定触发器的效用，迫使攻击者追加攻击成本，同时由于联邦学习的动态更新特征，局部中毒模型在参与全局聚合时，植入的触发器会在多次反复迭代中被消除痕迹，从而失去攻击效用。因此，本文认为只要中毒数据在攻击停止后的模型与当前模型的潜在特征表示相似，模型遗忘触发器特征的速度就会减缓。因为全局模型仍然保留对触发器特征的记忆，所以攻击效果仍然存在，从而增强了攻击停止后中毒的持续性。与此同时，我们学习了基于冗余神经元的指示机制[10]作为攻击反馈，具体来说，对于第 t 轮的攻击者 c 而言，攻击过程可以表述为以下 5 个步骤。

步骤 1　读取攻击反馈指标。如果 $epoch>1$，则读取指示机制的反馈并调整攻击策略，否则进入步骤 2。

步骤 2　优化触发器。使用本地数据集 $\boldsymbol{D}_c$ 训练获得良性模型，模拟攻击停止后的全局模型 $\boldsymbol{G}_{stop\_t}$。最小化中毒数据在当前全局模型 $\boldsymbol{G}_{stop\_t}$ 与 $\boldsymbol{G}_t$ 中的潜在表示差异，从而优化本轮触发器 $\boldsymbol{\xi}_t$。

步骤 3　后门训练。将已优化好的触发器植入本地数据以获得中毒数据 D_{train}，使用中毒数据 D_{train} 训练本地模型，获得初步的恶意模型。

步骤 4　添加噪声。结合投毒指示反馈，自适应地在输出层添加噪声，避免模型更新过度集中。

步骤 5　寻找或更新指示集。使用本地数据集 D_c 近似全局梯度更新，记录梯度小且曲率小的神经元索引并将其作为攻击指标。

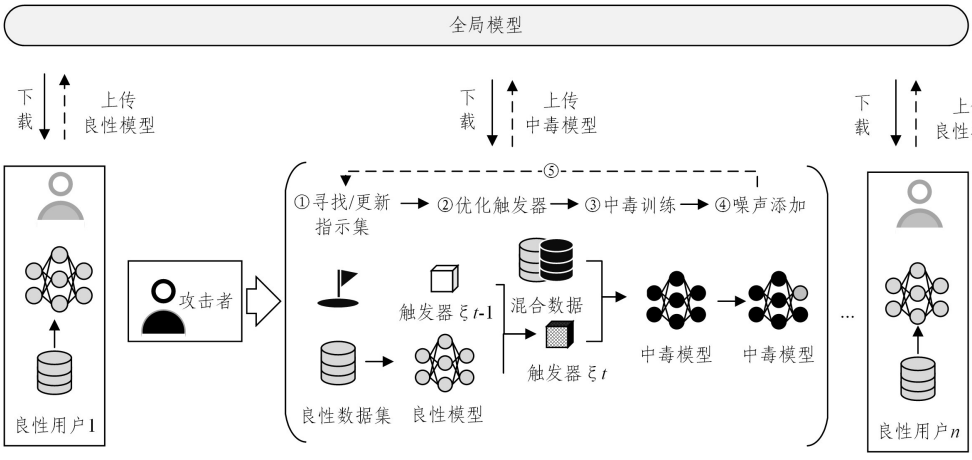


图 1　攻击方案的完整框架图

Fig. 1　Complete framework of attack scheme

3.3 攻击方案设计

本节首先介绍了如何具体地训练攻击模型，分为优化触发器和动态添加噪声两个阶段。其次，简要介绍了指示机制，以及如何依据指示机制动态调整本文的攻击策略。

3.3.1 训练攻击模型

首先介绍攻击者如何在本地训练一个攻击模型。具体来说,攻击主要分为两个步骤:1)在攻击者每一轮本地训练中动态更新触发器,使用已更新的触发器进行中毒训练;2)攻击者将添加噪声后的中毒模型作为最终模型上传到服务器。

1)优化触发器

一般来说,衡量后门攻击准确度的标准是评判模型对植入触发器的数据能否输出预设的标签。实际上,这是模型对植入触发器的数据与良性数据的特征学习结果不同,即两者的潜在数据表示存在明显区别,考验的是模型对触发器特征的记忆程度。传统的固定触发器只关注当下触发器的特征学习,当停止攻击后,触发器不再参与训练。神经网络具有"灾难性遗忘"的特性,模型会快速失去对植入触发器特征的记忆,导致固定触发器效用骤降。而现有的触发器优化方法关注的是局部触发器转移到全局后的效能影响,并没有讨论触发器完全不参与训练后的场景。

而本文延长后门触发器生命力的主要思想是,在 FL 动态更新中预测触发器不参与模型训练的场景,同步动态优化后门触发器,使得触发器提前适应攻击停止后的全局模型。但是,精准的预测无毒的全局模型是一件很困难的事情,一方面攻击者无法获取其他良性客户端的任何信息,另一方面攻击者无法得知全局模型的聚合与学习策略。因此,我们利用攻击者所掌握的资源进行近似预测。虽然攻击者无法明确是否被选中参与全局聚合,但是每一轮次攻击者可以接受上一轮下发的全局模型 $\boldsymbol{G}_t$,只要在本地使用无毒数据训练,即可模拟一个无毒的全局模型 $\boldsymbol{G}_{stop\_t}$。具体来说,利用当前的全局模型与攻击者掌握的本地数据集 D_{train} 模拟此时攻击停止后的全局模型 $\boldsymbol{G}_{stop\_t}$,然后优化触发器以保证在当前全局模型 $\boldsymbol{G}_t$ 攻击成功的同时最小化中毒数据在 $\boldsymbol{G}_t$ 和 $\boldsymbol{G}_{stop\_t}$ 的潜在表示差异。形式上,本文的目标函数如下:

$$\mathcal{L}=\alpha\cdot\mathcal{L}_{\text{backdoor}}+(1-\alpha)\cdot\mathcal{L}_{\text{trigger}} \tag{1}$$

其中,$\mathcal{L}_{\text{backdoor}}=\mathcal{L}_{\text{ce}}(x^*,y^*;G_t)$ 表示触发器在当前全局模型 $\boldsymbol{G}_t$ 的攻击任务损失,$x^*=x\oplus\xi$ 表示植入触发器的数据,y^* 表示目标中毒标签。此项用于训练中毒任务的有效性,即在攻击轮次可以完成后门投毒任务。

$$\begin{aligned}&\mathcal{L}_{\text{trigger}}=\|\sigma_1(G_{stop\_t}(x^*))-\sigma_1(G_t(x^*))\|^2\\&\text{s.t. } G_{stop\_t}=\arg\min_{G_t}E_{(x,y)\sim D_{\text{train}}}[\mathcal{L}_{\text{ce}}(x,y;G_t)]\end{aligned} \tag{2}$$

其中,$\sigma(\cdot)$ 是模型第一个激活层的激活函数,用于表示含触发器数据的潜在表示。本文将数据经过第一个激活层处理后的输出视为数据在模型中的潜在表示,因为第一个激活层位于模型的输入层之后,该层接收原始输入数据后产生新的表示,最直接地接触了原始数据特征。$\mathcal{L}_{\text{ce}}(\cdot)$ 表示交叉熵损失。$\boldsymbol{G}_{stop\_t}$ 是预测的攻击停止后的全局模型,这需要攻击者使用本地良性数据模拟一个无毒场景训练。

本文在每一轮攻击者投毒训练前进行触发器优化。一方面关注主要后门任务,使触发器根据最新接收的全局模型进行及时调整,从而在全局模型的变化中保持攻击任务的成功率;另一方面使得触发器提前适应停止攻击后的模型,促进全局模型在不断变化中保留对触发器特征的记忆,从而延长攻击寿命。

2)添加噪声

在优化触发器后,本文方法在绝大多数的防御场景下已经攻击成功。但是,对于文献[18,21]等中关注全连接层差异的防御场景而言,攻击者由于后门任务的引入在全连接层有显著的倾斜。文献[18]定义全连接层神经元的更新能量为 $\boldsymbol{Ups}$,即全局模型与当前模型权重和偏置项的差之和。因此,将噪声视为可优化目标,使用 $\boldsymbol{Ups}$ 表示恶意模型相对于全局模型的异常程度,在模型的全连接层添加噪声,降低恶意模型的 $\boldsymbol{Ups}$,避免全连接层过度集中的神经元更新。对于攻击者 $\boldsymbol{L}_c$ 而言,添加至 $\boldsymbol{L}_c$ 的噪声表示为 $\boldsymbol{n}_c$,输出层神经元更新能量集合表示为 $\boldsymbol{Ups}$。形式上,噪声添加的目标函数可以表示为:

$$\mathcal{L}_{\text{noise}}=\beta\cdot\mathcal{L}_{\text{var}}(Ups)+(1-\beta)\cdot\|\sum_1^m n_i\|^2 \tag{3}$$

其中,$\mathcal{L}_{\text{var}}(Ups)$ 表示输出层神经元的方差,第二项用于约束所有添加的噪声和满足范数 ℓ_2 为 0,m 表示输出层神经元的数量。通过添加噪声的方式,降低输出层神经元更新能量的不平衡。

3.3.2 调整攻击策略

本小节主要介绍指示机制的简要工作原理以及如何根据指示机制的反馈结果动态调整攻击策略。

1)指示机制工作原理

文献[10,22]研究发现,全局模型中的一些神经元数值不会发生太大变化,它们对大多数良性和恶意的梯度更新都不敏感,称为冗余神经元。在攻击过程中,如果前一轮攻击失败,那么全局模型会聚合而不会使用中毒客户端的更新,因此只要在前一轮赋予冗余神经元固定值,在当前轮次检查全局模型在冗余神经元上的变化就可以判断攻击是否成功。详细证明过程请参考文献[10]。

对于攻击者 $\boldsymbol{L}_c$,使用索引集 $\hat{\boldsymbol{I}}$ 表示神经元的位置,则 $\boldsymbol{\theta}_{L_c,I_c}$ 表示模型 $\boldsymbol{L}_c$ 的索引 $\boldsymbol{I}_c$ 处的参数值,$\boldsymbol{\theta}_{L_c-G_t,I_c}$ 表示模型 $\boldsymbol{L}_c$ 相较于模型 $\boldsymbol{G}_t$ 在索引 $\boldsymbol{I}_c$ 处的模型更新值,指示集合的大小由攻击客户端数量决定。假设存在 4 个攻击者,指示机制的执行过程可简述为 3 个步骤。

步骤 1 寻找冗余神经元。由于处在黑盒环境中,无法直接获取当前全局模型 $\boldsymbol{G}_t$ 具体的梯度更新,因此使用本地数据获得近似梯度和近似曲率。

步骤 2 赋值。选择绝对梯度与曲率最小的 4 个神经元作为指示神经元,并记录它们的索引集 $\hat{\boldsymbol{I}}=\{I_1,I_2,I_3,I_4\}$,将此处的参数值更改为 k 倍,$\boldsymbol{\theta}_{L_c,I_c}=k\cdot\boldsymbol{\theta}_{L_c,I_c}$。

步骤 3 检查指标并判断。若在下一轮接收的全局模型 $\boldsymbol{G}_{t+1}$ 中 $\boldsymbol{\theta}_{G_{t+1}-G_t,I_c}/\boldsymbol{\theta}_{L_c-G_t,I_c}<1/k$,则说明全局聚合中攻击者的更新未被接纳,即前一轮次投毒失败。

2)策略调整

在第一轮恶意训练中,本文的攻击者首先根据指示机制的步骤 1 寻找初始的冗余神经元作为攻击指示集,使用 *feedback* 反馈数组记录攻击成功的情况,a 表示攻击成功,r 表示攻击失败。在后续攻击中,攻击者首先根据接收的全局模型计算出 *feedback* 数组,然后根据 *feedback* 数组进行攻击策略的动态调整。由于本文实验设置允许服务器随机选择参与者,攻击者在每一轮训练中未必被选中,因此只要 *feedback*

数组中包含 a，即认为攻击成功。式(1)中的 α 代表触发器有效性与持续性任务两者的比例，它依据指示机制提供的反馈数组 *feedback* 进行调整。我们认为 *feedback* 包含 a 时表明攻击有效性方向满足，优化触发器任务应该注重于增强攻击的持续性，因此适当增加 a 的值。式(3)中的 β 用于平衡噪声任务的两种目标，即降低输出层更新能量的方差与降低噪声和，以避免影响主要任务。如果 *feedback* 包含太多 r(大于等于数量的一半)，说明存在攻击者未能攻击成功的情况，输出层可能仍然存在显著差异，因此噪声任务应该注重于减少差异性，可以适当增加 b 的值。

具体的算法实现过程如算法 1 所示。

算法 1 基于触发器优化的后门攻击算法

输入：全局模型 G_t，本地数据集 D_{train}，良性学习率 η，后门学习率 η_{adv}

输出：中毒模型 L_c，指示集 $\hat{I}$

1. 检查指示机制：如果 epoch＞1，则调整 α 与 β；

Repeat

2. 从 $\mathbf{D}_{train}$ 取样良性数据，训练获得无毒全局模型：$\mathbf{G}_{stop\_t} = \mathbf{G}_t - \eta \cdot \mathcal{L}_{ce}(x, y; G_t)$；

Until batch_size 结束；

Repeat

3. 根据式(1)优化触发器：$\xi = \xi - \eta_{adv} \cdot \mathcal{L}$；

Until 优化轮次结束；

4. 后门训练获得初步中毒模型 $\mathbf{L}_c$，；

5. 根据式(3)训练噪声 $\mathbf{n}_c$，获得最终中毒模型 $\mathbf{L}_c = \mathbf{L}_c + \mathbf{n}_c$；

6. 寻找或更新指示集 $\hat{I}$；

Return 中毒模型 $\mathbf{L}_c$，指示集 $\hat{I}$；

4 实验分析

为了评估本文方法的有效性与可行性，所有实验均在具有 32 GB RAM 和 Ubuntu 16.04LTS OS 平台的 NVIDIA Quadro P4000 GPU 的 RHEL7.5 服务器上完成，软件为 Window10 操作系统和 Pycharm。

4.1 数据集与实验设置

4.1.1 数据集介绍

在实验中，本文在 MNIST，CIFAR-10 和 CIFAR-100 3 个数据集上进行评估。MNIST 数据集包括从 0 到 9 的 10 类手写数字。它通常用于训练各种图像处理模型，共有 70000 张图像作为训练集(60000 张图像)和测试集(10000 张图像)。CIFAR-10 由 60000 张图像的训练集和 10000 张图像的测试集组成，包含 10 个类别中的 32×32 像素。CIFAR-100 是 CIFAR-10 数据集的扩展版本，包含 100 个不同的类别，每个类别都包含 600 张 32×32 像素的彩色图像。

4.1.2 实验相关设置

本文使用 Dirichlet[23] 分布对每个参与者的本地数据集进行采样，模拟非独立同分布数据集。无特殊声明时数据分布指数取值为 0.9，MNIST 有 20 名 FL 参与者，每轮训练全部参与全局聚合。CIFAR-10 和 CIFAR-100 有 100 名 FL 参与者，每轮训练随机选择 10 名参与者进行全局聚合。在接收到梯度更新后，中央聚合器应用有或没有后门防御的 FedAvg 来聚合这些更新。表 1 列出了 3 种数据集中每个参与者使用的模型架构。

表 1 数据集概述

Table 1 Dataset overview

数据集	规模(训练集/测试集)	类别数量	模型	特征	中毒学习率/训练轮数
MNIST	60000/10000	10	2 conv and 2 fc	28×28	0.1/5
CIFAR-10	50000/10000	10	ResNet18	32×32	0.01/5
CIFAR-100	50000/10000	100	ResNet18	32×32	0.01/5

假设攻击者在所有 N 个客户端破坏 P 个客户端。所有的攻击者只在有限轮次发起攻击。对于 MNIST，我们破坏 20%的客户端，每个客户端毒化全部本地数据集，攻击 40 轮。对于 CIFAR-10 和 CIFAR-100，我们破坏 5%的客户端，每个客户端毒化 25%的本地数据集，攻击 100 轮。本文实验在黑盒中，攻击者无法知道超出其受损设备的任何外部信息，并且与中央聚合器的聚合无关。

本文依据式(1)，使用梯度下降的方法实现触发器优化。在 MNIST 数据集中设置触发器学习率为 0.1，优化 10 个 epoch。在 CIFAR-10 和 CIFAR-100 数据集中设置触发器学习率为 0.01，优化 50 个 epoch。良性全局模型的学习率与表 1 中的中毒学习率保持一致。

4.1.3 评估指标

本文使用 3 种指标进行全面评估。1)主任务精确度(Main Task Accuracy，MTA)：模型在主要任务的准确度，即未中毒数据的准确度，衡量攻击隐蔽性。2)攻击成功率(Attack Success Rate，ASR)：模型在中毒数据上的准确度，衡量攻击有效性。3)生命周期(lifespan L(γ))：定义为停止攻击后 ASR 仍高于 γ 的轮数，数值越大说明触发器持久性越好，例如 $L(50\%)$ 代表 $ASR \geqslant 50\%$ 的轮数，$L(50\%)=100$ 即代表有 100 轮的攻击成功率高于 50%。

4.1.4 对比方法与防御方法

将本文方法与最新的攻击方法进行比较。在集中式触发器中，与模型替换攻击[9]、3DFed[10]、A3FL[15] 进行对比。在分布式触发器中，将本文方法与 DBA[12] 和 FCBA[17] 进行对比。

1)模型替换攻击[9]：使用重复并放大训练中毒示例的方法增强的局部训练数据计算模型更新，将该方法视为本文的基线攻击方法，用 Baseline 表示。

2)3DFed[10]：使用了 3 个正交规避模块，从不同的角度伪装后门模型。具体来说，通过约束项增强后门训练，以抵消范数裁剪防御；通过噪声掩码减轻了后门模型中过度集中的神经元更新和高成对余弦相似度；增加诱饵模型来混淆降维防御。

3)A3FL[15]：提出了对抗性适应损失，通过对抗适应动态全局模型来延长触发器的持久性。

模型替换攻击实验设置与 3DFed 一致，遵循 3DFed 原文设置；CIFAR-10 与 CIFAR-100 中为 100 个客户端，20 个恶意攻击者。MNIST 中为 20 个客户端，4 个恶意攻击者。A3FL 与本文的实验设置保持一致。

4)DBA[12]：以图像像素触发器为例，提出将触发器分解

为多个子像素块，并分别嵌入到不同的恶意客户端，如图2(a)所示，每个恶意客户端使用一个子矩阵。

5)FCBA[17]：首次使用组合数学理论来设计增强全局模型后门效应的触发策略，如图 2(a)所示，每个恶意客户端使用不同的子矩阵组合，包含一个子矩阵或多个子矩阵(少于 4 个)。

同时，在 5 种代表性的联邦学习后门防御下评估本文攻击的有效性和持久性，即 FedAvg[5]，Clip[20]，Foolsgold[21]，Rflbat[19]，Deepsight[18]。

(a) (b)

图 2 分布式触发器和集中触发器示例

Fig. 2 Examples of distributed centralized triggers

4.1.5 触发器设置

本文使用像素触发器进行实验分析。具体来说，对于集中式触发器，初始化为 5×5 的红色像素块，优化后的触发器不改变大小，仅改变数值。训练期间，攻击者每一轮都使用完整的 5×5 的像素块作为触发器发起投毒，测试期间，全局模型测试完整像素块的攻击成功率与持久性能。对于分布式触发器，保持与 DBA[12]和 FCBA[17]一致的触发器设置，即初始化为 4 个 1×6 的白色子像素块，组合为一个完整的触发器，总触发像素为 24。训练期间，DBA 和 FCBA 分别基于轮询与组合数学的思想选择一个(DBA)或多个子像素块(FCBA)组合发起投毒，测试期间，全局模型测试 4 个子像素块组合的完整触发器的攻击成功率与持久性能。为了公平比较，我们与 DBA 和 FCBA 攻击在同一轮中完成相同的触发像素的注入。以 MNIST 为例，图 2 中两张图片的左上角白色矩阵块即作为触发器的图像像素，图 2(a)为分布式触发器，每个恶意客户端仅掌握 4 个子矩阵块的 1 块或多块(小于等于 4 块)。集中式触发器可表示为图 2(a)的 4 块子矩阵的和或图 2(b)的单个完整矩阵块。在本文实验中，在与集中式触发器方法的对比实验中使用图 2(b)所示的触发器，在与分布式触发器方法的对比的实验中，使用图 2(a)所示的 4 块像素作为完整的触发器。

4.2 实验结果

4.2.1 攻击隐蔽性

首先，验证本文方法的隐蔽性。全局模型对未植入触发器的数据应该输出正确的标签，因此，本文衡量 FL 在主任务上的准确率以衡量本文攻击方法的隐蔽性。表 2 中，Avg 代表 FedAvg[5]方法，Fools 代表 Foolsgold[21]方法，Rfl 代表 Rflbat[19]方法，Dsight 代表 Deepsight[18]方法。由表 2 可见，在 3 个数据集上，FL 遭受本文方法的攻击后精度下降最大不超过 1%，在 CIFAR-10 和 CIFAR-100 的大部分场景下精度有所提升。实验证明，本文方法未损害 FL 的基础性能，FL 可以在非中毒数据中正常执行任务，证明本文的攻击方法是隐蔽的。

表 2 主要任务的准确度

Table 2 Accuracy of main task

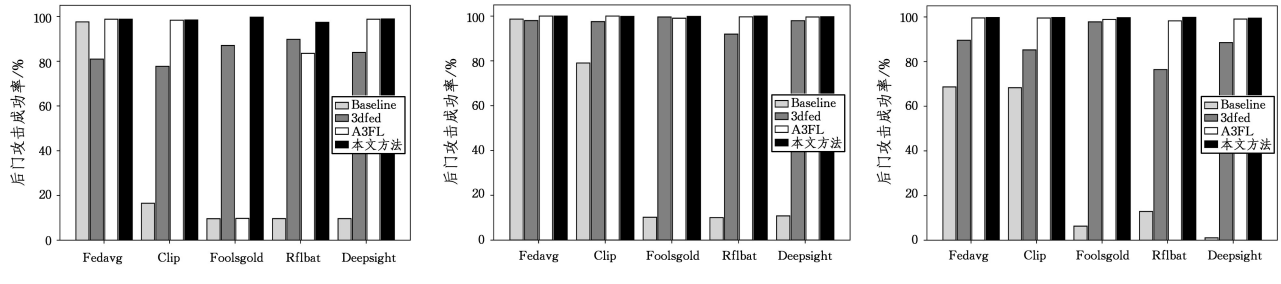
(%)

数据集	Avg[5]	Clip[20]	Fools[21]	Rfl[19]	Dsight[18]
MNIST	99.05*	99.18*	98.95*	99.12*	99.17*
	98.94	98.98	98.04	99.03	98.99
CIFAR-10	92.20*	92.23*	92.30*	92.56*	91.91*
	92.38	92.48	92.44	92.23	91.97
CIFAR-100	66.90*	66.87*	67.02*	66.75*	65.32*
	67.15	67.10	67.06	67.28	65.10

注：符号“*”表示无攻击场景的 MTA。

4.2.2 攻击有效性

图 3 给出了本文方法和对比的 3 种方法的攻击成功率，图中结果为攻击轮次的最后 10 轮的攻击成功率的平均值。



(a)MNIST (b)CIFAR-10 (c)CIFAR-100

图 3 本文方法的攻击成功率

Fig. 3 ASR of our method

本文方法在 3 种数据、5 种防御下均能攻击成功，其中在 CIFAR-10 与 CIFAR-100 上都达到接近 100%的 ASR。Baseline 方法仅在 FedAvg 方法下能取得不错的性能，因为其采用模型缩放攻击放大了中毒模型的参数，这虽然提高了中毒模型在全局聚合中的影响，但也加剧了中毒模型与其他良性模型的距离，无法应对基于距离的防御方法。在 MNIST 的 Foolsgold 和 Rflbat 场景下，本文方法攻击成功率显著高于其他对比方法，这是因为本文采用了双目标动态触发器，同时解决了模型输出神经元过度集中的问题，动态触发器在每轮都向后门任务的方向更新，噪声的添加掩盖了后门任务的统一目标，可以保证攻击的有效性。在具有 100 类的数据集 CIFAR-100 中，本文方法的 ASR 也高于 98%，相较于对比方法稳步提升。上述实验可以表明，本文方法在多种防御体系下仍然能保持高 ASR，并且适应于

多种分类任务和模型结构。

4.2.3 攻击持久性

1)集中式触发器

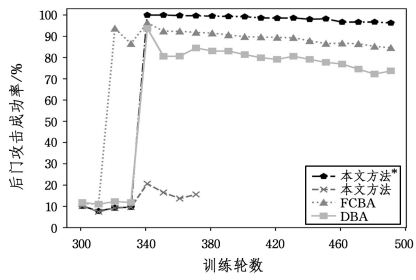
首先对比了本文方法的与常见的集中式触发器的攻击持久性指标。表3列出了集中式触发器的攻击持续性的实验结果。在3种数据下分别计算攻击停止后攻击成功率高于50%与90%的轮数，即$L(50\%)$与$L(90\%)$。对于MNIST数据集攻击40轮，统计停止攻击的100轮中ASR分别高于50%与90%的轮数。对于CIFAR-10和CIFAR-100攻击100轮，统计停止攻击的600轮中ASR分别高于50%与90%的轮数。由表3可见，本文方法在绝大多数联邦防御场景下都能保持高持久性，相较于其他攻击方法有显著提升。在CIFAR-10的4种防御方法下攻击成功率高于90%的轮数超过了600轮，而baseline和3dfed的攻击成功率则在攻击停止后迅速下降到50%以下。A3FL的持久性在MNIST和CIFAR-10的Foolsgold和Rflbat场景中显著低于本文方法。因为A3FL仅优化触发器，并没有考虑中毒任务带来的输出层参数不平衡。

表3 集中式触发器持久性对比

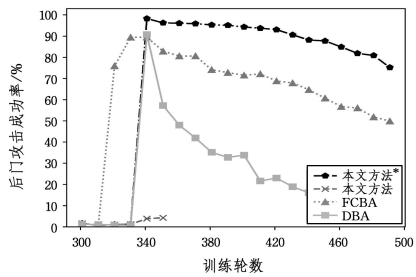
Table 3 Comparison of persistence of centralized triggers

数据集	防御方法	Baseline[9]		3DFed[10]		A3FL[15]		本文方法	
		$L(50\%)$	$L(90\%)$	$L(50\%)$	$L(90\%)$	$L(50\%)$	$L(90\%)$	$L(50\%)$	$L(90\%)$
MNIST	FedAvg[5]	7	3	6	—	84	13	**88**	**15**
	Clip[20]	—	—	7	—	60	9	**71**	**9**
	Foolsgold[21]	—	—	5	—	—	—	**93**	**32**
	Rflbat[19]	—	—	6	2	30	—	**81**	**11**
	Deepsight[18]	—	—	—	—	84	13	**90**	**15**
CIFAR-10	FedAvg[5]	163	29	97	10	600	597	**600**	**600**
	Clip[20]	65	—	112	14	600	594	**600**	**600**
	Foolsgold[21]	—	—	211	38	600	233	**600**	**600**
	Rflbat[19]	—	—	23	1	600	522	**600**	**599**
	Deepsight[18]	—	—	132	24	600	595	**600**	**600**
CIFAR-100	FedAvg[5]	9	—	10	—	288	62	**343**	**78**
	Clip[20]	9	—	18	—	296	52	**340**	**77**
	Foolsgold[21]	—	—	25	1	256	56	**290**	**73**
	Rflbat[19]	—	—	15	—	252	40	**311**	**57**
	Deepsight[18]	—	—	31	—	243	48	**271**	**56**

注：符号“—”表示未达到指定阈值的ASR。



(a)CIFAR-10数据集



(b)CIFAR-100数据集

图4 分布式触发器的攻击持久性

Fig. 4 Persistent attack of distributed triggers

2)分布式触发器

本文与FCBA中的设置保持一致，在CIFAR-10和CIFAR-100中使用总像素相同的触发器(24像素)，实验均在FedAvg场景下完成。DBA中指出只执行一次完整攻击时，攻击者需要将其恶意更新执行缩放，以压倒其他良性更新并确保后门在聚合步骤中生存。为了公平比较，与DBA和FCBA攻击在同一轮(轮数等于340)中完成一个完整的后门(触发像素相同)。图4中，符号“*”表示本文方法加入缩放系数，未加“*”表示本文方法未加入缩放系数。结果表明未加入缩放系数的本文方法无法取得优势，但加入与DBA和FCBA一致的缩放参数后，本文方法在单次攻击中显著优于先进的分布式触发器。这是因为，虽然分布式帮助全局模型学习多种组合的触发器特征，但本质依旧是固定触发器。上述实验结果表明，动态优化触发器对于攻击持久性具有重要意义。

4.3 讨论

本节改变了一些实验参数，并评估它们对攻击效果的影响。

4.3.1 攻击轮数

首先，在CIFAR-10数据集上的FedAvg场景下测试攻击开始的轮次对攻击效果的影响，攻击者数量与4.2节保持一致，设置分别从第0,320,600,900轮数开始发起100轮攻击。由图5可见，无论从何时发起攻击，攻击100轮后都可以达到接近于100%的ASR。如果从0个epoch发起攻击，由于全局模型在训练初期变化很大，因此后门触发器很容易被擦除，ASR在攻击停止之后很快降至50%以下。而从后期全局模型逐渐收敛后发起攻击，后门触发器的特征也被稳定记忆，因此在攻击停止且攻击轮次相同时也能保持相似的持久性。其次，设置攻击的轮次分别为40,80,100,120，图6给出了4种攻击轮次下ASR随着epoch的变化。由图可见，攻击轮次越多，全局模型将触发器的特征记忆得越深刻，触发器越持久。

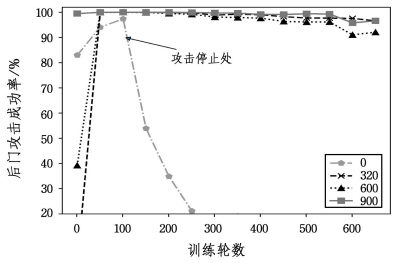


图 5 不同开始轮数的影响

Fig. 5 Impact of different start epochs of attack

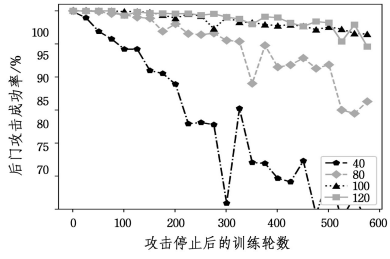
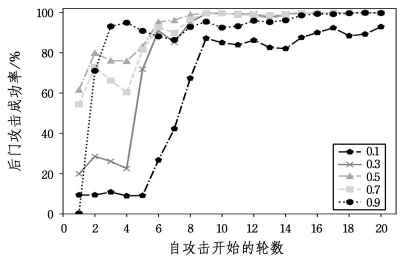


图 6 不同攻击轮数的影响

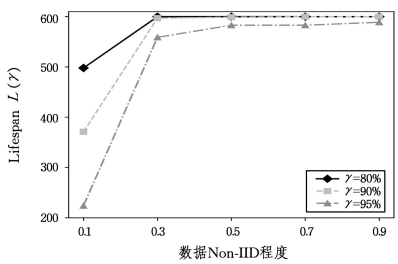
Fig. 6 Impact of different attack epochs

4.3.2 数据分布

本文研究了数据非独立同分布的程度对本文攻击的影响，如果攻击者掌握的本地数据集异质程度高，会导致数据分布远离全局数据分布，攻击难度加大。在 CIFAR-10 数据集上的 FedAvg 下设置 Dirichlet 数据分布参数分别为[0.1, 0.3, 0.5, 0.7, 0.9]。图 7(a)展示了不同数据分布下的 ASR，结果表明无论数据分布如何，前 20 轮的 ASR 都能迅速达到 80%以上。图 7(b)显示了攻击 100 轮，停止攻击的 600 轮中 ASR 分别高于 80%，90%与 95%的轮数，即 $L(80\%)$，$L(90\%)$与 $L(95\%)$。尽管高度异构的数据增加了植入后门的难度，但本文方法仍然取得了很高的成功率，即使在停止 600 轮期间也保持 ASR 高于 95%超过 200 轮次，高于 80%接近 500 轮次。



(a)不同数据分布的 ASR



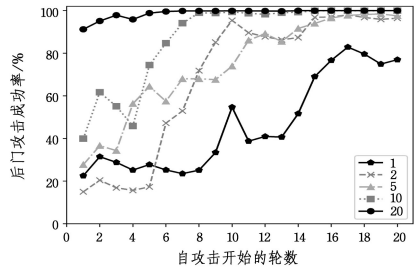
(b)不同数据分布的生命周期

图 7 不同数据分布的影响

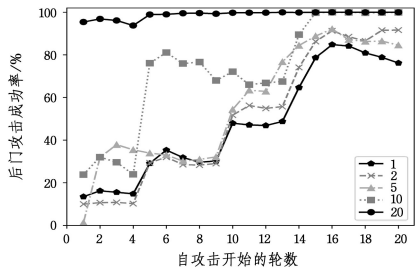
Fig. 7 Impact of different data distribution

4.3.3 攻击者数量

本文在 CIFAR-10 数据集上测试了攻击者数量对本文方法的影响。具体而言，本文中攻击者数量分别为[1, 2, 5, 10, 20]，在 Deepsight 与 Rflbat 场景下测试攻击成功率。无论攻击者数量是多少，全局模型仍然随机选择 10 个客户端参与全局聚合。图 8 给出了攻击前 20 轮的 ASR。实验结果表明，在常见的后门防御场景下，虽然减少攻击者的数量使 ASR 有所降低，但在大多数情况下，本文方法仍然能在 20 轮内达到 80%左右的攻击成功率；并且在只有 1 个攻击者时，ASR 仍然能在 15 轮内高于 50%。



(a)Deepsight



(b)Rflbat

图 8 不同攻击者数量的攻击成功率对比

Fig. 8 Comparison of attack success rates for different numbers of attackers

4.3.4 损失项消融

本文在 MNIST 和 CIFAR-10 数据集上评估了式(1)中第二项触发器损失的影响。依然设置 5 个攻击者，考虑 FedAvg, Rflbat 和 Deepsight 这 3 种后门防御场景。对于 MNIST 数据集，表 4 中符号“/”左表示停止攻击 100 轮中 ASR 高于 50%的轮数，“/”右表示攻击 40 轮的 ASR。对于 CIFAR-10 数据集，表 4 中符号“/”左表示停止攻击 600 轮中 ASR 高于 90%的轮数，“/”右表示攻击 100 轮的 ASR。实验证明了本文所提损失项的有效性，即使用模拟良性模型来优化触发器可以显著提高后门触发器的生命周期。

表 4 3 种防御下的损失消融

Table 4 Loss ablation under three defense method

数据集	FedAvg[5]	Rflbat[19]	Deepsight[18]
MNIST	34/95.39	39/91.60	34/95.21
CIFAR-10	452/99.99	278/99.92	341/99.99

结束语 本文提出了一种联邦学习中隐蔽且持续的后门攻击方法。在模拟的攻击停止后模型上优化触发器以及自适应添加噪声，使动态的全局模型保留对触发器特征的记忆。实验结果表明，本文的攻击方法在主要任务精度、攻击成功率和攻击持久性方面都保持领先，尤其在停止攻击后，CIFAR-

10 数据集上的攻击成功率高于 50%超过 600 轮。在未来的工作中，我们计划研究如何动态优化分布式触发器，实现多目标投毒，同时思考如何构建更好的防御来保护 FL 免受动态触发器的攻击。

参 考 文 献

[1] ZENG X，CAO K，ZHANG M. MobileDeepPill：A small-footprint mobile deep learning system for recognizing unconstrained pill images[C]// Proceedings of the 15th Annual Internation Conference on Mobile Systems，Applications，and Services. New York：ACM，2017：56-67.

[2] RAN X K，CHEN H L，ZHU X D，et al. Deepdecision：A mobile deep learning framework for edge video analytics[C]// Proceedings of the 37th IEEE Conference on Computer Communications. Piscataway，NJ：IEEE，2018：1421-1429.

[3] LIU L Y，LI H Y，MARCO G. Edge assisted real-time object detection for mobile augmented reality[C]// Proceedings of the 25th Annual Internation Conference on Mobile Computing and Networking. New York：ACM，2019：1-16.

[4] KONEČNÝ J，MCMAHAN B，RAMAGE D. Federated optimization：Distributed optimization beyond the datacenter[J]. arXiv：1511.03575，2015.

[5] MCMAHAN B，MOORE E，RAMAGE D，et al. Communication-efficient learning of deep networks from decentralized data[C]// Proceedings of the 20th Internation Conference on Artificial Intelligence and Statistics. New York：PMLR，2017：1273-1282.

[6] BONAWITZ K，EICHNER H，GRIESKAMP W，et al. Towards federated learning at scale：system design[J]. arXiv：1902.01046，2019.

[7] LIU Y，FAN T，CHEN T J，et al. FATE：an industrial grade platform for collaborative learning with data protection[J]. Journal of Machine Learning Research，2021，22(226)：1-6.

[8] SONG M K，WANG Z B，ZHANG Z F，et al. Analyzing user-level privacy attack against federated learning[J]. IEEE Journal on Selected Areas in Communications，2020，38(10)：2430-2444.

[9] BAGDASARYAN E，VEIT A，HUA Y Q，et al. How to backdoor federated learning[C]// Proceedings of Internation Conference on Artificial Intelligence and Statistics. Cambridge，MA：MIT Press，2020：2938-2948.

[10] LI H Y，YE Q Q，HU H B，et al. 3dfed：Adaptive and extensible framework for covert backdoor attack in federated learning[C]// Proceedings of the 44th IEEE Symp on Security and Privacy. Piscataway，NJ：IEEE，2023：1893-1907.

[11] BHAGOJI A，CHAKRABORTY S，MITTAL P，et al. Analyzing federated learning through an adversarial lens[C]// Proceedings of Internation Conference on Artificial Intelligence and Statistics. Cambridge，MA：MIT Press，2019：634-643.

[12] XIE C L，HUANG K L，CHEN P Y，et al. DBA：Distributed backdoor attacks against federated learning[C]// Proceedings of the 7th Internation Conference on Learning Representations. 2019.

[13] WANG H Y，SREENIVASAN K，RAJPUT S，et al. Attack of the tails：Yes，you really can backdoor federated learning[C]// Proceeding of the 34th Annual Conference on Neural Information Prosessing Systems. Massachusetts：MIT Press，2020：16070-6084.

[14] FANG P，CHEN J H. On the Vulnerability of Backdoor Defenses for Federated Learning[C]// Proceedings of the 37th AAAI Conference on Artificial Intelligence. Palo Alto，CA：AAAI Press，2023：11800-11808.

[15] ZHANG H F，JIA J Y，CHRN J H，et al. A3fl：Adversarially adaptive backdoor attacks to federated learning[C]// Proceedings of the 36th Annual Conference on Neural Information Prosessing Systems. Massachusetts：MIT Press，2023：61213-61233.

[16] QIAO Y Q，LIU D Z，CHEN C W，et al. FTA：Stealthy and Adaptive Backdoor Attack with Flexible Triggers on Federated Learning[J]. arXiv：2309.00127，2023.

[17] LIU T，ZHANG Y H，FENG Z，et al. Beyond traditional threats：A persistent backdoor attack on federated learning[C]// Proceedings of the 38th AAAI Conference on Artificial Intelligence. Palo Alto，CA：AAAI Press，2024：21359-21367.

[18] RIEGER P，NGUYEN D，MIETTINEN M，et al. Deepsight：Mitigating backdoor attacks in federated learning through deep model inspection[J]. arXiv：2201.00763，2022.

[19] WANG Y K，ZHAI D H，ZHAN Y G，et al. Rflbat：A robust federated learning algorithm against backdoor attack[J]. arXiv：2201.03772，2022.

[20] SUN Z T，PETER K，ANANDA T，et al. Can you really backdoor federated learning? [J]. arXiv：1911.07963，2019.

[21] FUNG C，YOON C J M，BESCHASTNIKH I. The limitations of federated learning in sybil settings[C]// Proceedings of the 23rd Internation Symposium on Research in Attacks，Intrusions and Defenses. Berlin：Springer，2020：301-316.

[22] ZHOU X C，XU M，WU Y M，et al. Deep model poisoning attack on federated learning[J]. Future Internet，2021，13(3)：73.

[23] HSU T，QI H，BROWN M. Measuring the effects of non-identical data distribution for federated visual classification[J]. arXiv：1909.06335，2019.

JIANG Yufei，born in 2000，postgraduate. Her main research interests include backdoor attack and federated learning.

ZHAO Yanchao，born in 1985，Ph.D supervisor，is a member of CCF（No. 24833S）. His main research interests include edge computing and processing，wireless sensing and optimization.

（责任编辑：喻藜）