

面向联邦学习的可学习触发器后门攻击

刘涛, 苟大鹏, 许晨, 吕继光, 冯驻, 曾凡一, 周雪, 杨武

(哈尔滨工程大学计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘要: 传统的干净标签后门攻击通常无法在触发器和目标类之间建立有效的强连接关系, 导致攻击成功率 (ASR) 降低。大量实验结果表明, 这种失效现象在联邦学习 (FL) 中更加严重。攻击失败的主要原因在于触发器的随机选择使其缺少和目标类的直接联系。为此, 面向 FL 设计一种可学习触发器后门攻击, 充分利用中心服务器下发的任务信息和共享模型训练与目标类强相关的触发器, 将该训练过程形式化为双目标优化问题并求解。在约束条件下找到最佳扰动尽可能模糊图像的原始特征, 从而最大化模型对触发器的学习能力; 然后对这些模糊图像添加规定范围内允许的触发器作为输入进行训练, 最小化其图像分类损失并利用小批量投影梯度下降 (PGD) 的优化方法快速生成最优触发器。利用该触发器激活的后门攻击在 FL 中仍能保证出色的攻击性能。在 3 个数据集上的实验结果表明, 所提方法在 FL 中攻击成功率远高于现有的各类干净标签后门攻击, 尤其在 CIFAR-10 上其 ASR 相较于基线方法提升了约 82 个百分点, 所提出的攻击方法对 FL 的安全性提出了新的挑战。

关键词: 机器学习; 联邦学习; 后门攻击; 干净标签攻击; 可学习触发器

源代码链接: <https://github.com/PhD-TaoLiu/LTBA-code>. git

中图分类号: TP393

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0252239

Learnable Trigger Backdoor Attacks for Federated Learning

LIU Tao, MAN Dapeng, XU Chen, LÜ Jiguang, FENG Zhu, ZENG Fanyi, ZHOU Xue, YANG Wu

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, Heilongjiang, China)

【Abstract】 Conventional clean label backdoor attacks often fail to establish a strong link between the trigger and target classes, resulting in a low Attack Success Rate (ASR). Extensive experimental evidence shows that this failure is even more severe in Federated Learning (FL). The main reason for attack failure is that the random selection of the trigger causes it to lack a direct connection with the target class. To this end, a learnable trigger backdoor attack is designed for FL, which fully utilized the task information and shared model issued by the central server to train a trigger that is strongly correlated with the target class. This training process is formalized into a dual-objective optimization problem and solved. The optimal perturbation under constraint conditions is found to blur the original features of the image as much as possible, thereby maximizing the model's learning ability for the trigger. Subsequently, these blurred images are trained by adding the triggers allowed within the specified range as inputs, minimizing their image classification loss and generating the optimal trigger quickly using the optimization method of small-batch Projected Gradient Descent (PGD). The backdoor attack activated by this trigger guarantees excellent attack performance in FL. The experimental results on three datasets show that the ASR of the proposed method in FL is much higher than that of all types of existing clean-label backdoor attacks, especially on CIFAR-10, which shows an improvement of more than 82 percentage points compared to the baseline method. The proposed attack method presents new challenges for the security of FL.

【Key words】 machine learning; Federated Learning (FL); backdoor attack; clean label attack; learnable trigger

0 引言

随着大数据时代的到来, 机器学习技术在计算机视觉、自然语言处理、自动驾驶等多个领域得到了广泛应用。这种应用趋势催生了对数据融合的迫切

需求, 而数据孤岛问题成为重大挑战。在数据共享过程中对用户敏感信息和数据隐私的保护也至关重要。在这一背景下, 联邦学习^[1] (FL) 作为一种全新的分布式学习范式被提出, 用于解决数据隔离问题, 特点是采用模型更新共享代替数据共享的

基金项目: 国家自然科学基金面上项目 (62272127); NSFC-企业联合基金重点项目 (U20B2048, U21B2019); NSFC-区域联合基金重点项目 (U22A2036); 黑龙江省自然科学基金研究团队项目 (TD2022F001); 中央高校基本科研业务费专项资金 (3072024XX0606)。

作者简介: 刘涛 (CCF 学生会员), 男, 博士研究生, 主研方向为分布式人工智能安全; 苟大鹏 (通信作者), 教授、博士; 许晨, 讲师、博士; 吕继光, 副教授、博士; 冯驻, 硕士研究生; 曾凡一、周雪, 博士研究生; 杨武, 教授、博士。

收稿日期: 2025-03-18

修回日期: 2025-05-19

E-mail: mandapeng@hrbeu.edu.cn

方式协作训练模型以确保数据隐私^[2-3]。然而,其分布式特性和数据的异质性^[4-5]为后门攻击创造了有利条件。后门攻击如 WANG 等^[6]所述,在训练样本中嵌入特制的“触发器”并修改标签,训练此类样本可将后门植入模型。携带后门的模型^[7]表现为:当输入干净的测试样本时模型能正确分类;当输入含有触发器的测试样本时模型会将其错分成指定的类别。

由于标签一致性检测技术的出现,标签异常的样本被轻易检测和去除,使得基于标签反转的后门攻击失效,因此广大学者开始关注和研究一种名为干净标签攻击的新型后门攻击^[8]。该攻击只对目标类样本投毒,并且不修改标签,以牺牲一部分攻击成功率(ASR)为代价换取攻击的隐蔽性^[9]。ASR 下降的原因主要在于携带后门的模型无法在触发器和目标类之间建立强连接^[10]。干净标签的限制使攻击者只能对目标类样本添加触发器,导致用于训练的后门样本数量明显减少,模型对该连接的学习能力减弱。BARNI 等^[11]试图通过增加投毒比例来解决这一问题,但一方面破坏了原始数据分布,另一方面损害了目标类干净样本的预测性能。标签一致性(LC)方法^[12]分别用基于生成对抗网络(GAN)的潜在空间插值和对抗性扰动的思想来模糊原始图像特征,引导模型将更多注意力放在对触发器特征的学习上。然而,这些方法并未从根本上解决触发器与目标类之间的弱连接问题。主要是因为这些攻击方法随机选择了与目标类缺少直接联系的静态触发器(如像素块)。

在联邦学习中,由于中心聚合和灾难性遗忘对后门的弱化,因此干净标签攻击的表现会更差。经实验发现一些现有的攻击方法在联邦学习中 ASR 几乎为 0,说明攻击完全失效。因此,攻击者更应该专门设计一种与目标类强相关的触发器来激活后门,大幅提升联邦学习的后门表现。

为此,本文设计了一种可学习触发器。攻击者利用中心服务器的训练任务和模型信息创建附加的本地任务^[13],学习与目标类强相关的触发器。其核心思路是将学习触发器的过程形式化为双目标优化问题进行求解。第 1 个优化目标是生成最佳扰动,通过尽可能模糊原始图像特征,尽可能提升模型对触发器的学习能力;第 2 个优化目标是生成最优触发器,使添加该触发器的模糊图像能以最大概率被预测为目标类。待触发器生成后,将其嵌入目标类干净样本中,通过训练这些样本,将后门植入攻击者控制的本地模型中^[14]。最终,攻击者提交恶意的模

型更新并参与中心聚合,成功污染全局模型。在测试阶段,全局模型除正常工作以外,还表现为该后门上的高 ASR。本文主要贡献如下:

1) 本文提出一种针对联邦学习范式的新型后门攻击方法——可学习触发器后门攻击(LTBA)。该方法利用联邦学习共享任务和模型信息的特性,通过攻击者的本地学习得到和目标类强相关的触发器,将其植入本地训练数据中来损害本地模型,并通过中心聚合最终污染全局模型。整个框架能适应性地调整攻击者的本地训练细节来提升攻击表现。

2) 本文将求解最优触发器的过程形式化为一个带约束的双目标优化问题。内层优化通过最大化损失值尽可能消除图像原始特征的影响,生成最佳模糊样本;再以这些样本作为输入进行外层优化,通过最小化损失提高目标类的预测置信度,生成最优触发器。上述过程显著增强了触发器和目标类的联系,提高了攻击性能,同时满足干净标签的设定条件,具有良好的隐蔽性。

3) 在数据集 CIFAR-10、GTSRB 和 PTSD 上,该攻击的 ASR 相比各基线方法分别提高了 82、68 和 37 百分点。本文还通过参数实验和不可见性评估,展示了一些关键因素对攻击性能的影响以及该方法出色的隐蔽性。

1 相关工作

1.1 干净标签后门攻击

传统后门攻击因无法躲避标签一致性检测而失效,使干净标签后门攻击的研究得到重视和发展。此类攻击特点是只针对目标类样本投毒,且不修改标签,因此用于训练的后门样本不存在标签异常的现象。BARNI 等^[11]提出一种最基本的方法,通过大幅提高投毒比例来提升 ASR。TURNER 等^[12]通过基于 GAN 的特征插值和添加对抗性扰动的方法模糊图像的原始特征,强化模型对触发器的学习能力,一定程度上降低了投毒比例。SAHA 等^[15]和文献^[16]将扰动后的目标类样本和含随机触发器的其他类样本进行特征碰撞,实现干净标签攻击,但方法的局限性在于只能实现单源类样本的错误预测。NING 等^[17]用自编码器将原始触发器转换为人眼不可见的噪声图像,同时确保其特征不变性以实现等效的后门注入。QUIRING 等^[18]通过缩放带有触发器的图像使其看似目标类图像,从而在干净标签条件下向数据集注入有毒数据。

1.2 面向联邦学习的后门攻击

面向联邦学习的后门攻击是新兴且活跃的研究

领域。BAGDASARYAN 等^[19]首次在联邦学习中提出后门攻击,通过放大恶意更新实现模型替换,提升全局模型中后门效应。BHAGOJI 等^[20]提出在模型未收敛时提高攻击者的本地学习率以实现高效攻击,并提出交替最小化策略增强攻击的隐蔽性。XIE 等^[21]首次提出分布式后门攻击,将全局触发器划分成多个单独训练的本地触发器,然后利用联邦学习的分布式特性聚合分散的后门信息,以提高攻击的持久性和隐蔽性。

综上所述,这些攻击方法存在显著缺陷:攻击者随机选择与目标类无关的触发器,导致模型无法通过训练少量后门样本建立两者间的强连接关系。联邦学习的中心聚合和灾难性遗忘会进一步弱化该连接,导致全局模型中后门攻击表现不佳。为此,本文提出一种与目标类强相关的可学习触发器。在联邦学习中,即使用该触发器激活干净标签攻击,也能保证理想的攻击性能。

2 本文方法

2.1 威胁模型

2.1.1 攻击场景

本文主要考虑联邦学习中单个攻击者发起后门攻击的场景^[22]。该攻击者控制单客户端,作为恶意参与者参加训练和聚合,首先利用联邦学习自身特性创建附加的本地任务,训练并生成可学习触发器;然后将其嵌入部分训练数据,通过训练主任务实现本地模型的后门化;最后通过提交恶意更新并参与中心聚合的方式,将后门成功植入全局模型。攻击者的最终目标是使携带后门的全局模型将含有触发器的测试样本预测为目标类 t ,同时不影响主任务的性能。本文采用干净标签攻击的设置^[23]进行数据投毒,确保训练阶段后门样本的标签是正常的。

2.1.2 攻击者知识和能力

遵循 Kerckhoffs 原则^[24],本文采用 XIE 等^[21]提出的假设,假定攻击者是完全知情的,并且能够完全控制客户端的本地训练过程,包括管理本地数据、模型更新以及每次迭代过程中对超参数的自适应微调。这一假设并不影响其他参与者或中心服务器,非常适用于联邦学习场景。

在联邦学习中,除中心服务器发布的训练任务(即主任务,称为“任务 1”)以外,攻击者还会控制恶意客户端执行附加的学习任务(称为“任务 2”),以生成可学习触发器。关于“任务 2”有以下 3 点说明:1)“任务 2”使用收敛后全局模型的副本进行训练;2)“任务 2”使用攻击者抽取的小样本集 D_t 进行

训练;3)“任务 2”和“任务 1”使用相同的模型进行训练,且训练量很小,不会显著增加计算代价。上述设定符合联邦学习范式的实际情况。

2.1.3 攻击指标

在现有的研究中,后门攻击通常通过 2 个关键指标来衡量:ASR^[25]和主任务上的准确率(ACC)^[26]。与脏标签攻击^[27-28]不同,干净标签攻击仅针对目标类样本进行投毒,如果设计不当,会显著降低目标类干净样本的预测精度,从而引起检测人员的警觉,增加后门被揭露的风险。第 3.1 节强调了当前干净标签攻击中的这一问题,并引入目标类准确率(Tar-ACC)作为补充指标。总体而言,优秀的干净标签后门攻击应同时实现较高的 ACC、Tar-ACC 和 ASR。

2.2 可学习触发器

现有干净标签攻击方法的局限性在于缺少触发器和目标类之间的强连接,这主要是由于随机选择的固定触发器。随机性使这些触发器在低投毒比例、不同目标类和模型结构的情况下容易受到影响,从而导致攻击效果不理想。基于联邦学习的特性,攻击者利用服务器发布的任务信息和共享的全局模型,在恶意客户端上创建了附加的训练任务,用于生成可学习触发器。这一过程涉及使用收敛后的全局模型副本和攻击者抽取的目标类小样本集,以求解双目标优化问题。目标 1 是生成模糊图像集,以增强模型对触发器的学习能力。目标 2 是以该模糊图像集作为输入训练模型副本,从而生成可学习触发器。

如第 2.1 节所述,攻击者在恶意客户端上创建“任务 2”用于生成可学习触发器。首先,攻击者根据攻击目标和中心服务器发布的任务信息,从本地训练数据中抽取小样本集 D_t ,仅含少量目标类干净样本,这里以交通标志识别任务^[29]为例,攻击者明确服务器发布的任务后,先确定目标类(一般是该任务中的某常见标签,如“最大限速 50 km/h”);然后,在本地训练集中抽取该类的少量干净样本构成 D_t ,待 T 轮聚合后,攻击者从服务器获得收敛后的全局模型 GM_T ,拷贝其副本 GM'_T 用于训练“任务 2”并在触发器生成后丢弃该副本,从而减少了对存储资源的需求,特别适用于存储容量有限的客户端环境。

这一阶段的总体目标是利用收敛的全局模型副本 GM'_T 和目标类小样本集 D_t 生成最优的可学习触发器。该触发器应满足如下 2 个条件:1)以更高的置信度将携带该触发器的测试样本预测为目标类;2)确保置信度的提高是由触发器特征而非图像

原始特征主导。故这里将上述问题形式化为 1 个双目标优化问题进行求解,如式(1)所示:

$$\delta^* = \operatorname{argmin}_{\delta} \sum L(f(\operatorname{argmax}_{x'} L(x', t, \theta) + \delta), t) \quad (1)$$

式中:外层优化对应条件 1),即通过最小化损失值来尽可能提高触发样本预测为目标类的置信度;内层优化对应条件 2),即通过最大化损失值生成扰动,来尽可能模糊图像的原始特征。在优化过程中,先执行内层优化生成最佳的模糊样本,再以这些模糊样本作为输入执行外层优化生成最优的触发器。2 个优化问题的具体细节描述如下。

首先,执行内层优化。本文采用对抗性扰动技术^[30]模糊图像原始特征,使模型将更多注意力放在对触发器的学习上,涉及下述具体优化(优化目标 1):

$$x_j^{\text{adv}} = \operatorname{argmax}_{\|x'_j - x_j\|_{\infty} \leq \epsilon} L(x'_j, t, \theta_{\text{GM}_T}) \quad (2)$$

式中: θ_{GM_T} 是副本 GM_T 的模型参数; L 是图像的分类损失,沿用经典的交叉熵损失。对 D_i 中的每个输入 x_j ,生成其扰动样本 x'_j ,扰动受 L_{∞} 范数和微小常数 ϵ 的约束。受投影梯度下降(PGD)攻击^[31]思想的启发,每个 x_j^{adv} 被直观地认为是 x_j 在指定约束条件 $\|x'_j - x_j\|_{\infty} \leq \epsilon$ 下的最佳扰动,因为添加这些扰动的图像被以最小置信度预测为目标类 t 。所有 x_j^{adv} 形成新的集合,称之为模糊图像集 D_i^{adv} 。

然后,执行外层优化。以 D_i^{adv} 作为输入,攻击者希望这些模糊图像在添加触发器后能以尽可能高的置信度被预测为目标类,这涉及下述具体优化(优化目标 2):

$$\delta^* = \operatorname{argmin}_{\delta \in \Delta} \sum_{(x_j^{\text{adv}}, t) \in D_i^{\text{adv}}} L(f_{\theta_{\text{GM}_T}}(x_j^{\text{adv}} + \delta), t) \quad (3)$$

式中: Δ 表示允许的触发器集合; L 用于计算将含触发器的扰动样本 $(x_j^{\text{adv}} + \delta)$ 预测为目标类 t 的损失。最优触发器 δ^* 被视为目标类最鲁棒和最具代表性的特征,添加后能最大化 x_j^{adv} 被预测为目标类的概率。因此, δ^* 被认为是用于激活后门最有效的触发器。

以上为基于双目标优化的可学习触发器完整的生成过程,具体的算法实现如算法 1 所示。该算法通过迭代更新触发器以优化学习效果。首先,初始化触发器状态为 0;随后,在每次迭代中根据全局模型参数的梯度信息,结合学习率和损失函数,求解新的触发器。同时,在每次迭代中应将新求出的触发器投影到允许的集合 Δ 中。迭代过程持续进行多

轮,直至所有步骤完成,最终获得最优的触发器。

算法 1 可学习触发器生成算法

输入 全局模型副本 $f_{\theta_{\text{GM}_T}}$,模糊图像集 D_i^{adv} ,允许的触发器集合 Δ ,总迭代次数 I ,学习率 η

输出 可学习触发器 δ^*

```

1.  $\delta_0 \leftarrow 0^{1 \times d}$  //初始化
2. for each iteration  $i \in (1, I-1)$  do
  //更新触发器
3.  $\delta_{i+1} \leftarrow \delta_i - \eta * \sum \nabla_{\delta} L(f_{\theta}(x_i^{\text{adv}} + \delta), t)$ 
  //对所有  $(x_i^{\text{adv}}, t) \in D_i^{\text{adv}}$  求和
  // $\theta$  是  $\text{GM}_T$  的模型参数
4.  $\delta_{i+1} \leftarrow \text{Proj}_{\Delta}(\delta_{i+1})$  //投影结束
5. end for
6. return  $\delta^*$ 

```

2.3 攻击执行流程

本节先概述 LTBA 的工作流程。需要重点关注的是基于双目标优化的可学习触发器生成过程。攻击者利用联邦学习的特性,在中心服务器下发主任务信息和全局模型后,控制恶意客户端,执行 5 个本地操作步骤,将后门植入本地模型。然后,通过该模型提交的恶意更新污染全局模型,导致全局模型在推理阶段发生预期的误分类行为。该攻击具体步骤如下。

1) 创建附加的本地训练任务。

如图 1 所示,攻击者根据中心服务器下发的主任务信息和全局模型,在某个被控制的恶意客户端(客户端 N)上部署额外的训练任务,目的是生成可学习触发器。然后,中心服务器将收敛的全局模型分发到各客户端。良性用户使用本地数据 $D_i (i \in \{1, 2, \dots, N-1\})$ 对接收到的全局模型进行训练并更新,这一过程称为“任务 1”。同时,客户端 N 开始执行“任务 2”,旨在生成最优的后门触发器。此过程包括:1)复制全局模型 GM_T 用于“任务 2”的训练;2)根据聚合器设定的学习任务(如交通标志识别)和预定的目标类(如靠右行驶)自主构建训练集。攻击者从外部来源(例如互联网)收集少量目标类的干净样本,并进行预处理,以此创建小型目标类样本集 D_i ,用于“任务 2”的训练。

2) 生成触发器。

通过训练步骤 1) 创建的“任务 2”,采用两阶段 PGD 方法生成最优触发器,以应对式(2)和式(3)提出的双目标优化问题。首先,采用 PGD 攻击方式对 D_i 中的各样本生成对抗样本,旨在削弱目标类样本 x_i 的特征表达。具体来说,对抗样本初始化后,每次迭代过程中执行以下步骤:(1)计算损失函数的梯度;(2)依据梯度符号更新样本;(3)将更新后的样本

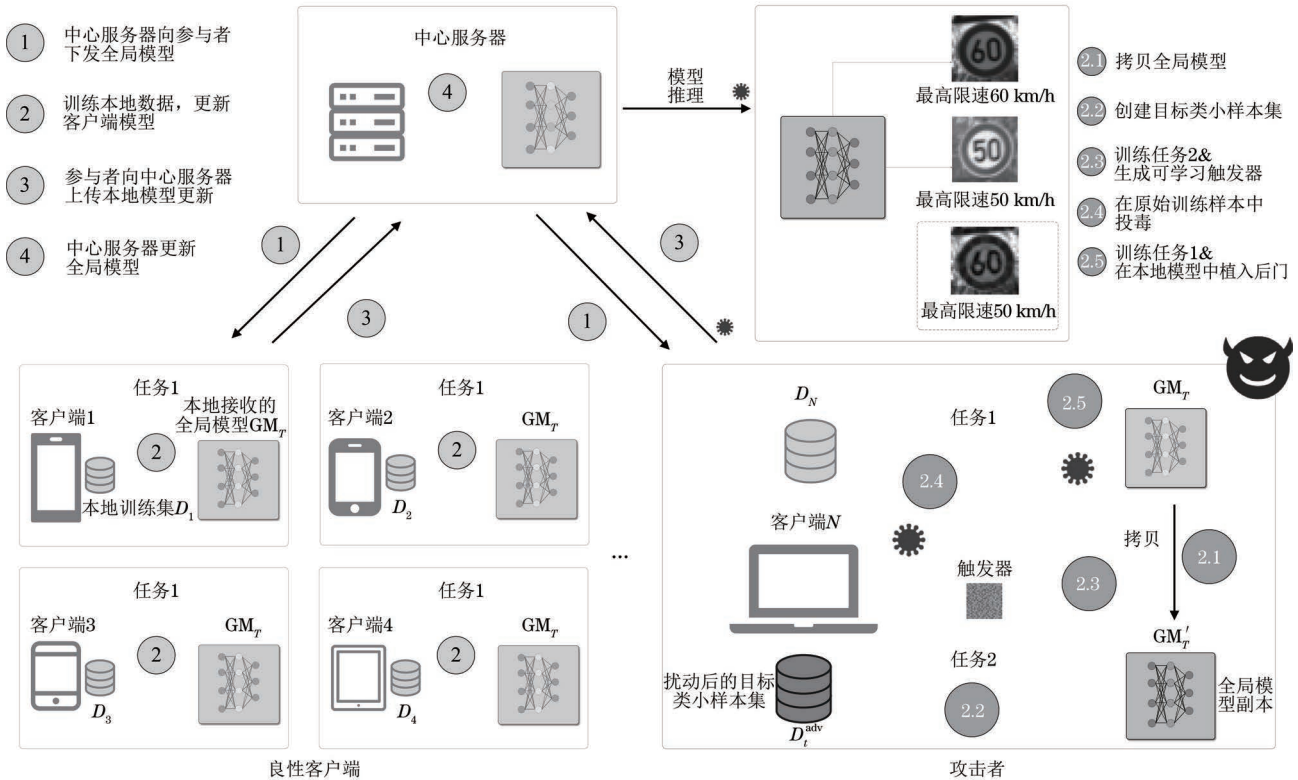


图 1 联邦学习中基于可学习触发器的后门攻击整体流程

Fig. 1 Overall process of the learnable trigger-based backdoor attack in federated learning

投影到有效输入范围,确保所引起的扰动在允许范围内。随后,利用小批量 PGD 优化方法生成最优触发器。选择该优化方法的原因首先是确保生成的最优解在上述允许的触发器集合内,通过投影的方式实现约束条件。其次,由于上述生成的扰动样本间存在较大差异,可能会因为单个样本带来的梯度噪声导致梯度更新不稳定,因此通过小批量的优化方式计算多个样本的平均梯度来减少这种噪声,使训练过程更稳定,同时通过加大更新步伐来加快收敛速度。最后,由于目标函数是高度非线性的且可能包含多个局部最小值,故利用小批量优化方式的随机性来跳出局部最小值,探索更大的解空间。该优化方法中的每次迭代包括:(1)计算每个小批量中目标函数的平均梯度;(2)更新触发器 δ ;(3)将更新后的触发器投影到允许的集合 Δ 中。在非自适应攻击的设置下, Δ 定义为以 L_∞ 范数为界的球形空间,即 $D = \{\delta: \|\delta\|_\infty \leq \epsilon'\}$ 。投影到此球形空间就是将 δ 的每个维度限制在 $[-\epsilon', +\epsilon']$ 的范围内。

3) 毒化本地模型。

在成功生成最优的可学习触发器之后,从本地数据集 D_N 中随机选取仅包含目标类样本的子集,并将触发器嵌入这些样本中,同时保持它们的原始标签不变。这些经过修改的样本随后用于“任务 1”的训练。通过这种方式,训练集被毒化,从而在本地

模型中嵌入后门^[32]。

4) 污染全局模型。

在本地模型植入后门之后,其在训练过程中向聚合器提交包含后门信息的恶意更新。这些恶意更新经中心服务器聚合后,全局模型将被污染。这样的全局模型在保持主任务性能的同时,也展现出良好的后门效果^[33]。

5) 推理阶段增强触发器。

为了提升攻击效果,攻击者在推理阶段可增强测试样本中的触发器强度,例如将触发器强度放大 3 倍,这是在先前的研究中探讨过的方法。这种增强方法能显著提升攻击表现,特别适合于像自动驾驶系统^[34]这一类实际应用。这些类型的场景通常需要实时预测,而且测试样本往往是在线生成的,所以测试样本难以保证像训练样本那样经过严格审查。

与现有的触发器策略相比,增强后的触发器在扰动幅度更小的情况下展现出更强的攻击性能和隐蔽性。这意味着,基于可学习触发器的后门攻击性能更强大,对 FL 范式的安全性发起了新的挑战。

3 实验与结果分析

3.1 实验设置

3.1.1 整体设置

实验中使用了一台服务器,该服务器配备了

4 块 NVIDIA A100 GPU 的服务器,并选用 PyTorch^[35]作为软件框架。测试工作覆盖了 3 个数据集,分别是 CIFAR-10^[36]、GTSRB^[37]和 PTSD,模型架构均采用经典的 ResNet-18^[38]。其中,GTSRB 和 PTSD 分别是源自德国和伊朗的交通标志数据集。详细的数据集和注毒轮次信息如表 1 所示。本次实验设置为“单发攻击”,即攻击者在特定轮次进行 1 次后门注入,从而实现攻击效果。

表 1 数据集和注毒轮次规范

Table 1 Specifications of datasets and injection round

数据集	标签数/个	图像尺寸	图像数量/幅		注毒轮次/轮
			训练集	测试集	
CIFAR-10	10	32×32×3	50 000	10 000	302
GTSRB	43	32×32×3	21 312	5 328	102
PTSD	43	64×64×3	14 405	2 421	502

3.1.2 FL 设置

在 CIFAR-10、GTSRB 和 PTSD 3 个数据集上,为模拟非独立同分布(non-IID)的数据环境,本文通过狄利克雷分布^[39](超参数 $\alpha=0.5$)在 100 个客户端中分配训练图像。在训练干净样本时,采用随机梯度下降(SGD)优化器^[40],本地学习率设置为 0.1,批次大小为 64,进行 2 次迭代。而在训练毒样本时,学习率调整至 0.005,迭代 6 次,并实施 90%的学习率衰减,旨在提高联邦学习中后门攻击的持久性。通常情况下,投毒比例设置为 1/64,每轮随机选择 10 个客户端提交更新。此外,各数据集的目标标签分别设定为 CIFAR-10 的“鸟”,GTSRB 的“最大限速 50 km/h”和 PTSD 的“靠右行驶”。

3.1.3 可学习触发器设置

第 2 章介绍生成可学习触发器的过程,该过程包括优化扰动和触发器 2 个阶段。根据早期研究和大量实验经验,这 2 个阶段生成的图像主要受 3 个参数的影响:优化算法中的迭代次数 I ,更新过程中使用的学习率 lr 以及图像像素的扰动幅度,后者通过 L_∞ 范数球的半径(radius)来定义。除超参数敏感性分析之外,在其他实验中这些超参数取值都遵循现有工作的标准设置,如表 2 所示。

表 2 双目标优化阶段的超参数设置

Table 2 Hyperparameter settings for the dual-objective optimization stage

阶段	扰动半径	迭代次数/轮	学习率	样本规模
扰动生成	16/255	20	0.005	50
触发器构建	16/255	100	0.010	50

3.2 攻击表现

本文将提出的 LTBA 与现有的干净标签攻击

进行对比,特别是与 LC 和隐蔽式触发器后门攻击(HTBA)。这些攻击在联邦学习范式中部署,相应被记为 FL-LC 和 FL-HT。通过仅对目标类样本进行投毒并保留其原始标签,实现了 BadNets^[41]干净标签版本的攻击,记为 FL-BN-c。此外,为了进行全面的比较,实验中还展示了 BadNets 传统脏标签版本的攻击效果,该方法记为 FL-BN-d。

由于联邦学习后门攻击中触发器的多样性及参数设置的复杂性,与基线方法的比较存在一定的不公平性。这种比较之所以进行,是因为目前尚无研究专门针对本文提出的更为现实的威胁模型。此外,HTBA 支持的是“一对一攻击”,即误导模型仅将某个源类的输入错误地分类为目标类,其攻击成功率的计算仅基于单一源类的测试示例。相较而言,其他基线方法及 LTBA 采用的是“所有对一攻击”,即使用所有非目标类的测试示例来评估攻击成功率。

3.2.1 ASR 评估

表 3~表 5 所示为在 CIFAR-10、GTSRB 和 PTSD 数据集上本文的攻击方法与基线方法的比较结果,使用了 ACC、Tar-ACC 和 ASR 等指标。其中,带有“*”的数据表示 FL-HT 方法的 ASR。不同于其他方法,FL-HT 属于“一对一攻击”设置,其 ASR 是基于单源类测试样本来计算和评估得到的结果。

在联邦学习中,恶意客户端的注毒轮次采用了 1/64 的投毒比例(毒样本占有用户数据的不足 0.02%),远低于现有工作中通常提及的 0.5%~5%。此外,为确保实验的可复现性,用于生成触发器的训练样本从整个数据集中随机采样 5 次(使用的随机种子^[42]值为 0、42、55、123、2 023),用这些采样得到的数据生成可学习触发器,并激活 LTBA,最终计算攻击性能的均值。实验结果表明,随机种子的变化对攻击性能影响不大。除非特别说明,后续实验均采用默认随机种子值为 55 的配置。

表 3 在 CIFAR-10 数据集上的 ASR 比较

Table 3 ASR comparisons on CIFAR-10 dataset %

指标	FL-HT	FL-BN-c	FL-BN-d	FL-LC	LTBA
ACC	79.62	79.70	79.61	79.62	79.65
Tar-ACC	69.10	69.00	69.00	68.90	68.94
ASR	4.00 *	3.52	3.53	3.44	86.65

从表 3~表 5 可以看出,在联邦学习环境中,基线方法的攻击性能较弱,ASR 的最高值仅为 4%,主要是因为所选触发器与目标类之间联系不够紧密。与此相比,本文提出的 LTBA 方法在效果上明显优于这些基线方法,尤其是在 CIFAR-10 数据集上,其

ASR 高出最佳基线方法约 82 百分点,同时对主任务性能几乎没有影响。在所有数据集上,本实验均采用了低投毒比例(1/64)以最小化对 Tar-ACC 的影响,并采用 non-IID 数据来模拟联邦学习参与者的真实情况。为确保实验结果更具代表性、可靠性和统计学意义,本实验进行 5 次并取平均值作为结果。

表 4 在 GTSRB 数据集上的 ASR 比较

Table 4 ASR comparisons on GTSRB dataset %

指标	FL-HT	FL-BN-c	FL-BN-d	FL-LC	LTBA
ACC	95.91	95.80	95.81	95.95	95.84
Tar-ACC	97.33	97.33	97.33	97.33	97.33
ASR	0.38 *	0.18	0.18	0.20	68.83

表 5 在 PTSD 数据集上的 ASR 比较

Table 5 ASR comparisons on PTSD dataset %

指标	FL-HT	FL-BN-c	FL-BN-d	FL-LC	LTBA
ACC	87.40	87.44	87.36	87.48	87.43
Tar-ACC	92.80	92.80	92.80	93.60	92.80
ASR	0 *	0	0	0	37.82

如表 6 所示,LTBA 在训练和测试阶段均实施了最小的扰动,故可学习触发器的视觉效果没有基线方法中的明显。“Poison”一栏展示了在投毒(训练)阶段用于生成触发器的 L_∞ 范数球半径,“Test”一栏则表示测试阶段触发器对应的扰动半径。因此,LTBA 比各基线方法攻击性能更强大,即使用更小的扰动预算取得更高 ASR。

表 6 不同攻击中触发器的扰动预算比较

Table 6 Comparison of perturbation budgets for triggers in different attacks

扰动	FL-HT	FL-BN-c	FL-BN-d	FL-LC	LTBA
Poison	16/255	255/255	255/255	16/255	16/255
Test	255/255	255/255	255/255	255/255	48/255

3.2.2 攻击持久性评估

如近期的综述文献[43-44]所述,在联邦学习范式中,后门攻击的持久性与 ASR 同等重要。成功的后门攻击应在停止投毒后一段时间内,在全局模型中保持一定的后门效应,显示其抵抗中心聚合和后门遗忘的能力。本实验通过追踪 LTBA 在后门注入前后几轮内 ASR 的变化,尤其是关注 ASR 衰减速度,以此评估攻击的持久性。

第 2 章提到,在本研究中,恶意客户端使用第 T 轮的全局模型来生成可学习触发器,其中 T 在不同数据集上取值不同。CIFAR-10 中 $T=300$ 轮, GTSRB 中 $T=100$ 轮, PTSD 中 $T=500$ 轮。LTBA 采用了由 BAGDASARYAN 等^[19]提出的“单发攻击”策略,即通过聚合一轮恶意更新即可有效植入

后门。为避免被基于异常梯度的检测技术识别,本实验没有对恶意更新方法进行处理。实验结果显示,即便如此,LTBA 仍能实现令人满意的攻击效果。

如表 7~表 9 所示,攻击者在第 $(T+2)$ 轮启动后门攻击后,LTBA 在所有数据集上展示了稳定的攻击性能。具体而言,在注毒后的 50 轮内,ASR 几乎未见明显下降。值得注意的是,在第 $(T+1)$ 轮,即植入后门的前一轮,全局模型已经能以较高的 ASR 将含有触发器的测试样本错误地分类为目标类。这表明嵌入的可学习触发器使得测试样本在像素空间上虽然接近源类,在特征空间上却更接近目标类,从而误导模型将外观上类似于源类的图像预测为目标类,因此可以将该触发器视作一种具有对抗性的触发器。

表 7 在 CIFAR-10 数据集上 LTBA 的指标变化

Table 7 Change in metrics of LTBA on CIFAR-10 dataset %

指标	($T+1$)轮	($T+2$)轮	($T+10$)轮	($T+20$)轮	($T+50$)轮
ACC	79.64	79.52	79.87	79.35	78.87
Tar-ACC	66.80	67.20	65.10	70.00	69.80
ASR	86.66	86.26	83.40	86.72	89.89

表 8 在 GTSRB 数据集上 LTBA 的指标变化

Table 8 Change in metrics of LTBA on GTSRB dataset %

指标	($T+1$)轮	($T+2$)轮	($T+10$)轮	($T+20$)轮	($T+50$)轮
ACC	95.98	96.15	97.20	97.54	98.10
Tar-ACC	98.00	98.00	98.67	98.00	98.67
ASR	72.39	73.55	74.64	70.20	68.14

表 9 在 PTSD 数据集上 LTBA 的指标变化

Table 9 Change in metrics of LTBA on PTSD dataset %

指标	($T+1$)轮	($T+2$)轮	($T+10$)轮	($T+20$)轮	($T+50$)轮
ACC	87.44	87.86	86.66	88.60	88.52
Tar-ACC	92.80	93.60	92.00	92.80	92.80
ASR	35.45	35.67	38.55	36.62	35.58

3.2.3 可学习触发器对目标类的对抗性评估

触发器本质上是对图像中部分像素点取值的修改。为了直观展示这些修改带来的视觉差异,本文采用了残差图。同时,触发器的不可见性通过现有的图像质量指标进行量化评估。

为了证实前一节中关于“可学习触发器具有对抗性”的论断,本节还开展了一项补充实验,并进行了深入分析以支持该结论。首先,引入并解释了“具有对目标类对抗性的源类对抗样本”概念。通过具体实例阐述此类对抗样本的特性:1)原始图像属于

某一非目标类别(如 #1),其生成的对抗样本真实标签仍为 #1,因此称之为 #1 的对抗样本;2)添加对抗性扰动后,模型错误地将其预测为 #2,使得该样本对 #2 展现出对抗性,即对模型预测 #2 产生干扰,因此,将这种扰动后的图像称为“具有对 #2 对抗性的 #1 对抗样本”。这种对抗样本由攻击者设计,目的是实现 2 个主要目标:1)引起模型的误分类,降低模型的可用性;2)根据预设的目标操控模型输出,达到攻击者的特定目的。

从攻击者的视角分析,对抗样本可分为无目标对抗样本和有目标对抗样本两类。如前文所述,本文讨论的攻击携带有目标对抗样本,这种攻击相较于传统的无目标对抗样本攻击,具有更细的粒度和更强的针对性。其扰动的目的不仅在于降低样本被正确预测为其真实分类的可能性,还特意引导模型将样本预测为特定的错误类别,这构成了一种策略性的误导。此处提及的“源类”可以是非目标类中的任一类样本(这称为“一对一攻击”),也可涵盖非目标类的全部样本(这称为“所有对一攻击”)。

在补充实验中,通过将 LTBA($\epsilon=4$)与其他基线方法进行对比,本研究证实了嵌入可学习触发器的样本对目标类具有明显的对抗性,而这种对抗性在其他基线方法中几乎不存在。

本研究打破传统的实验模式,首次尝试直接使用残差图(仅含触发器,不含原始样本)作为输入,对已收敛的全局模型进行测试。实验结果表明,只包含可学习触发器的图像预测结果始终与攻击者设定的目标标签保持一致(这里以“#2”为例),而在其他基线方法中没有这种现象,说明从模型的视角来看,所提的触发器具有目标类的明显特征。在 CIFAR-10、GTSRB 和 PTSD 数据集上残差图分类结果与触发器隐蔽性比较如表 10~表 12 所示。从表 10~表 12 可以看出,LTBA 的触发器除有出色的攻击表现之外,对原始图像的修改也不明显,肉眼几乎不可见,显著提高了攻击的隐蔽性。

表 10 在 CIFAR-10 数据集上残差图分类结果与触发器隐蔽性比较

Table 10 Comparison of residual map classification results and trigger invisibility on CIFAR-10 dataset

方法	残差图	目标类	分类结果	SSIM	LPIPS
LTBA		#2	#2	0.94	0.16
BadNets		#2	#0	0.97	0.11
LC		#2	#4	0.88	0.24
HTBA		#2	#4	0.96	0.11

表 11 在 GTSRB 数据集上残差图分类结果与触发器隐蔽性比较

Table 11 Comparison of residual map classification results and trigger invisibility on GTSRB dataset

方法	残差图	目标类	分类结果	SSIM	LPIPS
LTBA		#2	#2	0.95	0.10
BadNets		#2	#18	0.97	0.10
LC		#2	#13	0.86	0.26
HTBA		#2	#18	0.89	0.20

表 12 在 PTSD 数据集上残差图分类结果与触发器隐蔽性比较

Table 12 Comparison of residual map classification results and trigger invisibility on PTSD dataset

方法	残差图	目标类	分类结果	SSIM	LPIPS
LTBA		#2	#2	0.97	0.05
BadNets		#2	#25	0.97	0.02
LC		#2	#28	0.85	0.14
HTBA		#2	#25	0.85	0.16

值得注意的是,从像素空间角度来看,这些只包含触发器的残差图并不属于任何已知类别,包括目标类。然而,从特征空间的结果显示,含可学习触发器的图像分类结果与攻击者指定的目标类高度对齐。这种“对齐”凸显了含可学习触发器测试样本对目标类强烈的对抗性,是 LTBA 方法攻击效果显著的核心原因。此发现在研究中建立了后门攻击与对抗样本攻击之间的新联系,为理解和抵御此类攻击提供了新的视角。

研究表明,含可学习触发器的样本所展现的目标类对抗性适用于任何目标类,具有普适性。本文在残差图分类实验中预定义了多个不同的目标类(#2, #3, #5, #7, #8)。实验结果表明,在所有基线实验中,用于测试的 20 个触发样本的分类结果最多只能与攻击者预设的目标类对齐 1 次。相对而言,通过预设目标类的少量样本与收敛后的全局模型训练,本研究提出 LTBA 方法生成的可学习触发器使得残差图分类结果几乎在每次实验中都能与目标类对齐。此外,可学习触发器不仅有出色的目标类对抗性,也有较优的隐蔽性。根据第 3.4 节中的不可见性指标数值可以得出这一结论。

3.3 超参数敏感性分析

本研究在多个数据集上展开对可学习触发器的超参数敏感性分析。根据经验,本节重点分析影响 LTBA 攻击性能的几个关键因素。这些因素包括:

1) 扰动后的目标类数据集规模 $|D_t^{\text{adv}}|$; 2) 扰动预算, 这里指的是以 L_∞ 范数球半径 ϵ' 定义的扰动半径; 3) 触发器生成阶段的迭代次数 I ; 4) 恶意客户端采用的投毒比例 r 。经过这些因素的综合分析, 有助于深入理解并优化可学习触发器的攻击效果。

3.3.1 $|D_t^{\text{adv}}|$ 的影响

表 13~表 15 所示为在 GTSRB、PTSD 和 CIFAR-10 上目标类小样本集规模对攻击性能的影响, 使用更多扰动后的目标类样本进行训练能在前期提升 LTBA 的攻击性能。然而, 样本数量的增加并不会继续提升攻击效果, 反而增加了训练触发器的成本。因此, 在训练可学习触发器的过程中应权衡训练样本的数量。

表 13 在 GTSRB 上目标类小样本集规模对攻击性能的影响

指标	$ D_t^{\text{adv}} =10$	$ D_t^{\text{adv}} =20$	$ D_t^{\text{adv}} =30$	$ D_t^{\text{adv}} =50$	$ D_t^{\text{adv}} =100$
ASR	65.29	68.28	70.94	72.77	77.31
ACC	95.91	95.81	95.76	95.78	95.83
Tar-ACC	97.33	97.33	97.33	97.33	97.33

表 14 在 PTSD 上目标类小样本集规模对攻击性能的影响

指标	$ D_t^{\text{adv}} =10$	$ D_t^{\text{adv}} =20$	$ D_t^{\text{adv}} =30$	$ D_t^{\text{adv}} =50$	$ D_t^{\text{adv}} =100$
ASR	32.53	32.62	32.06	34.54	49.74
ACC	87.44	87.40	87.44	87.44	87.44
Tar-ACC	92.80	92.80	92.80	92.80	92.80

表 15 在 CIFAR-10 上目标类小样本集规模对攻击性能的影响

指标	$ D_t^{\text{adv}} =10$	$ D_t^{\text{adv}} =20$	$ D_t^{\text{adv}} =30$	$ D_t^{\text{adv}} =50$	$ D_t^{\text{adv}} =100$
ASR	70.50	82.74	83.72	88.52	85.92
ACC	79.65	79.63	79.63	79.68	79.64
Tar-ACC	69.00	69.00	69.00	69.00	69.00

3.3.2 ϵ' 的影响

表 16~表 18 所示为在 GTSRB、PTSD、CIFAR-10 上扰动半径对攻击性能的影响。初始增加扰动半径 ϵ' 能显著提升 LTBA 的攻击性能。然而, 随着 ϵ' 进一步增加, 这种增强效果会逐渐减弱, 主要是因为 ASR 已接近上限。持续扩大扰动半径使得触发器更加明显, 从而易于被人类观察者发现。关于触发器不可见性的进一步讨论, 详见第 3.4 节。

表 16 在 GTSRB 上扰动半径对攻击性能的影响

指标	$\epsilon'=2$	$\epsilon'=4$	$\epsilon'=8$	$\epsilon'=12$	$\epsilon'=16$	$\epsilon'=20$
ASR	29.77	54.22	71.60	72.80	72.77	72.91
ACC	95.85	95.80	95.74	95.93	95.78	95.80
Tar-ACC	97.33	97.33	97.33	97.33	97.33	97.33

表 17 在 PTSD 上扰动半径对攻击性能的影响

指标	$\epsilon'=2$	$\epsilon'=4$	$\epsilon'=8$	$\epsilon'=12$	$\epsilon'=16$	$\epsilon'=20$
ASR	3.70	17.77	34.19	34.54	34.54	34.54
ACC	87.44	87.40	87.40	87.44	87.44	87.40
Tar-ACC	92.80	92.80	92.80	92.80	92.80	92.80

表 18 在 CIFAR-10 上扰动半径对攻击性能影响

指标	$\epsilon'=2$	$\epsilon'=4$	$\epsilon'=8$	$\epsilon'=12$	$\epsilon'=16$	$\epsilon'=20$
ASR	20.22	61.51	85.00	85.93	88.52	87.64
ACC	79.64	79.64	79.65	79.65	79.68	79.66
Tar-ACC	69.00	69.00	69.00	69.00	69.00	69.00

3.3.3 I 的影响

表 19~表 21 所示为在 GTSRB、PTSD、CIFAR-10 上迭代次数对攻击性能的影响。增加迭代次数 I 确实能够提升 LTBA 的攻击表现。然而, 考虑到触发器生成算法的训练成本, 这里并不提倡使用过高的迭代次数。尤其是当训练样本的数量较小时, 使用过大的迭代次数不仅延长生成触发器的时间, 还可能导致生成结果收敛至局部最优解, 从而削弱 LTBA 的攻击效果。因此, 在训练可学习触发器的过程中应选择大小适当的迭代次数, 以权衡攻击性能和攻击方法的计算成本。大量实验结果表明, 迭代次数取 100~150 攻击性能可以保证较理想的性能。

表 19 在 GTSRB 上迭代次数对攻击性能的影响

指标	$I=50$	$I=75$	$I=100$	$I=125$	$I=150$
ASR	59.86	68.70	72.77	75.34	77.01
ACC	95.95	95.83	95.78	95.89	95.80
Tar-ACC	97.33	97.33	97.33	97.33	97.33

表 20 在 PTSD 上迭代次数对攻击性能的影响

指标	$I=50$	$I=75$	$I=100$	$I=125$	$I=150$
ASR	20.73	30.23	34.54	37.98	41.07
ACC	87.44	87.40	87.44	87.40	87.44
Tar-ACC	92.80	92.80	92.80	92.80	92.80

表 21 在 CIFAR-10 上迭代次数对攻击性能的影响

Table 21 Impact of iteration count on attack performance on CIFAR-10 %

指标	$I=50$	$I=75$	$I=100$	$I=125$	$I=150$
ASR	69.15	84.63	88.52	88.50	89.21
ACC	79.63	79.64	79.68	79.65	79.68
Tar-ACC	69.00	69.00	69.00	69.00	69.00

3.3.4 r 的影响

表 22~表 24 所示为在 GTSRB、PTSD、CIFAR-10 上投毒比例对攻击性能的影响。在投毒比例为 1/64 时,LTBA 取得了高攻击成功率,进一步提高投毒比例并未显著增强攻击性能,反而因为异常样本占比增多可能导致攻击被揭露的风险增加。

表 22 在 GTSRB 上投毒比例对攻击性能的影响

Table 22 Impact of poisoning ratio on attack performance on GTSRB %

指标	$r=\frac{1}{64}$	$r=\frac{1}{32}$	$r=\frac{1}{16}$	$r=\frac{1}{8}$	$r=\frac{1}{4}$
ASR	72.77	72.75	72.93	72.77	72.67
ACC	95.78	95.83	95.89	95.91	95.85
Tar-ACC	97.33	97.33	97.33	97.33	97.33

表 23 在 PTSD 上投毒比例对攻击性能的影响

Table 23 Impact of poisoning ratio on attack performance on PTSD %

指标	$r=\frac{1}{64}$	$r=\frac{1}{32}$	$r=\frac{1}{16}$	$r=\frac{1}{8}$	$r=\frac{1}{4}$
ASR	34.54	34.54	34.54	34.54	34.54
ACC	87.44	87.40	87.40	87.40	87.40
Tar-ACC	92.80	92.80	92.80	92.80	92.80

表 24 CIFAR-10 上投毒比率对攻击性能的影响

Table 24 Impact of poisoning ratio on attack performance on CIFAR-10 %

指标	$r=\frac{1}{64}$	$r=\frac{1}{32}$	$r=\frac{1}{16}$	$r=\frac{1}{8}$	$r=\frac{1}{4}$
ASR	88.52	88.69	87.43	89.07	88.72
ACC	79.68	79.62	79.61	79.67	79.55
Tar-ACC	69.00	68.70	68.40	69.00	68.80

3.4 可学习触发器的不可见性评估

近期的后门攻击研究,如基于分散像素扰动的后门触发器研究,突出了触发器不可见性作为评估指标的重要性,其目的是尽量减少触发器对人眼的可见性。触发器的不可见性不仅可以通过残差图^[45]进行定性评估,还可以借助如结构相似性指数(SSIM)^[46]、学习感知图像块相似度(LPIPS)^[47]等指标进行定量评估。本节详细评估了可学习触发器的不可见性,扰动半径对不可见性的影响,以及 ASR 与不可见性之间的权衡。

3.4.1 可学习触发器的不可见性

表 25 对 3 个数据集上使用随机选取的 #2 和 #3 两类干净样本生成的可学习触发器的不可见性进行了评估,其中扰动半径 $\epsilon'=16$ 。从表 25 可以看出,可学习触发器具有良好的不可见性。干净图像与其对应的中毒图像之间的差异几乎不可见,这在几乎全黑的残差图上得到了体现。此外,SSIM 的值接近 1 和 LPIPS 的值接近 0,也表明两者在感知上具有较高的相似性。

表 25 在 3 个数据集上可学习触发器的不可见性评估

Table 25 Invisibility assessment of learnable triggers on three datasets

数据集	标签	原图	毒图	残差图	SSIM $\uparrow$	LPIPS $\downarrow$
CIFAR-10	#2				0.83	0.29
	#3				0.89	0.27
GTSRB	#2				0.92	0.14
	#3				0.90	0.21
PTSD	#2				0.94	0.09
	#3				0.91	0.15

3.4.2 ϵ' 的影响

可学习触发器的不可见性会受到生成过程中扰动半径 ϵ' 影响。在前期增加 ϵ' 会轻微降低触发器的不可见性,但继续增加后效果趋于平稳。即使 ϵ' 增加至 20,触发器的隐蔽性仍然保持较高水平,如表 26~表 28 所示。这是因为当扰动半径增加到一定程度时,含触发器的样本几乎不再有明显变化。因此,扰动越小,触发器的不可见性越好。这里额外说明,因图像尺寸不一致,在实验前将 PTSD 数据集所有样本预处理成统一规格 64×64 ,而其他数据集为 32×32 。显然,相似的扰动幅度在小尺寸图像上更易引起肉眼察觉。因此,相似触发器对 PTSD 图像产生的视觉差异不如其他数据集明显。从更大的 SSIM 值和更小的 LPIPS 值上,也能说明这一点。

表 26 在 PTSD 上扰动半径对可学习触发器不可见性的影响

Table 26 Effect of perturbation radius on learnable triggers' invisibility on PTSD

ϵ'	毒图	残差图	SSIM $\uparrow$	LPIPS $\downarrow$
2			0.99	0.01
4			0.97	0.05
8			0.94	0.09
12			0.94	0.09
16			0.94	0.09
20			0.94	0.09

表 27 在 CIFAR-10 上扰动半径对触发器不可见性的影响

Table 27 Effect of perturbation radius on triggers' invisibility on CIFAR-10

ϵ'	毒图	残差图	SSIM $\uparrow$	LPIPS $\downarrow$
2			0.98	0.06
4			0.93	0.16
8			0.85	0.28
12			0.83	0.30
16			0.83	0.29
20			0.83	0.29

表 28 在 GTSRB 上扰动半径对触发器不可见性的影响

Table 28 Effect of perturbation radius on triggers' invisibility on GTSRB

ϵ'	毒图	残差图	SSIM $\uparrow$	LPIPS $\downarrow$
2			0.99	0.04
4			0.95	0.10
8			0.92	0.14
12			0.92	0.14
16			0.92	0.14
20			0.92	0.14

3.4.3 ASR 与不可见性的权衡

表 29~表 31 展示了各数据集上各基线方法及 LTBA 的综合性能评估结果。各基线方法均遵循原文设置实现,如 BadNets 中透明度参数 $T=0$ (不透明),LC 和 HTBA 中扰动半径 $\epsilon=16$ 。通过观察 LTBA 在不同扰动半径 ϵ' 上各项性能指标的变化,发现 ASR 与不可见性间存在权衡。当 $\epsilon'=2$ 时,LTBA 的攻击性能和隐蔽性优于所有的基线方法,但 ASR 不是特别高。当 $\epsilon'=4$ 时,则能够同时兼顾较高的 ASR 和隐蔽性。

表 29 在 GTSRB 上 LTBA 与各基线方法综合性能比较

Table 29 Comprehensive performance comparison of LTBA with each baseline method on GTSRB

方法	参数	ACC/%	Tar-ACC/%	ASR/% $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FL-BN	$T=0$	95.80	97.33	0.18	0.97	0.10
FL-LC	$\epsilon=16$	95.95	97.33	0.20	0.86	0.26
FL-HT	$\epsilon=16$	95.91	97.33	0.38	0.89	0.20
LTBA	$\epsilon'=2$	95.85	97.33	29.77	0.99	0.04
	$\epsilon'=4$	95.80	97.33	54.21	0.95	0.10
	$\epsilon'=8$	95.74	97.33	71.60	0.92	0.14
	$\epsilon'=12$	95.93	97.33	72.79	0.92	0.14
	$\epsilon'=16$	95.78	97.33	72.77	0.92	0.14
	$\epsilon'=20$	95.80	97.33	72.91	0.92	0.14

表 30 在 CIFAR-10 上 LTBA 与各基线方法综合性能比较

Table 30 Comprehensive performance comparison of LTBA with each baseline method on CIFAR-10

方法	参数	ACC/%	Tar-ACC/%	ASR/% $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FL-BN	$T=0$	79.62	69.10	4.00	0.97	0.11
FL-LC	$\epsilon=16$	79.62	68.90	3.44	0.88	0.24
FL-HT	$\epsilon=16$	79.70	69.00	3.52	0.96	0.11
LTBA	$\epsilon'=2$	79.64	69.00	20.22	0.98	0.05
	$\epsilon'=4$	79.64	69.00	61.51	0.94	0.16
	$\epsilon'=8$	79.65	69.00	85.00	0.85	0.28
	$\epsilon'=12$	79.65	69.00	85.93	0.83	0.30
	$\epsilon'=16$	79.68	69.00	85.52	0.83	0.29
	$\epsilon'=20$	79.66	69.00	87.64	0.83	0.29

表 31 在 PTSD 上 LTBA 与各基线方法综合性能比较

Table 31 Comprehensive performance comparison of LTBA with each baseline method on PTSD

方法	参数	ACC/%	Tar-ACC/%	ASR/% $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
FL-BN	$T=0$	87.44	92.80	0	0.97	0.02
FL-LC	$\epsilon=16$	87.48	93.60	0	0.85	0.14
FL-HT	$\epsilon=16$	87.40	92.80	0	0.84	0.16
LTBA	$\epsilon'=2$	87.44	92.80	3.70	0.99	0.01
	$\epsilon'=4$	87.40	92.80	17.77	0.97	0.05
	$\epsilon'=8$	87.40	92.80	34.19	0.94	0.09
	$\epsilon'=12$	87.44	92.80	34.54	0.94	0.09
	$\epsilon'=16$	87.44	92.80	34.54	0.94	0.09
	$\epsilon'=20$	87.40	92.80	34.54	0.94	0.09

3.5 攻击有效性的广泛验证

本文提出的攻击方法通用性较强,即在不同模型和数据集上均有出色表现。为了证明该结论,本节给出了更广泛的验证。

3.5.1 在模型 LeNet-5 和 ResNet-34 上的攻击性能评估

除了 ResNet-18 以外,本小节分别在 1 个更简单的模型(LeNet-5)和更复杂的模型(ResNet-34)上评估了 LTBA 的性能,如表 32 所示。受篇幅限制,这里仅展示数据集 GTSRB 上的实验结果,其他设置与主实验对齐。

表 32 在模型 LeNet-5 和 ResNet-34 上的攻击性能评估

Table 32 Attack performance evaluation on models LeNet-5 and ResNet-34

Model	ASR	ACC	Tar-ACC
ResNet-18	68.83	95.84	97.33
LeNet-5	84.57	83.52	67.67
ResNet-34	75.97	96.58	98.67

通过对比这些数据和 ResNet-18 上的实验结果可知,LTBA 方法在不同复杂程度的模型上均能表现出良好的攻击性能,对模型具有较强的通用性。

3.5.2 数据集 CIFAR-100 上的攻击性能评估

本节在 CIFAR-100 上评估了 LTBA 的攻击表现,以探索该方法在更多标签分类任务中的有效性。这里使用模型 ResNet-18 进行实验,不修改基准设置,结果如表 33 所示。

表 33 在数据集 CIFAR-100 上的攻击性能评估

Dataset	ASR	ACC	Tar-ACC
CIFAR-10	86.65	79.65	68.94
GTSRB	68.83	95.84	97.33
PTSD	37.82	87.43	92.80
CIFAR-100	78.60	67.67	60.00

通过对比 CIFAR-100 和之前的实验结果可知,LTBA 方法在更多标签分类任务中的表现依然良好,对各种分类任务有较强的适用性。

3.5.3 独立同分布数据上的攻击性能评估

鉴于先前的实验一直在 non-IID 数据上进行验证,本节补充了 LTBA 在独立同分布(IID)数据上的性能评估,实验结果如表 34 所示。这里展示了在数据集 GTSRB 模型 ResNet-18 上的攻击评估结果,除狄利克雷参数改为 $\alpha=1$ 以外,其他设置不变。

表 34 独立同分布数据上的攻击性能评估

α	ASR/%	ACC/%	Tar-ACC/%
0.5 (non-IID)	68.83	95.84	97.33
1 (IID)	74.67	96.72	97.67

通过对比两组数据发现,LTBA 在 IID 数据上有类似的攻击表现甚至更加出色,这一结论与之前大量同类型工作得出的结论一致。

3.5.4 不同客户端数量上的攻击性能评估

本小节分别选取不同数量的客户端进行实验,来揭示客户端总数对 LTBA 攻击性能的影响。这里展示了客户端总数分别为 50、200 个时,LTBA 在数据集 GTSRB 模型 ResNet-18 上的攻击表现,如表 35 所示,不修改基准设置。

通过上述 3 组数据的对比可知,常规的客户端数量增减对 LTBA 的攻击性能影响并不明显。这是因为全局模型中的后门效应主要取决于参与聚合的客户端中是否包含攻击者,而与客户端总数并无

直接关系。基准设置中保证参与每轮聚合的攻击者数量一致,所以整体的攻击表现没有明显差异。

表 35 不同客户端总数的攻击性能评估

Table 35 Attack performance evaluation with different total number of clients

客户端总数/个	ASR/%	ACC/%	Tar-ACC/%
100	68.83	95.84	97.33
50	72.61	98.03	99.33
200	78.16	90.00	77.31

3.6 LTBA 的鲁棒性分析

本小节探讨了 LTBA 在实际的联邦学习场景中对后门防御的抵抗性。目前的后门防御方法主要分为两大类:一类是基于异常检测的防御方法;另一类则是基于鲁棒聚合的防御方法。

近期部分工作指出,在实际的联邦学习场景中,异常检测方法尚难以有效部署,主要原因如下:1)联邦学习的安全聚合机制限制了对更新的直接访问,配合先进的隐私保护技术,使得异常更新不易被识别;2)若联邦学习系统未采用同态加密、差分隐私等安全措施,将面临更多安全威胁,例如推理攻击和重建攻击,这些风险严重威胁到系统的机密性,违背了保护参与者数据隐私的初衷;3)从理论和实践的角度分析,异常检测方法可能导致模型产生显著偏差,在 non-IID 的数据环境中,可能会误筛选掉许多离群但良性的样本,从而引发严重的不公平现象。

与异常检测方法不同,基于鲁棒聚合的防御策略在训练过程中直接减轻后门攻击的影响。传统方法如拜占庭容错聚合^[48-49],由于对数据分布及攻击目标的错误假设^[50-51],导致其与联邦学习范式不兼容^[52]。同时,其他策略如随机平滑,差分隐私等^[53-54],其聚合机制不够完善^[55],无法全面防御后门攻击,并可能因削弱部分良性更新而导致全局模型的主任务性能下降,需要在主任务性能和后门防御能力之间做出权衡,限制其实际应用效果。

此外,实验还追踪了恶意客户端本地模型在测试阶段的损失变化情况。实验结果表明,在投毒前后模型损失值并未发生显著变化,这表明恶意更新不太可能表现为显著的异常值。因此,大多数基于异常检测的防御策略难以奏效。

总体而言,现有的后门防御方法在实际的联邦学习场景中难以为 LTBA 提供彻底和实际的保护,这进一步凸显了 LTBA 的强鲁棒性。

4 结束语

本文提出的 LTBA 是一种新型干净标签后门

攻击,在联邦学习范式中攻击效果显著。通过大量实验已证明 LTBA 的攻击性能远超过传统基线方法。此外,本文设计的可学习触发器具有较高的不可见性。攻击者利用 FL 聚合器发布的任务信息及共享的全局模型,针对预设的目标类生成该触发器。本文还构建了双目标优化问题,并采用两阶段投影梯度下降方法求解,旨在建立预设目标类与生成触发器之间的强连接关系,以抵御中心聚合和后门遗忘对攻击效果的削弱。实验结果表明,所提方法具有较优的攻击成功率。下一步将在更复杂的范式中开发和改进所提出的攻击方法,如纵向联邦学习或联邦迁移学习范式。

参考文献

- [1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1602.05629>.
- [2] 吴若岚, 陈玉玲, 豆慧, 等. 抗攻击的联邦学习隐私保护算法[J]. 计算机工程, 2025, 51(2): 179-187.
WU R L, CHEN Y L, DOU H, et al. Privacy preserving algorithm using federated learning against attacks [J]. Computer Engineering, 2025, 51(2): 179-187. (in Chinese)
- [3] 熊世强, 何道敬, 王振东, 等. 联邦学习及其安全与隐私保护研究综述[J]. 计算机工程, 2024, 50(5): 1-15.
XIONG S Q, HE D J, WANG Z D, et al. Review of federated learning and its security and privacy protection [J]. Computer Engineering, 2024, 50(5): 1-15. (in Chinese)
- [4] 张世文, 陈双, 梁伟, 等. 联邦学习中的攻击手段与防御机制研究综述[J]. 计算机工程与应用, 2024, 60(5): 1-16.
ZHANG S W, CHEN S, LIANG W, et al. Survey on attack methods and defense mechanisms in federated learning [J]. Computer Engineering and Applications, 2024, 60(5): 1-16. (in Chinese)
- [5] ZHAO Y, LI M, LAI L Z, et al. Federated learning with non-IID data [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1806.00582>.
- [6] WANG Y L, ZHAO M H, LI S H, et al. Dispersed pixel perturbation-based imperceptible backdoor trigger for image classifier models [J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3091-3106.
- [7] NING R, LI J, XIN C S, et al. Hibernated backdoor: a mutual information empowered backdoor attack to deep neural networks [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2022: 10309-10318.
- [8] LI S F, XUE M H, ZHAO B Z H, et al. Invisible backdoor attacks on deep neural networks via steganography and regularization [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(5): 2088-2105.
- [9] ZHAO S H, MA X J, ZHENG X, et al. Clean-label backdoor attacks on video recognition models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2020: 14431-14440.
- [10] GAN L L, LI J W, ZHANG T W, et al. Triggerless backdoor attack for NLP tasks with clean labels [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2111.07970>.
- [11] BARNI M, KALLAS K, TONDI B. A new backdoor attack in CNNs by training set corruption without label poisoning [C]//Proceedings of the IEEE International Conference on Image Processing (ICIP). Washington D.C., USA: IEEE Press, 2019: 101-105.
- [12] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1912.02771>.
- [13] LIN D Q, GUO Y C, SUN H, et al. FedCluster: a federated learning framework for cross-device private ECG classification [C]//Proceedings of the IEEE INFOCOM Conference on Computer Communications Workshops. Washington D.C., USA: IEEE Press, 2022: 1-6.
- [14] CAO X Y, GONG N Z. MPAF: model poisoning attacks to federated learning based on fake clients [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Washington D.C., USA: IEEE Press, 2022: 3395-3403.
- [15] SAHA A, SUBRAMANYA A, PIRSIYAVASH H. Hidden trigger backdoor attacks [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020: 11957-11965.
- [16] CHELLAPPA R, FOWL L, GOLDBLUM M, et al. Sleeper agent: scalable hidden trigger backdoors for neural networks trained from scratch [C]//Proceedings of the Advances in Neural Information Processing Systems 35. Louisiana, USA: Neural Information Processing Systems Foundation, Inc, 2022: 19165-19178.
- [17] NING R, LI J, XIN C S, et al. Invisible poison: a blackbox clean label backdoor attack to deep neural networks [C]//Proceedings of the IEEE Conference on Computer Communications. Washington D.C., USA: IEEE Press, 2021: 1-10.
- [18] QUIRING E, RIECK K. Backdooring and poisoning neural networks with image-scaling attacks [C]//Proceedings of the IEEE Security and Privacy Workshops (SPW). Washington D.C., USA: IEEE Press, 2020: 41-47.
- [19] BAGDASARYAN E, VEIT A, HUA Y Q, et al. How to backdoor federated learning [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1807.00459>.
- [20] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1811.12470>.
- [21] XIE C L, HUANG K L, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning [C]//Proceedings of International Conference on Learning Representations. New York, USA: [s.n.], 2019: 1-10.
- [22] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: system design [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1902.01046>.
- [23] ZENG Y, PAN M Z, JUST H A, et al. Narcissus: a practical clean-label backdoor attack with limited information [C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2023: 771-785.
- [24] SHANNON C E. Communication theory of secrecy systems [J]. Bell System Technical Journal, 1949, 28(4): 656-715.
- [25] LI Y M. Poisoning-based backdoor attacks in computer vision [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 16121-16122.
- [26] JIANG W B, LI H W, XU G W, et al. Color backdoor: a robust poisoning attack in color space [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2023: 8133-8142.
- [27] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1712.05526>.

- [28] ZHONG H T, LIAO C, SQUICCIARINI A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[C]//Proceedings of the 10th ACM Conference on Data and Application Security and Privacy. New York, USA: ACM Press, 2020: 97-108.
- [29] LI Z S, CHEN M M, HE Y F, et al. An efficient framework for detection and recognition of numerical traffic signs[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D.C., USA: IEEE Press, 2022: 2235-2239.
- [30] BAN Y H, DONG Y P. Pre-trained adversarial perturbations[C]//Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. 2022: 1196-1209.
- [31] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1706.06083>.
- [32] ZHANG Z, PANDA A, SONG L, et al. Neurotoxin: durable backdoors in federated learning[C]//Proceedings of International Conference on Machine Learning. New York, USA: [s.n], 2022: 26429-26446.
- [33] LYU X T, HAN Y F, WANG W, et al. Poisoning with cerberus: stealthy and colluded backdoor attack against federated learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 9020-9028.
- [34] ANTONANTE P, VEER S, LEUNG K, et al. Task-aware risk estimation of perception failures for autonomous vehicles [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2305.01870>.
- [35] PASZKE A, LERER A, KILLEN T, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Proceedings of Conference on Advances in Neural Information Processing Systems. New York, USA: [s.n], 2019, 32.
- [36] RECHT B, ROELOFS R, SCHMIDT L, et al. Do CIFAR-10 classifiers generalize to CIFAR-10? [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1806.00451>.
- [37] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German traffic sign recognition benchmark: a multi-class classification competition [C]//Proceedings of the 2011 International Joint Conference on Neural Networks. Washington D.C., USA: IEEE Press, 2011: 1453-1460.
- [38] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2016: 770-778.
- [39] 万志成, 郑静. 基于狄利克雷过程高斯混合模型的变分推断[J]. 杭州电子科技大学学报(自然科学版), 2021, 41(5): 54-61.
- WAN Z C, ZHEGN J. Variational inference for Gaussian mixture model based on Dirichlet process[J]. Journal of Hangzhou University (Natural Sciences), 2021, 41(5): 54-61. (in Chinese)
- [40] KESKAR N S, SOCHER R. Improving generalization performance by switching from Adam to SGD [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1712.07628>.
- [41] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1708.06733>.
- [42] COLAS C, SIGAUD O, OUDEYER P Y. How many random seeds? statistical power analysis in deep reinforcement learning experiments [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1806.08295>.
- [43] NGUYEN T D, NGUYEN T, LE NGUYEN P, et al. Backdoor attacks and defenses in federated learning: survey, challenges and future research directions [J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107166.
- [44] GONG X, CHEN Y, WANG Q, et al. Backdoor attacks and defenses in federated learning: state-of-the-art, taxonomy, and future directions [J]. IEEE Wireless Communications, 2022, 30(2): 114-121.
- [45] SWARTZ C A. Some geometrical properties of residual maps [J]. Geophysics, 1954, 19(1): 46-70.
- [46] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
- [47] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2018: 586-595.
- [48] BLANCHARD P, MHAMDI E M E, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent [EB/OL]. [2025-02-10]. https://proceedings.neurips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.
- [49] CHEN Y D, SU L L, XU J M. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent[C]//Proceedings of the ACM on Measurement and Analysis of Computing Systems. New York, USA: ACM Press, 2017: 1-25.
- [50] DAMASKINOS G, MHAMDI E M E, GUERRAOUI R, et al. Asynchronous Byzantine machine learning (the case of SGD) [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1802.07928>.
- [51] XIE C, KOYEJO O, GUPTA I. Zeno: distributed stochastic gradient descent with suspicion-based fault-tolerance [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1805.10032>.
- [52] YIN D, CHEN Y D, KANNAN R, et al. Byzantine-robust distributed learning: towards optimal statistical rates [EB/OL]. [2025-02-10]. <https://arxiv.org/pdf/1803.01498>.
- [53] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1712.07557>.
- [54] MCMAHAN H B, RAMAGE D, TALWAR K, et al. Learning differentially private recurrent language models [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1710.06963>.
- [55] SUN Z T, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning? [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1911.07963>.

文字编辑 薛晋栋
栏目编辑 赖玉玲