计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20210659
58(11):2364-2373,2021

基于生成式对抗网络的联邦学习后门攻击方案

陈大卫[1,2] 付安民[1,2] 周纯毅[1] 陈珍珠[1]
[1](南京理工大学计算机科学与工程学院 南京 210094)
[2](信息安全国家重点实验室(中国科学院信息工程研究所) 北京 100093)
(894346698@qq.com)

Federated Learning Backdoor Attack Scheme Based on Generative Adversarial Network

Chen Dawei[1,2], Fu Anmin[1,2], Zhou Chunyi[1], and Chen Zhenzhu[1]
[1](*School of Computer Science and Engineering, Nanjing University of Science & Technology, Nanjing* 210094)
[2](*State Key Laboratory of Information Security* (*Institute of Information Engineering, Chinese Academy of Sciences*), *Beijing* 100093)

Abstract Federated learning enables users to participate in collaborative model training while keeping their data in local, which ensures the privacy and security of users' data. It has been widely used in smart finance, smart medical and other fields. However, federated learning shows inherent vulnerability to backdoor attacks, where the attacker implants the backdoor by uploading the model parameters. Once the global model recognizes the input with the trigger, it will misclassify the input as the label specified by the attacker. This paper proposes a new federated learning backdoor attack scheme, Bac_GAN. By combining generative adversarial network, triggers are implanted in clean samples in the form of watermarks, which reduces the discrepancy between trigger features and clean sample features, and enhance the imperceptibility of triggers. By scaling the backdoor model, the problem of offsetting the contribution of the backdoor during parameter aggregation is avoided, so that the backdoor model can converge in a short time, thus significantly increasing the attack success rate. In addition, we conduct experimental tests on the core elements of backdoor attacks, such as trigger generation, watermark coefficient and scaling coefficient, and give the best parameters that affect the performance of backdoor attack. Also, we validate the attack effectiveness of the Bac_GAN scheme on MNIST and CIFAR-10.

Key words federated learning; generative adversarial network; backdoor attack; trigger; watermark

摘 要 联邦学习使用户在数据不出本地的情形下参与协作式的模型训练,降低了用户数据隐私泄露风险,广泛地应用于智慧金融、智慧医疗等领域.但联邦学习对后门攻击表现出固有的脆弱性,攻击者通过上传模型参数植入后门,一旦全局模型识别带有触发器的输入时,会按照攻击者指定的标签进行误分类.

收稿日期:2021-06-11;**修回日期**:2021-07-30
基金项目:国家自然科学基金项目(62072239);信息安全国家重点实验室开放基金项目(2021-MS-07);中央高校基本科研业务费专项资金(30920021129,30921013111)
This work was supported by the National Natural Science Foundation of China (62072239), the Open Foundation of the State Key Laboratory of Information Security of China (2021-MS-07), and the Fundamental Research Funds for the Central Universities (30920021129, 30921013111).
通信作者:付安民(fuam@njust.edu.cn)

因此针对联邦学习提出了一种新型后门攻击方案 Bac_GAN,通过结合生成式对抗网络技术将触发器以水印的形式植入干净样本,降低了触发器特征与干净样本特征之间的差异,提升了触发器的隐蔽性,并通过缩放后门模型,避免了参数聚合过程中后门贡献被抵消的问题,使得后门模型在短时间内达到收敛,从而显著提升了后门攻击成功率.此外,论文对触发器生成、水印系数、缩放系数等后门攻击核心要素进行了实验测试,给出了影响后门攻击性能的最佳参数,并在 MNIST,CIFAR-10 等数据集上验证了 Bac_GAN 方案的攻击有效性.

关键词 联邦学习;生成式对抗网络;后门攻击;触发器;水印

中图法分类号 TP391

联邦学习[1]将深度学习模型与分布式训练相结合,使得多方用户在不共享数据的情况下,协同参与训练全局模型,降低了传统集中式学习中的用户隐私泄露风险和通信开销[2],从技术层面可以打破数据孤岛,明显提高深度学习的性能,能够实现多个领域的落地应用,比如智慧医疗、智慧金融、智慧零售和智慧交通等[3-4].联邦学习作为大数据使用的新范式,是破解数据隐私保护与数据孤岛难题的新思路,一经提出就成为国际学术界和产业界关注的焦点.

海量的用户数据、丰富的应用场景促进了联邦学习技术的蓬勃发展,但联邦学习数以万计的用户中可能存在恶意用户,并且用户的本地训练过程对于服务器不可见,服务器无法验证用户更新的正确性[5-7],特别是服务器采用加权平均算法对参数进行更新,限制了异常检测的使用,这些缺陷的存在使得联邦学习框架极易遭受投毒攻击[8]、对抗样本攻击[9-11]和后门攻击[12].

后门攻击通过在携带触发器的数据上训练模型,使得模型能够在不包含触发器的场景下正常进行分类任务.但当模型识别到触发器的特征时,后门被激发,模型会定向分类攻击者指定的标签.后门攻击的本质在于模型对于触发器的特征过拟合,当神经网络中神经元同时遇到普通特征与触发器特征时,会优先分类为触发器特征对应的标签.Yao 等人[13]在网络上发布植入后门的预训练模型,并通过迁移学习感染用户模型,提高了后门攻击的隐蔽性.Liu 等人[14]研究不同触发器与神经网络中内在神经元之间的联系,建立起触发器与关键神经元之前的强联系,模型识别触发器时,所选神经元就会触发,导致伪装式输出.这使得后门攻击的隐蔽性与危害性得到了显著的提升.

后门攻击在集中式学习与联邦学习中的实现方式有所差异,集中式学习中攻击者一般在本地训练后门模型,将其上传至云端服务器,供用户下载,从而实现后门攻击.而在联邦学习中,攻击者在本地将触发器植入训练数据生成后门样本数据集,在每个训练批次内同时训练本地样本以及后门样本,从而训练后门模型.后门模型经过参数聚合过程,逐渐影响整个联邦环境下用户的本地模型,因此危害面更广.Bagdasaryan 等人[15]首次提出了针对联邦学习的后门攻击,攻击者上传植入后门的模型,并通过优化算法缩放本地模型权重,进而替换全局模型.当全局模型在执行分类任务时,会按照攻击者指定的目标标签进行分类.Xie 等人[16]在现有的攻击方式上,提出了分布式后门攻击(distributed backdoor attack, DBA),该方案采用多个触发器植入后门,多个局部触发器组合成全局触发器,提升了后门攻击的精度.

针对存在的后门攻击,研究人员也提出了不同防御手段[17],包括神经元检测、触发器重构技术等.Liu 等人[18]采用人工脑刺激(artificial brain stimulation, ABS)技术识别对特定标签有较高异常值的关键神经元,能成功检测出模型中是否存在隐藏的后门.Wang 等人[19]通过重构触发器来识别后门模型.针对所有输出标签类,使用反向工程重构触发器以识别后门样本,以修剪与触发器相关的神经元的方式防御后门攻击.由于目前后门攻击方案需要执行多轮后门训练才能使后门模型收敛,造成了额外的计算开销,并且后门方案中的触发器采用随机性图片或像素点,与训练数据集中样本存在较大的差异,使触发器容易被检测、重构,影响了后门攻击的效果.

因此,针对后门攻击中收敛速率较慢以及触发器与干净样本差异较大而易被检测等问题,本文提出了一种新型联邦学习后门攻击方案 Bac_GAN,通过采用生成式对抗网络(generative adversarial networks, GAN)技术[20-22]设计了 1 个触发器生成算法 Trig_GAN,能够降低触发器样本与训练样本间的差异,从而提升触发器的隐蔽性.并且通过良性特征混合训练以及缩放后门模型大幅缩短模型的收敛

速率,从而显著提升后门攻击成功率.本文的主要贡献有 3 个方面:

1) 设计了 1 个新的基于生成式对抗网络的触发器生成算法 Trig_GAN,从联邦学习样本中直接生成触发器,将触发器以水印[23]的形式植入干净样本,明显降低了触发器样本与训练样本的差异,从而提升了触发器的隐蔽性.

2) 基于设计的触发器算法,提出了一种新型联邦学习后门攻击方案 Bac_GAN,该方案通过良性特征混合训练保证了后门任务与正常分类任务的精度.特别是,通过缩放后门模型,避免了参数聚合过程中后门贡献被抵消的问题,使得后门模型能够短时间内达到收敛,进而提升后门攻击成功率.

3) 通过从触发器生成、水印系数、缩放系数等后门攻击核心要素进行了实验测试,给出了影响后门攻击性能的最佳参数,并与在 MNIST 与 CIFAR-10 数据集上对比现有典型后门攻击方案,实验证明 Bac_GAN 能够显著提升后门攻击的收敛速率与攻击成功率.

1 相关技术

1.1 联邦学习框架

图 1 显示了联邦学习的框架,用户首先下载初始的全局模型参数,根据服务器提供的学习算法在本地进行训练并上传本地模型参数,服务器对用户上传的参数采用加权平均算法更新全局模型,然后用户即可下载新的全局模型参数进行下一轮训练.

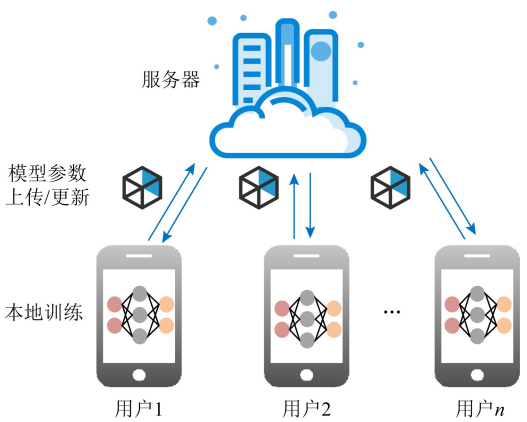


Fig. 1 The architecture of federated learning

图 1 联邦学习框架

具体来说,联邦学习目的在于将用户本地模型聚合成 1 个全局模型,将深度学习任务分配给 n 个用户.首先,在每轮 t 中,服务器随机挑选 m 个用户,并向他们分发初始的全局模型 G_t.每个被选中的用户对其本地数据进行训练,然后将该模型更新为新的本地模型 L_{t+1},并将差异 $L_{t+1}-G_t$ 返回服务器.服务器对接收到的更新进行聚合,以获得新的全局模型:

$$G_{t+1}=G_t+\frac{\eta}{n}\sum_{i=1}^{m}(L_{t+1}^{i}-G_t). \tag{1}$$

1.2 生成式对抗网络

生成式对抗网络 GAN 是 2014 年 Goodfellow 等人[24]首次提出,旨在解决如何从现有样本中训练新样本的问题,从而达到拓展数据集的目的.GAN 技术的关键在于学习训练样本的数据分布,从而发现随机变量演变到训练样本的映射函数.GAN 主要包含 2 部分,即生成器模型与判别器模型.生成器是 1 个由多层感知器组成的神经网络模型,其作用在于将随机的一维噪声转换为新的数据样本.判别器则是判断样本是真实样本还是生成网络产生的假样本的概率.首先,生成器通过输入满足先验概率分布的随机噪声 z,生成伪造的数据.其目的在于使伪造的数据与真实数据无差别,即判别器识别数据 $G(z,\theta_g)$为真的概率尽可能大.目标函数为

$$\max_{\theta_g} E_z[\log(F(H(z;\theta_g);\theta_d))], \tag{2}$$

θ_d,θ_g 分别为判别器与生成器的模型参数,x 为真实样本,$F(\cdot)$表示 x 来自于真实数据分布的概率,$H(\cdot)$表示将输入的噪声 z 映射成数据.判别器的功能是判断生成的样本是否为真,即使输出为真实数据 x 的概率尽可能高,而使输出为生成器生成的假样本 $H(z,\theta_g)$的概率尽可能低,其目标函数为

$$\max_{\theta_d} E_x[\log(F(x,\theta_d))]+ E_z[\log(1-F(H(z,\theta_g),\theta_d))], \tag{3}$$

2 基于生成式对抗网络的后门攻击

为了降低触发器特征与原样本特征之间的差异同时提高后门模型的收敛速率,我们利用 GAN 技术生成接近真实样本的触发器,进而实现了 Bac_GAN 后门攻击方案.下面我们先阐述联邦学习后门攻击模型,然后给出触发器算法 Trig_GAN 的设计,最后详细说明构建的后门攻击方案 Bac_GAN.

2.1 联邦学习后门攻击模型

图 2 显示了联邦学习场景下的后门攻击模型.攻击者伪装成良性用户参与联邦学习,在本地训练过程中,攻击者在携带触发器的数据集上训练后门

模型,上传的后门模型通过加权平均等算法进一步改变原始数据的原始分布和学习算法逻辑,从而试图控制联邦学习系统产生 1 个全局模型,使得全局模型在带有触发器的目标输入上实现较高的攻击成功率,同时在其主要分类任务上保持较高的精度.在分类任务的背景下,我们定义 3 个指标来评价后门攻击性能:

1) 后门攻击准确率.即攻击成功率,指带有触发器样本的分类置信度,使全局模型将带有触发器的图像分类为攻击者指定的标签;

2) 主要任务准确率.表示全局模型对干净样本应具有较高的分类精度,以防止全局模型被丢弃;

3) 收敛速率.达到相同准确精度,模型训练的轮数.

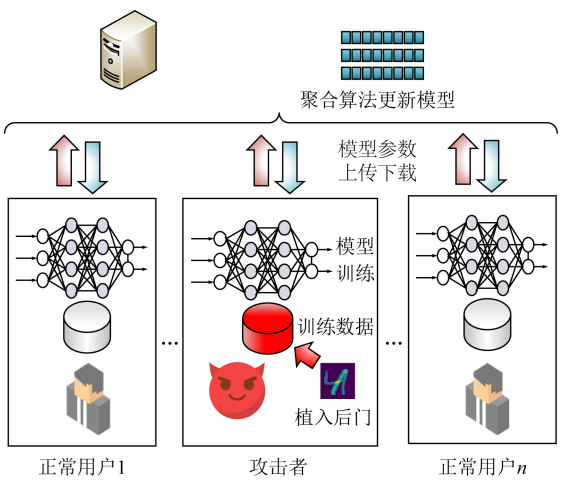


Fig. 2 The Backdoor attack model of federated learning

图 2 联邦学习后门攻击模型

通常,我们假设攻击者具有 4 种能力:

1) 攻击者控制本地训练数据;

2) 控制本地训练过程和模型结构以及超参数,如 *epoch* 数和学习率;

3) 在将生成的本地模型提交聚合之前,可以对本地模型参数进行修改;

4) 可以自适应地逐轮改变模型的局部训练.

2.2 触发器算法 Trig_GAN 设计

实施后门攻击的关键在于建立隐蔽、巧妙的触发器生成算法,而使用其他良性用户的真实训练样本作为触发器,不仅能增加触发器的隐蔽性,同时能使得后门模型在相同的训练轮数下更快收敛.为了使攻击者获得其他用户多个类的部分真实训练样本,同时判断出真实训练样本中的主要标签,本文提出了一种基于生成对抗网络的触发器生成算法 Trig_GAN.该算法利用 GAN 模型生成其他用户的真实样本作为触发器.

GAN 模型中需要真实样本作为判别器的输入,而在联邦学习模式下全局模型参数是通过在每个参与者的上传数据进行模型训练后加权平均生成,因此可以用来更新判别器的模型参数.这相当于直接在其他用户的真实训练样本上训练判别器.这种巧妙的方式使得生成器可以很容易地产生与真实训练样本相似的伪样本.图 3 为联邦学习模式下 GAN 模型示意图,其中判别器与全局模型有着相同的结构(网络层数相同,输出不同),并且随着用户与服务器迭代次数的增加而更新其网络参数.在联邦学习过程中,一方面本地用户模型的更新会促进全局模型的收敛,另一方面判别器模型作为全局模型的更新也会通过全局模型参数进行同步更新.同时,生成器以噪声 Z_{noise} 为输入,有条件地生成特定的样本.

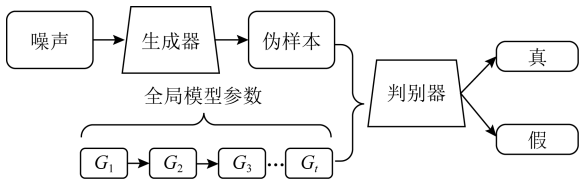


Fig. 3 GAN model of federated learning mode

图 3 联邦学习模式下 GAN 模型

本文将生成的伪样本反向输入全局模型中,选择准确率最高的一类作为触发器.算法 1 给出了触发器生成算法的形式化描述.在每轮迭代中,攻击者首先下载全局模型 G_t 并且初始化判别器的模型参数.生成器接受随机噪声后产生伪样本 x_{fake},并且将这些样本发送到判别器.如果判别器将 x_{fake} 的标签判断属于样本空间的一类,将 x_{fake} 赋值于 x_m,否则更新式(1).最后攻击者将选择的标签 y_m 分配给 x_m 并将其添加到触发器集合中.如果攻击者在联邦学习中不断加入,生成器可以生成大量的伪样本集合,进而生成大量的触发器集合.将触发器集合按照标签进行分类,反向输入模型进行预测,最后输出准确值最高的样本作为本文选择的触发器.

在整个联邦学习训练过程中,由于触发器的样本占所有用户样本的绝大部分,因此全局模型在某种程度上,对于触发器的特征过拟合.将触发器加入到攻击者的训练样本上,能够使得后门模型在短时间内达到收敛.同时触发器的特征接近真实样本的特征,避免触发器被检测、重构,提升了触发器的隐蔽性.

算法 1. 触发器生成算法 Trig_GAN.

输入:全局模型参数 G_t、噪声样本 Z_{noise}、标签分布 Y;

输出:触发器样本 D_{trigger}.

① 初始化判别器与生成器;

Repeat

② 使用 G_t 更新判别器的模型参数;

Repeat

③ 运行生成器生成伪样本 x_{fake},并将其发送至判别器;

④ 如果伪样本属于目标标签 y_m,使得 $x_m = x_{\text{fake}}$,否则基于式(2)更新生成器;

⑤ 如果 x_m 的标签等于 y_m,分配标签 y_m 给样本 x_m,并将样本(x_m,y_m)添加到 D_{trigger};

Until *epoch* 周期结束;

Repeat

⑥ 将触发器集合 D_{trigger} 按照标签分类,反向输模型,记录准确率较高的一类;

Until D_{trigger} 集合全部遍历;

Until 通信结束;

Return 触发器样本 D_{trigger}.

2.3 后门攻击 Bac_GAN 构建

基于 Trig_GAN 触发器生成算法,我们提出了一种新的后门攻击方案 Bac_GAN,图 4 描述了该后门攻击方案的架构图.攻击者首先下载全局模型参数更新判别器,进而生成触发器集合,然后将触发器以水印的方式植入训练样本生成后门训练数据,最后攻击者训练本地后门模型上传攻击全局模型.攻击者具体按照 5 个步骤实施后门攻击:

1) 攻击者从服务器下载初始全局模型参数,更新本地模型以及判别器模型.攻击者伪装成良性用户训练模型,并上传模型参数;

2) 攻击者正常参与联邦学习的过程中,根据触发器生成算法 Trig_GAN,生成触发器集合直到触发器准确率达到期望阈值;

3) 攻击者按照算法 2 将触发器以水印的方式植入本地样本,并修改其标签,生成后门训练集 D_{backdoor}.原样本的特征与触发器的特征进行重叠,同时保持视觉上的不同.触发器的特征会融入到本地训练样本中,即使经过再训练,也会使触发器保持在目标实例的特征空间附近;

4) 攻击者通过混合良性特征训练,训练后门模型,如算法 3 所示.其中良性特征混合训练即相同批次内在干净样本(正常训练的样本)、后门样本(触发器样本与干净样本的混合样本)和触发器样本(携带触发器的样本)中共同训练;

5) 攻击者采用后门模型 B 替换全局模型 G_t,并通过缩放系数 C 放大全局模型,使得后门模型能在加权平均期间保留后门模型的贡献.

$$B = C(B - G_t) + G_t. \tag{4}$$

在 Bac_GAN 后门攻击方案的构建过程中,利用 GAN 模型生成了接近真实训练样本的触发器集合,降低了触发器与训练数据之间的差异.在训练后门模型阶段,我们通过良性特征混合训练以及缩放后门模型,能够使得后门模型短时间内达到收敛.

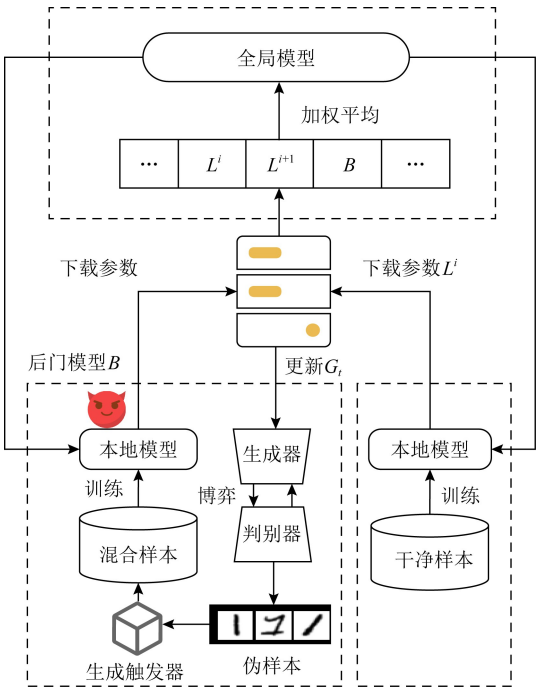


Fig. 4 The architecture diagram of Bac_GAN

图 4 Bac_GAN 方案架构图

算法 2. D_{backdoor} 生成算法.

输入:本地训练样本 D_{local}、触发器 D_{trigger}、水印系数 S、攻击标签 y_{attack}、触发器样本标签 y_{trigger};

输出:后门训练集 D_{backdoor}.

Repeat

① 如果 D_{local} 中样本的标签等于攻击者想要攻击的标签 y_{attack},将攻击样本加上触发器×水印系数 S,赋值于 x_{bac};

② 将(x_{bac},y_{trigger})添加到 D_{backdoor};

Until 集合遍历结束;

Return 后门训练集 D_{backdoor}.

算法 3. Bac_GAN 后门攻击方案训练算法.

输入:本地训练样本 D_{local}、后门训练样本

$D_{backdoor}$、触发器样本 $D_{trigger}$;

输出:上传的本地后门模型 B.

① 初始化后门模型 B 与损失函数

Repeat

② 判断后门模型 B 是否收敛,收敛则退出循环;

Repeat

③ 在训练批次中同时训练本地样本 D_{local}、后门样本 $D_{backdoor}$、触发器样本 $D_{trigger}$,并更新攻击模型 B;

Until batch_size 结束

④ 根据模型梯度更新学习率;

Until epoch 结束

⑤ 通过式(4)缩放本地模型参数;

Return 后门模型 B.

3 实验分析

我们在 MNIST[25],CIFAR-10[26]数据集上实现生成式对抗网络的后门攻击方案,从触发器生成、水印系数、缩放系数等后门攻击核心要素进行了实验测试,探讨和分析影响后门攻击性能的最佳参数,并与现有典型后门攻击方案进行对比,分析后门攻击方案 Bac_GAN 的有效性.

3.1 实验设置

我们模拟了真实联邦学习场景下的实验环境,通过 socket 技术实现服务器与用户端的模型参数上传下载过程.其中服务器端硬件为:Intel 的 i5-9300H(2.40 GHz)CPU 和 64 GB 内存.用户端硬件为:Intel 的 i5-6500CPU 和 16 GB 内存.服务器和用户端软件使用的是:Windows10 操作系统,Python 3.7.3 和 PyTorch 1.4.0+cu92.

本实验选择不同规模的数据集进行训练,表 1 中包含了 MNIST 数据集与 CIFAR-10 数据集的信息摘要.其中 MNIST:包含 10 个类别的手写数字数据集,本实验将采用 2 层卷积 2 层全连接的卷积神经网络[27]进行训练.CIFAR-10 数据集:包含 10 个类别普适物体的彩色图像数据集,该数据集较为复杂,本实验将采用 Rsnet18 模型进行训练.

Table 1 Dataset Summary

表 1 数据集摘要

数据集	数据集包含类别个数	特征	模型	训练轮数/测试轮数
MNIST	10	784	2 层卷积 2 层全连接	6 000/2 000
CIFAR-10	10	1 024	Resnet18	4 000/1 000

3.2 触发器生成实验

我们首先在 MNIST 与 CIFAR-10 数据集对 GAN 生成的样本进行可视化,按照触发器生成算法生成触发器集合.其中判别器的网络模型在每轮迭代过程中都通过全局模型进行更新.攻击者将在全局模型的准确度达到 85%时进行迭代,生成伪样本集合.在样本总和与参与用户总数不变的情况下,实验对于 2 种数据集分别迭代 200 次,从而利用生成器 GAN 生成伪样本.

图 5 分别显示了 MNIST 与 CIFAR10 数据集随联邦学习迭代生成的伪样本.由图 5 可以看出,100 轮左右重建结果已经逐渐靠近真实的 MNIST 与 CIFAR10 数据集样本,而在 150 轮到 200 轮左右,由于生成器的性能随着更新逐渐提升,因此产生的样本更加清晰.其中 MNIST 数据集由于图片简单、特征少,已经接近真实样本,而 CIFAR10 中的数据更加复杂,需要更多的训练轮数.

图 6 显示了将生成的伪样本集合反向输入模型得到的各种标签的准确率.由图 6 可知,利用 Trig_GAN 算法生成的图像分别在 MNSIT 与 CIFAR-10 分类模型中都有较高的准确率.特别是,从图 6 可以看出,得到“4”与“car”标签的准确率最高.原因在于联邦学习中“4”与“car”标签的样本在训练数据集中占绝大多数,模型对于这些标签的样本的分类逐渐达到收敛.因此,我们将在后续实验中分别选择“4”与“car”标签的样本作为触发器.

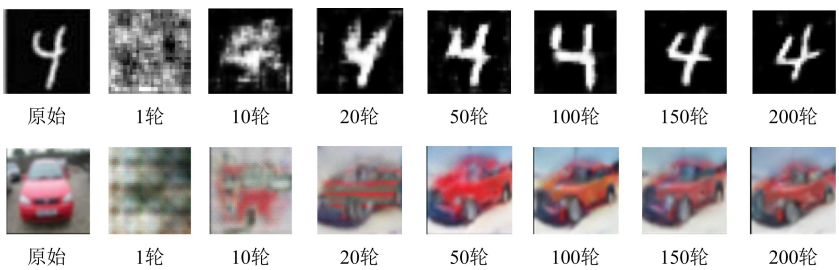
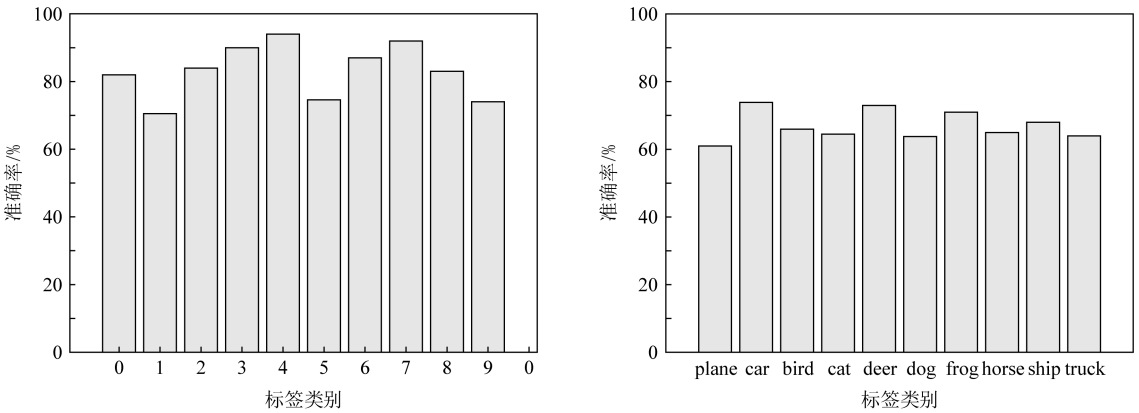


Fig. 5 Simulation of trigger generation

图 5 模拟触发器生成结果



(a) MNIST数据集 (b) CIFAR-10数据集

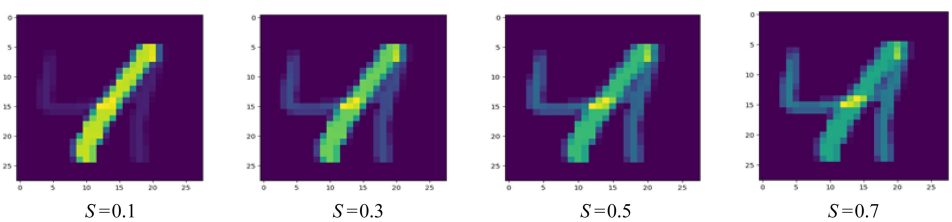
Fig. 6 Accuracy results of trigger collection

图 6 触发器集合准确率

3.3 水印系数实验分析

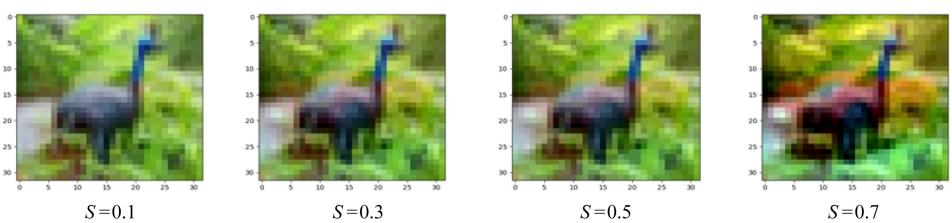
水印系数是影响触发器隐蔽的关键因素,本节将讨论不同水印系数 S 对于触发器的显示效果以及对于后门任务攻击准确率的影响.

图 7 显示了后门数据样本分别在水印系数 S 为 0.1,0.3,0.5,0.7 下的重构图像.以图 7(a)MNIST 数据集为例,我们重建了不同水印系数 S 情况下标签“4”做为触发器的后门数据集.其中在水印系数为 0.1 情形下的触发器几乎不可见,隐蔽性较高,但其特征不明显.而在水印为 0.5 以及 0.7 的情况下,触发器与原样本的特征开始重叠,此时触发器的特征会融入训练样本,从而达到后门植入的目的.



S=0.1 S=0.3 S=0.5 S=0.7

(a) MNIST数据集



S=0.1 S=0.3 S=0.5 S=0.7

(b) CIFAR-10数据集

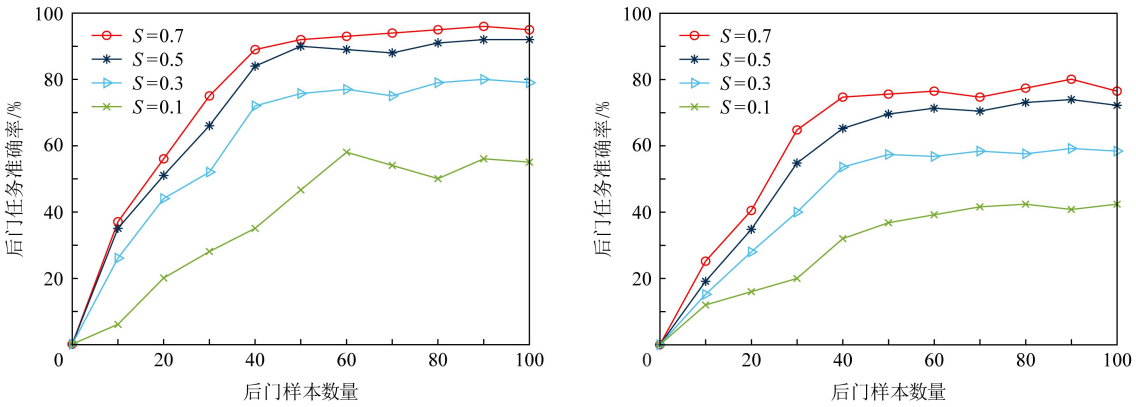
Fig. 7 Reconstructed samples under different watermark coefficients

图 7 不同水印系数下的重构样本

为进一步验证水印系数对于方案性能的影响,我们对不同水印系数下后门攻击准确率随训练样本的变化情况进行了测试.图 8 给出了实验在 MNIST 与 CIFAR-10 数据集中后门攻击方案准确率的结果.由图 8 可以看出,在水印系数 0.1 与 0.3 时,准确率较低,而在水印系数为 0.5 与 0.7 时准确率相差不大,这是因为触发器的特征在水印系数为 0.5 时就已经逐渐与普通样本重叠,此时提升水印系数的效果已经逐渐接近阈值.因此,我们在后续实验中采用水印系数为 0.5,以便同时保证后门准确率与隐蔽性.

3.4 缩放系数实验分析

在联邦学习中,由于服务器端采用加权平均算法导致后门模型的贡献被抵消,全局模型很快会遗忘后门,攻击者持续参与模型聚合才能够成功.因此攻击者会通过缩放后门模型,使得后门模型能在加权平均期间保留后门模型的贡献.



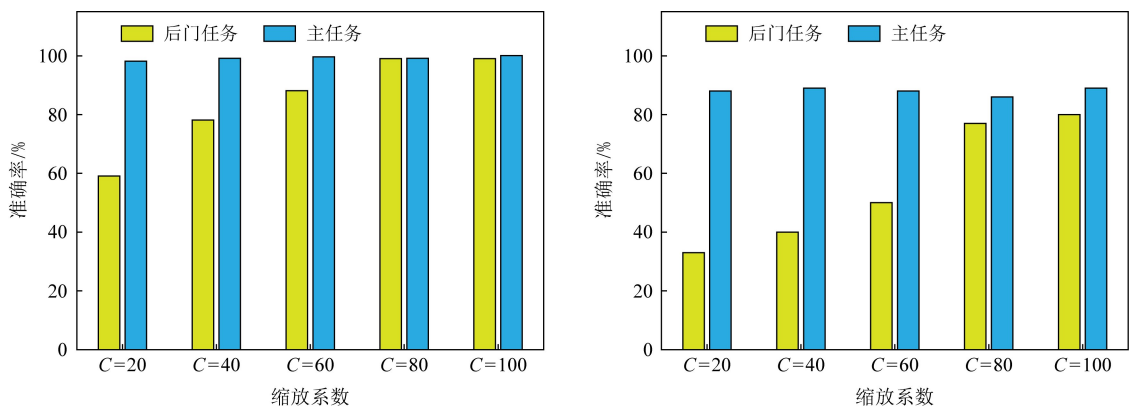
(a) MNIST数据集 (b) CIFAR-10数据集

Fig. 8 Accuracy results of backdoor attack schemes under different watermark coefficients

图 8 不同水印系数下的后门攻击方案准确率结果

图 9 显示了设置不同缩放系数对后门任务与主要任务的准确率影响.为更准确地说明实验结果,我们选取并运行 200 轮联邦学习迭代,并取不同缩放系数情况下($C=20,40,60,80,100$)准确率的平均值作为评判依据.

由图 9 可以看出,缩放系数越高对于后门任务的准确性有着相应地提升,特别是图 9(a)中在缩放系数为 80 时,后门模型性能的提升开始放缓,并基本接近 100%.而由于 CIFAR-10 数据集过于复杂,攻击者拥有着较少的训练样本,从而限制了后门的性能,但我们的 Bac_GAN 方案的攻击成功率也达到了 80%.



(a) MNIST数据集 (b) CIFAR-10数据集

Fig. 9 Accuracy results of backdoor attack schemes under different scaling factors

图 9 不同缩放系数下的后门攻击方案准确率结果

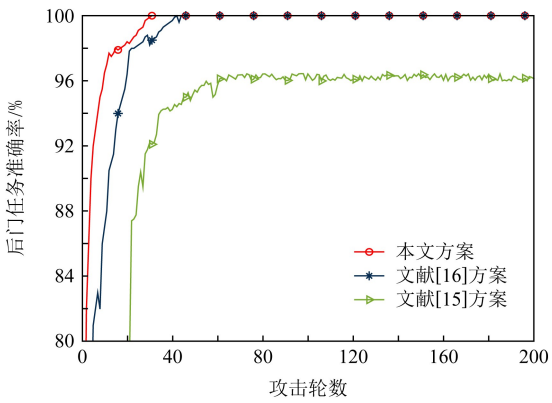
3.5 后门攻击方案对比分析

为了进一步评估方案 Bac_GAN 的有效性,我们与现有典型后门攻击方案进行了实验对比.为了显示后门攻击的效果,我们将在多轮迭代中,同时以缩放系数为 100 实施完整的后门样本训练.

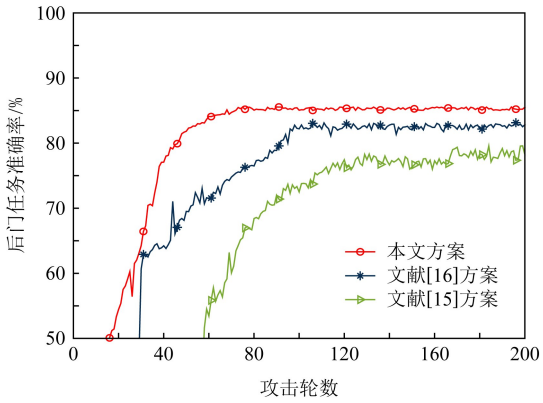
图 10 给出了 Bac_GAN 方案与现有典型后门攻击方案的准确率对比结果.由图 10 可以看出,Bac_GAN 攻击方案的成功率一直优于经典后门攻击方案.即使在 MNSIT 数据集中与方案[16]同时达到了 100%,方案 Bac_GAN 依然保证了更快的收敛速率,比方案[16]提高了 10 轮.而在 CIFAR-10 数据集上,方案 Bac_GAN 的成功率总是在所有情况下都优于现有后门攻击方案,同时保证更高的收敛速率.例如,方案 Bac_GAN 在 70 轮左右就达到收敛,而方案[16]需要 100 轮左右.

3.6 讨 论

经过在 MNIST 与 CIFAR-10 数据集中的多轮对比实验,我们发现在相同的后门样本数量下,水印系数越高攻击成功率越高,但此时的触发器隐蔽性逐渐降低.因此在实现后门攻击时,我们应权衡隐蔽



(a) MNIST数据集



(b) CIFAR-10数据集

Fig. 10 Comparison results of the accuracy of backdoor attack schemes under different rounds

图 10 不同轮数下的后门攻击方案准确率对比结果

性与成功率之间的取舍.在选择合适的水印系数状况下,模型的缩放系数成为实现攻击的关键因素.缩放系数越高后门攻击全局模型的速率越快,但这将导致与正常用户更新差异增加,增加被检测和识别的可能性.最后通过与现有典型后门攻击方案比较分析来看,本文所提方案 Bac_GAN 实现了更高的攻击成功率以及更快的模型收敛速率.归根结底在于 Trig_GAN 触发器来源于联邦学习中的绝大多数样本,全局模型对于采用的触发器已经从某种程度过拟合,从而能缩短后门模型的收敛速率.

4 结束语

本文提出一种基于生成式对抗网络的联邦学习后门攻击方案 Bac_GAN,方案使用 Trig_GAN 算法重构联邦学习中样本数据,作为攻击者的候选触发器,降低了触发器特征与干净样本特征之间的差异,并通过以水印的方式添加至干净样本,提升了触发器的隐蔽性.同时,通过缩放模型技术使得后门模型在短时间内达到收敛,从而提升了攻击成功率.与现有典型后门攻击方案相比,所提方案 Bac_GAN 在短时间内使得后门模型收敛,同时有效提升后门攻击的成功率.在未来工作中,我们将进一步研究触发器特征与样本特征之间的联系,从特征层面植入后门触发器,以便设计更加隐蔽的后门攻击方案.

参 考 文 献

[1] Qi Jia, Lin Keguo, Zhan Pengjin, et al. Preserving model privacy for machine learning in distributed systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(8): 1808-1822

[2] Chen Yufei, Shen Chao, Wang Qian, et al. Security and privacy risks in artificial intelligence systems [J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150)

[3] Zhou Chunyi, Chen Dawei, Wang Shang, et al. Research and challenge of distributed deep learning privacy and security attack [J]. Journal of Computer Research and Development, 2021, 58(5): 927-943 (in Chinese)
(周纯毅, 陈大卫, 王尚, 等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. 计算机研究与发展, 2021, 58(5): 927-943)

[4] Tan Zuowen, Zhang Lianfu. Survey on privacy preserving techniques for machine learning [J]. Journal of Software, 2020, 31(7): 2127-2156 (in Chinese)
(谭作文, 张连福. 机器学习隐私保护研究综述[J].软件学报, 2020, 31(7): 2127-2156)

[5] Fu Anmin, Zhang Xianglong, Xiong Naixue, et al. VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT [J]. IEEE Transactions on Industrial Informatics, arXiv preprint, arXiv:2007.13585, 2020

[6] Zhou Lei, Fu Anmin, Yang Guomin, et al. Efficient certificateless multi-copy integrity auditing scheme supporting data dynamics [J/OL]. IEEE Transactions on Dependable and Secure Computing, 2020 [2020-08-04]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9158490

[7] Zhou Chunyi, Fu Anmin, Yu Shui, et al. Privacy-preserving federated learning in fog computing [J]. IEEE Internet of Things Journal, 2020, 7(11): 10782-10793

[8] Cao Di, Chang Shan, Lin Zhijia, et al. Understanding distributed poisoning attack in federated learning [C] //Proc of the 25th IEEE Int Conf on Parallel and Distributed Systems (ICPADS). Piscataway, NJ: IEEE, 2019: 233-239

[9] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 372-387

[10] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2017: 3-18

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014 [2021-01-16]. https://iclr.cc/archive/2014/conference-proceedings

[12] Rakin A, He Zhezhi, Fan Deliang. TBT: Targeted neural network attack with bit trojan [C] //Proc of the 2020 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13195-13204

[13] Yao Yuanshun, Li Huiying, Zheng Haitao, et al. Latent backdoor attacks on deep neural networks [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 2041-2055

[14] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks [C/OL] //Proc of the 2018 Network and Distributed System Security Symp. Reston, VA: The Internet Society, 2018 [2021-01-16]. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf

[15] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning [C] //Proc of Int Conf on Artificial Intelligence and Statistics. Cambridge, MA: MIT Press, 2020: 2938-2948

[16] Xie Chulin, Huang Keli, Chen Pinyu, et al. DBA: Distributed backdoor attacks against federated learning [C/OL] //Proc of the 7th Int Conf on Learning Representations. 2019 [2021-01-16]. https://openreview.net/group?id=ICLR.cc/2019/Conference

[17] Xue Mingfu, He Can, Sun Shichang, et al. Robust backdoor attacks against deep neural networks in real physical world [J]. arXiv preprint, arXiv:2104.07395, 2021

[18] Liu Yingqi, Lee W C, Tao Guanhong, et al. ABS: Scanning neural networks for back-doors by artificial brain stimulation [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 1265-1282

[19] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2019: 707-723

[20] Chen Foji, Zhu Feng, Wu Qingxiao, et al. A survey about image generation with generative adversarial nets [J]. Chinese Journal of Computers, 2021, 44(2): 347-369 (in Chinese)
(陈佛计, 朱枫, 吴清潇, 等. 生成对抗网络及其在图像生成中的应用研究综述[J]. 计算机学报, 2021, 44(2): 347-369)

[21] Kurakin A, Goodfellow I J, Bengio S, et al. Adversarial examples in the physical worlds [J]. arXiv preprint, arXiv: 1804.00097, 2018

[22] Chen Zhenzhu, Fu Anmin, Zhang Yinghui, et al. Secure collaborative deep learning against GAN attacks in the Internet of things [J]. IEEE Internet of Things Journal, 2021, 8(7): 5839-5849

[23] Zhang Yingjun, Chen Kai, Zhou Geng, et al. Research progress of neural networks watermarking technology [J]. Journal of Computer Research and Development, 2021, 58(5): 964-976 (in Chinese)
(张颖君, 陈恺, 周赓, 等. 神经网络水印技术研究进展[J]. 计算机研究与发展, 2021, 58(5): 964-976)

[24] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv: 1412.6572, 2014

[25] Zhang Xianglong, Fu Anmin, Wang Huaqun, et al. A privacy-preserving and verifiable federated learning scheme [C] //Proc of the 2020 IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2020: 1-6

[26] Zhang Jiale, Chen Bin, Cheng Xiang, et al. PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems [J]. IEEE Internet of Things Journal, 2021, 8(5): 3310-3322

[27] Cao T D, Tram T H, Hien D T, et al. A federated learning framework for privacy-preserving and parallel training [J]. arXiv preprint, arXiv:2001.09782, 2020

Chen Dawei, born in 1997. Master candidate. His main research interests include machine learning security and privacy preserving.
陈大卫,1997 年生.硕士研究生.主要研究方向为机器学习安全和隐私保护.

Fu Anmin, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include cryptography, machine learning security and privacy preserving.
付安民,1981 年生.博士,教授,博士生导师.CCF 高级会员.主要研究方向为物联网安全、机器学习安全和隐私保护.

Zhou Chunyi, born in 1995. PhD candidate. His main research interest includes machine learning security and privacy preserving.
周纯毅,1995 年生.博士研究生.主要研究方向为机器学习安全与隐私保护.

Chen Zhenzhu, born in 1993. PhD candidate. Her main research interests include machine learning security and privacy preserving.
陈珍珠,1993 年生.博士研究生.主要研究方向为机器学习安全与隐私保护.