中国科学: 信息科学 2023 年 第 53 卷 第 3 期: 470–484
SCIENTIA SINICA Informationis
论文

面向联邦学习的对抗样本投毒攻击

王波[1], 代晓蕊[1], 王伟[2*], 于菲[1], 魏飞[3], 赵梦楠[1]

1. 大连理工大学信息与通信工程学院, 大连 116024, 中国
2. 中国科学院自动化研究所智能感知与计算研究中心, 北京 100190, 中国
3. Department of Electrical Engineering, Arizona State University, Tempe AZ85281, USA
* 通信作者. E-mail: wwang@nlpr.ia.ac.cn

收稿日期: 2022–03–24; 修回日期: 2022–06–29; 接受日期: 2022–08–18; 网络出版日期: 2023–03–13

国家自然科学基金 (批准号: U1936117, 62106037, 62076052)、大连市科技创新基金应用基础研究项目 (批准号: 2021JJ12GX018)、模式识别国家重点实验室开放课题基金 (批准号: 202100032) 和中央高校基本科研业务费 (批准号: DUT21GF303) 资助项目

摘要 为了解决传统的机器学习中数据隐私和数据孤岛问题, 联邦学习技术应运而生. 现有的联邦学习方法采用多个不共享私有数据的参与方联合训练得到了更优的全局模型. 然而研究表明, 联邦学习仍然存在很多安全问题. 典型地, 如在训练阶段受到恶意参与方的攻击, 导致联邦学习全局模型失效和参与方隐私泄露. 本文通过研究对抗样本在训练阶段对联邦学习系统进行投毒攻击的有效性, 以发现联邦学习系统的潜在安全问题. 尽管对抗样本常用于在测试阶段对机器学习模型进行攻击, 但本文中, 恶意参与方将对抗样本用于本地模型训练, 旨在使得本地模型学习混乱的样本分类特征, 从而生成恶意的本地模型参数. 为了让恶意参与方主导联邦学习训练过程, 本文进一步使用了 “学习率放大” 的策略. 实验表明, 相比于 Fed-Deepconfuse 攻击方法, 本文的攻击在 CIFAR10 数据集和 MNIST 数据集上均获得了更优的攻击性能.

关键词 联邦学习, 对抗样本, 投毒攻击

1 引言

机器学习技术已经广泛应用于各个领域 [1~3]. 然而数据隐私和数据孤岛问题仍然是阻碍其发展的两大挑战. 例如, 在医疗方面, 要训练一个性能良好的机器学习模型, 需要各个医疗机构或者部门提供大量可以描述患者症状的信息, 而医疗数据往往具有很强的隐私性和敏感性 [4]. 同样, 一个城市的应急、后勤和安保等信息部门会产生大量的异构数据, 这些数据往往以孤岛的形式存在, 难以整合利用 [5]. 为了解决上述问题, 联邦学习技术 [6] 应运而生. 不同于传统的机器学习, 联邦学习采用分布式的架构, 不需要将数据集中存储后再进行模型训练, 而是将该过程转移至本地训练参与方, 通过向中心服务器提交本地模型参数的方式保护用户隐私.

引用格式: 王波, 代晓蕊, 王伟, 等. 面向联邦学习的对抗样本投毒攻击. 中国科学: 信息科学, 2023, 53: 470–484, doi: 10.1360/SSI-2022-0116
Wang B, Dai X R, Wang W, et al. Adversarial examples for poisoning attacks against federated learning (in Chinese). Sci Sin Inform, 2023, 53: 470–484, doi: 10.1360/SSI-2022-0116

尽管联邦学习有效地解决了传统的机器学习中数据隐私和数据孤岛问题[7], 但是仍然存在很多安全问题. 研究针对联邦学习系统的攻击, 对于发现联邦学习的系统漏洞, 促进相应防御方法的研究, 进而构建更加安全鲁棒的联邦学习系统, 具有十分重要的意义. 目前, 对联邦学习系统攻击的研究主要集中在训练阶段, 恶意参与方通过在训练阶段对联邦学习系统进行攻击, 使得联邦学习全局模型失效和参与方隐私泄露. 根据恶意参与方的攻击目的不同, 在训练阶段针对联邦学习的攻击可分为推理攻击和投毒攻击[8,9]. 恶意参与方对联邦学习进行推理攻击, 旨在推理训练过程的信息, 例如, 其他参与方的训练样本和标签等. 而对于投毒攻击, 则可进一步分为模型投毒攻击和数据投毒攻击[10]. 模型投毒攻击是指恶意参与方通过控制其本地模型参数的更新, 使得联邦学习系统的全局模型预测错误. 而对联邦学习的数据投毒攻击则又可以分为两类. 一类为有目标攻击, 其目的是使全局模型实现有方向的特定预测, 典型地如标签反转攻击[11] 和后门攻击[12]. 本文重点关注数据投毒攻击的另外一类, 即无目标攻击, 其核心思想是在训练样本中添加精心设计的噪声, 以实现在训练阶段对联邦学习系统的攻击[13].

联邦学习系统中每个参与方的本地训练过程, 本质是一个传统的机器学习模型的训练过程. 对抗样本攻击[14] 常用于在测试阶段对机器学习系统进行攻击, 通过给测试样本添加一些人眼无法察觉的微小扰动, 就可以使得一个已经训练好的机器学习模型进行误判. Fowl 等[15] 将对抗样本与投毒攻击相结合, 通过在训练样本上添加对抗性噪声, 实现了在训练阶段攻击传统机器学习模型. 受该方法的启发, 本文旨在通过研究针对联邦学习系统的对抗样本投毒攻击, 发现联邦学习系统的脆弱性, 为下一步研究相应的防御提供方向. 具体地, 考虑联邦学习系统中的一个或者多个恶意参与方给本地训练样本添加对抗扰动生成 “有毒” 的对抗样本, 并基于这些对抗样本进行本地训练, 以生成恶意本地模型参数, 从而实现对联邦学习系统的投毒攻击.

然而, 联邦学习在两方面与传统的机器学习不同. 一方面, 联邦学习需要在服务器端对各个本地模型参数进行联邦平均 (FedAvg) 聚合[6], 这会缩小恶意参与方的模型参数, 从而削弱了恶意参与方的 “毒性”. 另一方面, 聚合过程中其他非恶意参与方参与聚合, 也会使得最终的模型参数偏移攻击者的模型参数, 进一步削弱攻击效果. 上述问题是造成针对联邦学习系统投毒攻击效果差的主要原因.

为了使恶意参与方主导联邦学习的训练过程, 从而获得更优的攻击性能, 本文进一步使用了 “学习率” 放大的攻击策略. 具体地, 恶意参与方在本地训练过程中提高学习率以加速梯度下降, 促进恶意本地模型参数快速地生成, 从而在每轮聚合过程中向服务器端提交更强的 “毒药”. 实验结果表明, 在只有少数的恶意参与方时, 本文的攻击方法可以以较高的攻击成功率实现对联邦学习系统的投毒攻击.

本文的其余部分安排如下: 第 2 节介绍了联邦学习和针对其训练阶段的攻击以及对抗样本的相关工作; 第 3 节详细介绍了本文的攻击方法和策略; 第 4 节通过大量实验讨论了本文攻击方法的性能; 最后进行了总结.

2 相关工作

联邦学习最早在 2016 年由谷歌提出[6], 它保证多个训练参与方在不暴露私有训练数据的同时能够联合训练一个更优的机器学习模型, 有效地解决了现实世界中数据不平衡和非独立同分布问题. 具体地, 在联邦学习系统中共有 m 个训练参与方, 每个参与方各自持有一部分数据 D_i, $|D_i| = l_i$, 所有参与方的数据总和为 $\sum_i l_i = L$.

联邦学习一次训练可以总结为如下过程.

(1) 在第 t 轮, 中心服务器向参与聚合的参与方 $n \in [N]$ 发送全局模型 W_G^t, 其中, $[N]$ 是本轮训练

参与方组成的集合 $\{1,2,\ldots,N\}$;

(2) 训练参与方基于第 t 轮全局模型和其训练样本进行 E 轮本地训练, 得到本地模型参数 W_i^{t+1};

(3) 参与方将本地参数发送给中心服务器;

(4) 最后, 在服务器端进行联邦平均聚合得到全局模型 W_G^{t+1}, 其中, $W_G^{t+1}=\frac{l_i}{L}\sum_{i\in[N]}W_i^{t+1}$.

联邦学习作为一项新的热门技术, 针对其安全问题也已经有了广泛的研究. 恶意参与方在训练阶段的攻击会对联邦学习系统造成巨大的威胁. 根据攻击目的不同, 可简单地分为推理攻击和投毒攻击.

恶意参与方执行推理攻击的目的是推理联邦学习系统中的其他信息. Fu 等 [16] 针对纵向联邦学习提出了主动和被动的标签推理攻击, 他们将最终训练好的全局模型加上额外的分类层构成一个 “完整的模型”, 然后用极少量的有标签的数据在该模型上进行半监督训练. Wang 等 [17] 针对服务器端的推理攻击, 提出了一个生成对抗网络 GAN 与多任务鉴别器相结合的框架, 可以实现同时区分输入样本的类别、客户端身份等信息. Geiping 等 [18] 利用输入图像的余弦相似度和对抗攻击从梯度信息中恢复出输入数据.

不同于推理攻击, 投毒攻击旨在通过控制本地模型行为来影响全局模型. 恶意参与方可以通过控制本地模型参数和本地训练数据实现针对训练阶段的投毒攻击. 具体地, 可分为模型投毒攻击和数据投毒攻击. 对于模型投毒攻击, Fang 等 [19] 提出了一种抗拜占庭式鲁棒联邦学习模型的攻击手段, 通过全局模型参数计算其他参与方的模型参数, 并设计本地模型参数以使得全局模型失效. Bagdasaryan 等 [20] 提出模型替换的概念, 他们首先使用添加了后门的训练数据训练本地模型, 然后通过全局模型计算其他参与方的模型参数, 进而设计本地参数使得聚合后的全局模型与目标模型一致. 目前, 针对联邦学习系统的数据投毒攻击主要集中在标签反转攻击和后门攻击. Bhagoji 等 [12] 在本地训练样本中添加后门, 并对梯度进行放大, 使得联邦全局模型对添加了后门的样本进行错误分类. Xie 等 [21] 进一步提出针对联邦系统的分布式后门攻击, 将后门分布于多个恶意参与方本地数据, 结果表明此类攻击性能优于集中式的后门攻击. Tolpegin 等 [22] 在联邦学习系统中引入标签反转, 即不改变本地样本的特征, 而是将其标签篡改为其他标签, 实现了针对联邦学习系统的有目标攻击. Cao 等 [23] 进一步提出分布式标签反转攻击, 相比集中式标签反转攻击, 该攻击方法的性能更优. 另一种可行的投毒攻击方法是在训练样本中加入精心设计的噪声. 冯霁等将针对传统机器学习的 DeepConfuse 数据投毒攻击方法 [24] 扩展到联邦学习系统中, 提出针对联邦学习系统的 Fed-DeepConfuse [13] 数据投毒攻击方法, 攻击者通过一个噪声生成器给训练参与方的本地训练样本中添加有害噪声, 影响本地训练模型的同时损害联邦学习全局模型的性能.

对抗样本攻击 [14] 常用于在测试阶段给样本添加人眼无法察觉的对抗扰动, 使得一个已经训练好的网络模型发生误判. 梯度是影响机器学习模型训练的重要因素, 基于这一发现, 研究人员提出了多种基于梯度的对抗攻击方法 [25~27], 主要通过单步或多步梯度更新, 并限制每次攻击的扰动大小. 对抗训练 [28] 是防御对抗样本攻击的方法之一, 其在每一个批次训练过程都生成对抗样本, 然后将对抗样本与正常样本一起参与训练. 表面上看, 对抗训练似乎与通过对抗攻击生成的 “有毒” 对抗样本在训练阶段对模型进行投毒攻击非常相似. 然而, 对抗训练是保证模型可以对处于训练样本所构成的球内的输入测试样本正确分类, 这是通过在整个训练过程中更新对输入的扰动来实现的, 这一过程使得模型对加了微小扰动的样本不敏感. 相比之下, 用 “有毒” 的对抗样本进行模型训练旨在使得模型更加拟合精心设计的对抗扰动, 而非原始样本信息.

表 1 本文涉及的重要符号以及释义

Table 1 Important symbols and definitions involved in the paper

Symbol	Definition
D_k, D_k^{adv}	Clean training dataset, adversarial sample training dataset of malicious participants
x_i, y_i, $x_{i,\text{adv}}$	Each training sample and label in D_k, x_i's corresponding adversarial sample
$\mathcal{S}$	Noise set meeting disturbance limit conditions
δ_i, ε, ω	Perturbation added to each x_i, perturbation upper limit, perturbation set
θ_k, Θ	The best model parameters trained by "poisonous" adversarial samples, the set of all best model parameters
F, θ, F_1, θ_1	Classifier and model parameters based on clean training samples, classifier and model parameters based on adversarial training samples
$g(y_i)$	Targeted label generation function
η, γ	Local training learning rate, amplification factor of malicious participants

3 联邦对抗样本投毒攻击

本文定义如下场景, 假设有 m ($m \geqslant 2$) 个训练参与方, 考虑分别有 1 到 m 个恶意参与方, 每个恶意训练参与方都是攻击者, 其仅可以操纵自己的本地私有训练数据和学习过程, 而不能访问或操作其他训练参与方的数据或模型学习过程, 如损失函数、梯度下降过程和服务器端的聚合过程等.

恶意参与方的攻击目标是其本地模型参数在参与聚合后, 使得联邦学习全局模型在测试集上的性能尽可能差. 首先, 恶意参与方给本地私有训练样本添加一些人眼无法察觉的对抗扰动生成 "有毒" 的对抗样本, 并基于这些样本进行本地训练. 其次, 为了主导全局模型的训练过程, 恶意参与方在本地训练过程中提高训练学习率以加速恶意模型参数的生成. 最后, 恶意参与方将其本地模型参数上传至服务器端参与聚合以影响全局模型. 接下来详细介绍本文的攻击方法.

为了便于讨论, 以下假设第 k 个参与方是攻击者, 其本地私有数据集为 D_k. 在联邦学习系统中, 每个训练参与方的本地训练都可以看作是一个传统的机器学习模型训练过程. 直观地, 攻击者通过解决以下双层优化问题可以训练得到恶意的本地模型参数, 进一步实现针对联邦学习系统的投毒攻击:

$$\max_{\omega \in \mathcal{S}} \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(F(x;\theta(\omega)),y)], \tag{1}$$

$$\text{s.t. } \theta(\omega) \in \underset{\Theta=\{\theta_1,\ldots,\theta_n\}}{\arg\min} \sum_{k=1}^{n} \sum_{(x_i,y_i)\in D_k} \mathcal{L}\left(F\left(x_i+\delta_i;\theta_k;\eta\right),y_i\right), \tag{2}$$

其中, x_i 和 y_i 分别表示 D_k 中的每个训练样本和对应的标签, $x=\{x_i\}$ 是所有训练样本的集合, $y=\{y_i\}$ 是标签集合. $\mathcal{S}$ 为符合扰动限制条件的可选噪声集合. 对 D_k 中每个样本 x_i 添加扰动 δ_i, 并限制扰动 $\|\delta_i\|_\infty \leqslant \varepsilon$, ε 为扰动上限. $\omega=\{\delta_i\}$ 为给所有训练样本添加的扰动集合. η 为恶意参与方的本地训练学习率, $\mathcal{L}$ 为损失函数, θ_k 为模型参数. 式 (2) 的优化过程会有多个扰动集合 ω, 对于每个 ω 都会有一个对应的在 "有毒" 的对抗样本下训练得到的最优模型参数 θ_k, Θ 为所有最优模型参数的集合. 攻击者的目标是在集合 $\mathcal{S}$ 中找到一组最优的扰动集合 ω, 使得其对应的最优分类器 $F(\theta(\omega))$ 在 D_k 的样本空间分布 $\mathcal{T}$ 上的泛化性能尽可能差. 本文涉及的重要符号和释义如表 1 所示.

然而直接解决上述非凸优化问题非常困难. 要实现式 (1) 的优化目标, 需要在式 (2) 上不断地寻找 ω 并执行梯度下降. 而 ω 的选取非常依赖 $\mathcal{S}$. 为了近似解决上述双层优化目标, 本文引入对抗样

本攻击, 将优化 (2) 中寻找 ω 生成投毒样本的问题转化为对抗样本生成问题. 利用经典的对抗样本生成方法给原始样本添加对抗扰动, 以打乱原始样本的分布. 恶意参与方用对抗样本进行本地模型训练, 使得模型学习混乱的样本分类特征, 以降低模型的测试准确率.

具体地, 本文通过以下过程实现针对联邦学习系统的对抗样本投毒攻击.

(1) 首先用原始干净样本训练对抗样本生成模型, 在对抗样本生成过程中模型参数固定不变.

$$\min_{\theta}\left[\sum_{(x_i,y_i)\in D_k}\mathcal{L}\left(F\left(x_i;\theta;\eta\right),y_i\right)\right], \tag{3}$$

其中, F 和 θ 分别表示基于干净样本训练的分类器和模型参数.

(2) 基于上述固定的训练好的模型, 生成对抗样本作为 "有毒" 的训练样本. 本文使用经典的对抗样本生成方法 Project Gradient Descent (PGD) 攻击 [27] 生成对抗样本. 该方法在生成对抗样本之前引入初始化噪声以打乱原始样本的分布, 然后进行多步梯度更新以生成对抗样本. 式 (4) 和 (5) 分别以无目标攻击和有目标攻击生成对抗样本, 后文分别用有目标对抗样本和无目标对抗样本表示这两种方法生成的对抗样本.

$$x_{i,\mathrm{adv}}^{t+1}=\Pi_{x_{i,\mathrm{adv}}}\left(x_{i,\mathrm{adv}}^{t}+\alpha\,\mathrm{sgn}\left(\nabla_{x_{i,\mathrm{adv}}^{t}}\mathcal{L}\left(F\left(x_{i,\mathrm{adv}}^{t};\theta\right),y_i\right)\right)\right), \tag{4}$$

其中, 在原始样本 x_i 中引入初始化噪声作为 $x_{i,\mathrm{adv}}$ 的初始值, $x_{i,\mathrm{adv}}^{t}$ 表示经过 t 轮迭代攻击后生成的对抗样本. α 表示单步扰动程度. $\Pi_{x_{i,\mathrm{adv}}}$ 表示在 ε-ball 球上的投影, 即当扰动幅度过大时就会将其限制在球内.

$$x_{i,\mathrm{adv}}^{t+1}=\Pi_{x_{i,\mathrm{adv}}}\left(x_{i,\mathrm{adv}}^{t}+\alpha\,\mathrm{sgn}\left(\nabla_{x_{i,\mathrm{adv}}^{t}}\left(-\mathcal{L}\left(F\left(x_{i,\mathrm{adv}}^{t};\theta\right)\right),g(y_i)\right)\right)\right), \tag{5}$$

其中, $g(y_i)$ 是目标标签生成函数, $g(y_i)$ 的选择不是固定的, 可以根据样本、模型和联邦学习数据分布有所不同, 本文定义 $g(y_i)=y_i+1$.

(3) 恶意参与方使用对抗样本进行本地训练, 得到恶意的本地模型参数. 将该参数上传至服务器端参与聚合实现对联邦学习系统的无目标攻击.

$$\min_{\theta_1}\left[\sum_{\left(x_i^{\mathrm{adv}},y_i\right)\in D_k^{\mathrm{adv}}}\mathcal{L}\left(F_1\left(x_i^{\mathrm{adv}};\theta_1;\eta\right),y_i\right)\right], \tag{6}$$

其中, $D_k^{\mathrm{adv}}=\{x_{i,\mathrm{adv}},y_i\}$ 为对抗样本集, F_1 为基于对抗样本训练的分类器, θ_1 为恶意模型参数.

特别地, 由于在联邦学习系统中, 各个参与方需要将本地模型参数上传至服务器端进行联邦平均聚合, 该过程会缩小恶意参与方的模型参数从而削弱其 "毒性". 此外, 其他非恶意训练参与方参与最终的全局模型聚合也会降低恶意参与方的攻击能力.

为了解决上述问题, 本文采用 "学习率放大" 的策略, 攻击者在本地训练过程中恶意提高本地训练学习率, 以加速其梯度下降, 促进本地恶意模型参数快速地生成, 从而在每轮聚合过程中向服务器端提交更强的 "毒药", 在每轮聚合过程中让全局模型更加依赖恶意参与方的本地模型. 在无目标和有目标对抗样本生成方式下, 式 (6) 被修改为式 (7), 以更好地适应于联邦学习对抗样本投毒攻击.

$$\min_{\theta_1}\left[\sum_{\left(x_i^{\mathrm{adv}},y_i\right)\in D_k^{\mathrm{adv}}}\mathcal{L}\left(F_1\left(x_i^{\mathrm{adv}};\theta_1;\gamma\times\eta\right),y_i\right)\right], \tag{7}$$

其中, γ 为恶意参与方本地训练学习率放大因子.

算法详细过程如算法 1 所示.

算法 1 Adversarial example poisoning attack against federated learning

Input: Malicious participant k, local training samples of malicious participants D_k, local training rounds T, total training batches m, the num of samples of every batch b, learning rate scaling factor γ;

Output: Local model parameter participating in the tth aggregation θ_t^k;

Malicious participant trains with local clean training data and get θ_k^*, which is fixed during poison generation;
for $i = 1$ to m **do**
  Generate adversarial examples $\{x_{i,1}^{\text{adv}}, x_{i,2}^{\text{adv}}, \ldots, x_{i,b}^{\text{adv}}\}$ using network $F(\theta_k^*)$ and PGD attack;
  $D_k^{\text{adv}}[i] \leftarrow \{(x_{i,1}^{\text{adv}}, y_{i,1}), (x_{i,2}^{\text{adv}}, y_{i,2}), \ldots, (x_{i,b}^{\text{adv}}, y_{i,b})\}$;
end for
Receive $t-1$ round global model G_{t-1};
$\theta_k^{t-1} \leftarrow G_{t-1}$; //Update the local model with G_{t-1}.
for $j = 1$ to T **do**
  for $i = 1$ to m **do**
    $[x_{i,n}^{\text{adv}}, y_{i,n}] \in D_k^{\text{adv}}[i]$; //$n \in [1, b]$.
    $\theta_{k,t-1}^i = \theta_{k,t-1}^{i-1} - \gamma \times \eta \nabla_{\theta_{k,t-1}^{i-1}} L(F_{\theta_{k,t-1}^{i-1}}(x_{i,n}^{\text{adv}}, y_i^n))$; //Malicious participants trains with adversarial examples and larger learning rate.
  end for
  $\theta_{k,t-1}^j = \theta_{k,t-1}^m$;
end for
$\theta_k^t = \theta_{k,t-1}^T$; //Update tth local model parameters of malicious participants.

4 实验

本节通过实验主要讨论以下几方面的问题.

(1) 本文投毒攻击方法的性能. 具体地, 讨论在联邦学习独立同分布 (IID) 与非独立同分布 (Non-IID) 场景下, 不同数量的恶意参与方参与训练时的攻击性能.

(2) 学习率对攻击性能的影响. 具体地, 分别讨论攻击者使用不同的学习率放大因子的攻击性能以及非恶意参与方的学习率对攻击性能的影响.

(3) 参与方的本地训练轮数对攻击效果的影响. 具体地, 讨论训练参与方的本地训练轮数分别为 1∼5 时的攻击性能.

(4) 对抗样本的泛化性能. 具体地, 讨论针对某个特定模型生成的对抗样本用于攻击其他联邦学习模型是否同样有效.

(5) 在本文的投毒攻击下联邦学习系统与传统的机器学习系统的鲁棒性比较.

(6) 对抗样本相对于随机噪声样本的攻击有效性.

4.1 实验设置

本文分别在 MNIST [29] 数据集和 CIFAR10 [30] 数据集上进行了实验. MNIST 数据集由美国国家标准与技术研究所 (National Institute of Standards and Technology, NIST) 发起整理, 共包含 10 类大小为 28×28 的灰度图, 其中训练集 60000 张, 测试集 10000 张. CIFAR10 数据集由 Hinton 团队整理, 包含了 10 类 60000 张 32×32 的彩色图, 其中训练集 50000 张, 测试集 10000 张.

CIFAR10 和 MNIST 数据集全局聚合次数分别为 60 次和 100 次, 以保证模型收敛. 扰动上限 ε 与本文对比算法 Fed-DeepConfuse [13] 保持一致, 分别设置为 0.032 和 0.3. 实验讨论了在 CIFAR10 和 MNIST 数据集上参与方的本地训练轮数 E, 即从 1∼5, 对本文攻击方法性能的影响. 同时还分别讨论了恶意参与方本地训练学习率为 0.001∼0.3, 非恶意参与方本地训练学习率为 0.0001∼0.1, 对攻击性

表 2　联邦学习 Non-IID 场景数据分布

Table 2　Data distribution of Non-IID federated learning

Total number of participants	Participant	MNIST	CIFAR10
2	Part-1	$0,1,2,3,4,5$	$0,1,2,3,4,5,6,7$
	Part-2	$4,5,6,7,8,9$	$2,3,4,5,6,7,8,9$
3	Part-1	$0,1,2,3,4$	$0,1,2,3,4$
	Part-2	$3,4,5,6,7$	$3,4,5,6,7$
	Part-3	$5,6,7,8,9$	$5,6,7,8,9$
4	Part-1	$0,1,2,3$	$0,1,2,3$
	Part-2	$2,3,4,5$	$2,3,4,5$
	Part-3	$4,5,6,7$	$5,6,7,8$
	Part-4	$6,7,8,9$	$6,7,8,9$
For MNIST: (hand-written digits: $0,1,2,3,4,5,6,7,8,9$)→($0,1,2,3,4,5,6,7,8,9$)			
For CIFAR10: (airplane, automobile, bird, cat, deer, frog, horse, ship, truck)→($0,1,2,3,4,5,6,7,8,9$)			

能的影响.

本文分别讨论了在联邦学习数据独立同分布 (IID) 与非独立同分布 (Non-IID) 场景下攻击的有效性. 考虑在训练参与方数量分别为 2, 3 和 4 的场景下, 不同数量的恶意参与方参与训练时的攻击效果. 在 IID 场景中, 每个参与方拥有全部类别, 每个类别被平均分配给各个参与方. 在 MNIST 与 CIFAR10 数据集上, 各个参与方均使用 ResNet18 作为分类网络. 在 Non-IID 场景中, 为了与 Fed-DeepConfuse 方法进行公平比较, 采用了与其相同的数据划分方式和网络结构. 数据划分如表 2 所示, 每个客户端拥有的样本标签和数量互不相同. CIFAR10 数据集上使用 ResNet18 进行训练, MNIST 数据集上使用具有两个卷积层, 通道数分别是 20 和 50 卷积神经网络 [13].

4.2　对抗样本投毒攻击性能

本小节对比分析了本文的攻击方法与同类方法 Fed-DeepConfuse [13] 的攻击性能, 并进一步讨论了在联邦学习 IID 与 Non-IID 场景下不同数量的恶意参与方对全局模型测试准确率的影响. 实验发现, 在 CIFAR10 数据集和 MNIST 数据集上, 恶意参与方与非恶意参与方的本地训练学习率分别为 0.3, 0.01 和 0.3, 0.001 时, 本文的攻击可以达到最好的攻击性能. 本小节在上述设置下进行攻击性能的研究, 其他学习率组合对本文攻击方法的性能影响分析见 4.3 小节.

图 1 展示了原始样本及在有目标攻击和无目标攻击下的对抗样本.

图 2 为在联邦学习数据 IID 分布场景下, 恶意参与方使用有目标对抗样本进行投毒攻击的实验结果. 可以发现, 即使只有一个恶意参与方, 全局模型的测试准确率也会显著降低. 具体地, 在共有 2, 3 和 4 个训练参与方且只有一个恶意参与方的联邦场景中, 全局模型的测试准确率在 CIFAR10 数据集上分别降低了 78.23%, 77.48% 和 65.54%, 在 MNIST 数据集上分别降低了 67.33%, 61.15% 和 48.56%. 值得注意的是在 CIFAR10 数据集上, 所有参与方均为恶意参与方时, 全局模型的准确率会低于随机猜测.

可以发现相比 MNIST 数据集, 本文的攻击方法对基于 CIFAR10 数据集训练的联邦学习系统的威胁更大. 如在共有 4 个训练参与方且只有一个恶意参与方的联邦系统中, 在 CIFAR10 数据集上训练

图 1 (网络版彩图) 第一行: 原始样本. 第二行: 有目标攻击下对抗样本. 第三行: 无目标攻击下对抗样本

Figure 1 (Color online) First row: original examples. Second row: targeted attack adversarial examples. Third row: untargeted attack adversarial examples. (a) CIFAR10 dataset; (b) MNIST dataset

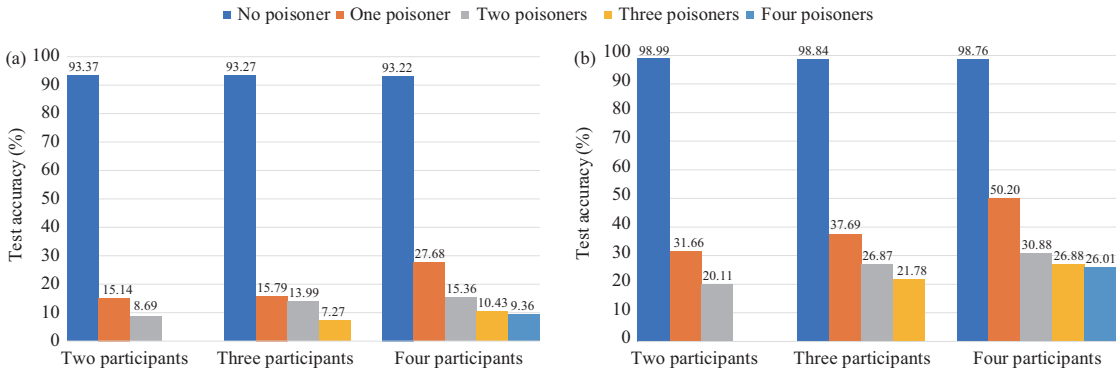

图 2 (网络版彩图) 在 IID 场景中不同数量的恶意参与方参与训练时全局模型的测试准确率, 这里恶意参与方使用有目标对抗样本进行投毒攻击, 恶意参与方学习率 $\mathrm{lr}_{\mathrm{mal}} = \mathbf{0.3}$, 非恶意参与方学习率 $\mathrm{lr}_{\mathrm{CIFAR10}} = \mathbf{0.01}$, $\mathrm{lr}_{\mathrm{MNIST}} = \mathbf{0.001}$, $E_{\mathrm{CIFAR10}} = \mathbf{5}$, $E_{\mathrm{MNIST}} = \mathbf{1}$

Figure 2 (Color online) In IID federated learning, the test accuracy of the global model when different numbers of malicious participants that generate local adversarial training samples with targeted attacks participate in the training, malicious participants learning rate $\mathrm{lr}_{\mathrm{mal}} = 0.3$, non-malicious participants learning rate $\mathrm{lr}_{\mathrm{CIFAR10}} = 0.01$, $\mathrm{lr}_{\mathrm{MNIST}} = 0.001$, $E_{\mathrm{CIFAR10}} = 5$, $E_{\mathrm{MNIST}} = 1$. (a) Experimental results of CIFAR10; (b) experimental results of MNIST

的全局模型的测试性能比基于 MNIST 数据集上训练的全局模型测试性能多下降 16.98%. 本文分析认为, 针对联邦系统的对抗样本投毒攻击成功率取决于: (1) 恶意参与方本地模型的 “恶意” 程度. 使用有目标对抗样本进行投毒攻击, 在 CIFAR10 数据集上恶意参与方的本地模型测试准确率会低至 6%. 而在 MNIST 数据集上, 恶意参与方的本地模型测试准确率在 25% 左右. (2) 在不同数据集上模型的收敛速度. 本文希望通过学习率放大加速恶意模型参数的生成, 使得全局模型 “更早” 且 “更多” 地依赖恶意参与方的本地模型. 然而, 在相同的设置下, 对于 CIFAR10 数据集, 在 5 个 epoch 内正常训练参与方的本地模型测试准确率为 69.47%, 而在 MNIST 数据集上 5 个 epoch 内正常训练参与方的本地模型测试准确率已经达到 97.89%. 这也意味着, 非恶意参与方在每轮聚合时对基于 MNIST 数据集上训练的全局模型测试准确率影响更大.

图 3 所示是在联邦 IID 场景下, 使用无目标对抗样本对联邦学习系统进行投毒攻击的实验结果. 对比图 2 可以看出, 其攻击性能比使用有目标对抗样本的攻击性能弱. 然而使用无目标攻击对抗样本仍然使得联邦学习系统性能显著下降. 只有一个恶意参与方时, 在 CIFAR10 数据集上全局模型的

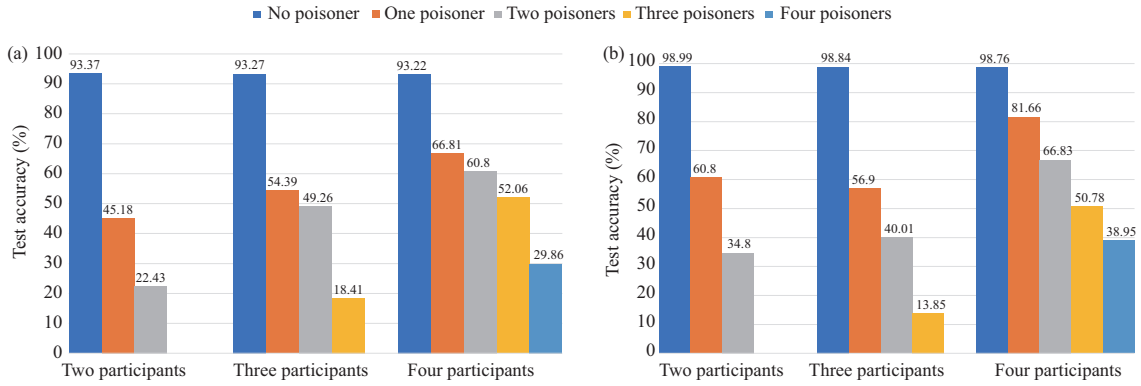

图 3 (网络版彩图) 在 IID 场景中不同数量的恶意参与方参与训练时全局模型的测试准确率, 这里恶意参与方使用无目标对抗样本进行投毒攻击, 恶意参与方学习率 $\mathrm{lr}_{\mathrm{mal}}=0.3$, 非恶意参与方学习率 $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$

Figure 3 (Color online) In IID federated learning, the test accuracy of the global model when different numbers of malicious participants that generate local adversarial training samples with untargeted attacks participate in the training, malicious participants learning rate $\mathrm{lr}_{\mathrm{mal}}=0.3$, non-malicious participants learning rate $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$. (a) Experimental results of CIFAR10; (b) experimental results of MNIST

测试准确率分别降低了 48.19%, 38.88% 和 26.14%, 在 MNIST 数据集上分别降低了 38.04%, 41.52% 和 17.10%. 采用无目标对抗样本进行投毒攻击具有不稳定性, 这与被攻击模型和攻击时 “毒药” 的初始化有关系.

在如表 2 所示的联邦学习 Non-IID 场景下, 使用有目标对抗样本对联邦学习系统进行投毒攻击的实验结果如图 4 所示, 对 CIFAR10 数据集而言, 当只有一个恶意参与方时, 全局模型的测试准确率下降至接近随机猜测. 对 MNIST 数据集而言, 全局模型的测试准确率分别降低了 69.78%, 72.77% 和 57.45%.

同时, 本文还讨论了在 10 个训练参与方的联邦学习场景下, 恶意参与方分别为 1~5 时的攻击性能. 实验结果如表 3 所示, 随着恶意参与方的比例增加, 攻击性能如预期的提高. 当有 50% 的恶意参与方时, 在 CIFAR10 和 MNIST 数据集上联邦学习全局模型的测试准确率相比未被攻击时分别降低了 67.29% 和 67.04%.

在相同的联邦学习场景, 即共有 2, 3 和 4 个参与方且只有 1 个恶意参与方的联邦系统下, 本文与 Fed-DeepConfuse 算法 [13] 进行了全局模型测试准确率下降程度的对比分析. 由表 4 结果可知, 本文的方法对联邦学习系统的攻击效果优于 Fed-DeepConfuse. 如在有 2 个参与方的联邦学习系统中, 对 CIAFR10 数据集, 在 Fed-DeepConfuse 攻击下全局模型测试性能下降了 18.99%. 本文使用有目标对抗样本进行投毒攻击全局模型的测试准确率下降了 72.43%, 使用无目标对抗样本进行投毒攻击全局模型测试准确率下降了 60.23%.

此外, 值得注意的是, 相比 Fed-DeepConfuse 本文的方法可以更快速地训练对抗样本生成模型. 分别用 m 和 T 表示训练最大批次和迭代轮数, 本文训练对抗样本生成模型时间复杂度为 $\mathrm{O}(2mT)$, Fed-DeepConfuse 训练噪声生成模型的时间复杂度为 $\mathrm{O}((5m)+2)\times T$. 在 CIFAR10 数据集上, 本文的 m 和 T 分别取值为 390 和 150, Fed-DeepConfuse 的 m 和 T 分别取值为 390 和 500. 在 CIFAR10 数据集上, 本文中恶意参与方利用本地未经扰动的样本直接训练 ResNet18 作为对抗样本生成模型. Fed-DeepConfuse 训练一个自动编码器 U-Net 作为噪声生成器, 该方法首先利用自动编码器生成的对抗训练样本训练分类器 ResNet18, 并收集分类器的更新轨迹, 然后再利用收集到的轨迹对分类器计算

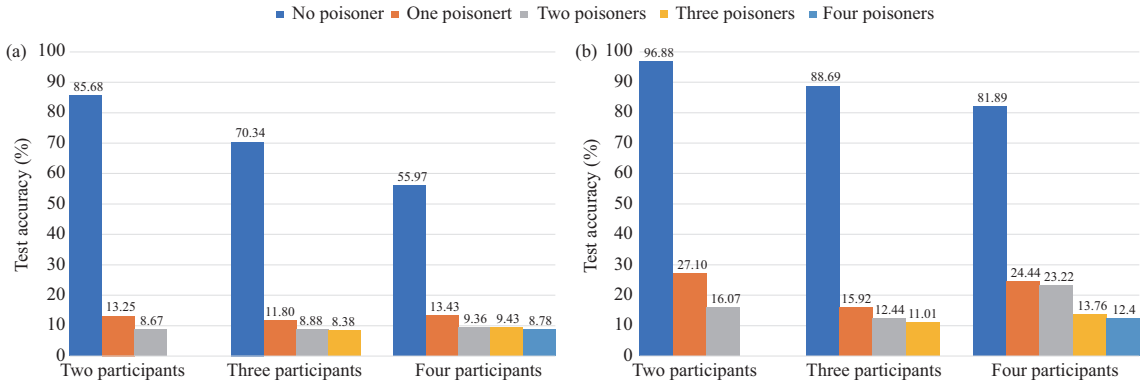

图 4 (网络版彩图) 在 **Non-IID** 场景中不同数量的恶意参与方参与训练时全局模型的测试准确率, 这里恶意参与方使用有目标对抗样本进行投毒攻击, 恶意参与方学习率 $\mathrm{lr}_{\mathrm{mal}}=0.3$, 非恶意参与方学习率 $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$

Figure 4 (Color online) In Non-IID federated learning, the test accuracy of the global model when different numbers of malicious participants that generate local adversarial samples with targeted attacks participate in the training, malicious participants learning rate $\mathrm{lr}_{\mathrm{mal}}=0.3$, non-malicious participants learning rate $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$. (a) Experimental results of CIFAR10; (b) experimental results of MNIST

表 3 在 10 个训练参与方下, 分别有 1~5 个恶意参与方时全局模型的测试准确率 (%)

Table 3 Test accuracy (%) of the global model with 1 to 5 malicious participants with 10 training participants

	No poisoner	One poisoner	Two poisoners	Three poisoners	Four poisoners	Five poisoners
CIFAR10	93.17	61.94	45.90	37.94	30.69	21.86
MNIST	98.40	89.80	73.90	58.78	44.63	31.36

表 4 只有一个恶意参与方时, 不同的攻击方法下全局模型测试准确率下降程度 (%). **Untarget** 和 **Target** 分别代表本文的对抗样本生成方法

Table 4 Only one malicious participant, the accuracy (%) of the global model test decreases under different attacks. Untarget and Target represent the adversarial sample generation method, respectively

	Fed-DeepConfuse		Ours			
			Target		Untarget	
	CIFAR10	MNIST	CIFAR10	MNIST	CIFAR10	MNIST
Two parts	−18.99	−30.69	−72.43	−69.78	−60.23	−72.62
Three parts	−18.21	−14.03	−58.54	−72.77	−53.68	−59.39
Four parts	−7.88	−8.05	−42.84	−57.45	−32.61	−57.90

一个 “伪更新” 从而更新自动编码器, 整个过程重复 T 次直到自动编码器收敛.

4.3 学习率对攻击性能的影响

实验发现学习率是影响攻击成功率的一个重要因素, 本小节讨论本文攻击方法在不同学习率组合下的攻击性能.

首先讨论非恶意参与方的本地训练学习率对攻击成功性能的影响. 在共有 2 个参与方且 1 个为恶意参与方的联邦学习系统上进行了实验, 实验中固定恶意参与方的本地训练学习率, 在 CIFAR10 和 MNIST 数据集上均为 0.3, 非恶意参与方的本地训练学习率研究范围均为 0.0001~0.1. 实验结果如

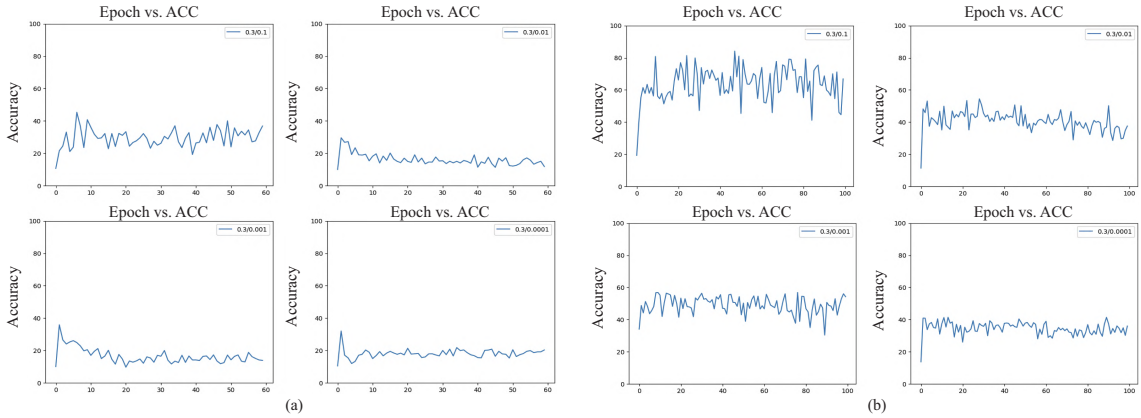

图 5 (网络版彩图) 非恶意参与方使用不同的学习率进行本地训练时全局模型的测试准确率. 恶意参与方学习率 $\mathrm{lr}_{\mathrm{mal}}=0.3$, $E_{\mathrm{MNIST}}=1$, $E_{\mathrm{CIFAR10}}=5$

Figure 5 (Color online) Test accuracy of the global model when normal participants use different learning rates for local training, malicious participants learning rate $\mathrm{lr}_{\mathrm{mal}}=0.3$, $E_{\mathrm{MNIST}}=1$, $E_{\mathrm{CIFAR10}}=5$. (a) Experimental results of CIFAR10; (b) experimental results of MNIST

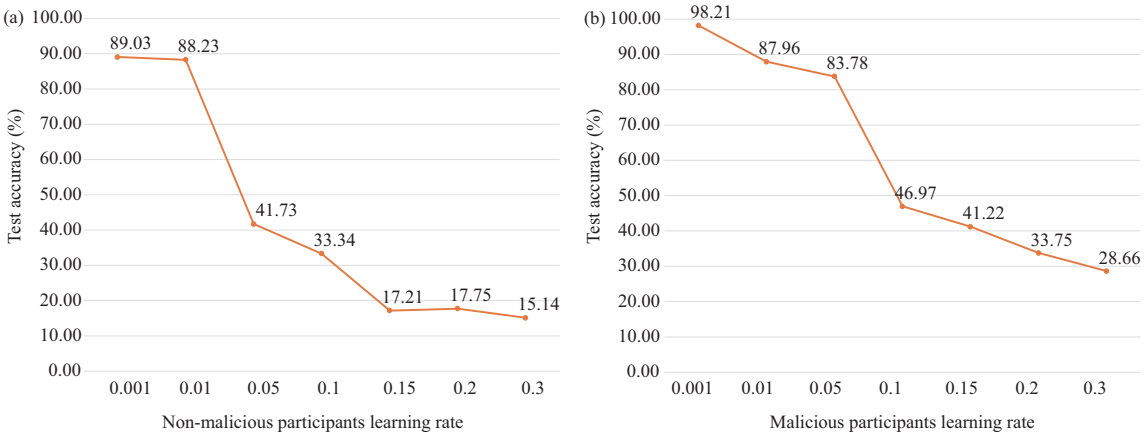

图 6 (网络版彩图) 恶意参与方使用不同的学习率进行本地训练时全局模型的测试准确率. 非恶意参与方学习率 $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$

Figure 6 (Color online) Test accuracy of global model when malicious participants use different learning rates for local training, non-malicious participants learning rate $\mathrm{lr}_{\mathrm{CIFAR10}}=0.01$, $\mathrm{lr}_{\mathrm{MNIST}}=0.001$, $E_{\mathrm{CIFAR10}}=5$, $E_{\mathrm{MNIST}}=1$. (a) Experimental results of CIFAR10; (b) experimental results of MNIST

图 5 所示, 从实验结果可以发现, 当非恶意参与方本地学习率较大时, 全局模型会有比较大的震荡, 且在相同的学习率设置下 MNIST 数据集上的震荡程度更大.

其次讨论恶意参与方本地训练学习率对攻击性能的影响. 同样在共有 2 个参与方且 1 个为恶意参与方的联邦学习系统上进行了实验, 实验中固定非恶意参与方的本地训练学习率, 在 CIFAR10 和 MNIST 数据集中分别设置为 0.01 和 0.001, 恶意参与方的本地训练学习率研究范围均为 0.001~0.3. 实验结果如图 6 所示, 随着恶意参与方本地训练学习率放大程度的增加, 攻击成功率会不断提高, 并且攻击成功率在某个学习率区间内趋于稳定.

表 5 训练参与方的本地训练轮数分别为 1∼5 时全局模型测试准确率 (%)

Table 5 Test accuracy (%) of the global model when the participants trained with different local training rounds

	1	2	3	4	5
CIFAR10	25.14	18.13	17.22	17.53	15.14
MNIST	31.66	48.01	46.42	46.14	50.50

表 6 采用不同结构的本地训练模型时联邦学习全局模型测试性能 (%)

Table 6 Test performance (%) when using different local models

	CIFAR10			MNIST		
	ResNet18	VGG19	MobileNetV2	CNN	$\text{CNN}_{\text{small}}$	$\text{CNN}_{\text{large}}$
No poisioner	93.37	90.84	93.24	96.88	98.21	98.80
One poisioner	15.14	15.81	17.21	27.10	20.50	28.50

4.4 参与方的本地迭代轮数对攻击性能的影响研究

联邦学习中训练参与方的本地训练轮数是一个重要参数, 本小节讨论了各个参与方使用不同的本地训练轮数时本文攻击方法的有效性. 在共有 2 个参与方且其中 1 个为恶意参与方的联邦场景中进行了实验, 实验结果如表 5 所示. 对于 CIFAR10 数据集, 恶意参与方的本地训练轮数对于攻击性能的影响很小. 而当本地训练轮数增加时, 在 MNIST 数据集上的攻击性能有所下降. 当本地训练轮数为 5 时, 全局模型测试准确率提高至 50.5%. 如 4.2 小节中分析, 在 MNIST 数据集上 5 个 epoch 内正常训练参与方的测试性能已经达到 97.89%, 这也意味着非恶意参与方在每轮聚合时对基于 MNIST 数据集上训练的全局模型测试准确率影响更大.

4.5 泛化性能研究

本小节讨论本文攻击方法的泛化性能, 即相同的对抗样本对于模型结构不同的联邦学习系统的攻击性能如何? 考虑共有 2 个参与方且其中 1 个为恶意参与方的联邦学习场景. 在 CIFAR10 数据集上使用 ResNet18 作为对抗样本生成网络, 在 VGG19 和 MobileNetV2 上进行泛化性能测试. 在 MNIST 数据集上使用 Non-IID 场景中的 CNN 网络作为对抗样本生成网络, 在 $\text{CNN}_{\text{large}}$ 和 $\text{CNN}_{\text{small}}$ (对其通道数分别加倍和减半 [13]) 上进行泛化性能测试. 如表 6 结果所示, 本文的攻击方法具有良好的泛化性能.

4.6 联邦学习系统与机器学习系统的鲁棒性对比

本小节讨论本文攻击方法的鲁棒性, 即在相同比例的训练样本被 “毒害” 时, 联邦学习与机器学习的鲁棒性对比. 实验中考虑共有 10 个参与方的联邦学习系统. 结果如表 7 所示, 从实验结果来看, 联邦学习系统更加脆弱, 当有 20% 的对抗样本参与训练时, 在 CIFAR10 数据集上, 机器学习模型测试准确率有 91.39%, 而联邦学习全局模型测试准确率仅有 45.90%.

4.7 对抗样本相对于随机噪声样本的攻击有效性研究

本小节讨论了在训练样本中添加精心设计的对抗噪声相对于添加高斯随机噪声的攻击有效性. 在共有 2, 3 和 4 个参与方且只有一个恶意参与方的联邦场景中进行了实验, 实验中随机噪声的大小与本文的对抗噪声大小一样, 在 CIFAR10 和 MNIST 数据集上分别为 0.032 和 0.3.

表 7　联邦学习系统与机器学习系统的鲁棒性对比 (%). 0.2, 0.5, 0.8 和 0.9 分别代表对抗样本的比例

Table 7　Robustness comparison (%) between federated learning systems and machine learning systems. 0.2, 0.5, 0.8, and 0.9 represent the proportion of adversarial examples, respectively

	0.2	0.5	0.8	0.9
Machine learning	92.42	90.20	86.56	84.19
Federated learning	45.90	21.86	20.33	15.46

表 8　在 CIFAR10 数据集上, 恶意参与方分别使用对抗样本与随机噪声样本进行本地训练时全局模型的测试准确率 (%)

Table 8　In the CIFAR10 dataset, the test accuracy (%) of the global model when malicious participants trained with adversarial samples and random noise samples, respectively

	No attack	Random noise poisoning attack		Adversarial examples poisoning attack	
		Normal learning rate	Larger learning rate	Normal learning rate	Lager learning rate
Two parts	93.37	92.44	85.56	88.21	15.14
Three parts	93.27	92.95	86.95	89.34	15.79
Four parts	93.22	92.29	86.88	90.57	27.68

表 9　在 MNIST 数据集上, 恶意参与方分别使用对抗样本与随机噪声样本进行本地训练时全局模型的测试准确率 (%)

Table 9　In the MNIST dataset, the test accuracy (%) of the global model when malicious participants trained with adversarial samples and random noise samples, respectively

	No attack	Random noise poisoning attack		Adversarial examples poisoning attack	
		Normal learning rate	Larger learning rate	Normal learning rate	Larger learning rate
Two parts	98.99	98.32	98.24	95.41	31.66
Three parts	98.84	98.45	98.36	96.55	37.69
Four parts	98.76	98.65	98.57	97.12	50.20

表 8 和 9 分别为在 CIFAR10 数据集与 MNIST 数据集上的实验结果. 可以发现, 当恶意参与方使用添加高斯随机噪声的样本进行投毒攻击时, 无论是否采用学习率放大策略其攻击能力都非常有限, 学习率放大时在 CIFAR10 数据集上全局模型的测试准确率分别下降了 7.81%, 6.32% 和 6.32%, 在 MNIST 数据集上全局模型的测试准确率分别下降了 0.75%, 0.48% 和 0.21%. 当恶意参与方使用对抗样本作为 "有毒" 样本进行本地训练时, 在不采用学习率放大策略时其攻击能力有限, 但是在使用学习率放大策略后攻击性能显著提升. 在 CIFAR10 数据集上全局模型的测试准确率分别下降了 78.23%, 77.48% 和 65.54%, 在 MNIST 数据集上全局模型的测试准确率分别下降了 67.33%, 61.15% 和 48.56%.

5　总结

本文讨论了针对联邦学习系统的对抗样本的投毒攻击, 证明对抗样本不仅在测试阶段可以攻击联邦学习系统, 在训练阶段也会对联邦学习系统造成巨大的威胁. 此外, 本文还发现学习率是影响攻击成功率的一个重要的因素. 实验表明恶意参与方使用对抗样本和较大的学习率进行本地训练, 可以有效地攻击联邦学习系统, 使得全局模型的测试准确率显著下降. 同时, 实验结果显示, 本文的攻击方法具有很好地泛化性能, 当训练参与方的模型发生改变时仍然具有很好的攻击效果.

联邦学习越来越广泛地应用于各个领域, 针对联邦学习的投毒攻击仍然是一个重要的研究课题. 在本文的攻击方法基础上, 未来还可以从不同的攻击策略展开研究, 以测试、分析和评估联邦学习系统的安全性. 与此同时, 针对此类攻击的有效防御方法也将是联邦学习安全的研究方向之一.

参考文献

1 Zhao M C, Bo A, Kiekintveld C. Optimizing personalized email filtering thresholds to mitigate sequential spear phishing attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016. 30

2 Deng H, Qin Z, Sha L, et al. A flexible privacy-preserving data sharing scheme in cloud-assisted IoT. IEEE Internet Things J, 2020, 7: 11601–11611

3 Dong J, Cong Y, Sun G, et al. What can be transferred: unsupervised domain adaptation for endoscopic lesions segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 4022–4031

4 Chen Y, Qin X, Wang J, et al. FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell Syst, 2020, 35: 83–93

5 Zhou C X, Sun Y, Wang D G, et al. Survey of federated learning research. Chin J Network Inform Secur, 2021, 7: 77–92 [周传鑫, 孙奕, 汪德刚, 等. 联邦学习研究综述. 网络与信息安全学报, 2021, 7: 77–92]

6 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Artificial Intelligence and Statistics, 2017. 1273–1282

7 Zhang C, Xie Y, Bai H, et al. A survey on federated learning. Knowledge-Based Syst, 2021, 216: 106775

8 Mothukuri V, Parizi R M, Pouriyeh S, et al. A survey on security and privacy of federated learning. Future Generation Comput Syst, 2021, 115: 619–640

9 Rahman K M J, Ahmed F, Akhter N, et al. Challenges, applications and design aspects of federated learning: a survey. IEEE Access, 2021, 9: 124682

10 Zhou J, Fang G Y, Wu N. Survey on security and privacy-preserving in federated learning. J Xihua Univ (Nat Sci Edition), 2020, 39: 9–17 [周俊, 方国英, 吴楠. 联邦学习安全与隐私保护研究综述. 西华大学学报 (自然科学版), 2020, 39: 9–17]

11 Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems. In: Proceedings of European Symposium on Research in Computer Security, 2020. 480–501

12 Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens. In: Proceedings of International Conference on Machine Learning, 2019. 634–643

13 Feng J, Cai Q Z, Jiang Y. Towards training time attacks for federated machine learning systems. Sci Sin Inform, 2021, 51: 900–911 [冯雾, 蔡其志, 姜远. 联邦学习下对抗训练样本表示的研究. 中国科学: 信息科学, 2021, 51: 900–911]

14 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples, In: Proceedings of International Conference on Learning Representations, 2015. 120–128

15 Fowl L, Goldblum M, Chiang P, et al. Adversarial examples make strong poisons. In: Proceedings of Neural Information Processing Systems, 2021

16 Fu C, Zhang X, Ji S, et al. Label inference attacks against vertical federated learning. In: Proceedings of USENIX Security, 2022

17 Wang Z, Song M, Zhang Z, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning. In: Proceeding of IEEE Conference on Computer Communications, 2019. 2512–2520

18 Geiping J, Bauermeister H, Dröge H, et al. Inverting gradients—How easy is it to break privacy in federated learning? In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020. 33: 16937–16947

19 Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning. In: Proceedings of the 29th USENIX Security Symposium, 2020. 1605–1622

20 Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2020. 2938–2948

21 Xie C, Huang K, Chen P Y, et al. DBA: distributed backdoor attacks against federated learning. In: Proceedings of International Conference on Learning Representations, 2019

22 Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems. In: Proceedings of European Symposium on Research in Computer Security, 2020. 480–501
23 Cao D, Chang S, Lin Z, et al. Understanding distributed poisoning attack in federated learning. In: Proceedings of IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), 2019. 233–239
24 Feng J, Cai Q Z, Zhou Z H. Learning to confuse: generating training time adversarial data with auto-encoder. In: Proceedings of Neural Information Processing Systems, 2019
25 Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. In: Proceedings of the International Conference on Learning Representations, 2017. 332–340
26 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of IEEE Symposium on Security and Privacy, 2017. 39–57
27 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Representation Learning, 2018
28 Koh P W, Liang P. Understanding black-box predictions via influence functions. In: Proceedings of the International Conference on Machine Learning, 2017. 1885–1894
29 Deng L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Process Mag, 2012, 29: 141–142
30 Krizhevsky A. Learning Multiple Layers of Features From Tiny Images. Technical Report TR-2009, 2009

Adversarial examples for poisoning attacks against federated learning

Bo WANG[1], Xiaorui DAI[1], Wei WANG[2*], Fei YU[1], Fei WEI[3] & Mengnan ZHAO[1]

1. *School of Information and Communication Engineering, Dalian University of Technology, Dalian* 116024*, China*;
2. *Intelligent Perception and Computing Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China*;
3. *Department of Electrical Engineering, Arizona State University, Tempe* AZ85281*, USA*
* Corresponding author. E-mail: wwang@nlpr.ia.ac.cn

Abstract Federated learning was developed to solve the data privacy and data island in traditional machine learning. Existing federated learning methods use multiple participants who do not share private data to jointly train a better global model. However, research shows that security problems in federated learning remain numerous. Typically, federated learning is attacked by malicious participants during training, resulting in the failure of the global model and the leakage of the private data of the participants. This paper studies the effectiveness of adversarial example poisoning attacks on federated learning and further finds potential security problems in federated learning. Although adversarial examples are often used to attack machine learning models during testing, in this paper, malicious participants use adversarial examples for training the local models, aiming to make the local model learn chaotic sample classification features, thereby generating malicious local model parameters. To let the malicious participants dominate the federal learning and training process, we further use a strategy of "learning rate amplification." Experiments show that compared with the Fed-Deepconfuse attack method, the attacks in this paper achieve better attack performance on the CIFAR10 and MNIST datasets.

Keywords federated learning, adversarial example, poisoning attack