王悦. 纵向联邦学习中的无目标中毒攻击[J]. 智能计算机与应用,2026,16(1):103-109. DOI:10.20169/j.issn.2095-2163.24040102

纵向联邦学习中的无目标中毒攻击

王 悦
(东华大学 计算机科学与技术学院, 上海 201620)

摘 要: 纵向联邦学习用于解决多个数据持有方想要合作建立机器学习模型但又不希望共享原始数据的情况。然而,纵向联邦学习很容易受到恶意参与者的中毒攻击,这些参与者的目的是通过在联合训练过程中发送恶意向量来阻碍纵向联邦学习模型的准确性。本文提出了一个针对纵向联邦学习模型中毒攻击的通用框架,最大化恶意向量和良性嵌入向量之间的距离,使模型训练方向发生偏移,影响模型整体性能。为评估该攻击方法,本文在 MNIST 数据集、CIFAR10 数据集和 Yahoo Answers 数据集的基础上,综合 2 种基础攻击方法中攻击效果最好的一种与其进行对比。实验结果表明,该攻击方法比基础攻击方法的攻击效果更好,攻击影响度提高了 47.89%、42.90%和 32.63%。

关键词: 联邦学习; 纵向联邦学习; 模型安全; 无目标中毒攻击; 机器学习
中图分类号: TP309 **文献标志码**: A **文章编号**: 2095-2163(2026)01-0103-07

Untargeted poisoning attack against Vertical Federated Learning

WANG Yue
(College of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Vertical Federated Learning is used to solve the situation where multiple data holders want to cooperate to build machine learning models but do not want to share the original data. However, Vertical Federated Learning is vulnerable to poisoning attacks by malicious participants who aim to hinder the accuracy of Vertical Federated Learning models by sending malicious embedding vectors during the federated training process. This paper proposes a general framework for poisoning attacks on Vertical Federated Learning models, which maximizes the distance between malicious vectors and benign embedding vectors, causing the model training direction to shift and affecting the overall performance of the model. In order to evaluate this attack method, the paper uses the one with the best attack effect among the two basic attack methods to compare it based on the MNIST dataset, the CIFAR10 dataset and Yahoo Answers dataset. Experimental results show that this attack method is more effective than the basic attack method, and the attack impact is increased by 47.89%, 42.90% and 32.63%.

Key words: federated learning; Vertical Federated Learning; model security; untargeted poisoning attack; machine learning

0 引 言

在当今的大数据时代,每天都有海量、多样和高速增长的数据产生,这些数据提供了许多机遇,但也面临着不小挑战。传统的机器学习模型需要收集不同来源的数据在中心化服务器上进行训练,这就涉及到数据隐私的问题[1]。为了充分利用这些分散的数据资源并保护数据隐私,联邦学习应运而生。纵向联邦学习[2](Vertical Federated Learning)是联邦学习[3]的一种方式,适用于各参与方所持用户重叠较多但是特征重叠较少的情况[4]。

在联邦学习的发展过程中,诸多安全问题也日益凸显[5]。其中,中毒攻击会在一定程度上对模型造成损害,对联邦学习系统产生严重威胁[6-8]。目前,目标中毒攻击已在横向联邦学习和纵向联邦学习进行了广泛研究,而无目标中毒攻击只在横向联邦学习中开展了相关研究。纵向联邦学习中的目标中毒攻击通常先通过标签推理攻击推断出样本标签,然后再进行攻击[9-11]。Fu 等学者[12]通过梯度符号判断、补全本地模型以及增加本地模型权重来实现标签推理。Naseri 等学者[13]通过标签推理攻击重建数据样本的标签,从而发起后门攻击。Gu 等学者[14]利用推理阶段的本地潜在表示输出在纵向联邦学习中注入后门,微调其底部模型实现中毒攻

作者简介: 王 悦(1997—),女,硕士,主要研究方向:联邦学习,信息安全。Email:littlewangyue@outlook.com。

收稿日期: 2024-04-01

击。横向联邦学习中的无目标攻击都需要访问全局模型参数,并通过修改模型参数来实现[15-17]。然而,在纵向联邦学习中,只有运行服务器的参与者拥有标签信息,因此攻击者无法访问样本标签,并且攻击者只拥有自己的一部分数据样本。此外,攻击者无法访问顶部模型,也无法访问其他参与者的底部模型。因此,探讨研究针对纵向联邦学习的中毒攻击更具有挑战性。

1 纵向联邦学习系统实现

纵向联邦学习系统的主要架构如图 1 所示。该架构中包含 K 个参与者 ($K \geqslant 2$) 和 1 个可信的第三方服务器[18]。由于在现实世界的应用程序中,完全可信的第三方服务器很难实现,因此纵向联邦学习中的服务器通常由其中一个参与者来进行管理。其中,每个参与者只拥有垂直切分的数据,只有服务器拥有样本标签。本文假设各个参与方已通过安全实体对齐找到公共标识符[19-21],各方使用对应的数据样本进行后续模型训练。

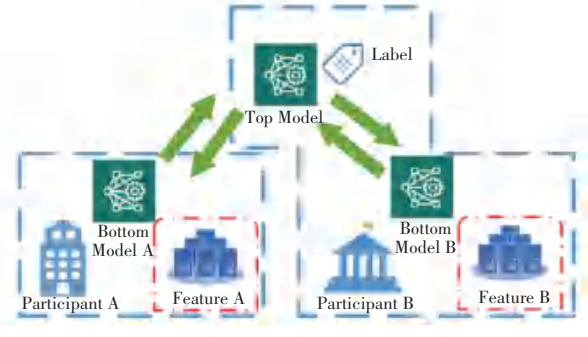

图 1 纵向联邦学习系统架构

Fig. 1 System architecture of Vertical Federated Learning

纵向联邦学习模型的具体训练过程如算法 1 所示。

算法 1 纵向联邦学习原型系统实现算法

输入 顶部模型 θ_{top},K 个参与者的底部模型 $\theta_1,\theta_2,\cdots,\theta_K$,样本标签为 y,学习率为 η

输出 VFLModel

1. 初始化 $\theta_{\text{top}},\theta_1,\theta_2,\cdots,\theta_K$
2. While stopping epoch not met do
3. For each batch b of sample IDs do
4. For $k = 1$ to K do
5. $o_k \leftarrow$ ParticipantForwordProp (θ_k, b) //每个参与者前向传播
6. $o_{\text{all}} \leftarrow$ ConcatBottomModels $(o_1,\cdots,o_k)$
7. $o_{\text{final}} \leftarrow \theta_{\text{top}}(o_{\text{all}})$
8. $L \leftarrow$ LossFunc (o_{final}, y)
9. $\theta_{\text{top}} \leftarrow \theta_{\text{top}} - \gamma \cdot \dfrac{\partial L}{\partial \theta_{\text{top}}}$
10. For $k = 1$ to K do //每个参与者进行后向传播更新其底部模型
11. $\theta_k \leftarrow \theta_k - \gamma \cdot \dfrac{\partial L}{\partial o_k} \cdot \dfrac{\partial o_k}{\partial \theta_k}$

其中,每次训练迭代可以分为 2 个步骤。

(1)前向传播。所有参与者使用本地数据和底部模型在本地进行前向传播,并将本地训练得出的特征嵌入向量提交给服务器,服务器使用来自所有参与者的特征嵌入向量来计算最终预测结果以及相应的损失值。

(2)反向传播。服务器进行反向传播,并根据每个参与者的输出计算损失的梯度,该梯度被发送给每个参与者。基于此,每个参与者继续向后传播并更新其底部模型。

2 最大化距离度量攻击

2.1 攻击概述

本文假设攻击者可以在训练过程中任意修改数据样本或上传的特征嵌入向量,即攻击者通过不同的数据变换方式以及设计特殊的触发向量代替其局部嵌入,在不知道样本标签和其他参与者特征嵌入的情况下,也可以提高样本错误分类的概率。首先,在纵向联邦学习的设置中,每个参与方都会拥有所有数据样本的一部分,并且每个参与方都是独立的,这意味着即使有一方不参与训练,全局模型也可以从已有的数据中学习到有用的信息;其次,在传统的中毒攻击中,最常见且有效的方法通常需要修改样本标签,而在纵向联邦学习中是不可能的,因为攻击者无权访问或修改服务器私有的标签。针对以上 2 个挑战,本文提出了一种面向纵向联邦学习的无目标中毒攻击,攻击示意如图 2 所示。该攻击通过最大化恶意嵌入向量和良性嵌入向量之间的距离,在空间内找到局部最优攻击点,有效提高攻击效率。

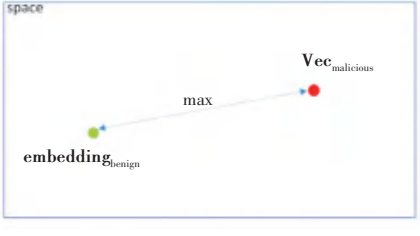

图 2 攻击示意图

Fig. 2 Attack diagram

2.2　攻击流程

2.2.1　确定扰动向量

攻击者的目标是尽可能地破坏模型训练结果，而没有目的地操纵数据样本和特征嵌入只能在有限范围内使模型偏离原先的训练方向。因此，本文提出了最大化距离度量攻击（MaxD）。该方法通过设计一种特殊的偏移特征嵌入向量，可以使模型训练方向发生最大偏移，从而显著削弱模型精度。由于从一个无限大的搜索空间中找到一个恶意向量的代价非常大，因此需要初始化一个恶意向量，通过不断优化得到最优解。本文为了设计初始化扰动，该扰动称为正反向系数扰动，表示为：

$$\boldsymbol{x}_p = \pm \mathbf{sign}(\mathbf{embedding}_b) \tag{1}$$

其中，$\mathbf{embedding}_b$ 表示每个样本的良性嵌入向量，$\mathbf{sign}(\mathbf{vec})$ 表示向量 $\mathbf{vec}$ 中每个元素对应的符号所组成的向量。

2.2.2　距离最大化优化函数

在纵向联邦学习中，增大原始特征嵌入向量与恶意向量之间的距离不仅可以使模型预测边界更加模糊、不易被服务器察觉，而且可以使攻击更有欺骗性，从而更容易混淆模型并实施攻击以提高攻击成功的概率。因此本文设计了距离最大化优化函数。

为了保证恶意向量和原始特征嵌入向量的分布不会相差太大，本文在原始特征嵌入向量上添加了扰动向量 $\boldsymbol{\delta} \cdot \boldsymbol{x}_p$ 之后，得到初始化恶意向量，即：

$$\mathbf{Vec}_m = \mathbf{embedding}_b + \boldsymbol{\delta} \cdot \boldsymbol{x}_p \tag{2}$$

然后，在训练过程中，通过求解以下优化函数，使恶意向量和良性特征嵌入向量之间的距离最大化，得到最终的攻击向量，即：

$$\underset{\mathbf{Vec}_m}{\operatorname{argmax}} \operatorname{Distance}(\mathbf{embedding}_b, \mathbf{Vec}_m) \tag{3}$$

其中，Distance（·）使用的是欧氏距离度量、即 L_2，用来衡量空间中两点之间的距离。假设两点的坐标分别为 $X(x_1, x_2, \cdots, x_n)$ 和 $Y(y_1, y_2, \cdots, y_n)$，则两者之间的欧式距离定义为：

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

2.2.3　次优解系数

本文攻击的目标是找到最佳系数 δ，计算最终的恶意向量。由于无法在一个无限大的空间里搜索到最优解，因此本文根据研究得到的实验结果，分析得出一个有限的搜索空间，将 δ 的搜索距离控制在 $(0,30]$，以步长为 0.2 的速度进行搜索，得到最终的攻击向量。本文将距离度量函数封装成一个函数接口 F，F 的输入为良性嵌入向量 $\mathbf{embedding}_b$ 和系数 δ 函数 $F(\mathbf{embedding}_b, \delta)$ 返回良性特征嵌入和恶意向量之间的距离，在搜索完成后找到距离良性特征嵌入向量最远的向量作为恶意攻击向量。求解过程如算法 2 所示。

算法 2　系数求解算法

输入　δ_{init}，F，$\mathbf{embedding}_b$，τ

输出　δ_{res}

1. step←0.2，$\delta \leftarrow \delta_{\text{init}}$，$\tau \leftarrow 30$，$d_{\max} \leftarrow 0$
2. while $\delta < \tau$ do
3.　　if $F(\mathbf{embedding}_b, \delta) > d_{\max}$ then
4.　　$\delta_{\text{res}} \leftarrow \delta$
5.　　$d_{\max} \leftarrow F(\mathbf{embedding}_b, \delta)$
6.　　end if
7.　　$\delta = \delta$ + step
8. end while

首先将初始值 δ 设为 0，步长 step 设为 0.2，阈值 τ 设为 30，当函数 F 计算得出的距离大于当前最大距离时，记录当前的 δ 和最大距离，直到找到当距离度量最大化时 δ 的解。

3　实验结果与分析

3.1　实验设置

本文使用 3 种数据集来评估攻击效果，即 MNIST、CIFAR10 和 Yahoo Answers。其中，MNIST 数据集包含了 70 000 张灰度图像，每张图像大小为 28×28 像素。CIFAR10 数据集包含 60 000 张彩色图像，每张图像大小为 32×32 像素。Yahoo Answers 数据集是一个文本分类任务的数据集，涉及健康、社会与文化、科学与数学等 10 个主题，每个主题包含 140 000 个训练样本和 6 000 个测试样本。本文默认包含 2 个参与者，如图 3 所示，每个参与者拥有一半的样本。对于 MNIST 和 CIFAR10 数据集，底部模型采用 ResNet-18 网络结构；对于 Yahoo Answers 数据集，底部模型采用 Bert 网络。所有顶部模型均采用 4 层全连接层。

（a）完整样本　（b）攻击者　（c）良性参与者

图 3　纵向联邦学习中样本分割方式

Fig. 3　Sample segmentation methods in Vertical Federated Learning

3.2 评估标准

由于该方法为纵向联邦学习的无目标攻击,攻击者的目标是损害模型整体的预测精度,因此攻击效果评价标准设定为攻击影响度,即攻击后所导致的全局模型精度的降低。用如下公式进行定义:

$$\mathrm{IoA(Impact\ of\ Attack)} = \mathrm{Acc}_b - \mathrm{Acc}_m \quad (5)$$

为了评估 MaxD 攻击方法的攻击效率,本文设计了 2 种基础攻击,即添加噪声攻击和噪声替换攻击。其中,添加噪声攻击是指攻击者在数据样本或特征嵌入向量上根据不同方差添加随机高斯噪声;噪声替换攻击是指攻击者将数据样本或特征嵌入向量替换成不同方差的随机高斯噪声。本文通过实验得到评估基准,详见 3.3.1 节。

3.3 攻击结果与分析

3.3.1 评估基准选择

本文使用不同方差的高斯噪声进行评估,选择合适的评估基准。在表 1 中,噪声设置为随机高斯噪声,噪声方差分别设置为 1~10 的整数。此外,为了观察在大方差情况下攻击是否对模型影响的效果更为显著,本文也将方差设置为 100 以内 10 的倍数。不同 σ^2 样本噪声扰动下的模型准确率见表 1、表 2,不同 σ^2 特征噪声扰动下的模型准确率见表 3、表 4。

表 1~表 4 中,AN 表示样本加噪(Add Noise),RN 表示噪声替换(Replace Noise)。对于 MNIST、CIFAR10、Yahoo Answers 三个数据集而言,模型精度最多分别下降到 84.47%、71.43%、59.08%。实验结果表明,攻击者使用样本噪声扰动和特征噪声扰动可以在有限范围内对模型造成损害。然而,即使噪声方差逐渐增大也无法使模型受到更严重的损害。本文最终选择样本噪声扰动和特征噪声扰动中效果最好的结果作为 Baseline 攻击方法。

表 1 不同 σ^2 样本噪声扰动下的模型准确率

Table 1 Model accuracy under different σ^2 sample noise perturbations %

数据集	扰动方式	σ^2								
		1	2	3	4	5	6	7	8	9
MNIST	AN	93.69	93.28	**86.77**	93.44	88.98	92.02	92.69	93.08	92.24
	RN	93.14	89.67	92.20	92.80	90.28	88.98	92.32	93.20	**84.47**
CIFAR10	AN	74.15	73.69	74.39	73.81	74.79	74.09	73.66	**74.43**	**73.38**
	RN	74.05	72.47	74.32	74.02	73.98	73.57	**72.23**	73.15	74.20

表 2 不同 σ^2 样本噪声扰动下的模型准确率($\sigma^2 \in [10, 100] \cap \mathrm{Mod}(\sigma^2, 10) = 0$)

Table 2 Model accuracy under different σ^2 sample noise perturbations ($\sigma^2 \in [10, 100] \cap \mathrm{Mod}(\sigma^2, 10) = 0$) %

数据集	扰动方式	σ^2									
		10	20	30	40	50	60	70	80	90	100
MNIST	AN	94.59	**93.62**	92.98	93.41	93.88	93.69	93.61	93.84	93.12	92.69
	RN	**88.40**	89.96	90.75	90.70	90.38	91.32	91.63	91.09	90.43	91.54
CIFAR10	AN	73.73	74.45	73.67	73.71	73.64	74.06	74.78	74.13	**73.32**	74.91
	RN	71.78	72.23	72.32	**71.43**	72.67	73.49	73.73	74.19	73.70	73.92

表 3 不同 σ^2 特征噪声扰动下的模型准确率

Table 3 Model accuracy under different σ^2 embedding noise perturbations %

数据集	扰动方式	σ^2								
		1	2	3	4	5	6	7	8	9
MNIST	AN	92.63	92.44	90.32	92.72	**89.42**	91.33	91.42	92.88	90.85
	RN	91.43	90.62	92.90	91.55	90.36	**87.32**	91.66	95.49	88.30
CIFAR10	AN	73.75	72.54	73.79	72.29	73.78	**72.03**	74.23	73.55	72.68
	RN	73.59	73.38	72.43	73.85	72.39	72.50	**71.49**	72.38	73.04
Yahoo Answers	AN	60.45	61.94	60.99	61.77	62.33	**60.01**	61.93	60.93	61.88
	RN	**59.38**	60.48	60.63	60.89	60.37	59.99	60.70	60.44	60.93

表 4　不同 σ^2 特征噪声扰动下的模型准确率($\sigma^2 \in [10, 100] \cap \text{Mod}(\sigma^2, 10) = 0$)

Table 4　Model accuracy under different σ^2 embedding noise perturbations ($\sigma^2 \in [10, 100] \cap \text{Mod}(\sigma^2, 10) = 0$)　%

数据集	扰动方式	σ^2									
		10	20	30	40	50	60	70	80	90	100
MNIST	AN	93.42	**91.88**	93.23	93.92	94.82	93.96	92.67	92.34	92.23	92.88
	RN	89.83	**88.52**	90.40	90.93	91.82	90.42	89.98	90.91	91.13	91.64
CIFAR10	AN	73.43	73.95	74.17	73.92	**72.04**	73.83	73.84	73.73	72.98	72.38
	RN	72.28	72.53	71.92	**71.83**	71.99	72.37	72.38	73.39	72.09	72.34
Yahoo Answers	AN	60.43	61.92	60.40	61.23	**60.11**	61.02	62.77	61.83	60.23	60.91
	RN	60.18	59.29	59.23	60.29	61.25	**59.08**	60.28	61.14	61.90	60.63

3.3.2　优化系数阈值选择

本文的目标是通过最大化恶意向量和良性特征向量之间的距离,找到扰动向量的最优系数,从而计算出最优恶意攻击向量。本文通过对比分析不同系数 δ 对模型精度的影响,得出优化系数的阈值 τ,将搜索范围尽可能缩小,以便快速找到攻击效果最好的恶意向量。2.2.3 小节得到的搜索空间即是基于这里的研究成果。从图 4 中可以看出,当 $\delta < 0$ 时,恶意攻击向量对所有模型的精度影响都非常小,由此可以判断,有效的 δ 应该大于 0。当 $\delta \in [0,20]$ 时,3 个数据集模型精度都发生显著下降,当 $\delta > 25$ 后,模型精度都逐渐上升。由此可以得出,对于模型影响最大的 δ 在区间 [0,30] 内,因此本文将系数阈值 τ 设为 30。

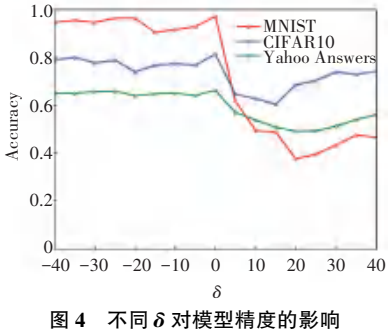

图 4　不同 δ 对模型精度的影响

Fig. 4　The impact of different δ on model accuracy

3.3.3　攻击影响评价

攻击影响度对比见表 5。在表 5 中,Baseline-SN 表示样本噪声扰动,Baseline-EN 表示特征噪声扰动,MaxD 为本文所提出的最大化距离度量攻击。分析可知,Baseline-SN 攻击影响对于 MNIST 数据集和 CIFAR10 数据集分别只有 12.69% 和 9.73%;Baseline-EN 攻击影响对于 MNIST 数据集、CIFAR10 数据集和 Yahoo Answers 数据集则分别为 9.84%、9.77%和 7.09%。当攻击者以攻击效果最佳的方差对样本和特征嵌入向量进行扰动时也无法对模型产生显著影响。而本文所提出的最大化距离度量攻击的攻击影响对 3 个数据集可以达到 60.58%、52.63% 和 39.72%,有效改变纵向联邦学习模型训练方向,损害其全局模型训练精度。

3.3.4　攻击效果可视化

为了可视化 MaxD 的攻击效果,本文使用 Grad-CAM 对模型进行可视化。Grad-CAM (Gradient-weighted Class Activation Mapping)是指梯度加权类激活映射,用来可视化模型在预测时更加关注样本的哪个区域,提升网络的可解释性。MNIST 数据集、CIFAR10 数据集的 Grad-CAM 可视化结果如图 5、图 6 所示。图 5、图 6 中,原始模型为没有攻击时模型的 GradCAM 可视化结果,σ^2_{AN} 表示样本加噪时设置的方差,σ^2_{RN} 表示噪声替换时设置的方差,其中所设置的 σ^2_{AN} 和 σ^2_{RN} 为 3.3.1 小节中所提及的 Baseline。本文假设攻击者所持有左半边样本,服务器方、即良性方拥有右半边样本。从图 5、图 6 中可以看到,原始模型会同时关注所有参与者的样本。当使用 Baseline 攻击时,模型只关注良性方的样本。而当使用本文所提出的 MaxD 攻击时,可以使模型无法学习到攻击方的特征,同时也极大地阻碍了模型学习良性方的特征,有效使模型整体训练方向发生偏移,从而使模型精度显著下降。

表 5　攻击影响度对比

Table 5　Comparison of attack impact　%

Datasets	No Attack	Baseline-SN		Baseline-EN		MaxD	
		Accuracy	IoA	Accuracy	IoA	Accuracy	IoA
MNIST	97.16	84.47	12.69	87.32	9.84	36.58	**60.58**
CIFAR10	81.16	71.43	9.73	71.49	9.77	28.53	**52.63**
Yahoo Answers	66.17	–	–	59.08	7.09	26.45	**39.72**

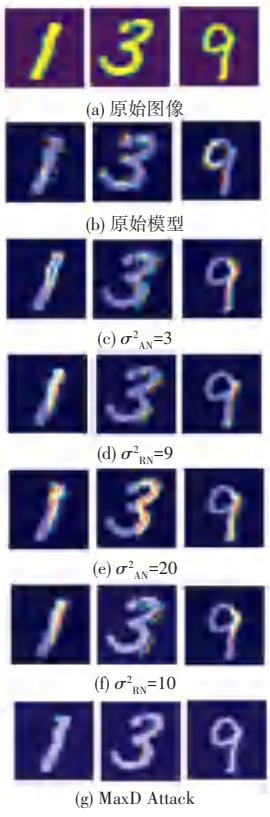

图 5 MNIST 数据集 Grad-CAM 可视化结果

Fig. 5 Grad-CAM visualization results of MNIST dataset

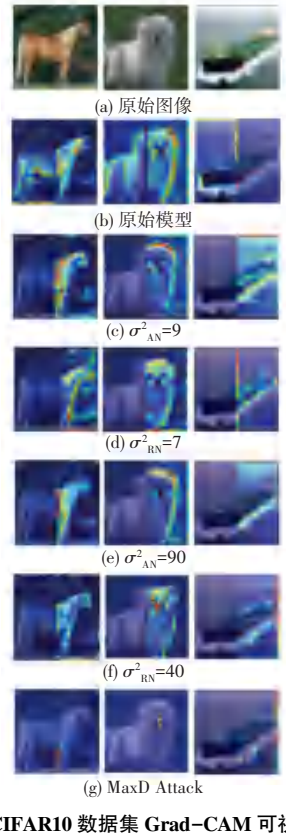

图 6 CIFAR10 数据集 Grad-CAM 可视化结果

Fig. 6 Grad-CAM visualization results of CIFAR10 dataset

4 结束语

本文针对纵向联邦学习系统的安全问题,通过研究纵向联邦学习的安全漏洞,提出了 2 种噪声扰动基础攻击,并通过实验证明基础攻击在一定程度上能够损害模型。同时,本文设计了一种最大化距离度量无目标中毒攻击,该攻击在给定优化阈值范围内,通过使良性特征嵌入向量和恶意向量之间的距离最大化,使模型对于分类边界逐渐模糊,对模型预测结果产生重大影响。通过观察实验效果发现,本文提出的最大化距离度量无目标中毒攻击的攻击影响比基础扰动攻击更显著。最后,本文使用 GradCAM 对图像数据集进行攻击可视化,使攻击效果更具可解释性。

参考文献

[1] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning [C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2015: 1310-1321.

[2] LIU Yang, ZHANG Xiong, WANG Libin. Asymmetrical vertical federated learning[J]. arXiv preprint arXiv,2004.07427, 2020.

[3] LI Qinbin, WEN Zeyi, WU Zhaomin, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 3347-3366.

[4] ROMANINI D, HALL A J, PAPADOPOULOS P, et al. Pyvertical: A vertical federated learning framework for multi-headed splitnn[J]. arXiv preprint arXiv,2104.00489, 2021.

[5] LI Li, FAN Yuxi, TSE M, et al. A review of applications in federated learning[J]. Computers & Industrial Engineering, 2020, 149: 106854.

[6] LIU Pengrui, XU Xiangrui, WANG Wei. Threats, attacks and defenses to federated learning: Issues, taxonomy and perspectives [J]. Cybersecurity, 2022, 5(1): 4.

[7] RODRÍGUEZ-BARROSO N, JIMÉNEZ-LÓPEZ D, LUZÓN M V, et al. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges[J]. Information Fusion, 2023, 90: 148-173.

[8] SIKANDAR H S, WAHEED H, TAHIR S, et al. A detailed survey on federated learning attacks and defenses[J]. Electronics, 2023, 12(2): 260.

[9] WEI Kang, LI Jun, MA Chuan, et al. Vertical federated learning: Challenges, methodologies and experiments [J]. arXiv preprint arXiv,2202.04309, 2022.

[10] WENG H, ZHANG J, XUE F, et al. Privacy leakage of real-world vertical federated learning[J]. arXiv preprint arXiv,2011.09290, 2020.

[11] CHEN H, FU C, RUAN N. Steal from collaboration: Spy attack by a dishonest party in vertical federated learning [C]// Proceedings of International Conference on Applied Cryptography and Network Security. Cham: Springer, 2023: 583-604.

[12] FU Chong, ZHANG Xuhong, JI Shoujing, et al. Label inference attacks against vertical federated learning[C]//Proceedings of the 31st USENIX security symposium (USENIX Security 22).

Boston：USENIX Association，2022：1397-1414.

[13]NASERI M，HAN Yufei，CRISTOFARO D E. Badvfl：Backdoor attacks in vertical federated learning[J]. arXiv preprint arXiv，2304.08847，2023.

[14]GU Y，BAI Y. LR-BA：Backdoor attack against vertical federated learning using local latent representations[J]. Computers & Security，2023，129：103193.

[15]BARUCH G，BARUCH M，GOLDBERG Y. A little is enough：Circumventing defenses for distributed learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York：ACM，2019：775.

[16]FANG Minghong，CAO Xiaoyu，JIA Jinyuan，et al. Local model poisoning attacks to Byzantine-Robust federated learning[C]// Proceedings of the 29th USENIX Conference on security symposium (USENIX Security 20). New York：ACM，2020：1605-1622.

[17]SHEJWALKAR V，HOUMANSADR A. Manipulating the byzantine：Optimizing model poisoning attacks and defenses for federated learning[C]// Proceedings of the Network and Distributed System Security Symposium (NDSS). The Internet Society，2021：1-18.

[18]CEBALLOS I，SHARMA V，MUGICA E，et al. Splitnn-driven vertical partitioning[J]. arXiv preprint arXiv，2008.04137，2020.

[19]LU Linpeng，DING Ning. Multi-party private set intersection in vertical federated learning[C]//Proceedings of 2020 IEEE 19th International Conference on Trust，Security and Privacy in Computing and Communications (TrustCom). Piscataway，NJ：IEEE，2020：707-714.

[20]SUN Jiankai，YANG Xin，YAO Yuanshun，et al. Vertical federated learning without revealing intersection membership[J]. arXiv preprint arXiv，2106.05508，2021.

[21]ZHOU Zhou，TIAN Youliang，PENG Changgen. Privacy-preserving federated learning framework with general aggregation and multiparty entity matching[J]. Wireless Communications and Mobile Computing，2021，2021：6692061.