面向纵向联邦学习的对抗样本生成算法

陈晓霖[1,2]，咎道广[1,2]，吴炳潮[1,2]，关贝[2,3]，王永吉[2,3]

（1. 中国科学院软件研究所协同创新中心，北京 100190；2. 中国科学院大学计算机科学与技术学院，北京 100049；
3. 中国科学院软件研究所集成创新中心，北京 100190）

摘 要：为了适应纵向联邦学习应用中高通信成本、快速模型迭代和数据分散式存储的场景特点，提出了一种通用的纵向联邦学习对抗样本生成算法 VFL-GASG。具体而言，构建了一种适用于纵向联邦学习架构的对抗样本生成框架来实现白盒对抗攻击，并在该架构下扩展实现了 L-BFGS、FGSM、C&W 等不同策略的集中式机器学习对抗样本生成算法。借鉴深度卷积生成对抗网络的反卷积层设计，设计了一种对抗样本生成算法 VFL-GASG 以解决推理阶段对抗性扰动生成的通用性问题，该算法以本地特征的隐层向量作为先验知识训练生成模型，经由反卷积网络层产生精细的对抗性扰动，并通过判别器和扰动项控制扰动幅度。实验表明，相较于基线算法，所提算法在保持高攻击成功率的同时，在生成效率、鲁棒性和泛化能力上均达到较高水平，并通过实验验证了不同实验设置对对抗攻击效果的影响。

关键词：机器学习；纵向联邦学习；对抗样本；对抗攻击；深度卷积生成对抗网络

中图分类号：TP309.2

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023149

Adversarial sample generation algorithm for vertical federated learning

CHEN Xiaolin[1,2], ZAN Daoguang[1,2], WU Bingchao[1,2], GUAN Bei[2,3], WANG Yongji[2,3]

1. Collaborative Innovation Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, School of Computer Science and Technology, Beijing 100049, China
3. Integrated Innovation Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

Abstract: To adapt to the scenario characteristics of vertical federated learning (VFL) applications regarding high communication cost, fast model iteration, and decentralized data storage, a generalized adversarial sample generation algorithm named VFL-GASG was proposed. Specifically, an adversarial sample generation framework was constructed for the VFL architecture. A white-box adversarial attack in the VFL was implemented by extending the centralized machine learning adversarial sample generation algorithm with different policies such as L-BFGS, FGSM, and C&W. By introducing deep convolutional generative adversarial network (DCGAN), an adversarial sample generation algorithm named VFL-GASG was designed to address the problem of universality in the generation of adversarial perturbations. Hidden layer vectors were utilized as local prior knowledge to train the adversarial perturbation generation model, and through a series of convolution-deconvolution network layers, finely crafted adversarial perturbations were produced. Experiments show that VFL-GASG can maintain a high attack success while achieving a higher generation efficiency, robustness, and generalization ability than the baseline algorithm, and further verify the impact of relevant settings for adversarial attacks.

Keywords: machine learning, VFL, adversarial sample, adversarial attack, DCGAN

收稿日期：2023−06−14；**修回日期**：2023−07−25

通信作者：关贝，guanbei@iscas.ac.cn

基金项目：国家自然科学基金资助项目（No.61762062）

Foundation Item: The National Natural Science Foundation of China (No.61762062)

0 引言

作为人工智能训练的基础，数据在推动机器学习的发展中扮演着核心角色。据预测，互联网数据量在 2025 年将暴增至 175 ZB[1]。而伴随全球范围内公众对数据安全的日益关注以及数据安全相关法律法规的实施，数据在各机构与企业间的流通受到严格约束[2-4]。因此，研究合规的数据获取和使用方法成为人工智能领域的重要挑战。自 2016 年谷歌公司首次将联邦学习[5]作为一种解决方案以来，它便受到了广泛的关注。

根据 Yang 等[6]的划分方式，联邦学习应用可以按照数据分布分为横向联邦学习与纵向联邦学习（VFL）2 种类型。横向联邦学习针对的是参与方数据样本空间不同而数据特征空间相同的情况。纵向联邦学习针对的是参与方数据特征空间不同而数据样本空间相同的情况，这种建模方式适用于不同业务属性的参与方协同建模。纵向联邦学习在国内不同领域得到了广泛应用。例如，微众银行[7]使用本地持有的用户授信数据和合作公司持有的用户消费信息构建了风控领域的纵向联邦神经网络，提高了风险识别能力；字节跳动公司[8]通过纵向联邦树模型引入外部数据源构建了一个推荐领域的纵向联邦树模型，广告投放增效 209%；此外，纵向联邦模型还应用到智能制造[9]和智能电网[10]等领域，在保护数据隐私的同时处理预测任务。

随着纵向联邦学习的广泛应用，其安全性问题逐渐受到研究人员关注。Zhu 等[11]提出了梯度泄露问题，在该问题中，建模过程的中间传输变量或传输梯度会泄露一部分的原始数据信息，恶意参与方据此可以推理原始数据。此后，许多研究人员分别在模型训练阶段和推理阶段设计了不同的攻击方案。在训练阶段，Weng 等[12]针对逻辑回归模型和集成树模型通过反向乘法攻击和反向求和攻击的方式重构原始数据；Fu 等[13]针对纵向联邦学习网络提出了 3 种标签推理攻击的方式，包括补全本地模型的被动标签推理攻击、增加本地模型权重占比的主动标签推理攻击以及梯度标签的直接标签推理攻击。在推理阶段，Luo 等[14]提出了一种模型逆向攻击的通用推理框架，该框架借助生成模型并通过观察模型的输出来反向更新恢复样本的原始数据；Jin 等[15]则利用样本对齐提供的样本空间信息，采用逐层迭代的方法来恢复原始数据；Yang 等[16]利用零阶梯度估计的方式计算模型参数构造推理模型，从而对原始数据进行特征推理攻击。这些方法的攻击目标是纵向联邦数据的隐私性，而纵向联邦学习模型同样可能受到恶意参与方数据篡改，引发模型安全性威胁，例如本文研究的对抗攻击。

联邦学习的对抗攻击指的是恶意参与方在推理阶段利用持有的样本部分特征构建对抗样本，削弱全局模型的预测能力。纵向联邦学习场景具备高通信成本、快速模型迭代和数据分散式存储的特点，具体如下。1)纵向联邦学习框架在实现模型的协同训练或推理过程中，各参与节点需要进行多轮次的参数和数据交换，这造成了极大的通信成本[17]和计算负担[18]。2)纵向联邦学习模型在应用及上线过程中会定期进行版本迭代以适应数据的变化并提升模型性能。③在纵向联邦学习场景中，数据被分散存储于不同的参与方本地[6]，恶意参与方仅能获得本地特征空间权限，并通过勾结或推理攻击[14-15]来获取部分样本空间的其他特征数据，而获取全局样本空间是有挑战性的。基于上述场景特点，纵向联邦场景下生成的对抗样本需要具备较高的生成效率、鲁棒性及泛化能力。

为了解决上述问题，本文首先对纵向联邦学习场景中的对抗攻击、集中式机器学习的对抗样本生成算法和生成对抗网络进行理论介绍；其次，提出了纵向联邦学习的对抗样本生成框架并扩展 6 种集中式机器学习对抗样本生成算法作为基线算法；再次，在上述框架下提出了一种生成算法 VFL-GASG；最后，通过实验验证所提算法相比于其他算法在效率、鲁棒性和泛化性方面的能力，并进一步验证实验参数对攻击效果的影响。本文贡献主要包括以下 3 个方面。

1) 本文在纵向联邦学习场景下提出了一种具备高扩展性的白盒对抗样本生成框架，恶意参与方通过局部样本特征构造对抗样本从而影响模型的准确性，该框架可以适用于基于有限内存的拟牛顿求解（L-BFGS）[19]、符合梯度法（FGSM）[20]等多种集中式机器学习对抗样本生成策略。

2) L-BFGS、FGSM 等对抗样本生成算法均基于输入迭代构造对抗样本，而纵向联邦学习环境中获取全局样本空间是有挑战性的。本文在上述框架下提出了一种基于生成对抗网络的对抗样本生成算法 VFL-GASG，该算法将本地特征的隐层向量作为先验知识训练生成模型，经由反卷积网络产生精细的对抗性扰动，然后借助生成模型从少量样本中学习到对抗性扰动的通用生成算法，从而有较好的泛化能力。

3) 本文在 MNIST 和 CIFAR-10 等数据集上通过实验验证了 VFL-GASG 算法在效率、泛化性、鲁棒性方面的表现。此外，实验发现，即使在只有恶意参与方数量和所持特征数量较少的情况下，对抗攻击也会显著影响模型的性能。最后，通过实验验证不同实验参数对于对抗攻击的影响。

1　理论基础

1.1　纵向联邦学习的对抗攻击

1.1.1　纵向联邦学习架构

纵向联邦学习的参与方数据集在样本维度上有重叠，但在特征维度上分布不同，即不同特征空间的参与方在中央服务器的协调下进行联合建模。图 1 是一种典型的纵向联邦学习训练流程。①中央服务器和各参与方进行样本对齐，使不同参与方的样本空间同步；②参与方依据本地特征和本地模型进行模型计算，并将结果上传作为全局模型的输入；③中央服务器收到上传数据后，依据全局模型进行模型计算；④中央服务器计算梯度和损失，并将梯度进行反向传播；⑤中央服务器和参与方分别更新全局模型与本地模型；⑥重复②～⑤，直到达到收敛条件。纵向联邦学习推理架构和图 1 相似，由于不涉及模型更新，通过①～③就能完成数据推理。多种纵向联邦学习模型均在上述框架下进行部署，如 SecureBoost[21]、Fed-VGAN[22] 和 SplitNN[23]等。

1.1.2　威胁模型

本文假设纵向联邦学习网络中存在某个恶意的参与方，该参与方利用训练完全的目标模型构造对样本进行攻击。纵向联邦学习场景下的威胁模型通过以下 3 个部分进行刻画。

1) 攻击目标

恶意参与方进行对抗攻击的目标是通过在本地所持推理样本的部分特征上添加对抗性扰动，从而使模型在推理样本上分类错误。因此，本文所提算法的攻击效果可以由目标模型在推理集合上分类准确率的下降幅度来度量。

2) 攻击知识

在纵向联邦学习架构中，恶意参与方能够获取其本地数据以及本地模型的详细信息，并且具有训练及推理过程的模型输出和传递参数的完整知识。然而，对于其他参与方的信息，由于存在数据分布和模型架构的差异性，恶意参与方仅能获取其他参与方的嵌入层信息，而无法获知其他参与方的本地数据及模型结构等具体信息。

3) 攻击能力

在纵向联邦学习攻击过程中，恶意参与方可能执行如下操作。①恶意参与方篡改本地数据，以此扰乱模型的训练和推理过程。②恶意参与方修改本地模型和参数信息，并调用全局模型。③恶意参与方通过观测每个通信轮次的模型输出及嵌入层信息来推测其他参与方的信息。

在横向联邦学习攻击过程中，恶意参与方的攻击能力主要聚焦于训练过程的操控，包括在训练阶段篡改数据、模型和参数等信息，以及通过观测信息推测数据隐私。然而，在推理过程中，参与方使用服务器分发的全局模型对本地数据进行独立推理，其推理过程不易受其他参与方的干扰。因此，横向联邦学习中的恶意参与方在推理阶段的攻击威胁性较低[24]，如本文中的对抗攻击。尽管目前的研究中许多学者采用对抗样本的生成算法在横向联邦学习的训练过程中实现数据投毒[25]或增强模型性能[26]，但这并不属于推理过

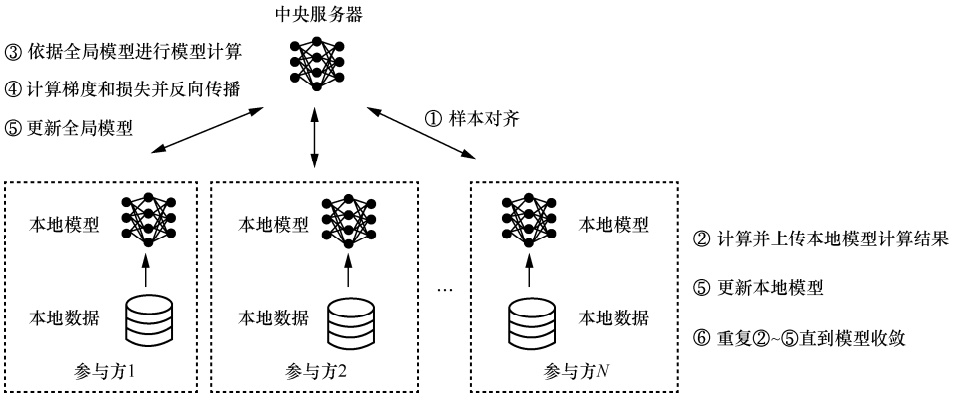


图 1　纵向联邦学习训练流程

程的对抗攻击问题。相比之下，纵向联邦学习需要依靠多参与方进行联合推理，恶意参与方可以通过注入对抗性扰动至本地特征，干扰联合推理的结果。相较于横向联邦学习，纵向联邦学习更容易受到不可信来源的参与方对抗攻击的威胁。

1.2 集中式机器学习对抗样本生成算法

集中式机器学习与联邦学习的分布式策略有所不同，其中心节点通常需要汇总并处理所有数据以训练和推理机器学习模型。Szegedy 等[19]首次在集中式机器学习环境中提出了对抗攻击的概念，并向中心节点的真实样本添加微小扰动，尽管这些扰动并未显著改变样本的整体特性，却可能导致机器学习模型分类错误。集中式机器学习的对抗样本生成流程如图 2 所示。中心节点将训练好的分类模型作为对抗攻击的目标模型；然后，通过向原始图像添加扰动生成对抗样本，并通过目标模型的输出进行扰动迭代；直至满足设定条件，从而生成最终的对抗样本。

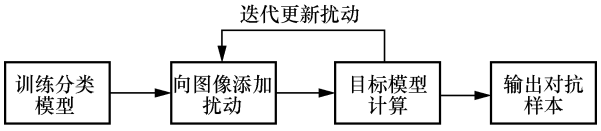


图 2 集中式机器学习的对抗样本生成流程

在集中式机器学习环境中，已经发展出了多种对抗样本生成策略，其中代表性方法包括基于优化求解策略的 L-BFGS[19]和 C&W（Carlini&Wagner）[27]，基于梯度迭代策略的 FGSM[20]、I-FGSM[28]、MI-FGSM[29]，以及基于像素级扰动策略的显著图攻击（JSMA）[30]。下面介绍不同对抗样本生成算法的理论部分。

1) L-BFGS

Szegedy 等[19]最早提出了集中式机器学习的对抗样本生成算法 L-BFGS，该方法将对抗样本的扰动计算转化为如下问题并通过拟牛顿法求解

$$\min_{\boldsymbol{r}} c|\boldsymbol{r}|+\ell(\boldsymbol{x}+\boldsymbol{r},l)$$
$$\text{s.t. } \boldsymbol{x}+\boldsymbol{r}\in[0,1]^m \tag{1}$$

其中，$\boldsymbol{x}$ 是输入的原始图像特征，l 是错误的分类标签，$\boldsymbol{r}$ 是需要加入的对抗性扰动向量，ℓ 是损失函数。

2) FGSM

Goodfellow 等[20]提出了 FGSM 方法，该方法通过计算模型损失相对于输入数据的梯度，在有限扰动的情况下沿着损失函数最大化的方向进行更新。对抗性扰动表示为

$$\boldsymbol{r}=\epsilon\ \text{sign}(\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x},y)) \tag{2}$$

其中，ϵ 用于控制扰动大小，$\boldsymbol{x}$和y 分别是输入图像特征和真实标签。

3) I-FGSM

Kurakin 等[28]采用迭代的策略，将增大分类器损失函数的过程分解为多个迭代步骤，迭代过程为

$$\boldsymbol{x}^{t+1}=\text{Clip}_{\boldsymbol{x},\epsilon}\{\boldsymbol{x}^t+\alpha\text{sign}(\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}^t,y))\} \tag{3}$$

其中，α 表示每次迭代的步长；Clip 表示裁剪函数，其对生成样本的每个像素进行裁剪，使迭代数据保持在原始图像的 ϵ 邻域范围内。

4) MI-FGSM

Dong 等[29]为了提升攻击的成功率，通过动量迭代来替代 I-FGSM 中的梯度迭代信息策略，同时引入历史扰动来规避迭代过程可能产生的局部最优问题，提高迭代过程的稳定性。该过程可表示为

$$\boldsymbol{r}^{t+1}=\mu\boldsymbol{r}^t+\frac{\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}^t,y)}{\left\|\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x}^t,y)\right\|}$$
$$\boldsymbol{x}^{t+1}=\boldsymbol{x}^t+\alpha\text{sign}(\boldsymbol{r}^{t+1}) \tag{4}$$

其中，$\boldsymbol{r}^{t+1}$ 和 $\boldsymbol{x}^{t+1}$ 分别表示第 $t+1$ 次迭代产生的扰动和对抗样本，μ 表示衰减因子。

5) C&W

Carlini 等[27]则在 L-BFGS 的优化问题基础上，通过映射变换解决像素溢出的问题，并引入置信度函数来控制错误概率。优化问题表示为

$$\min_{\boldsymbol{w}}\left\|\frac{\tanh(\boldsymbol{w})+1}{2}-\boldsymbol{x}\right\|_2^2+cf\left(\frac{\tanh(\boldsymbol{w})+1}{2}\right)$$
$$f(\boldsymbol{h})=\max(\max\{Z(\boldsymbol{h})_i:i\neq t\}-Z(\boldsymbol{h})_t,-k) \tag{5}$$

其中，$\frac{\tanh(\boldsymbol{w})+1}{2}-\boldsymbol{x}$ 表示添加的扰动，k 用于控制分类错误置信度，$Z(\boldsymbol{h})_i$ 表示未经过 softmax 层输出结果的第 i 类值。

6) JSMA

Papernot 等[30]提出一种通过类别梯度生成像素显著图来筛选干扰像素对的方法。显著性映射为

$$S(\boldsymbol{x},l)[i]=\begin{cases}0, & \dfrac{\partial Z_l(\boldsymbol{x})}{\partial\boldsymbol{x}_i}<0 \text{ 或 } \displaystyle\sum_{j\neq t}\frac{\partial Z_j(\boldsymbol{x})}{\partial\boldsymbol{x}_i}>0\\ \dfrac{\dfrac{\partial Z_l(\boldsymbol{x})}{\partial\boldsymbol{x}_i}}{\displaystyle\sum_{j\neq t}\frac{\partial Z_j(\boldsymbol{x})}{\partial\boldsymbol{x}_i}}, & \text{其他}\end{cases} \tag{6}$$

其中，$Z_l(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 在分类层 l 类上的输出。该方法利用显著图来筛选出对目标模型分类输出影响最大的特征对，并在对应像素点上施加扰动。

1.3 生成对抗网络

生成对抗网络（GAN, generative adversarial network）由 Goodfellow 等[31]提出，它由生成器 $\mathcal{G}$ 和判别器 $\mathcal{D}$ 两部分组成。GAN 的核心思想是通过生成器和判别器的对抗训练，学习数据的隐含分布，从而生成与真实数据相似的样本。具体来说，生成器 $\mathcal{G}$ 接收低维随机噪声作为输入并输出伪造样本，其目标是生成能够欺骗判别器的样本，使判别器难以区分真实样本和生成的伪造样本。判别器 $\mathcal{D}$ 的任务是对输入样本进行判断，输出其作为真实样本的概率。综合考虑生成器和判别器的优化目标，可以表示为

$$\min_{\mathcal{G}}\max_{\mathcal{D}}\mathbb{E}_{\boldsymbol{x}}[\log\mathcal{D}(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{z}}[1-\log\mathcal{D}(\mathcal{G}(\boldsymbol{z}))] \quad (7)$$

自 Goodfellow 首次提出生成对抗网络以来，许多研究者对其进行了改进。Arjovsky 等[32]采用 Wasserstein 距离作为损失函数以缓解训练过程的梯度消失问题。Radford 等[33]提出了深度卷积生成对抗网络（DCGAN, deep convolutional generative adversarial network），这种方法将 GAN 中的生成器全连接层替换为反卷积层，显著降低了模式坍塌现象出现的频率。在 DCGAN 方法中，卷积层通过滑动窗口机制有效地提取高维图像特征，反卷积层则利用这些特征以逆卷积的操作进行图像重建。该方法显著地提高了在图像生成任务中生成对抗网络的训练稳定性。

2 纵向联邦学习对抗样本生成

2.1 问题定义

假设纵向联邦学习系统中有 $N+1$ 个参与方，分别是 N 个诚实参与方和一个恶意参与方。每个诚实参与方持有特征数据 $\boldsymbol{X}_{\text{hon}}$，特征维度为 d_{hon}；恶意参与方持有特征数据 $\boldsymbol{X}_{\text{adv}}$，特征维度为 d_{adv}。数据总和 $\boldsymbol{X}=\{\boldsymbol{X}_{\text{hon}},\boldsymbol{X}_{\text{adv}}\}=\{\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N,\boldsymbol{X}_{N+1}\}$，特征总维度为 $d=d_{\text{adv}}+\sum_{i=1}^{N}d^i$。依据本地模型 $f_i(\cdot)$、全局模型 $G(\cdot)$，纵向联邦学习环境中的对抗样本生成问题形式化表示为

$$\min_{\boldsymbol{r}}\|\boldsymbol{r}\|_2^2$$
$$\text{s.t.}\quad G\big(f_{\text{hon}}(\boldsymbol{X}_{\text{hon}}),f_{\text{adv}}(\boldsymbol{X}_{\text{adv}}+\boldsymbol{r})\big)=l \quad (8)$$

即最小化恶意参与方对抗性扰动的同时，使中央服务器全局模型在对抗样本上分类错误。表 1 定义了纵向联邦学习对抗样本生成的相关参数。

表 1　相关参数

参数	含义
N	纵向联邦学习诚实参与方数量
S	中央服务器
C_i	第 i 个参与方
$G(\cdot)$	中央服务器全局模型
$f_i(\cdot)$	第 i 个参与方的本地模型
$f_{\text{hon}}(\cdot)$	诚实参与方本地模型
$f_{\text{adv}}(\cdot)$	恶意参与方本地模型
$\boldsymbol{X}_i$	第 i 个参与方持有的本地数据
$\boldsymbol{X}_{\text{hon}}$	诚实参与方持有的本地数据
$\boldsymbol{X}_{\text{adv}}^t$	恶意参与方第 t 次迭代后的对抗样本
$\boldsymbol{X}_{\text{adv}}$	恶意参与方持有的本地数据
d_i	第 i 个参与方的本地数据维度
d_{hon}	诚实参与方的本地数据维度
d_{adv}	恶意参与方的本地数据维度
$\boldsymbol{r}$	恶意参与方添加噪声向量
$\boldsymbol{r}^t$	恶意参与方第 t 次迭代后的噪声向量
l	目标标签
ϵ	扰动超参数，用于控制扰动大小
$\mathcal{G}$	生成对抗网络生成器
$\mathcal{D}$	生成对抗网络判别器
ℓ_{adv}	目标模型分类损失
ℓ_{GAN}	对抗网络损失

2.2 纵向联邦学习对抗样本生成框架

如图 3 所示，本文提出的纵向联邦学习对抗样本生成框架的流程如下：首先，每个参与方根据其本地模型和数据生成全局模型的输入；接着，联邦学习服务器根据全局模型执行计算，产生用于扰动的更新参数，并反馈给各个参与方，恶意参与方根据回传梯度等参数更新对抗样本；之后，对抗样本被作为恶意参与方的本地模型输入，经迭代更新，直至达到预设迭代次数或其他条件；最终，输出完成的对抗样本。详细过程如算法 1 所示。

算法 1　纵向联邦学习对抗样本生成框架

输入　中央服务器 S，诚实参与方 $C_1,\cdots,C_N$，恶意参与方 C_{adv}，本地样本推理 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N,\boldsymbol{X}_{N+1}$

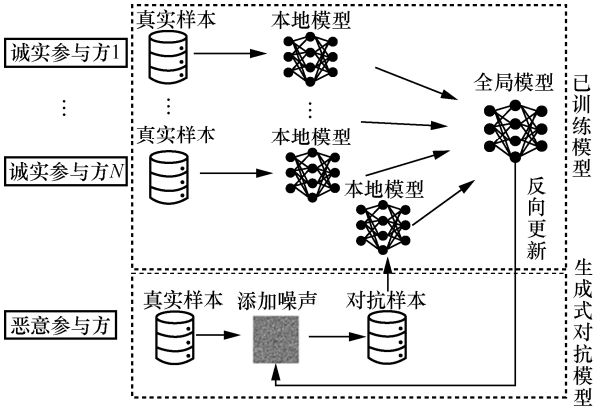


图 3　纵向联邦学习对抗样本生成框架

输出　恶意参与方对抗样本 $\boldsymbol{X}_{\text{adv}}^{T}$

1)　for　C_i $(i \leqslant N)$　do

2)　　基于本地模型计算 $f_i(\boldsymbol{X}_i)$ 并上传至服务器 S

3)　end for

4)　for　$t=0,1,\cdots,T-1$

5)　　for　C_{adv}　do

6)　　基于本地模型计算 $f_{\text{adv}}\left(\boldsymbol{X}_{\text{adv}}^{t}\right)$ 并上传至中央服务器 S

7)　　收集中央服务器回传参数更新对抗样本，并进行裁剪 $\boldsymbol{X}_{\text{adv}}^{t+1}=\text{Clip}_{\boldsymbol{x},\epsilon}\left\{\boldsymbol{X}_{\text{adv}}^{t}+\boldsymbol{r}^{t}\right\}$

8)　　end for

9)　　for　S　do

10)　　收集所有参与方上传数据 $f_i(\boldsymbol{X}_i)$

11)　　全局模型计算 $G\left(f_{\text{hon}}(\boldsymbol{X}_{\text{hon}}),\ f_{\text{adv}}\left(\boldsymbol{X}_{\text{adv}}^{t+1}\right)\right)$

12)　　计算损失函数和梯度等回传参数，并将回传参数发送至恶意参与方 C_{adv}

13)　　end for

14)　end for

15)　return　$\boldsymbol{X}_{\text{adv}}^{T}$

在该框架下，集中式机器学习对抗样本生成策略可以扩展至纵向联邦学习场景中，本文扩展了多种生成算法，如表 2 所示，其中，VFL-LBFGS 和 VFL-C&W 通过优化策略生成对抗样本，VFL-JSMA 采用像素级扰动方法生成对抗样本，VFL-FGSM、VFL-IFGSM 和 VFL-MIFGSM 则是通过梯度迭代更新的方式产生扰动，这些算法将作为基准算法参与后续实验验证。

2.3　基于 GAN 的对抗样本生成

为了解决对抗性扰动的通用性问题，本文在算法 1 框架下提出了一种基于 GAN 的对抗样本生成算法 VFL-GASG。其中，生成器负责生成噪声并将其注入真实样本形成对抗样本；判别器则对真实样本和对抗样本进行分类，以便更精确地辨别这 2 类样本。该过程可分为模型训练和对抗样本生成 2 个阶段。模型训练阶段指的是利用训练样本完成生成器和判别器的训练，以达到预定的生成目标；而对抗样本生成阶段则是指恶意参与方利用训练好的生成器生成对抗样本的过程。

图 4 是基于 GAN 的对抗样本生成模型的训练过程，其中，生成模型由多层卷积网络和反卷积网络组成。在该模型中，卷积网络负责处理输入图像，提取深层特征；反卷积网络采用一种与卷积过程相反的运算，从特征空间映射回原始图像空间，它利用由卷积网络提取的特征，加入一定随机噪声，重建出具有对抗性扰动的图像。这种方法有效地避免

表 2　　**纵向联邦学习对抗样本生成算法**

生成算法	目标函数	回传参数	更新方式
VFL-LBFGS	$c\lvert\boldsymbol{r}\rvert+\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},l\right)$	$\nabla\boldsymbol{x}_{\text{adv}}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},l\right)$	$\boldsymbol{x}_{\text{adv}}^{t+1}=\boldsymbol{x}_{\text{adv}}^{t}+\alpha^{t}(\beta^{t})^{-1}\left(c\dfrac{\boldsymbol{r}}{\lvert\boldsymbol{r}\rvert}+\nabla\boldsymbol{x}_{\text{adv}}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},l\right)\right)$
VFL-FGSM	$\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}},y\right)$	$\nabla\boldsymbol{x}_{\text{adv}}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}},l\right)$	$\boldsymbol{x}_{\text{adv}}=x+\nabla\boldsymbol{x}_{\text{adv}}\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}},y\right)$
VFL-IFGSM	$\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},y\right)$	$\nabla\boldsymbol{x}_{\text{adv}}^{t}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},y\right)$	$\boldsymbol{x}_{\text{adv}}^{t+1}=\boldsymbol{x}_{\text{adv}}^{t}+\nabla\boldsymbol{x}_{\text{adv}}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},y\right)$
VFL-MIFGSM	$\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},y\right)$	$\boldsymbol{g}^{t}=\nabla\boldsymbol{x}_{\text{adv}}^{t}\ell\left(\boldsymbol{x}_{\text{hon}},\boldsymbol{x}_{\text{adv}}^{t},y\right)$	$\boldsymbol{r}^{t+1}=\mu\boldsymbol{r}^{t}+\dfrac{\boldsymbol{g}^{t}}{\left\lVert\boldsymbol{g}^{t}\right\rVert},\boldsymbol{x}_{\text{adv}}^{t+1}=\boldsymbol{x}_{\text{adv}}^{t}+\text{sign}(\boldsymbol{r}^{t+1})$
VFL-C&W	$\ell_{\text{CW}}\left\lVert\boldsymbol{x}_{\text{adv}}^{t}-\boldsymbol{x}\right\rVert_2^{2}+cf\left(\boldsymbol{x}_{\text{adv}}^{t}\right)$	$\nabla_{\boldsymbol{w}}\ell_{\text{CW}}$	$\boldsymbol{w}=\boldsymbol{w}+\nabla_{\boldsymbol{w}}\ell_{\text{CW}},\boldsymbol{x}_{\text{adv}}^{t+1}=\dfrac{\tanh(\boldsymbol{w})+1}{2}$
VFL-JSMA	—	$(p_1,p_2)=\arg\max_i S\left[\boldsymbol{x}_{\text{adv}},l\right]_i$	$\boldsymbol{x}_{\text{adv},p_1}=\boldsymbol{x}_{\text{adv},p_1}+\epsilon$ $\boldsymbol{x}_{\text{adv},p_2}=\boldsymbol{x}_{\text{adv},p_2}+\epsilon$

了由于低维特征重建引发的模式坍塌现象。这种对抗样本对目标模型形成有效的干扰，从而在维持原有识别精度的同时，提高了对抗攻击的有效性。判别器则由卷积网络组成，对输入的真实特征和伪造特征进行二分类。

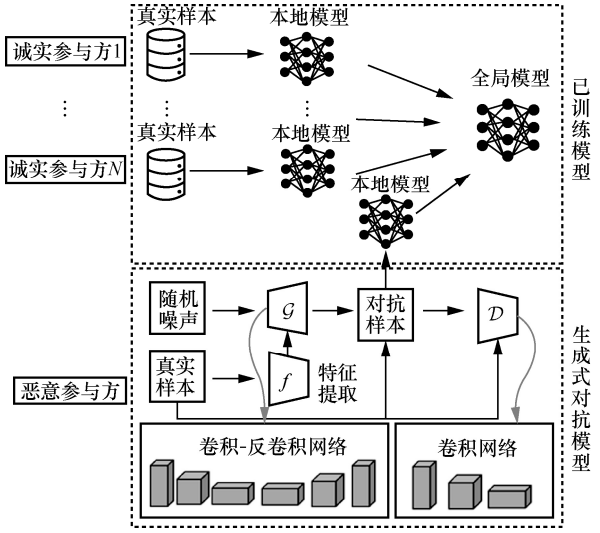


图 4　基于 GAN 的对抗样本生成模型的训练过程

在训练过程中，诚实参与方根据其本地模型和样本特征执行本地计算，并将结果作为全局模型输入上传。恶意参与方通过生成器对真实特征进行编码来产生对抗性扰动，然后将这些扰动添加到真实样本上形成对抗样本。接着，恶意参与方将对抗样本输入到本地模型中进行计算，并将结果上传至服务器。服务器根据全局模型计算模型输出及损失 ℓ_{adv}。恶意参与方将真实样本和对抗样本输入判别器，计算对抗损失 ℓ_{GAN}。计算完成后，判别器和生成器根据全局损失 ℓ 分别更新模型参数。恶意参与方重复上述过程，直到收敛后保存模型 $\mathcal{G}$、$\mathcal{D}$。具体过程如算法 2 所示。

算法 2　VFL-GASG 模型的训练过程

输入　中央服务器 S，诚实参与方 $C_1,\cdots,C_N$ 和恶意参与方 C_{adv}，本地训练样本 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_N,\boldsymbol{X}_{\text{adv}}$

输出　训练完成的生成器 $\mathcal{G}$ 和判别器 $\mathcal{D}$

1) for C_i $(i\leqslant N)$ do
2) 　基于本地模型计算 $f_i(\boldsymbol{X}_i)$ 并上传服务器 S
3) end for
4) for $t=0,1,\cdots,T-1$
5) 　for C_{adv} do
6) 　　$t=0$ 时，初始化生成器 $\mathcal{G}$ 和判别器 $\mathcal{D}$
7) 　　添加随机噪声并通过生成器生成扰动 $\mathcal{G}\left(\boldsymbol{X}_{\text{adv}}^{t},Z\right)$
8) 　　更新对抗样本 $\boldsymbol{X}_{\text{adv}}^{t+1}=\mathcal{G}\left(\boldsymbol{X}_{\text{adv}}^{t},Z\right)+\boldsymbol{X}_{\text{adv}}^{t}$
9) 　　本地计算 $f_{\text{adv}}\left(\boldsymbol{X}_{\text{adv}}^{t+1}\right)$ 并上传至服务器 S
10) 　　判别器 $\mathcal{D}$ 对真实样本和对抗样本进行分类
11) 　　计算对抗损失 $\ell_{\text{GAN}}=\mathbb{E}_{\boldsymbol{x}}[\log\mathcal{D}(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{x}}\left[\log\left(1-\mathcal{D}\left(\boldsymbol{x}+\mathcal{G}(\boldsymbol{x})\right)\right)\right]$
12) 　　$t\neq 0$ 时，收集服务器回传参数，判别器和生成器根据总损失函数 ℓ 更新，并裁剪 $\ell=\ell_{\text{adv}}+\lambda_{\text{GAN}}\ell_{\text{GAN}}+\lambda_r\mathbb{E}_{\boldsymbol{x}}[\mathcal{G}(\boldsymbol{x})]$
13) 　end for
14) 　for S do
15) 　　收集所有参与方上传数据
16) 　　全局模型计算 $G\left(f_{\text{hon}}(\boldsymbol{X}_{\text{hon}}),f_{\text{adv}}\left(\boldsymbol{X}_{\text{adv}}^{t+1}\right)\right)$
17) 　　计算损失 $\ell_{\text{adv}}=\mathbb{E}_{\boldsymbol{x}}[\ell\left(\boldsymbol{x}+\mathcal{G}(\boldsymbol{x}),l\right)]$ 并发送至恶意参与方 C_{adv}
18) 　end for
19) end for
20) return $\mathcal{G},\mathcal{D}$

图 5 是基于 GAN 的对抗样本生成模型的生成过程。在该过程中，恶意参与方将给定推理样本集合 $\boldsymbol{X}'_{\text{adv}}$ 输入生成模型产生扰动 $\mathcal{G}(\boldsymbol{X}'_{\text{adv}},Z)$，然后这些扰动被添加到推理样本上形成对抗样本 $\boldsymbol{X}'_{\text{adv}}+\mathcal{G}(\boldsymbol{X}'_{\text{adv}},Z)$，恶意参与方将其作为本地模型输入，参与后续推理。

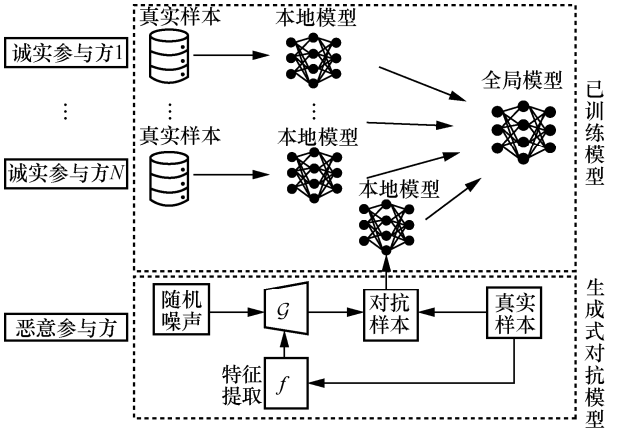


图 5　基于 GAN 的对抗样本生成模型的生成过程

3　实验评估

本节首先详述实验设定，包括实验环境、数据集和模型结构；其次展示不同对抗样本生成算法所产生的对抗样本及其攻击效果；再次通过实验验证 VFL-GASG 在效率、鲁棒性和泛化性上的能力；最后在纵向联邦环境中，通过实验分析不同参数对对抗攻击的影响。

3.1 实验设置

1) 实验环境

本文实验运行于以下环境：Intel(R) Xeon(R) Gold，64 内核，2.3 GHz，内存 128 GB，V100 GPU，Ubuntu 16.04 操作系统。本文提出的对抗样本生成算法基于 Pytorch 框架和 foolbox 库实现。

2) 数据集

本文数据集选用了 MNIST、CIFAR-10 和 ImageNet-100。其中，MNIST 数据集是一个广泛应用于图像分类任务的手写数字图像数据库，包含 60 000 个训练样本和 10 000 个测试样本，每个样本都是 28 像素×28 像素的灰度图像；CIFAR-10 数据集也常用于图像分类任务，是一个包含 10 个不同类别的彩色图像数据集，每个图像为 32 像素×32 像素，总计包含 50 000 个训练样本和 10 000 个测试样本；ImageNet-100 数据集是 ImageNet 大规模视觉识别任务的子数据集，包括 100 个类别的样本数据，共计 50 000 个训练图像和 10 000 个测试图像，经处理后每个图像为 214 像素×214 像素。

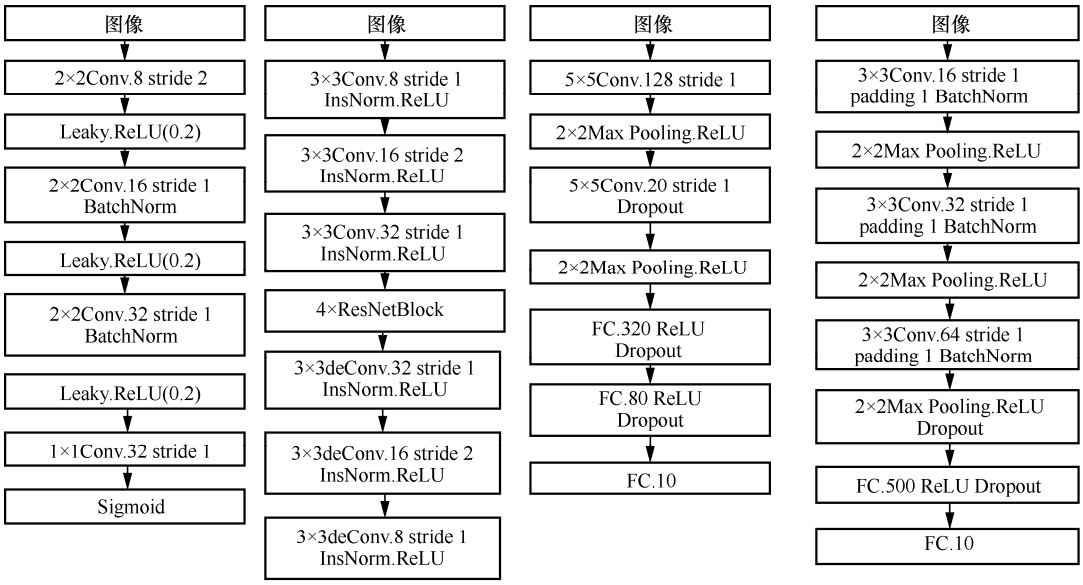
本文构建了一个包含 3 个参与方的仿真纵向联邦学习环境。其中，2 个参与方被设定为诚实参与方，主要负责本地模型的计算和数据上传；另一个参与方被设定为恶意参与方，其任务是生成对抗样本。在基于特征维度的数据划分过程中，诚实参与方和恶意参与方所持有的特征数量的比例保持在 1:1，这种设计能够在相对平衡的初始环境中评估和比较算法性能。

3) 模型架构

本文将目标模型在测试集上的分类准确率下降幅度作为对抗攻击的度量，首先训练目标模型，然后在测试样本上添加扰动生成对抗样本，统计对抗样本在目标模型上的准确率。对抗样本在目标模型的准确率越低，那么对抗攻击的效果越好。在构建本地模型和全局模型的过程中，本文设计了不同的分类模型（图 6 的目标模型 1 与目标模型 2），经过训练后，该模型在 MNIST 和 CIFAR-10 测试集上的准确率分别为 97.27%和 79.91%，而选用 ResNet 模型作为 ImageNet-100 的目标模型，分类准确率为 65.09%。本文采用深度卷积生成对抗网络来改善训练的稳定性。判别器包括 4 层卷积网络，生成器包括 3 层卷积网络、4 层 ResNetBlock 和 3 层反卷积网络。

3.2　对抗样本生成

为了直观展示对抗样本的生成效果，本文在上述实验环境中进行重建数据的可视化，分别在 MNIST 和 CIFAR-10 数据集上生成对抗样本，当恶



(a) 判别器　(b) 生成器　(c) 目标模型1　(d) 目标模型2

图 6　GAN 及目标模型网络结构

意参与方完成计算后，将不同参与方的样本特征进行整合来直观展示对抗样本生成的效果。如图 7 和图 8 所示，尽管样本在局部位置出现了微小的扰动，然而总体上并未影响图像的辨认。

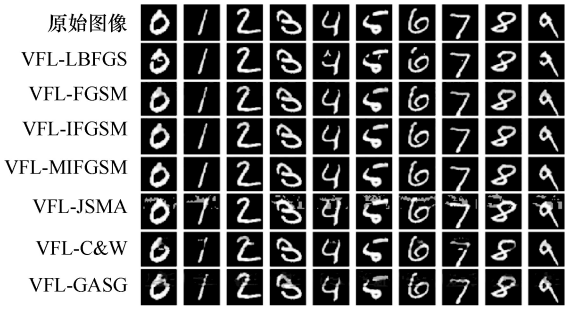


图 7　在 MNIST 上生成的对抗样本

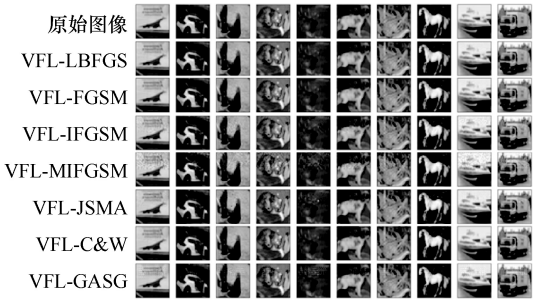


图 8　在 CIFAR-10 上生成的对抗样本

本文进一步在 ImageNet-100 数据集上生成对抗样本，如图 9 所示。相比于 MNIST 和 CIFAR-10 上生成的对抗样本，ImageNet-100 上样本的轮廓更清晰，视觉效果与原始图像更相似，噪声的直观感受更小。

图 9　在 ImageNet-100 上生成的对抗样本

表 3 统计了不同对抗样本生成算法在目标模型上的分类准确率。结果表明，所有的对抗攻击都导致了模型分类准确率的大幅下降，这证明了对抗样本生成框架的有效性。然而，不同生成算法的效果存在差异，VFL-C&W 生成的对抗样本使模型的分类准确率下降最多，其能够在引入细小对抗扰动的同时保持较高的攻击效果。VFL-JSMA 通过选择对分类结果影响最大的像素对来生成对抗样本。尽管这种算法改变的像素较少，但个别像素值的变动较大，因此，在实际的对抗攻击中，这种算法生成的对抗样本更容易被识别。VFL-FGSM、VFL-IFGSM、VFL-MIFGSM、VFL-LBFGS 算法在攻击效果上相近。而本文提出的 VFL-GASG 算法在攻击效果上接近于 VFL-C&W 算法，这证明该算法在纵向联邦学习场景中可以保持较高的攻击成功率。

3.3　效果对比

在效率方面，本文在 3 个数据集的测试图像上分别生成了 10 000 个对抗样本，并统计了所需时间，如表 4 所示。在使用 VFL-GAN 生成对抗样本的过程中，恶意参与方仅需将原始图像输入生成器，生成器就可以计算出相应扰动，这种方式减少了参与方之间的数据传输次数，因此在计算复杂度上具有明显优势。基于表 4，本文提出的 VFL-GASG 算法在对抗样本的生成效率上明显优于其他算法，生成样本耗时分别为 5.51 s、9.99 s 和 167.89 s，这甚至少于 VFL-FGSM 算法。虽然在表 2 中 VFL-C&W 算法展示了较好的攻击效果，但其代价是极高的计算复杂度，VFL-C&W 的计算耗时远超其他方法。而本文提出的 VFL-GASG 算法不仅保持了高攻击成功率，同时在生成效率上也表现出显著优势。

为了探究对抗样本对模型的鲁棒性，本文调整了 3 个数据集上目标模型的参数和网络结构，得到新的目标模型 A、B 和 C。接着，将原对抗样本输入这些模型，通过观察对抗样本在新目标模型上的准确率变化来评估对抗样本对于模型的鲁棒性。如表 4 所示，即使调整了目标模型，这些对抗样本仍能显著降低模型的准确率，特别是 VFL-GASG 生成的对抗样本依然使不同模型分类准确率大幅下降，这证明联邦模型迭代后，VFL-GASG 生成的对抗样本仍能有效干扰模型。

表 5 展示了基于不同训练样本比例所训练的生成模型所产生的扰动对目标模型准确性的影响，其中极差反映了度量指标在各比例间的变动幅度。结

表 3 不同对抗样本生成算法在目标模型上的分类准确率

对抗样本生成算法	MNIST		CIFAR-10		ImageNet-100	
	Top1	Top3	Top1	Top3	Top1	Top3
VFL-FGSM	92.13%	99.26%	48.22%	84.91%	49.60%	69.93%
VFL-IFGSM	85.42%	98.72%	41.25%	82.86%	40.42%	67.46%
VFL-MIFGSM	80.32%	91.51%	43.04%	89.45%	37.51%	53.42%
VFL-LBFGS	82.29%	95.86%	45.89%	82.54%	49.65%	63.73%
VFL-JMSA	87.00%	97.72%	57.65%	89.65%	37.93%	51.98%
VFL-C&W	41.24%	**55.45**%	**20.45**%	76.37%	**19.14**%	**40.62**%
VFL-GASG	**31.24**%	92.89%	43.00%	**75.21**%	22.36%	47.41%
目标模型	97.27%	99.78%	79.91%	95.36%	65.09%	75.74%

表 4 不同对抗样本生成算法的计算耗时和鲁棒性对比

对抗样本生成算法	MNIST			CIFAR-10			ImageNet-100		
	计算耗时/s	目标模型 1	模型 A	计算耗时/s	目标模型 2	模型 B	计算耗时/s	目标模型 3	模型 C
VFL-FGSM	19.64	92.13%	94.25%	12.36	48.22%	61.34%	197.92	49.60%	56.34%
VFL-IFGSM	56.23	85.42%	91.83%	41.25	41.25%	59.84%	471.50	40.42%	55.41%
VFL-MIFGSM	57.75	80.32%	86.26%	51.89	43.04%	67.63%	563.23	37.51%	49.92%
VFL-LBFGS	935.92	82.29%	89.24%	269.93	45.89%	60.93%	2669.86	49.65%	52.53%
VFL-JSMA	1771.82	87.00%	91.52%	1524.34	57.65 %	68.42%	6404.60	37.93%	51.69%
VFL-C&W	9360.65	41.24%	59.25%	8248.43	**20.45**%	**51.14**%	33849.15	**19.14**%	42.82%
VFL-GASG	**5.51**	**31.24**%	**52.21**%	**9.99**	43.00%	58.57%	**167.89**	22.36%	**38.84**%

表 5 不同训练样本比例的攻击效果

训练样本比例	MNIST		CIFAR-10		ImageNet-100	
	Top1	Top3	Top1	Top3	Top1	Top3
5%	33.16%	92.13%	42.90%	74.39%	23.87%	48.27%
10%	30.90%	92.94%	41.79%	73.39%	24.92%	47.48%
20%	30.73%	92.50%	43.51%	76.01%	25.83%	49.45%
40%	31.24%	92.86%	43.48%	76.65%	23.65%	48.18%
80%	30.96%	92.88%	42.14%	75.05%	22.83%	48.32%
100%	31.24%	92.89%	43.00%	75.21%	22.36%	47.41%
极差	2.43%	0.81%	1.72%	3.26%	3.47%	2.04%

果显示，即使训练样本的比例存在差异，VFL-GASG 算法都能显著降低目标模型的 Top1 分类准确率，这证实了 VFL-GASG 在部分样本上对抗性扰动生成算法的泛化能力。在纵向联邦学习的环境中，即使泄露了少量样本信息，该算法也能够利用这些信息构建噪声生成网络，进而构造对抗攻击。相比之下，其他对抗样本生成算法需要对每个输入样本进行独立的计算与迭代，其泛化能力相对较差。

3.4 不同实验设置的影响

在对抗攻击过程中，攻击者利用其持有的特征创建对抗样本以混淆目标模型，恶意参与方特征数量可能影响对抗攻击的效果。为了探究恶意参与方特征数量对攻击效果的影响，本节在不同恶意参与方特征数量条件下评估了不同对抗样本

生成算法的攻击效果，实验结果如图 10 所示。本节实验分别在 MNIST、CIFAR-10、ImageNet-100 数据集上根据不同的敌手特征数量对图像划分，例如，在 Imagenet 数据集上，敌手客户端分别持有 0～224 维的不同高度像素，其余像素则均匀分配给诚实参与方。随着恶意参与方特征数量的增加，模型的分类准确率在一定范围内呈线性下降趋势。值得注意的是，即使敌手只持有较少的特征，其攻击也能显著影响模型性能，例如当特征数量分别为 8、8 和 56 时，VFL-GASG 在 3 个数据集上产生的对抗攻击使模型分类准确率分别下降 30.80%、22.36%和 17.62%；当恶意参与方特征数量超过总体特征的 50%时，模型分类准确率均保持在较低水平。

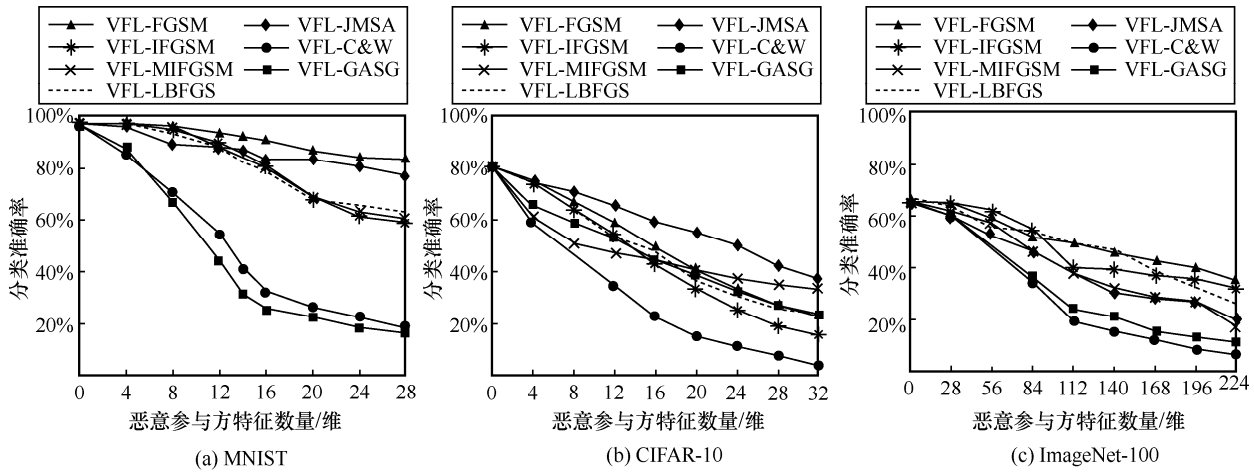


图 10　恶意参与方特征数量对于对抗攻击的影响

本节探究了不同恶意参与方占比对纵向联邦对抗攻击的影响，结果如图 11 所示。实验设置在 8 个参与方中引入不同比例（0%、12.5%、25%、50%）的恶意参与方，评估其对模型性能的影响。实验表明恶意客户端数量的增加引发了对抗攻击效果的显著提升，并导致目标模型分类准确率大幅降低。即使恶意参与方比例仅为 12.5%，模型分类准确率也表现出显著下滑。当恶意参与方比例达到 50%时，模型分类准确率进一步降低至 31.38%、42.89%和 22.36%。以上结果表明，保证联邦学习环境的安全性以防止恶意参与方的参与，对于维持模型的高分类准确率具有至关重要的意义。

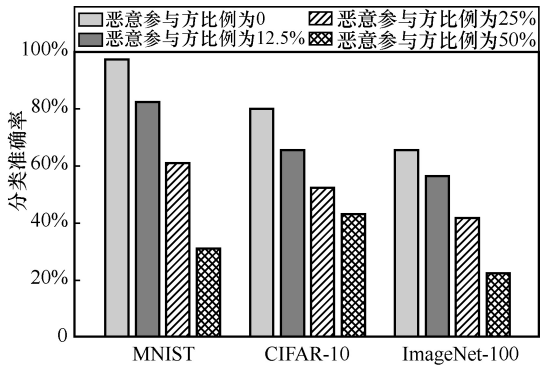


图 11　恶意参与方占比对纵向联邦对抗攻击的影响

本节针对 MNIST、CIFAR-10 和 ImageNet-100 数据集，探究了参与方数量和实验参数配置如何影响 VFL-GASG 对抗攻击的效果，如图 12 所示。在将参与方总数设为 2、4、6、8 的不同场景中，设定其中一个参与方作为恶意参与方并进行特征均匀分配。图 12 的实验结果表明，随着参与方数量的增长，对抗攻击的攻击准确率显著降低，这是因为受扰动的特征数量随着参与方数量的增加而减少，而其他参与方提供的建模特征则会缓解恶意参与方对攻击的影响。另一方面，实验参数设定可以影响损失优化方向中各项的相对重要性。当扰动参数 λ_r 较大时，模型的攻击准确率降低较慢，意味着模型优化过程更注重于减小扰动，而不是攻击效果。反之，当生成对抗网络的参数 λ_{GAN} 增大时，生成模型则更多地倾向于降低目标模型的性能，以提高对抗攻击效果。值得注意的是，在 MNIST 和 CIFAR 数据集上 10～12 轮的部分准确率曲线有上升趋势，这是因为 λ_r 相比于 λ_{GAN} 过小，导致模型更倾向于牺牲对抗性扰动而注重提高攻击效果，但经裁剪后的图像使攻击效果反而下降。这些实验结果表明，合适的模型参数对于对抗攻击 VFL-GASG 尤其关键。

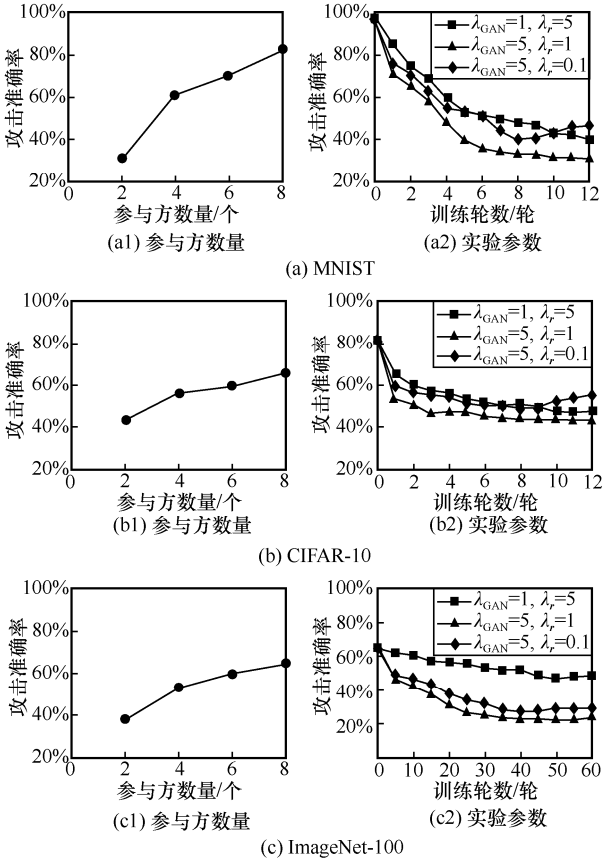


图 12　参与方数量和实验参数配置对 VFL-GASG 攻击效果的影响

4　结束语

本文探讨了纵向联邦学习环境下的对抗样本生成问题。首先，在纵向联邦学习架构的背景下详细分析了对抗攻击的威胁模型，进而为纵向联邦学习环境构建了一种具有高扩展性的对抗样本生成框架，并扩展了不同机器学习对抗样本生成策略作为基准算法。鉴于生成算法的泛化能力的要求，提出了一种基于 GAN 的对抗样本生成算法 VFL-GASG，该算法在生成模型中通过卷积−反卷积网络构建精细化扰动。通过在多个数据集上的实验证明，该算法在生成效率、鲁棒性和泛化性方面表现出优良的性能，符合纵向联邦学习环境的需求。即便在恶意参与方数量较少或其持有的特征数量较少的情况下，模型的准确性也会受到对抗攻击的显著影响。进一步，通过实验分析了可能影响对抗攻击效果的各种因素，为后续研究提供有意义的借鉴。

参考文献：

[1] JOHN R, DAVID R, JOHN G. Data age 2025: the digitization of the world from edge to core[R]. 2018.

[2] VOIGT P, BUSSCHE A V D. The EU general data protection regulation (GDPR)[R]. 2017.

[3] PIPER D L A. Data protection laws of the world: full handbook[R]. 2017.

[4] 第十三届全国人民代表大会. 中华人民共和国数据安全法[Z]. 2021. The 13th National People's Congress. Data security law of the People's Republic of China[Z]. 2021.

[5] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint, arXiv: 1602.05629, 2016.

[6] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[7] WANG G. Interpret federated learning with shapley values[J]. arXiv Preprint, arXiv: 1905.04519, 2019.

[8] CAI F. ByteDance breaks federal learning: open source fedlearner framework, 209% increase in advertising efficiency[R]. 2020.

[9] GE N, LI G H, ZHANG L, et al. Failure prediction in production line based on federated learning: an empirical study[J]. Journal of Intelligent Manufacturing, 2022, 33(8): 2277-2294.

[10] LIU H, ZHANG X, SHEN X, et al. A federated learning framework for smart grids: securing power traces in collaborative learning[J]. arXiv Preprint, arXiv: 2103.11870, 2021.

[11] ZHU L, LIU Z, HAN S. Deep leakage from gradients[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Piscataway: IEEE Press, 2019: 14774-14784.

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

[13] FU C, ZHANG X, JI S, et al. Label inference attacks against vertical federated learning[C]//31st USENIX Security Symposium. Berkeley: USENIX Association, 2022: 1397-1414.

[14] LUO X J, WU Y C, XIAO X K, et al. Feature inference attack on model predictions in vertical federated learning[C]//Proceedings of 2021 IEEE 37th International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2021: 181-192.

[15] JIN X, CHEN P Y, HSU C Y, et al. CAFE: catastrophic data leakage in vertical federated learning[J]. arXiv Preprint, arXiv: 2110.15122, 2021.

[16] YANG R K, MA J F, ZHANG J Y, et al. Practical feature inference attack in vertical federated learning during prediction in artificial Internet of things[J]. IEEE Internet of Things Journal, 2023: doi. 10.1109/JIOT.2023.3275161.

[17] ZHANG C, LI S, XIA J, et al. Batchcrypt: efficient homomorphic encryption for cross-silo federated learning[C]//Proceedings of the 2020 USENIX Annual Technical Conference. Berkeley: USENIX Association, 2020:

[18] LIU Y, ZHANG X W, KANG Y, et al. FedBCD: a communication-efficient collaborative learning framework for distributed features[J]. IEEE Transactions on Signal Processing, 2022, 70: 4277-4290.

[19] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[20] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.

[21] CHENG K W, FAN T, JIN Y L, et al. SecureBoost: a lossless federated learning framework[J]. IEEE Intelligent Systems, 2021, 36(6): 87-98.

[22] NI X, XU X, LYU L, et al. A vertical federated learning framework for

graph convolutional network[J]. arXiv Preprint, arXiv: 2106.11593, 2021.
[23] CEBALLOS I, SHARMA V, MUGICA E, et al. SplitNN-driven vertical partitioning[J]. arXiv Preprint, arXiv: 2008.04137, 2020.
[24] 陈晋音, 李荣昌, 黄国瀚, 等. 纵向联邦学习方法及其隐私和安全综述[J]. 网络与信息安全学报, 2023, 9(2): 1-20.
CHEN J Y, LI R C, HUANG G H, et al. Survey on vertical federated learning: algorithm, privacy and security[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 1-20.
[25] 王波, 代晓蕊, 王伟, 等. 面向联邦学习的对抗样本投毒攻击[J]. 中国科学(信息科学), 2023, 53(3): 470-484.
WANG B, DAI X R, WANG W, et al. Adversarial examples for poisoning attacks against federated learning[J]. Scientia Sinica (Informationis), 2023, 53(3): 470-484.
[26] 冯霁, 蔡其志, 姜远. 联邦学习下对抗训练样本表示的研究[J]. 中国科学: 信息科学, 2021, 51(6): 900-911.
FENG J, CAI Q Z, JIANG Y. Towards training time attacks for federated machine learning systems[J]. Scientia Sinica (Informationis), 2021, 51(6): 900-911.
[27] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.
[28] KURAKIN A, GOODFELLOW I J, BENGIO S. Artificial intelligence safety and security[M]. Boca Raton: CRC Press, 2018.
[29] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.
[30] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.
[31] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[32] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. arXiv Preprint, arXiv: 1701.07875, 2017.
[33] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[作者简介]

陈晓霖（1996− ），男，山东潍坊人，中国科学院软件研究所博士生，主要研究方向为机器学习、联邦学习、隐私计算。

訾道广（1997− ），男，山东济宁人，中国科学院软件研究所博士生，主要研究方向为自然语言处理、代码生成。

吴炳潮（1994− ），男，浙江绍兴人，中国科学院软件研究所博士生，主要研究方向为人工智能、推荐系统。

关贝（1986− ），男，山西运城人，博士，中国科学院软件研究所高级工程师，主要研究方向为人工智能和大数据、网络安全技术、虚拟化技术、操作系统技术、云计算。

王永吉（1962− ），男，辽宁营口人，博士，中国科学院软件研究所研究员、博士生导师，主要研究方向为人工智能、大数据分析、智能制造、云计算、隐蔽信道、高可信网络技术。