

基于生成对抗网络的隐私增强联邦学习方案

余锋^{1,2}, 林庆新³, 林晖^{1,2}, 汪晓丁^{1,2}

(1. 福建师范大学计算机与网络空间安全学院, 福建 福州 350017;
2. 网络安全与教育信息化福建省高校工程研究中心(福建师范大学), 福建 福州 350017; 3. 福州大学至诚学院, 福建 福州 350002)

摘要: 联邦学习作为一种分布式机器学习范式, 其具有隐私保护能力和异构协作等特性, 引起了研究者极大的关注。然而, 研究工作表明通过梯度可以确定一个确切的数据记录或一个具有特定属性的数据记录是否包含在其他参与者的批处理中, 甚至揭露参与者的训练数据, 通常称之为“梯度泄露”。同时, 当前隐私增强联邦学习方法的工作可能存在准确率下降或者计算通信开销增加等问题, 甚至引发新的不安全因素。因此, 提出一种差分隐私增强的生成对抗网络模型, 该模型向 vanilla GAN 中引入了识别器, 通过生成器与鉴别器、生成器与识别器两个博弈过程, 生成器合成的数据尽可能接近输入数据的同时满足差分隐私的约束。将此模型应用到联邦学习框架中, 在一定程度上保证了模型准确率, 并且提高了联邦学习框架的隐私保护能力。仿真实验验证了所提方案在客户端/服务器联邦学习架构下的有效性, 相比 DP-SGD 方法, 所提方案平衡了数据隐私性与实用性而不是以牺牲准确率为代价来增强隐私保护能力。从理论上分析了所提模型在点对点(P2P, peer-to-peer)架构下的可用性, 并讨论了未来研究工作。

关键词: 联邦学习; 梯度泄露; 隐私增强; 生成对抗网络; 差分隐私

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023043

Privacy-enhanced federated learning scheme based on generative adversarial networks

YU Feng^{1,2}, LIN Qingxin³, LIN Hui^{1,2}, WANG Xiaoding^{1,2}

1. College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350017, China
2. Engineering Research Center of Cyber Security and Education Informatization, Fujian Province University, Fuzhou 350117, China
3. Zhicheng College, Fuzhou University, Fuzhou 350002, China

Abstract: Federated learning, a distributed machine learning paradigm, has gained a lot of attention due to its inherent privacy protection capability and heterogeneous collaboration. However, recent studies have revealed a

收稿日期: 2022-06-05; 修回日期: 2022-08-30

通信作者: 林庆新, lqx@fdzcxy.edu.cn

基金项目: 国家自然科学基金(U1905211, 61702103); 福建省自然科学基金(2020J01167, 2020J01169)

Foundation Items: The National Natural Science Foundation of China (U1905211, 61702103), The Natural Science Foundation of Fujian Province (2020J01167, 2020J01169)

引用格式: 余锋, 林庆新, 林晖, 等. 基于生成对抗网络的隐私增强联邦学习方案[J]. 网络与信息安全学报, 2023, 9(3): 113-122.

Citation Format: YU F, LIN Q X, LIN H, et al. Privacy-enhanced federated learning scheme based on generative adversarial networks [J]. Chinese Journal of Network and Information Security, 2023, 9(3): 113-122.

potential privacy risk known as “gradient leakage”, where the gradients can be used to determine whether a data record with a specific property is included in another participant’s batch, thereby exposing the participant’s training data. Current privacy-enhanced federated learning methods may have drawbacks such as reduced accuracy, computational overhead, or new insecurity factors. To address this issue, a differential privacy-enhanced generative adversarial network model was proposed, which introduced an identifier into vanilla GAN, thus enabling the input data to be approached while satisfying differential privacy constraints. Then this model was applied to the federated learning framework, to improve the privacy protection capability without compromising model accuracy. The proposed method was verified through simulations under the client/server (C/S) federated learning architecture and was found to balance data privacy and practicality effectively compared with the DP-SGD method. Besides, the usability of the proposed model was theoretically analyzed under a peer-to-peer (P2P) architecture, and future research work was discussed.

Keywords: federated learning, gradient leakage, privacy enhancement, generative adversarial network, differential privacy

0 引言

在万物互联的世界中,各种终端设备产生的海量数据使得机器学习(ML, machine learning)进入人们的日常生活以及许多服务中。传统的ML方法以集中式方式工作,并且具有依赖于中心服务提供者收集大量用户数据的特点。然而,用户数据敏感性、用户对个人隐私保护的关注度日益增长以及法律法规(如欧盟颁布的一般数据保护条例等)的监管需求,使得这种传统ML不再符合各方的需求。

联邦学习(FL, federated learning)作为一种分布式机器学习范式,能够在不暴露参与方本地数据的情况下,使得参与方协同训练一个全局模型^[1]。FL的训练过程:①所有参与方从聚合服务器下载全局模型;②参与方将全局模型作为本地模型,在本地数据上训练并更新模型;参与方将本地模型的更新参数或梯度参数上传给聚合服务器;③服务器聚合所有本地更新。重复执行以上步骤直到全局模型收敛。在第③步中,参与方只需上传本地模型的更新参数或梯度参数这一过程又称共享模型或共享梯度,通常认为参与方没有离开本地,并且共享梯度不会泄露用户信息,因此这可以减少对用户隐私泄露的影响。

文献[2-4]指出通过梯度可以确定一个确切的数据记录或一个具有特定属性的数据记录是否包含在其他参与者的批处理中,甚至揭露参与者

的训练数据。因此,如何设计一种隐私增强的联邦学习方法引起了研究者的极大兴趣。文献[5]通过融合差分隐私(DP, differential privacy)和FL来降低车辆众包应用的隐私风险和通信成本。文献[6]引入了一个基于链式安全多方计算(SMC, secure multiparty computing)的隐私保护FL框架。它能够达到与基线FL算法相似的性能,同时优于基于差分隐私的FL算法。文献[7]表明SMC容易受到推断攻击,同时DP导致准确率较低,因此利用SMC和DP的优势提出一种混合方法。

通过以上技术实现隐私增强的联邦学习方案在实际应用中可能存在不同的问题,如基于同态加密的方法会由于交互式解密增加很大的通信开销;基于安全多方计算的方法因为参与者的数据被要求在服务器之间的秘密共享,而受到推理攻击;基于差分隐私的方法如果加入过多的噪声可能存在不可忽略的精度损失,甚至导致模型无法收敛,如果噪声不够,又可能无法起到隐私增强的作用。

因此,本文提出了一种差分隐私增强的生成对抗网络(DP-enabled GAN, differential privacy-enabled generative adversarial network)模型,其中包含两个对抗过程,一个是生成器(G, generator)与鉴别器(D, discriminator)的博弈,另一个是生成器与识别器(I, identifier)的博弈。识别器的引入使生成器合成的数据尽可能接近输入数据的同时满足差分隐私的约束。因此,本文

提出的方案能够保证模型准确率的同时起到增强隐私的效果。

1 相关工作

1.1 联邦学习中的隐私保护

联邦学习是一种使参与方（设备或组织）能够协同训练机器学习模型而不暴露客户端本地数据的分布式机器学习范式。文献[1]提出了联邦学习的概念。然而，FL 的发展，面临许多严重的挑战。特别是没有任何隐私保护机制的“裸”FL 方法被证明很容易受到深度泄露（DLG, deep leakage from gradient）^[2-4]等数据恢复攻击，文献[2]发现从公开共享梯度获得私有训练数据（即来自梯度的深度泄露）是可能的。联邦学习中的隐私保护方法可以分为加密类方法、数据扰动类方法、差分隐私以及其他类隐私保护方法^[8-9]。其中常使用的是安全多方计算、同态加密和差分隐私。文献[10]提出了一种支持两名参与者在半诚实假设下进行机器学习的 SMC 协议；文献[11]设计了一种新的 SMC 协议，它允许在完全不揭露数据的情况下实现模型训练与验证。然而，属于加密方法的 SMC 和同态加密，往往将明文编码为密文，只允许特定人员解码，但这通常需要较大的计算开销，比较难应用于实际场景；而直接应用 DP 于联邦学习，常常随着噪声的加入，导致基于差分隐私的联邦学习方法存在不可忽略的精度损失。

1.2 联邦学习中的生成对抗网络

在 FL 背景下使用生成对抗网络（GAN, generative adversarial network）的研究工作主要集中在两方面。一方面是利用 GAN 执行数据恢复攻击。文献[12]假设有一个恶意客户端，它利用共享模型作为鉴别器来训练 GAN 中的生成器；使用训练好的生成器模拟受害者客户端的训练样本。文献[13]假定服务器是恶意的。恶意服务器学习受害者客户端的数据表征，DLG 利用这些表征训练生成器，生成器最终可以生成受害者的私人数据。另一方面是在隐私、效率或异构性约束下跨分布数据源训练高质量的 GAN。Hardy 等^[14]提出了一个服务器托管生成器，每个客户端托管一个识别器的系统。这些鉴别器以点对点的方式

相互通信，以提高计算效率。FedGAN^[15]以一种有效的通信方式在 non-IID 数据源上训练 GAN，但可能产生有偏差的数据。文献[16]通过使用合成源数据微调联邦 GAN 生成无偏差的合成数据集。文献[17]提出 Fed-TGAN 用于学习不相同客户端上的复杂表格 GAN。

2 预备知识

2.1 联邦学习

联邦学习是一种分布式学习，而且客户端能够将私有数据保留在本地并进行模型训练，所以与直接从客户端访问私有数据的传统训练方法相比，在隐私保护方面具有显著优势。C/S 架构的经典 FL 框架如图 1 所示，有 N 个客户端，每个客户端拥有自己的私有数据集。在每次迭代过程中，客户端从服务器下载当前全局模型到本地；客户端对每个模型进行本地训练；客户端将这些本地更新发送回服务器；聚合服务器将收到的本地更新累积到当前全局模型中。重复执行以上步骤直到全局模型收敛。全局模型的更新可以用式(1)表示。

$$X_{t+1} = X_t + \frac{1}{N} \sum_{k=1}^N \theta_t^k \quad (1)$$

其中， X_t 表示第 t 轮迭代后的全局模型， θ_t^k 表示第 t 轮迭代中第 k 个客户端的本地模型更新。文献[1]论证了只有当客户端接收到具有相同的初始化模型时，才可以通过平均更新来得到有效模型，而联邦学习满足此条件。

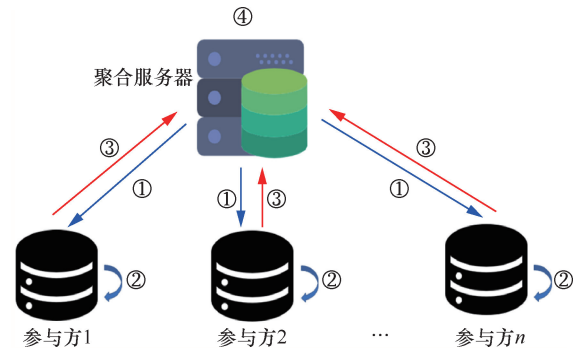


图 1 FL 框架
Figure 1 Architecture of FL

2.2 差分隐私

差分隐私^[18-19]由 Dwork 在 2006 年提出，

目前已经成为统计数据库中保护个人隐私的主要方法之一。简单来说，它保证了算法的输出分布不会因为任何单独记录的存在或不存在而发生显著变化。对手无法通过观察输出了解更多关于个体的信息，因此，隐私受到了保护。当满足以下条件时，两个数据库 D_1 和 D_2 是相邻的。

$$\exists x \in D_1, \text{s.t. } D \setminus \{x\} = D_2 \quad (2)$$

定义 1 差分隐私：假设随机机制 M 满足 (ϵ, δ) -DP^[18]，对于任何输出集合 S 和任何相邻数据库 D_1 和 D_2 满足

$$P(M(D_1) \in S) \leq e^\epsilon P(M(D_2) \in S) + \delta \quad (3)$$

其中，参数 ϵ 和 δ 是非负实数， ϵ 用于限制 D_1 和 D_2 中 M 的输出概率差异。 ϵ 越小，表示隐私性越强， ϵ 越大，表示隐私性越弱。 δ 表示失败概率，当 $\delta = 0$ 时，该定义与 ϵ -DP 等价^[19]。即使在 ϵ -DP 失败的情况下，为了提供个人隐私保护，推荐的 δ 值应该小于数据库大小的倒数，即 $1/|D_1|$ 。

2.3 生成对抗网络

GAN^[20]由 Goodfellow 提出，它是一个通过对抗过程估计生成模型的新框架。GAN 架构如图 2 所示。实际上，GAN 由生成器和鉴别器组成。生成器通过映射隐分布来学习原始数据分布 p_{data} 。具体来说，一方面，生成器将从隐分布 p_z （如高斯分布）中采样的随机噪声 z 作为输入并生成合成数据。另一方面，鉴别器学习区分从 p_{data} 中抽样的样本 x 和 G 生成的样本。生成器将样本作为输入并返回一个分数，表示它是真实的或合成的。生成器通过生成看似来自原始数据分布的样本来达到欺骗鉴别器的目标，让其认为生成样本是真实样本。生成器和鉴别器通过对抗过程同时训练：生成器越提高生成数据的质量，鉴别器就越难区分原始样本和生成样本。这个问题可以由具有以下目标的表达式来描述。

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (4)$$

交替训练两个模型 G 和 D 直至博弈达到纳什

均衡，此时 G 生成的样本很难与已读的样本区分开。理论上， $p_{\text{data}} = p_g$ 表示达到全局最优状态， p_g 表示生成样本的分布。

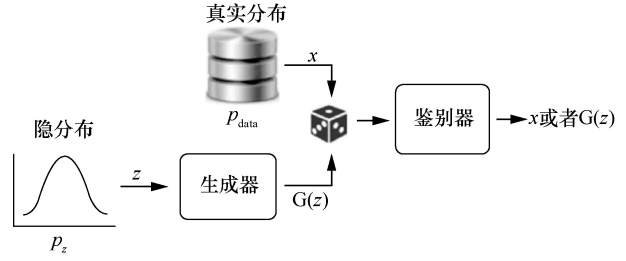


图 2 GAN 架构
Figure 2 GAN architecture

3 本文方案

3.1 联邦学习梯度泄露攻击

联邦学习多阶段框架可以分为 3 个阶段，包括数据和行为审计、训练及预测。该框架在联邦学习执行的每个阶段都面临着不同的安全和隐私威胁。因此建立一个可信联邦学习框架需要在每个阶段采取有效措施，以充分减少安全和隐私的威胁。本文关注联邦学习训练阶段模型推理攻击下的梯度泄露攻击（如图 3 所示）。从关于联邦学习的预备知识介绍中容易得知，联邦学习的核心思想是“数据不动，模型动”。一些研究表明，通过梯度会泄露训练数据的属性特征，如通过属性分类器能够判别是否具有某种属性的样本^[21]。而在联邦学习训练阶段，模型训练经常被视为数据的高级表示，这使得基于梯度参数的隐私推断成为可能。“梯度泄露攻击”过程可以描述为：随机初始化生成一堆伪输入和标签，执行前向和反向传播；从伪数据推导出伪梯度后，优化伪输入和标签来最小化伪梯度与真实梯度之间的距离，而不是像典型训练中优化模型权重；匹配梯度使伪数据接近原始数据；当以上的优化过程完成以后，私有数据（包括样本和标签）就会被恢复出来^[2]。

3.2 威胁模型

根据参与方的可信程度，可以建立 3 种安全模型：理想模型、半诚实模型和恶意模型。其中半诚实模型又被称为诚实且好奇模型。

1) 理想模型：在协议执行过程中，每个参与

方都是可信的，一方将其信息发送给另一方后，另一方不会查看此信息，只会根据约定的协议继续执行流程。

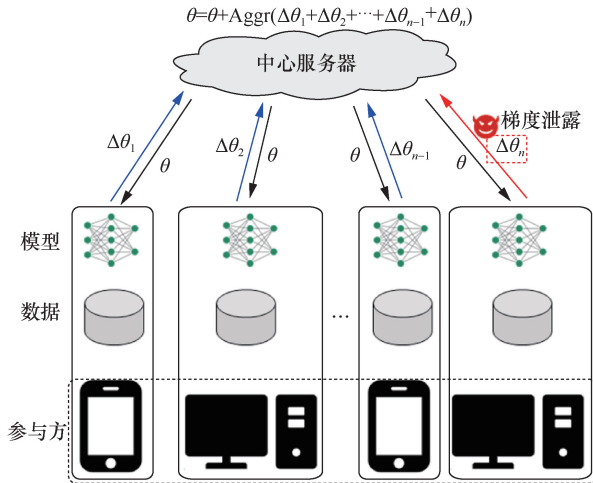


图3 FL梯度泄露攻击模型
Figure 3 FL gradient leak attack model

2) 半诚实模型：在协议执行过程中，参与方按照协议规定的流程执行，但它可以根据其他参与方的输入或者中间结果来推导出额外的信息。

3) 恶意模型：在协议执行过程中，恶意攻击者可能不会诚实地执行协议，它可以通过其控制下的参与方进行不合法的输入或者恶意篡改输入等来分析诚实参与方的隐私信息，甚至可以通过提前终止和拒绝参与等方式终止协议。

本文假设威胁模型是半诚实模型，具体为服务器是诚实且好奇的，同时假设服务器具有无限的计算资源。服务器拥有的知识包含模型的结构、权重参数和本地参与者每个阶段的梯度参数。理论上，诚实且好奇的服务器不会修改网络结构，但可能会从本地梯度推理出训练样本或其他隐私信息从而造成梯度泄露攻击。

3.3 DP-enabled GAN 设计

在原始 GAN 模型中，有两个组成部分——生成器和鉴别器，它们建立在博弈理论的基础上，使得生成器生成的数据尽量符合真实数据的分布。受 GAN 模型原理启发，以及差分隐私的应用，本文提出一种差分隐私增强的生成对抗网络模型并应用到联邦学习中（如图4所示），来解决联邦学习的梯度泄露问题。

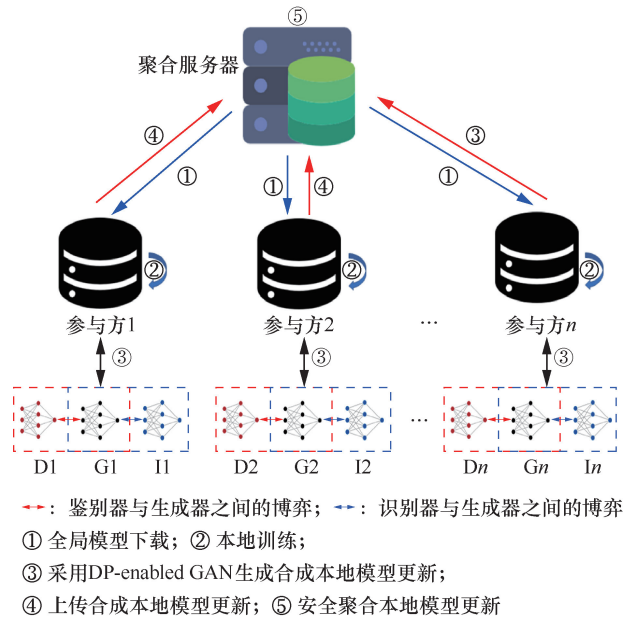


图4 基于生成对抗网络的隐私增强联邦模型
Figure 4 Privacy-enhancing federated model based on generative adversarial networks

如图4所示，生成器的功能是通过学习本地更新的特征，在鉴别器与识别器的指导下，将加入噪声后的合成模型更新尽可能拟合为本地更新的真实分布，而鉴别器的功能是尽可能区分本地更新与合成本地更新，另外，加入了一个基于差分隐私的识别器 I，它的功能是尽可能地区分满足差分隐私约束的合成模型更新。因此，如果鉴别器和识别器都不能将合成模型与本地模型区分开，那么差分隐私约束的合成模型的生成过程就完成。鉴别器、生成器和识别器这3个感知器组成的模型被称为 DP-enabled GAN。对比图1与图4可知，此模型应用在参与方（即客户端）上，具体为上传本地更新之前、本地训练步骤之后。因此，在联邦训练过程中，本地模型更新（可以是梯度参数）作为 DP-enabled GAN 的输入，对应的输出为满足差分隐私的合成本地模型更新。结合以上分析，可以解决模型上传过程的梯度泄露问题，同时避免直接添加差分隐私噪声导致的准确率骤降甚至无法收敛的结果。

3.3.1 对抗网络：生成器、鉴别器和识别器

定义一个先验输入噪声分布 $p_z(z)$ 以学习在训练数据 x 上生成器的分布 p_g ， $G(z; \theta_g)$ 表示生成器到数据空间的一个映射，其中 $G()$ 是一个带

有参数 θ_g 的多层感知器表征的可微函数。 $D(x; \theta_d)$ 表示鉴别器输出一个标量, 其中 $D(x)$ 表示 x 来自数据而不是 p_g 的概率。训练鉴别器使其最大限度地正确标签分配给训练样本和来自生成器样本的概率, 训练生成器最小化输入数据来源被发现的概率。将上述对抗过程建模为最小最大化问题。从分布为 $p_{\text{data}}(x)$ 的数据集中采样 N 个训练 (真实) 样本 $\{x_1, x_2, \dots, x_N\}$, 同时从先验噪声分布 $p_g(z)$ 中采样 N 个噪声样本 $\{z_1, z_2, \dots, z_N\}$, 并且通过生成器获得 N 个生成样本 $\{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N\}$; 然后固定生成器, 使用梯度上升策略训练鉴别器使其能够更好地判断某样本是真实样本还是生成样本, 如式(5)所示。

$$\nabla_{\theta_d} \frac{1}{N} \sum_{i=1}^N [\log D(x_i) + \log(1 - D(G(z_i)))] \quad (5)$$

对鉴别器训练循环多次后, 使用较小的学习率来对生成器进行优化, 生成器使用梯度下降策略优化, 如下。

$$\nabla_{\theta_g} \frac{1}{N} \sum_{i=1}^N \log(1 - D(G(z_i))) \quad (6)$$

$I(z; \theta_i)$ 表示识别器利用 GAN 产生差分隐私噪声, 用于识别合成数据是否服从拉普拉斯分布。 $I(z; \theta_i)$ 、 $G(z; \theta_g)$ 和 $D(x; \theta_d)$ 服从对抗过程。因此, 定义 $I(D; I)$ 为鉴别器和识别器的博弈结构, 这个更新过程可以描述为

$$\nabla_{\theta_N} \frac{1}{N} \sum_{i=1}^N [\log I(x_i | D, I) + \log(1 - I(G(z_i) | D; I))] \quad (7)$$

基于式(5)~式(7), 这个最小最大化问题可以描述为

$$\min_G \max_I V(I, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log I(x_i | D; I)] + \mathbb{E}_{z \sim p_g(z)} [\log(1 - I(G(z_i) | D; I))] \quad (8)$$

其中, $\max_I$ 用于寻求识别效果最佳的识别器, $\min_G$ 用于当识别器最佳时, 使得鉴别器 D 区分概率最小。

基于生成对抗网络的联邦学习 (GAN-based FL) 如算法 1 所示。其中的步骤 3)~步骤 6)即隐私对抗模型应用于参与方的过程。GAN-based FL 具体流程如下: 参与方从服务器下载模型, 然后

在各自数据集上做本地训练; 参与方将本地模型更新作为 DP-enabled GAN 的输入, 得到合成满足差分隐私约束的本地模型更新; 参与方上传上一步的合成本地模型更新; 服务器聚合并更新全局模型。重复迭代以上步骤直至全局模型准确率达到某个阈值。

算法 1 GAN-based FL

- 1) 参数初始化;
- 2) **while** 全局准确率 < 阈值 **do**:
- 3) 参与方本地训练;
- 4) 参与方通过 DP-enabled GAN 合成满足差分隐私约束的本地模型更新;
- 5) 参与方上传合成本地模型更新;
- 6) 服务器聚合并更新全局模型, 然后参与方下载该模型;
- 7) **end while**
- 8) **return** 最终的全局模型

3.3.2 理论分析

联邦学习的主要目标之一是保护本地设备数据的隐私, 同时由于本地数据的敏感性, 联邦学习通常需要考虑数据隐私泄露问题。一些研究表明上传的梯度参数可能会泄露数据隐私^[2]。为了增强本地数据隐私保护, 本文提出并设计了差分隐私增强的生成对抗网络模型, 并将其部署到每一台本地设备上。本文方案主要利用生成对抗网络的学习能力以及差分隐私的特性来解决本地模型的利用和本地数据隐私保护之间的平衡问题。而对于聚合服务器而言, 考虑到仅通过本地模型参数不能破坏本地数据的隐私性以及模型准确率的要求, 本文暂不在服务器端添加隐私保护机制。

本文利用差分隐私的传递型特性, 即对于满足 (ϵ, δ) -DP 的随机机制 M , 对其产生的结果进行任何形式的新算法同样满足 (ϵ, δ) -DP, 在传统 GAN 中添加识别器, 通过识别器与生成器的对抗过程, 客户端本地训练过程中的中间计算结果满足差分隐私约束; 同时通过鉴别器与生成器的对抗过程, 获取高质量的合成样本, 提高分类准确率。因为 DP-enabled GAN 模型作用于联邦学习的客户端, 所以本文方案能够抵抗来自服务器或其他半诚实敌手发起的梯度推理攻击。

4 实验分析

本文在 MNIST 和 Fashion MNIST 数据集上实现了基于生成对抗网络的隐私增强联邦学习方案，同时为了证实本文方案的合理性与有效性，本文将当前常用的隐私增强方案，即加高斯噪声的差分隐私方案（DP-SGD, differentially private SGD）^[22]和 GAN^[23]作为对比方案。

(1) DP-SGD

联邦学习训练与推断流程可以简述为：构建联邦学习模型并初始化模型参数；输入训练数据并计算梯度；用梯度下降法更新模型直到模型收敛；将收敛的模型用于数据预测。DP-SGD 利用差分隐私的传递型特性，即对于满足 (ϵ, δ) -DP 的随机机制 M ，对其产生的结果进行任何形式的新算法同样满足 (ϵ, δ) -DP。因此 FL 训练过程中，DP-SGD 在客户端每轮迭代计算好的模型参数上直接添加噪声，以达到隐私保护的目的。

(2) GAN

先通过 GAN 方法生成合成数据，然后将合成数据用于 FL 训练过程，将训练好的模型用于数据预测。

4.1 实验设置

本文采用 Python 3.8 和 PyTorch 1.11.0，装有 Ubuntu 20.04 操作系统以及单卡 GeForce RTX 3090 的服务器来做仿真实验。本文实验使用两种不同规模的数据集 MNIST 和 Fashion-MNIST 进行训练，对图像和对应标签的联合分布进行建模，也就是将标签提供给生成器和鉴别器，并根据输入生成图像。同时，对于分类准确率实验中的所有方法，设置隐私预算参数： $\epsilon = 10$ ， $\delta = 10^{-5}$ 。接下来，阐述实验中使用的评估指标。

本文方案的目的是在联邦学习训练中增强隐私的同时能够维持模型准确率，所以实验从隐私性（由 ϵ 决定）和实用性两个角度来评估 DP-enabled GAN 方法。以往的工作通常采用 IS（inception score）^[24]和 FID（fréchet inception distance）^[25]来评估 GAN 生成样本的质量，即生成样本的分类准确率。

(1) IS

IS 的定义如下。

$$IS = \exp(\mathbb{E}_{x \sim G(z)} D_{KL}(P(y|x) \| P(y))) \quad (9)$$

其中， $P(y|x)$ 和 $P(y)$ 都是预先训练好的分类器将生成样本作为输入时的输出分布。具体来说，如果 $P(y|x)$ 的熵值越低（即生成的图像越清晰）以及 $P(y)$ 的熵值越高（生成的样本覆盖所有不同类并且具有较高的多样性），那么 IS 的值越大。

(2) FID

FID 的定义如下。

$$FID = \|\mu_r - \mu_g\|^2 + \text{tr}(\sum_r + \sum_g - 2(\sum_r \sum_g)^{1/2}) \quad (10)$$

其中， $x_r \sim N(\mu_r, \sum_r)$ 和 $x_g \sim N(\mu_g, \sum_g)$ 分别是真实样本和生成样本在 Inception Net-v3 输出的 2 048 维的激活函数。FID 值越低，说明真实样本与生成样本的差异越小，也就说明样本质量和多样性越好。

(3) 分类准确率

在仿真实验中，考虑多层感知器（MLP, multi-layer perceptron）和卷积神经网络（CNN, convolutional neural network）这两种模型。

4.2 数值分析

4.2.1 训练迭代次数对隐私性与实用性的影响

为了评估 DP-enabled GAN 方案的有效性，根据隐私-实用表现中的重要影响指标（训练迭代次数），通过调整训练迭代次数进行仿真实验，分别计算并保存 IS 和 FID 以及不同迭代次数对应的隐私预算 ϵ 。

图 5 给出了在 MNIST 数据集上，DP-enabled GAN 与 DP-SGD 在不同训练迭代次数时，IS 和 FID 的数值变化情况。从图 5(a)可以看出，训练迭代次数越多，IS 越大，即样本可用性越高，但这是以更高的隐私预算 ϵ 为代价的。从图 5(b)可以看出，训练迭代次数越多，FID 越低，即样本质量越高，同样也是以更高的隐私预算 ϵ 为代价的。进一步地，在 Fashion-MNIST 数据集上通过调整训练迭代次数进行仿真实验，结果如图 6 所示。从图 6 可以得到与图 5 一致的结论。通过仔细观察图 5 与图 6，在隐私预算 $\epsilon = 10$ 时，DP-SGD 与本文方案都可以在尽可能少的训练

迭代次数的同时达到几乎最佳的样本质量, 这从侧面印证了分类准确率实验中使用隐私预算 $\varepsilon = 10$ 的合理性。

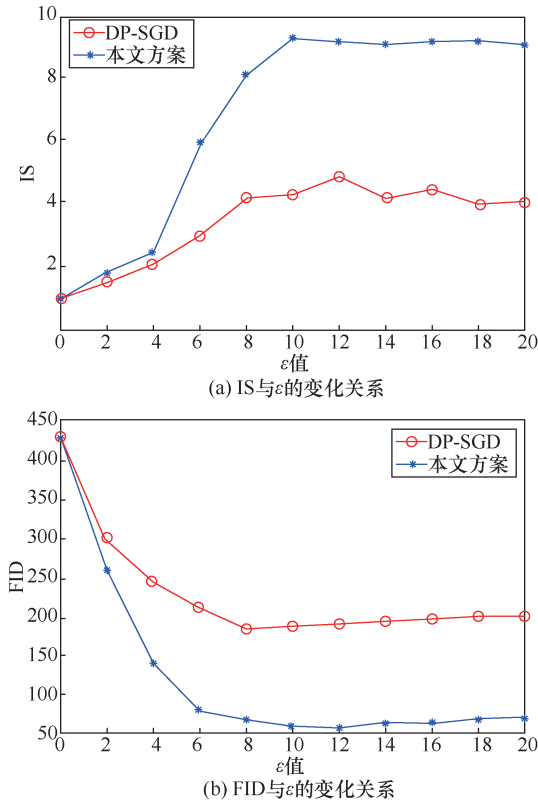


图5 MNIST 数据集上的隐私性与实用性 ($\delta = 10^{-5}$)

Figure 5 Privacy and utility on MNIST dataset ($\delta = 10^{-5}$)

4.2.2 分类准确率实验

为了进一步量化 DP-enabled GAN 在 C/S 架构联邦学习训练过程的有效性, 分别在 MNIST 和 Fashion-MNIST 数据集上采用 GAN、DP-SGD 作为增强隐私对比方案。分类准确率表现如表 1 和表 2 所示, 其中 No 表示直接使用真实样本数据作为训练的输入而不是经过隐私处理的样本 (即加噪或合成样本)。从表 1 和表 2 可以看出, 本文方案比 DP-SGD 的分类准确率高, 这是因为 DP-enabled GAN 利用了 GAN 中的对抗特性 (即能够使识别器无法将合成样本与真实样本区分开来), 但比直接使用 GAN 的分类准确率低, 这是因为本文方案采用了差分隐私, 牺牲了一定的准确率。但是相比 DP-SGD, DP-enabled GAN 更好地平衡了数据的隐私性与可用性。

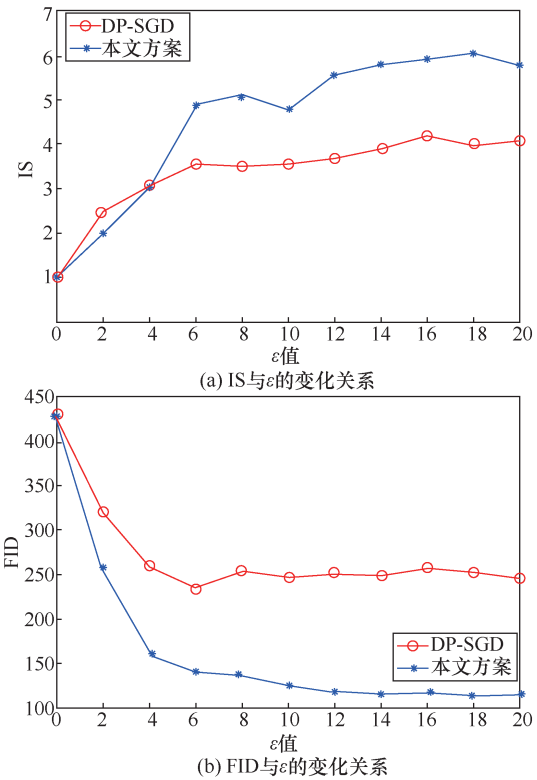


图6 Fashion-MNIST 数据集上的隐私性与实用性 ($\delta = 10^{-5}$)

Figure 6 Privacy and utility on Fashion-MNIST dataset ($\delta = 10^{-5}$)

表1 MNIST 数据集上各方案在不同模型的分类型准确率
Table 1 Classification accuracy of each method in different scheme on the MNIST dataset

模型	分类准确率			
	No	GAN	DP-SGD	DP-enabled GAN
MLP	97%	82%	60%	78%
CNN	99%	82%	63%	79%

表2 Fashion-MNIST 上各方案在不同模型的分类型准确率

Table 2 Classification accuracy of each scheme in different models on Fashion-MNIST

模型	分类准确率			
	No	GAN	DP-SGD	DP-enabled GAN
MLP	86%	75%	50%	65%
CNN	89%	72%	45%	64%

5 未来的研究工作

根据仿真实验可知, 在隐私预算相同的情况下, 与 DP-SGD 相比, DP-enabled GAN 合成的样本质量更高, 并且样本更多样, 因此对应的 IS 更大、FID 更低以及分类准确率更高, 而与 GAN 相

比, 由于噪声的存在, DP-enabled GAN 合成样本的质量更差, 因此其分类准确率更低。根据以上理论分析以及实验结论可知, DP-enabled GAN 能够在 C/S 架构联邦学习训练过程中起到增强隐私的作用, 并且相比 DP-SGD 的分类准确率更好。然而, C/S 架构的可信中心服务器这一假设条件在实际中可能无法满足, 并且存在单点故障问题。针对上述缺陷, 研究人员开展了相关工作, 如文献[26]指出区块链与联邦学习的结合能够有效地解决前述问题。DP-enabled GAN 主要应用在联邦学习客户端(参与方), 理论上来说, 它也可以应用于联邦学习 P2P 架构, 进一步地, 它能在基于区块链的联邦学习框架中工作并发挥其增强隐私作用。

本文下一步工作是进一步研究 DP-enabled GAN 模型在区块链的联邦学习框架中的实际效果。同时, DP-enabled GAN 模型需要在每个客户端进行对抗网络训练过程, 而从既有的工作以及本实验可知, 这一训练过程耗费时间, 同时, 这与模型可扩展性及其实用性紧密相关, 因此, 如何平衡模型训练时间开销以及隐私增强程度是值得深入研究的主题。

6 结束语

本文提出一种基于生成对抗网络的隐私增强联邦学习方案, 方案使用了差分隐私增强的生成对抗网络模型, 通过生成器与鉴别器、生成器与识别器两个博弈过程, 生成器合成的数据尽可能接近输入数据的同时满足差分隐私的约束; 然后将此模型应用到联邦学习框架中, 从而在一定程度上保证了模型准确率, 并且提高了联邦学习框架的隐私保护能力, 平衡了数据隐私性与实用性的关系。在 C/S 网络架构下, 通过仿真实验验证了所提方案的有效性。进一步, 从理论上分析了 DP-enabled GAN 模型在联邦学习 P2P 网络架构框架中的可用性, 以及模型可扩展性与实用性。

参考文献:

- [1] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint arXiv:1602.05629, 2017.
- [2] ZHU L, LIU Z, HAN S. Deep leakage from gradients[C]//Advances in Neural Information Processing Systems. 2019: 32.
- [3] ZHAO B, MOPURI KR, BILEN H. IDLG: improved deep leakage from gradients[J]. arXiv Preprint arXiv: 2001.02610, 2020.
- [4] GEIPING J, BAUERMEISTER H, DRÖGE H, et al. Inverting gradients: how easy is it to break privacy in federated learning[J]. arXiv Preprint arXiv: 2003.14053, 2020.
- [5] ZHAO Y, ZHAO J, YANG M, et al. Local differential privacy based federated learning for internet of things[J]. arXiv Preprint arXiv: 2004.08856, 2020.
- [6] LI Y, ZHOU Y P, JOLFAEI A, et al. Privacy-preserving federated learning framework based on chained secure multiparty computing[J]. IEEE Internet of Things Journal, 2021, 8(8): 6178-6186.
- [7] TRUEX S, BARACALDO N, ANWAR A, et al. A hybrid approach to privacy-preserving federated learning[J]. Informatik Spektrum, 2019, 42(5): 356-357.
- [8] 杨庚, 王周生. 联邦学习中的隐私保护研究进展[J]. 南京邮电大学学报(自然科学版), 2020, 40(5): 204-214.
YANG G, WANG Z S. Survey on privacy preservation in federated learning[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2020, 40(5): 204-214.
- [9] 陈兵, 成翔, 张佳乐, 等. 联邦学习安全与隐私保护综述[J]. 南京航空航天大学学报, 2020, 52(5): 675-684.
CHEN B, CHENG X, ZHANG J L, et al. Survey of security and privacy in federated learning[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(5): 675-684.
- [10] MOHASSEL P, ZHANG Y. SecureML: a system for scalable privacy-preserving machine learning[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 19-38.
- [11] KILBERTUS N, GASCON A, KUSNER M, et al. Blind justice: fairness with encrypted sensitive attributes[C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 2630-2639.
- [12] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the gan: information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 603-618.
- [13] WANG Z B, SONG M K, ZHANG Z F, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]//Proceedings of IEEE INFOCOM 2019-IEEE Conference on Computer Communications. 2019: 2512-2520.
- [14] HARDY C, LE MERRER E, SERICOLA B. MD-GAN: multi-discriminator generative adversarial networks for distributed datasets[C]//Proceedings of 2019 IEEE International Parallel and Distributed Processing Symposium. 2019: 866-877.
- [15] RASOULI M, SUN T, RAJAGOPAL R. FedGAN: federated generative adversarial networks for distributed data[J]. arXiv Preprint arXiv: 2006.07228, 2020.
- [16] MUGUNTHAN V, GOKUL V, KAGAL L, et al. Bias-free FedGAN:

- a federated approach to generate bias-free datasets[J]. arXiv Preprint arXiv: 2103.09876, 2021.
- [17] ZHAO Z, BIRKE R, KUNAR A, et al. Fed-TGAN: federated learning framework for synthesizing tabular data[J]. arXiv: 2108.07927, 2021.
- [18] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our data, ourselves: privacy via distributed noise generation[C]//Advances in Cryptology - Eurocrypt 2006. 2006: 486-503.
- [19] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography. 2006: 265-284.
- [20] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [21] MELIS L, SONG C, DE CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]//2019 IEEE Symposium on Security and Privacy (SP). 2019: 691-706.
- [22] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016: 308-318.
- [23] AUGENSTEIN S, MCMAHAN H B, RAMAGE D, et al. Generative models for effective ML on private, decentralized datasets[J]. arXiv:1911.06679,B2020.
- [24] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[J]. arXiv Preprint arXiv: 1606.03498, 2016.
- [25] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[J]. arXiv Preprint arXiv: 1706.08500, 2018.
- [26] NGUYEN D C, DING M, PHAM Q V, et al. Federated learning meets blockchain in edge computing: opportunities and challenges[J]. IEEE Internet of Things Journal, 2021, 8(16): 12806-12825.

[作者简介]



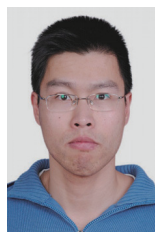
余锋（1998—），男，湖北大悟人，福建师范大学硕士生，主要研究方向为联邦学习、隐私保护。



林庆新（1979—），男，福建安溪人，福州大学副教授，主要研究方向为计算机网络安全。



林晖（1977—），男，福建福州人，博士，福建师范大学教授、博士生导师，主要研究方向为机器学习、移动边缘计算和无线网络信息安全。



汪晓丁（1982—），男，福建福州人，福建师范大学副教授，主要研究方向为网络优化与无线通信网络。