自适应差分隐私的高效深度学习方案

王玉画，高 胜，朱建明，黄 晨
(中央财经大学 信息学院，北京 100081)

摘要：深度学习在诸多领域取得巨大成功的同时，也逐渐暴露出一系列严重的隐私安全问题。作为一种轻量级隐私保护技术，差分隐私通过对模型添加噪声使得输出结果对数据集中的任意一条数据都不敏感，更适合现实中个人用户隐私保护的场景。针对现有大多差分隐私深度学习方案中迭代次数对隐私预算的依赖、数据可用性较低和模型收敛速度较慢等问题，提出一种自适应差分隐私的高效深度学习方案。首先，基于沙普利加性解释模型设计一种自适应差分隐私机制，通过对样本特征加噪使得迭代次数独立于隐私预算，再利用函数机制扰动损失函数，从而实现对原始样本和标签的双重保护，同时增强数据可用性。其次，利用自适应矩估计算法调整学习率来加快模型收敛速度。并且，引入零集中差分隐私作为隐私损失统计机制，降低因隐私损失超过隐私预算带来的隐私泄露风险。最后，对方案的隐私性进行理论分析，并在 MNIST 和 Fashion-MNIST 数据集上通过对比实验验证了所提方案的有效性。

关键词：深度学习；差分隐私；自适应；隐私损失；模型收敛
中图分类号：TP309 **文献标识码**：A

Efficient deep learning scheme with adaptive differential privacy

WANG Yuhua, GAO Sheng, ZHU Jianming, HUANG Chen
(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract: While deep learning has achieved a great success in many fields, it has also gradually exposed a series of serious privacy security issues. As a lightweight privacy protection technology, differential privacy makes the output insensitive to any data in the dataset by adding noise to the model, which is more suitable for the privacy protection of individual users in reality. Aiming at the problems of the dependence of iterations on the privacy budget, low data availability and slow model convergence in most existing differential private deep learning schemes, an efficient deep learning scheme based on adaptive differential privacy is proposed. First, an adaptive differential privacy mechanism is designed based on the Shapley additive explanation model. By adding noise to the sample features, the number of iterations is independent of the privacy budget, and then the loss function is perturbed by the function mechanism, thus achieving the dual protection of original samples and labels while enhancing the utility of data. Second, the adaptive moment estimation algorithm is used to adjust the learning rate to accelerate the model convergence. Additionally, zero-centralized difference privacy is introduced as a statistical mechanism of privacy loss, which reduces the risk of privacy leakage caused by the privacy loss exceeding the privacy budget. Finally, a theoretical analysis of privacy is made, with the effectiveness of the proposed scheme verified by comparative experiments on the MNIST and Fashion-MNIST datasets.
Key Words: deep learning; differential privacy; self-adaptation; privacy loss; model convergence

1 引 言

近年来，深度学习技术作为机器学习研究的前沿领域，凭借着对文本、声音和图像等数据的强大处理和理解能力，在社会网络分析、物联网和无线通信等诸多领域任务中表现出优越的性能。它的巨大成功主要依赖于高性能的计算、大规模的数据以及各种深度学习框架的开源。深度学习

收稿日期：2023-01-12
基金项目：国家自然科学基金(62072487)；北京市自然科学基金(M21036)
作者简介：王玉画（2000—），女，中央财经大学大学硕士研究生，E-mail:wyh1921352947@163.com
高 胜（1987—），男，教授，博士，E-mail: sgao@cufe.edu.cn
黄 晨（1997—），男，中央财经大学大学硕士研究生，E-mail: ichuang12@163.com
通信作者：朱建明（1965—），男，教授，博士，E-mail: zjm@ cufe.edu.cn

技术主要分为两个阶段：首先是模型训练阶段，用收集到的海量数据对深度神经网络模型进行迭代训练，直到模型收敛便获得了目标模型；其次是模型推理阶段，利用训练好的目标模型对目标数据集执行分类和预测等任务。

然而，由于攻击手段不断的演进，深度学习模型所存在的隐私泄露风险也随之增加。常见的攻击方式有模型反演攻击和成员推理攻击。模型反演攻击在模型训练和推理阶段都有可能发生，敌手通过截取模型参数和测试模型输出来重建出训练数据集。SONG 等[1]根据模型参数重构出原始的训练数据，窃取特定个体数据的敏感信息；成员推理攻击主要发生在模型推理阶段，敌手通过目标模型的输出差异来推断给定样本是否属于模型的训练集[2]。SALEM 等[3]证明了敌手可以在没有任何背景信息的情况下，根据目标模型的输出规律判断出样本是否参与过训练。本质上，这些隐私问题的产生很大程度上归因于深度神经网络独特的学习和训练方法，通过大量的隐藏层不断提取高维数据特征，模型将不经意地记住某些数据细节，甚至是整个数据集[4]。

针对深度学习潜在的隐私威胁，现有的方案通过结合一些经典的隐私保护机制来增强隐私，主要分为加密机制和扰动机制[5]。加密机制目的在于保护数据交换的过程，常用同态加密和安全多方计算实现。其中，同态加密允许第三方无需解密就可以直接在加密域上执行计算，保证了模型参数的精度[6-7]；安全多方计算允许当不可信多方参与到模型的训练和推理过程时，通过秘密共享或不经意传输等来实现数据的安全性[8-9]。相比于同态加密方法，基于安全多方计算的方案虽然不需要大量计算开销，但却增加了通信成本。扰动机制目的在于保护数据内容本身，通过差分隐私[10] (Differential Privacy, DP) 技术在模型训练过程中添加噪声来扰动，使得某条数据是否参与训练对最终的输出结果影响微乎其微。这是一种轻量级隐私保护技术，计算效率高，通信开销低，且具有后处理性。基于差分隐私的方案关键在于模型效用和隐私保护之间的权衡[11-15]。ABADI 等[16]设计了一种差分隐私随机梯度下降（Differential Private Stochastic Gradient Descent, DPSGD）算法，将多个数据批分为一组，对每组的累积梯度添加噪声，还引入矩会计 (Moment Accountant, MA) 来追踪隐私损失，从而获得更紧致的整体隐私损失估计。然而，该算法以等量的隐私预算加噪会导致原始梯度出现较大失真，显著降低了数据的可用性。ZHANG 等[17]提出一种自适应衰减噪声的隐私保护算法，每次迭代中向梯度加入通过线性衰减率调整的噪声，以减少负噪声的添加，但此方案对于线性衰减率并没有很好的计算方法，只能通过实验调试，实用性较弱。所提两种方案都是对梯度进行二范数裁剪来控制梯度的敏感度，可实际应用中对高维梯度的裁剪范围是很难把握的，而且每轮训练中每个批次的迭代都需要加噪，使得隐私损失严重依赖于迭代次数，当需要较多次迭代来保证模型准确性时，其训练效果会受到影响。PHAN 等[18]提出了一种自适应拉普拉斯机制，通过逐层相关传播 (Layer-wise Relevance Propagation，LRP) 算法衡量深度神经网络中输入与输出的相关性，再根据相关性对第一个隐藏层加入拉普拉斯噪声，真正实现了从样本特征的角度来自适应确定噪声大小。可是，在使用 LRP 算法时可能会泄露隐私。作为改进，ZHANG 等[19]设计了一种自适应动态隐私预算分配的差分隐私方案 (adaptive allocation dynamic privacy budget differential privacy, ADDP)，对 LRP 算法输出的相关性也进行了加噪处理。LIU 等[20]引入随机化隐私保护调整技术，直接对相关性超过设定阈值的输入特征进行扰动，未超过阈值的特征由随机因子决定是否被扰动。然而，不同预定阈值和随机因子的选取会对模型效用造成不同的影响。以上三种方案都采用拉普拉斯机制加噪太过严格，且没有很好地考虑到相关性衡量算法与数据可用性之间的关系，较精确的相关性衡量才能获得较好的数据可用性。更多地，它们都没有在设计时兼顾到模型的收敛速度，而在实际应用中这往往也是至关重要的。

为解决现有深度学习差分隐私保护方案中所存在的迭代与隐私预算之间依赖、数据可用性较低和收敛速度较慢等问题，本文提出了一种自适应差分隐私的高效深度学习 (adaptive differential privacy-based efficient deep learning, ADPE) 方案。本文的主要贡献如下。

1）设计一种自适应差分隐私机制，通过 Shapley 加性解释模型在特征维度对原始样本进行自适应扰动，使得迭代次数独立于隐私预算，并结合函数机制来保护样本的真实标签，从而实现对原始样本及其标签提供隐私保护的同时，保证数据的可用性。

2）将自适应矩估计算法与指数衰减函数相结合，利用先验知识优化梯度，针对不同的参数调整学习率，加快模型收敛速度，并增强后期模型训练的稳定性。

3）引入零集中差分隐私中的组合机制对整个方案的隐私损失进行更清晰更紧凑的统计，从而降低因隐私损失超过隐私预算带来的隐私泄露风险，更好地平衡隐私和效用之间的关系。

4）给出了详细的隐私分析，并在 MNIST 和 Fashion-MNIST 数据集上通过衡量模型的分类准确率进行了对比实验，与其他方案相比，本文所提方案效果更优。

2 预备知识

2.1 差分隐私

差分隐私的提出是为了解决查询数据库中的隐私信息泄露问题，其主要基于扰动的思想，让敌手无法根据查询结果来判断出单条数据记录的更改或增删，即输出结果对于数据集中的任何一条特定记录都不敏感。差分隐私的形式化定义如下。

定义 1 （(ε,δ) - DP[10]）设有隐私机制 M，其定义域为 $Dom(M)$，值域为 $Ran(M)$。若隐私机制 M 对于任意两个仅相差一条记录的相邻数据集 D 和 $D'\subseteq Dom(M)$，$O\subseteq Ran(M)$ 满足

$$\Pr[M(D)\in O]<e^{\varepsilon}\Pr[M(D')\in O]+\delta \tag{1}$$

则称隐私机制 M 满足 (ε,δ) - DP。其中，$\Pr[x]$ 表示数据 x 泄露的概率；参数 ε 称为隐私预算，用来衡量隐私保护的程度，ε 越小，隐私保护程度越高；参数 δ 为违反隐私机制 M 的概率，$\delta=0$ 时意味着隐私机制 M 满足严格差分隐私，即 ε - DP。

定义 2 （全局敏感度[10]）给定数据集 D 上的一个查询函数 $f:D\to R^d$，f 的全局敏感度是指删除数据集中任何一条记录所引起查询结果的最大变化，定义为

$$S_f(D)=\max_{D,D'}\|f(D)-f(D')\|_l \tag{2}$$

其中，D 和 D' 是任意两个相邻数据集，l 表示度量距离的向量范数，通常为 1 或 2 范数距离。

定理 1（高斯机制[10]）设 $\varepsilon,\delta\in(0,1)$，$f(D)$ 是 l_2 敏感度为 S_f 的查询函数，当 $S_f\sigma\geq\dfrac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$ 时，隐私机制 $M(D)=f(D)+N(0,{S_f}^2\sigma^2)$ 满足 (ε,δ) - DP。

2.2 零集中差分隐私

在训练深度神经网络模型时，由于迭代次数较多，对隐私损失的估计至关重要。零集中差分隐私[21]（zero-concentrated differential privacy，zCDP）是一种新的差分隐私松弛形式，与 (ε,δ) - DP 相比，对多次迭代计算的隐私损失提供了更清晰和更严格的分析。zCDP 的定义如下。

定义 3（zCDP[21]）对于任意 $\alpha>1$，若隐私机制 M 对于任意两个仅相差一条记录的相邻数据集 D 和 D' 满足

$$D_\alpha(M(D)\|M(D'))=\frac{1}{\alpha-1}\log E[e^{(\alpha-1)L^{(O)}}]\leq\rho \tag{3}$$

则称该隐私机制满足 ρ - zCDP。其中，$D_\alpha(M(D)\|M(D'))$ 表示 $M(D)$ 和 $M(D')$ 间的 $\alpha-Renyi$ 距离，$L^{(O)}$ 表示输出结果为 O 时，算法在两个数据集之间产生的隐私损失，即

$$L^{(O)}_{(M(D)\|M(D'))}=\ln\frac{\Pr[M(D)\in O]}{\Pr[M(D')\in O]} \tag{4}$$

本文使用到的 zCDP 一些性质和定理如下。

性质 1[21] 高斯机制返回 $f(D)+N(0,S_f^2\sigma^2)$ 时满足 $(1/2\sigma^2)$ - zCDP。

性质 2[21] 假设两种机制满足 ρ_1 - zCDP 和 ρ_2 - zCDP，那么它们的组合满足 $(\rho_1+\rho_2)$ - zCDP。

性质 3[21] 若机制 M 满足 ρ - zCDP，那么对于任意 $\delta>0$，M 满足 $(\rho+2\sqrt{\rho\log(1/\delta)},\delta)$ - DP。

定理 2 [22] 设 M 由一系列自适应隐私机制 $M_1,M_2,...,M_k$ 组成，其中每个机制 $M_i:\prod_{i=1}^{k}R_i\times D\to R_i$ 满足 ρ_i - zCDP。当数据集 D 被随机拆分为 $D_1,D_2,...,D_k$ 时，机制 $M(D)=(M_1(D\cap D_1),...,M_k(D\cap D_k))$ 满足 $\{\max_i\rho_i\}$ - zCDP。

2.3 SHAP

SHAP[23] (Shapley additive explanations) 是一种对黑箱模型进行解释的方法。SHAP 基于 Shapley 值被解释为一种加性特征归因方法，以此衡量出每个输入特征对最终预测结果的贡献程度。模型的预测结果被解释为二元变量的线性函数如下：

$$g(z)=\phi_0+\sum_{i=1}^{|M|}\phi_i \tag{5}$$

其中，g 表示解释模型，M 表示输入特征集合，ϕ_0 表示平均模型的预测，ϕ_i 为每个特征 i 的 Shapley 值，其计算公式为

$$\phi_i=\sum_{V\subseteq\{M\setminus x_i\}}\frac{(|M|-|V|-1)!|V|!}{|M|!}\{f(x_{V\cup\{i\}})-f(x_V)\} \tag{6}$$

其中，V 表示 $\{M\setminus x_i\}$ 的子集合，分式表示不同特征组合对应的概率，$f(x_{V\cup\{i\}})$ 与 $f(x_V)$ 分别表示不同特征组合下 x_i 入模和不入模时的预测结果。

3 方案设计

本文基于差分隐私的思想，在模型训练过程中，首先利用 SHAP 模型衡量每个输入特征对模型预测结果的贡献度，再根据贡献比例对每条原始数据在特征维度进行自适应扰动，解除迭代次数与隐私预算之间的依赖；其次，基于函数机制原理，对损失函数进行泰勒展开获取近似多项式并对其系数进行扰动，确保每条样本的真实标签信息也不会被泄露。在每次参数更新时，通过自适应矩估计算法来优化梯度和调整学习率，从而加快模型收敛速度。此外，还引入了零集中差分隐私的组合机制对整个训练过程中隐私损失进行了更严格更清晰的度量。最终，本文的 ADPE 方案在保护了整个深度学习模型隐私的同时，极大地保证了模型训练的准确性和实用性。具体系统设计图如图 1 所示。

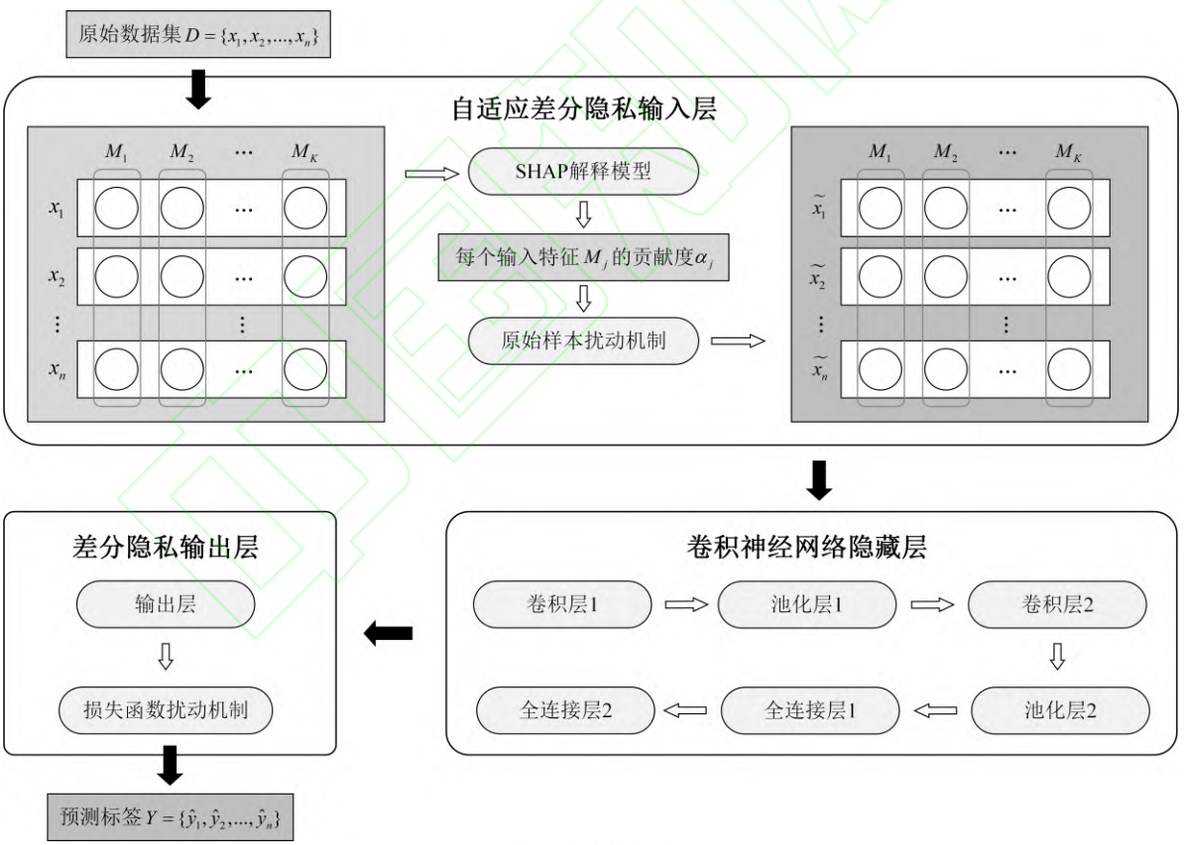


图 1 系统设计图

3.1 模型框架

本文以卷积神经网络作为基础网络结构，每个隐藏层神经元的转换过程可表示为 $h=a(\mathbf{x}W^T+b)$，其中 $\mathbf{x}$ 为输入向量，h 为输出，b 为偏置项，W 为权重矩阵。$\mathbf{x}W^T+b$ 表示线性变换部分，$a(\cdot)$ 为激活函数。给定一个模型参数为 θ 的损失函数 $L(\theta)$，通过在 N_{epoch} 轮训练中应用 Adam 算法优化数据集 D 上的损失函数 $L(\theta)$ 来训练卷积神经网络。其中，每个训练轮次进行 $N_{iteration}$

次迭代，每批次训练样本 B 是 D 中大小为 $|B|$ 的随机集合。

ADPE 方案主要考虑是白盒攻击[24]，即敌手拥有该深度学习模型的全部知识，包括模型结构和参数，可以访问发布的模型而不只是训练过程。此时主要存在以下两种隐私泄露的威胁。1）敌手依据模型参数获取敏感信息甚至原始数据。2）敌手试图通过目标模型推断出某条目标样本是否参与过训练。

3.2 具体流程

接下来，将介绍 ADPE 方案的具体流程，如算法 1 所示，主要分为五个阶段。需要注意的是，总迭代次数为训练轮数和每轮迭代次数的乘积。

1）自适应噪声尺度的分配。从输入特征的角度来看，不同的输入特征对预测结果影响程度是不同的，较重要的特征往往对预测结果起到决定性作用，而不重要的特征无论如何扰动都不会对结果产生太大影响。因此，可以将每个特征的贡献度作为分配噪声尺度的依据。

首先，读取批量数据进行特征维度上的贡献度计算，记作 Ctr_j，$j\in[1,k]$，表示第 j 个特征对预测标签的贡献度。对于每个输入特征 M_j，计算每个样本中该输入特征的 SHAP 值，将所有样本中该特征 SHAP 值累加求均值便得到了该特征的贡献度。其次计算每个输入特征对于预测结果的重要性，即贡献比例 $\alpha_j=\dfrac{K|Ctr_j|}{\sum_{i=1}^{K}|Ctr_j|}$，$\sum_{i=1}^{K}|Ctr_j|$ 表示 K 个特征的贡献度之和。最后，基于该贡献比例决定每个特征的噪声尺度 $\sigma_j=\dfrac{1}{\alpha_j}\sigma_1$。

2）原始样本的扰动。考虑到原始样本只作为神经网络的输入被使用，在样本被输入神经网络时，构造自适应差分隐私机制对每条数据添加高斯噪声来实现扰动，无需在每次迭代中都对模型梯度或权重加噪。这能够让隐私损失不再受迭代次数的影响，提高了模型的准确性和实用性。具体而言，对样本集合 B 的每个样本 x_i 中第 j 个输入特征值添加的噪声如下：

$$x_i(M_j)=x_i(M_j)+\frac{1}{|B|}N(0,\Delta s_1^{\,2}\sigma_j^{\,2}I) \tag{7}$$

其中，Δs_1 表示原始数据的敏感度。假设两个相邻样本集合 B 和 B' 中只有最后一个样本 x_n 和 x_n' 不同，且 $x_i(M_j)$ 被归一化到 $[0,1]$，则敏感度 Δs_1 计算如下：

$$\begin{aligned}\Delta s_1&=\sum_{j=1}^{K}\Big\|\sum_{x_i\in B}x_i(M_j)-\sum_{x_i'\in B'}x_i'(M_j)\Big\|\\&=\sum_{j=1}^{K}\|x_n(M_j)-x_n'(M_j)\|\\&\le 2\max_{x_i\in B}\sum_{j=1}^{K}\|x_i(M_j)\|=2K\end{aligned} \tag{8}$$

通过式(7)可以看出，某输入特征对预测结果的贡献度越小，所分配的隐私预算越少，添加的噪声尺度就越大，因为对于它们而言，加太多噪声对预测结果的影响并不大。该过程衡量了隐私与效用之间的关系：在提供隐私保护的同时，尽可能保证数据的可用性。

3）损失函数的扰动。由现有损失函数的定义可知，原始样本的真实标签值 $\{y_1,\dots,y_d\}$ 参与了损失函数的计算，因此，为保护原始样本中的标签，可以根据函数机制原理[25]来处理损失函数。本文采用经典的 sigmoid 作为激活函数，交叉熵作为损失函数，表示如下：

$$\begin{aligned}L(\theta)&=-\sum_{l=1}^{d}\sum_{x_i\in B}\Big(y_{il}\log y_{il}+(1-y_{il})\log(1-y_{il})\Big)\\&=-\sum_{l=1}^{d}\sum_{x_i\in B}\left(y_{il}\log\left(\frac{1}{1+e^{-H_{x_i}W^T}}\right)+(1-y_{il})\log\left(1-\left(\frac{1}{1+e^{-H_{x_i}W^T}}\right)\right)\right)\\&=-\sum_{l=1}^{d}\sum_{x_i\in B}y_{il}\log(1+e^{-H_{x_i}W^T})+(1-y_{il})\log(1+e^{H_{x_i}W^T})\end{aligned} \tag{9}$$

其中，$H_{x_i}W^T$ 为最后一个隐藏层的输出。通过泰勒展开将损失函数在 0 处展开到二阶：

$$L(\theta)=\sum_{l=1}^{d}\sum_{x_i\in B}\sum_{R=0}^{2}\left(\frac{f_{1l}^{(R)}(0)}{R!}+\frac{f_{2l}^{(R)}(0)}{R!}\right)(H_{x_i}W^T)^R \tag{10}$$

其中，令多项式系数 $P_{x_il}^{(R)}=\frac{f_{1l}^{(R)}(0)}{R!}+\frac{f_{2l}^{(R)}(0)}{R!}$。

此时，真实的标签值 y_{il} 只与多项式系数 $P_{x_il}^{(R)}$ 有关，可以对每个系数项添加噪声：

$$P_{x_il}^{(R)}\leftarrow P_{x_il}^{(R)}+\frac{1}{|B|}N(0,\Delta s_2{}^2\sigma_2{}^2I) \tag{11}$$

其中，Δs_2 表示近似多项式系数的敏感度。同理，假设两个相邻样本集合 B 和 B' 中只有最后一个样本 x_n 和 x_n' 不同，则有[18]：

$$\begin{aligned}\Delta s_2&=\sum_{l=1}^{d}\sum_{R=0}^{2}\|\sum_{x_i\in B}P_{x_il}^{(R)}-\sum_{x_i'\in B'}P_{x_i'l}^{(R)}\|\\&=\sum_{l=1}^{d}\sum_{R=0}^{2}\|P_{x_nl}^{(R)}-P_{x_n'l}^{(R)}\|=\sum_{l=1}^{d}\sum_{R=1}^{2}\|P_{x_nl}^{(R)}-P_{x_n'l}^{(R)}\|\\&\le 2\max_{x_n}\sum_{l=1}^{d}\sum_{R=0}^{2}\|P_{x_nl}^{(R)}\|\le d(|H_{x_n}|+\frac{1}{4}|H_{x_n}|^2)\end{aligned} \tag{12}$$

式中，$R=0$ 时 $P_{x_nl}^{(0)}=P_{x_n'l}^{(0)}=\log 2$，$|H|$ 为最后一个隐藏层的神经元个数。

4）参数更新。自适应矩估计（adaptive moment estimation, Adam）算法[26]通过引入二次梯度矫正，为不同参数设计自适应的学习率，每次迭代的学习率都能有确定范围，使得参数比较平稳。模型训练的目标是为了使损失函数不断减小，直到得到全局最优解。首先对前一步的扰动近似损失函数 $L(\theta_t)$ 估计梯度值 $g_t\leftarrow\frac{1}{|B|}\nabla L(\theta_t)$，其次计算梯度的一阶矩估计和二阶矩估计：

$$s_1=\gamma_1 s_1+(1-\gamma_1)g_t \tag{13}$$

$$s_2=\gamma_2 s_2+(1-\gamma_2)g_t^2 \tag{14}$$

其中，γ_1 和 γ_2 表示指数衰减率。为防止 s_1 和 s_2 趋向 0，通过计算偏差来修正：

$$s_1^*=\frac{s_1}{1-\gamma_1^t},\quad s_2^*=\frac{s_2}{1-\gamma_2^t} \tag{15}$$

最后用优化的梯度更新参数：

$$\theta_{t+1}\leftarrow\theta_t-\frac{\eta}{\sqrt{s_2^*}+\xi}s_1^* \tag{16}$$

式中，ξ 是为了维持数值稳定性而添加的常数。不难看出，Adam 算法将历史梯度作为先验知识，利用历史梯度的指数衰减平均值更新当前梯度，加快了模型收敛速度；利用历史梯度平方的指数衰减平均值更新学习率，使得模型收敛过程更稳定。

5）隐私损失统计。引入差分隐私机制为深度学习训练过程提供更高的隐私保证，但不可避免地造成一定的隐私损失，当超出预定的隐私预算时，被保护的数据就会有泄露的风险。本文将隐私机制 M 的隐私损失记为随机变量 $Z\sim L_{(M(D)\|M(D'))}^{(O)}$。在严格的差分隐私定义中 $\Pr[Z>\varepsilon]=0$ 表示不允许该隐私机制被破坏，灵活性较差。(ε,δ)- DP 可以被看作一种松弛的差分隐私，允许 $\Pr[Z>\varepsilon]<\delta$，但用来衡量隐私损失上界时是比较宽泛的。

在深度学习训练过程中，模型收敛需要通过不断迭代来完成，并且通常需要多次迭代。将对同一组数据集的多轮重复训练当作是对数据集的多次查询，隐私损失会随着查询次数的增加而累积。因此，本文使用更加严密的差分隐私形式 zCDP 来统计隐私损失，降低隐私泄露的风险，更好地平衡隐私和模型效用之间的关系。其中，每一轮训练中的隐私损失为 $\rho=\frac{(\sigma_{\max}{}^2+\sigma_2{}^2)}{2\sigma_{\max}{}^2\sigma_2{}^2}$，$\sigma_{\max}=\max_{j\in K}\sigma_j$。

算法 1　ADPE

输入：总迭代次数 T，每轮迭代次数 $N_{iteration}$，批次训练样本 B，输入特征集合 $M=\{M_1,...,M_K\}$，超参数学习率 η，损失函数 $L(\theta)$，噪声尺度 σ_1 和 σ_2，全局敏感度 Δs_1 和 Δs_2

输出：目标模型参数 θ_T 和总体隐私损失 ρ_{total}

1) 初始化模型参数 θ_0 和隐私损失统计量 ρ
2) **//确定自适应噪声尺度**
3) **for** $j\to 1$ to K **do**
4) 计算输入特征 M_j 对预测标签的贡献度 Ctr_j
5) 计算贡献比例 $\alpha_j \leftarrow \dfrac{K|Ctr_j|}{\sum_{i=1}^{K}|Ctr_j|}$
6) 获取自适应的噪声尺度 $\sigma_j \leftarrow \dfrac{1}{\alpha_j}\sigma_1$
7) **end for**
8) **for** $t\leftarrow 1$ to T **do**
9) 获取批次训练集 B 中的每个样本 x_i
10) **//扰动原始样本**
11) 对于 x_i 中第 j 维度输入特征 $x_i(M_j)$
12) $x_i(M_j)=x_i(M_j)+\dfrac{1}{|B|}N(0,\Delta s_1^{\,2}\sigma_j^{\,2}I)$
13) **//扰动损失函数**
14) 计算损失函数 $L(\theta_t,x_i)$ 及其近似多项式 $L(P_{x_i l}^{(R)})$
15) $P_{x_i l}^{(R)} \leftarrow P_{x_i l}^{(R)}+\dfrac{1}{|B|}N(0,\Delta s_2^{\,2}\sigma_2^{\,2}I)$
16) 损失函数更新 $L(\theta_t)\leftarrow \sum_{x_i\in B}L(P_{x_i l}^{(R)})$
17) 梯度大小 $g_t \leftarrow \dfrac{1}{|B|}\nabla L(\theta_t)$
18) **//更新参数**
19) 使用 Adam 算法优化梯度 $\overline{g_t}\leftarrow \dfrac{s_1^*}{\sqrt{s_2^*}+\xi}$
20) 参数更新 $\theta_{t+1}\leftarrow \theta_t-\eta\overline{g_t}$
21) **//统计隐私损失**
22) **If** $0\leftarrow t \bmod N_{iteration}$ **do**
23) $\rho \leftarrow \dfrac{(\sigma_{\max}^{\,2}+\sigma_2^{\,2})}{2\sigma_{\max}^{\,2}\sigma_2^{\,2}}$
24) $\rho_{total}\leftarrow \rho_{total}+\rho$
25) **end if**
26) $t\leftarrow t+1$
27) **end for**
28) 输出 θ_T，ρ_{total}

3.3 隐私性分析

由 3.1 节可知，本文主要存在两种隐私泄露的威胁，二者本质上都是由于敌手可以从模型本身获取到隐私数据。首先，针对威胁 1），在训练之前直接对原始数据进行加噪处理，从而在训练过程中减弱中间参数与原始数据的关联性，让敌手无法反推出真正准确的数据信息。其次，针对威胁 2），通过加入满足差分隐私定义的噪声，使得相邻的两条数据样本无法区分，敌手就无法判断目标样本是否真实存在于训练数据集。因此，通过证明算法 1 满足差分隐私来论证对上述两种威

胁的抵抗。

定理 3 算法 1 满足 $N_{epoch}\rho_0$ - zCDP，即 $(\rho_1+2\sqrt{\rho_1\log(1/\delta)},\delta)$ - DP。

证明：在所提方案中，一个数据集被分为多个不重复的数据批在每轮训练中训练一次，每个数据批在一轮训练中进行了两次加噪处理。第一次是在原始样本输入时，为每条样本的每个特征添加满足 $N(0,\Delta s_1^2\sigma_j^2 I)$ 的高斯噪声，第二次是在损失函数输出时添加满足 $N(0,\Delta s_2^2\sigma_2^2 I)$ 的高斯噪声，由 zCDP 性质 1 和定理 2 可知，两次处理分别满足 $\frac{1}{2\sigma_{\max}^2}$ - zCDP 和 $\frac{1}{2\sigma_2^2}$ - zCDP，其中 $\sigma_{\max}=\max_{j\in K}\sigma_j$。结合 zCDP 性质 2 得出，该数据批每轮训练结束后所添加噪声满足 ρ_0 - zCDP，其中 $\rho_0=\frac{(\sigma_{\max}^2+\sigma_2^2)}{2\sigma_{\max}^2\sigma_2^2}$。又因为每轮训练中每次迭代的数据批互不相交，由差分隐私并行组合性知，整个数据集在一轮训练后仍然满足 ρ_0 - zCDP，此时发现本文方案的隐私损失与每轮的迭代次数是无关的。当进行 N_{epoch} 轮训练时，整个过程便满足 ρ_1 - zCDP，其中 $\rho_1=N_{epoch}\rho_0$，如算法 1 在每轮训练结束时计算隐私损失。此外，通过性质 3 可以实现 zCDP 到 DP 的转换，即算法 1 提供了 $(\rho_1+2\sqrt{\rho_1\log(1/\delta)},\delta)$ - DP。值得注意的是，上述分析是以整个训练过程都输入相同数据集为前提展开的，如果都在不相交的数据集上运行，那么实际的隐私损失累计将会更小。

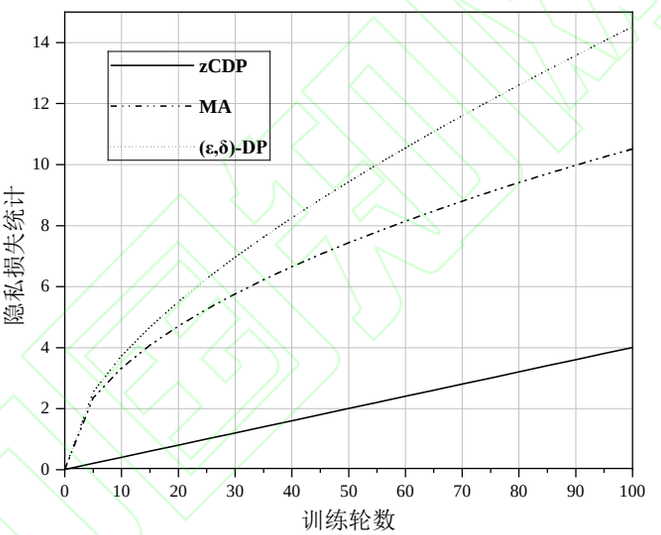


图 2 隐私损失与训练轮数的关系

本文的隐私损失统计部分可以扩展为隐私损失的动态监测机制，即给定 zCDP 的总隐私预算 ρ_{tatol}，每轮训练之前都先判断：剩余的隐私预算减去本轮所需的隐私预算后是否大于 0，只有大于 0 才继续执行训练，从而保证整个训练的运行都满足 ρ_{tatol} -zCDP。图 2 展示了当 $\sigma_{\max}=\sigma_1=\sigma_2=5$ 时，随着训练轮数的增加，分别采用 zCDP、 MA 和 (ε,δ)-DP 来统计隐私损失的变化情况，其中 $\delta=1\times10^{-2}$。

4 实验与分析

4.1 实验设置

本文使用 MNIST 和 Fashion MNIST 两种数据集进行实验验证。其中，MNIST 数据集包含 10 种类别的手写数字图片，有 60000 个训练样本和 10000 个测试样本，每个样本由 28×28 个像素点的灰度图像构成。Fashion MNIST 数据集由 10 种类别的服装正面图片组成，分为 60000 个训练图像和 10000 个测试图像，每个样本图像包含 28×28 像素。

实验部署在操作系统为 Windows 11 64 位、CPU 为 12th Gen Intel(R) Core(TM) i7-12700H @2.30 GHz、GPU 为 Nvidia GeForce GTX2050 4GB 和内存 16GB 的工作站，基于 Python 3.8 仿真实验。具体来说，预训练时使用 DeepSHAP 衡量输入特征对输出的贡献度，采用 Tensorflow1.5.0 训练深

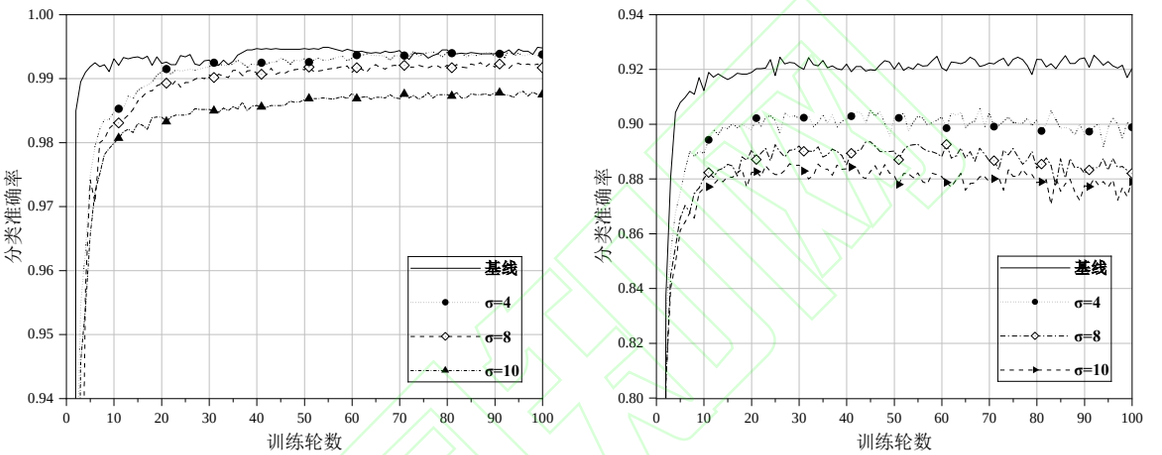
度学习模型，网络结构为卷积神经网络，包含 2 个 32 和 64 个特征、卷积核大小为 5×5、步长为 1 的卷积层，2 个 2×2 的最大池化层，以及 2 个神经元个数均为 30 的全连接层。利用 Adam 算法进行模型训练时基本参数设置为 $\xi=1\times10^{-8}$，$\gamma_1=0.9$，$\gamma_2=0.999$，并结合指数衰减法优化学习率使得模型在后期训练中更加稳定，所选择的批次样本大小为 600。

4.2 实验结果

在上述实现环境下，主要进行两个实验，实验一是验证 ADPE 方案的有效性，实验二是将所提 ADPE 方案与现有方案在模型准确性上进行对比。

4.2.1 有效性验证

该实验通过对比模型引入自适应差分隐私机制前后的模型准确率，来验证 ADPE 方案的有效性。引入差分隐私机制前以常规的方式训练本文的基础网络结构模型，称作基线模型，引入后在 $\delta=1\times10^{-5}$ 的情况下分别设置 $\sigma_1=\sigma_2=4$，8，10。在训练轮数 $N_{epoch}=100$ 时结果如图 3 所示。



(a) MNIST 数据集　　(b) Fashion-MNIST 数据集

图 3 ADPE 方案有效性验证实验

观察图 3 可以得出以下结论。

1）引入 ADPE 方案的隐私保护机制对模型进行扰动时，不会明显降低模型的准确率。

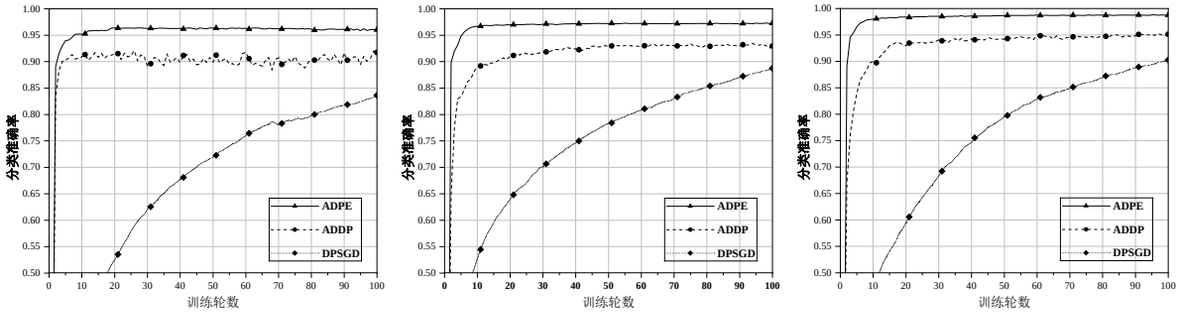
2）在 MNIST 数据集上，第 10 轮训练时，三种噪声条件下的模型准确率都达到 98%以上，第 50 轮训练后趋于稳定，尤其是当 $\sigma=4$ 时中后期的训练效果与基线几乎一致，说明模型较好的可用性。

3）由于 Fashion-MNIST 数据集的图像比 MNIST 数据集更复杂，因此模型的准确率没有 MNIST 数据集那么高，且模型中后期训练包括基线模型在内也没有那么稳定，会在 1%上下波动，但总体的训练效果依然在 87%以上，说明该方案的有效性。

4）对于不同的噪声尺度，噪声尺度越小，模型的分类准确率就越高，说明用户可以根据个性化需求调整噪声尺度实现该方案隐私和效用之间的平衡。

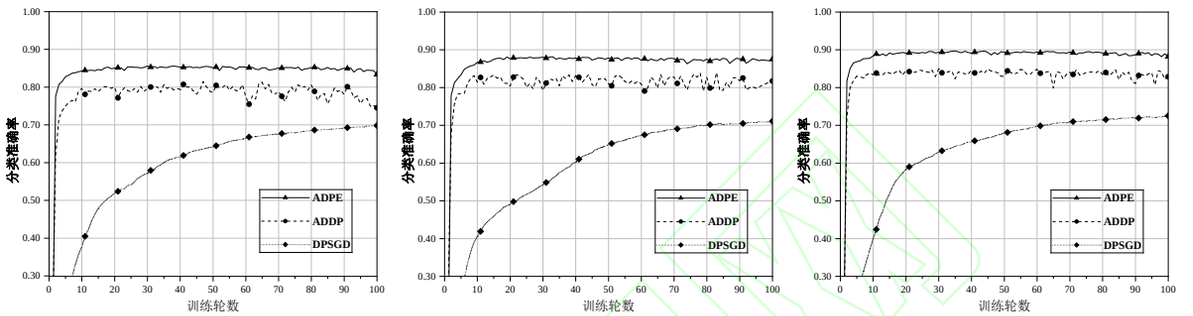
4.2.2 对比分析

该实验探究本文所提方案 ADPE 与经典方案 DPSGD[13]和较为先进的方案 ADDP[19]对模型提供隐私保护时的对比情况。对于 3 种方案，设置 $N_{epoch}=100$，$\delta=1\times10^{-4}$，当取不同隐私预算时，三种方案在 2 种数据集上的分类准确率随训练轮数的变化情况分别如图 4 和图 5 所示。 其中，ADPE 方案采用 3.3 节的公式得出隐私预算和噪声参数的关系。



(a) $\varepsilon = 0.5$ (b) $\varepsilon = 2$ (c) $\varepsilon = 4$

图 4 MNIST 数据集



(a) $\varepsilon = 0.5$ (b) $\varepsilon = 2$ (c) $\varepsilon = 4$

图 5 Fashion-MNIST 数据集

观察图 4 和图 5 可得出以下结论。

1）随着隐私预算的增大，意味着所添加的噪声逐渐变少，三个方案的模型分类准确率都呈上升趋势。

2）隐私预算相同的情况下，所提 ADPE 方案在 2 种数据集上的模型分类准确率都高于其他两个方案，说明 ADPE 方案的性能较优。具体地，在较大的隐私预算 $\varepsilon = 4$ 时，该方案的模型在 MNIST 数据集上能达到 98.7%的准确率，在 Fashion-MNIST 数据集上准确率后期最高为 89.7%。

3）在 MNIST 数据集上，当隐私预算 $\varepsilon = 0.5$ 时，ADDP 方案难以达到收敛状态，波动较为剧烈，而 ADPE 和 DPSGD 方案表现较为平和，说明加入高斯噪声更有利于模型的稳定。随着隐私预算增大，ADPE 和 ADDP 方案在 20 轮训练后都趋于稳定，甚至 ADPE 方案在 10 轮左右就基本收敛，而 DPSGD 方案在训练 100 轮后还未达到明显的收敛状态。

4）在 Fashion-MNIST 数据集上，由于其样本结构的复杂性，ADPE 方案和 ADDP 方案会出现一定程度的波动，但前者的波动范围是在小区域内，后者的波动范围较大，而 DPSGD 方案仍然收敛较慢。综上可见，ADPE 方案在加快了模型收敛的同时具有一定的稳定性。

5 结 论

本文提出了一种基于自适应差分隐私的高效深度学习方案，有效平衡了模型的隐私性和可用性。该方案基于沙普利加性解释模型设计了一种自适应差分隐私机制，用于保护原始数据样本，并利用函数机制扰动原始标签，增强了深度模型训练的隐私性。同时，引入零集中差分隐私的组合机制度量整个训练过程的隐私损失，使得方案有更好的隐私保证。通过在两个经典数据集 MNIST 和 Fashion-MNIST 上的实验分析表明，所提方案能够在保护隐私的前提下尽可能实现较高的模型准确率，并且加快了模型收敛速度以及保证了模型中后期训练的稳定。

参考文献

[1] SONG M K, WANG Z B, ZHANG Z F, et al. Analyzing user-level privacy attack against federated learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430-2444.

[2] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]// 2017 IEEE Symposium on Security and Privacy. IEEE, 2017: 3-18.

[3] SALEM A, ZHANG Y, HUMBERT M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[C]//The 26th Annual Network and Distributed System Security Symposium. California: NDSS, 2019: 24-27.

[4] Yu L, Liu L, Pu C, et al. Differentially Private Model Publishing for Deep Learning[C]//2019 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2019: 332-349.
[5] 康海燕, 冀源蕊. 基于本地化差分隐私的联邦学习方法研究[J]. 通信学报, 2022, 43(10): 94-105.
KANG H, JI Y. Research on federated learning approach based on local differential privacy[J]. Journal on Communications, 2022, 43(10): 94-105.
[6] PODSCHWADT R, TAKABI D, HU P, et al. A Survey of Deep Learning Architectures for Privacy-Preserving Machine Learning with Fully Homomorphic Encryption[J]. IEEE Access, 2022, 10: 117477-117500.
[7] CHEN J, LI K and YU P. Privacy-Preserving Deep Learning Model for Decentralized VANETs Using Fully Homomorphic Encryption and Blockchain[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 11633-11642.
[8] Resende A, Railsback D, Dowsley R, et al. Fast privacy-preserving text classification based on secure multiparty computation[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 428-442.
[9] FENG Q, HE D, SHEN J, et al. PpNNT: Multi-Party Privacy-Preserving Neural Network Training System[J]. IEEE Transactions on Artificial Intelligence, 2023.
[10] DWORK C. Differential privacy[C]//The 33rd International Conference on Automata, Languages and Programming - Volume Part II, ICALP'06. Berlin: Springer, 2006: 1–12.
[11] 徐花, 田有亮. 差分隐私下的权重社交网络隐私保护[J]. 西安电子科技大学学报, 2022, 49(1): 17-25.
Xu H, Tian Y. Protection of privacy of the weighted social network under differential privacy[J]. Journal of Xidian University, 2022, 49(1): 17-25.
[12] 晏燕, 董卓越, 徐飞, 等. 一种Hilbert编码的本地化位置隐私保护方法[J]. 西安电子科技大学学报, 2022: 1-15.
YAN Y, DONG Z, XU F, et al. Localized location privacy protection method using the Hilbert encoding[J]. Journal of Xidian University, 2022: 1-15.
[13] Hu Y, Tan Z, Li X, et al. Adaptive Clipping Bound of Deep Learning with Differential Privacy[C]//2021 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2021: 428-435.
[14] FU J, CHEN Z, and HAN X. Adap DP-FL: Differentially Private Federated Learning with Adaptive Noise[C]//2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Wuhan, China: IEEE, 2022: 656-663.
[15] WANG F, XIE M, TAN Z, et al. Preserving Differential Privacy in Deep Learning Based on Feature Relevance Region Segmentation[J]. IEEE Transactions on Emerging Topics in Computing, 2023.
[16] ABADI M, CHU A, GOODFELLOW I, et al. 2016. Deep Learning with Differential Privacy[C]//The 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16). New York: ACM, 2016: 308–318.
[17] ZHANG X, DIING J, WU M, et al. Adaptive Privacy Preserving Deep Learning Algorithms for Medical Data[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2021: 1168- 1177
[18] PHAN N, WU X, HU H. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning[C]//2017 IEEE International Conference on Data Mining (ICDM). Lafayette: IEEE, 2017: 385-394.
[19] ZHANG Y, BAI S. An Improved LRP-Based Differential Privacy Preserving Deep Learning Framework[C]//2021 17th International Conference on Computational Intelligence and Security (CIS). IEEE, 2021: 484-488.
[20] LIU X, LI H, XU G, et al. Adaptive privacy-preserving federated learning[J]. Peer-to-Peer Networking and Applications, 2020, 13: 2356–2366.
[21] BUN M, STEINKE T. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds[J]. Theory of Cryptography, Berlin: Springer, 2016, 9985.
[22] Yu L, Liu L, Pu C, et al. Differentially private model publishing for deep learning[C]//2019 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2019: 332-349.
[23] LI C, LOU J, LIU S, et al. Shapley Explainer-An Interpretation Method for GNNs Used in SDN[C]//GLOBECOM 2022-2022 IEEE Global Communications Conference. IEEE, 2022: 5534-5540.
[24] 纪守领, 杜天宇, 李进锋, 等.机器学习模型安全与隐私研究综述[J]. 软件学报, 2021,32(1): 41-67.
JI S, DU T, LI J, et al. Research on Security and Privacy of Machine Learning Models[J]. Journal of Software, 2021, 32(1): 41-67.
[25] ZHANG J, ZHANG Z, XIAO X, et al. Functional Mechanism: Regression Analysis under Differential Privacy[J]. Proceedings of the VLDB Endowment, 2012, 5(11) : 1364-1375.
[26] KUMAR G, PRIYA G, DILEEP M, et al. Image Deconvolution using Deep Learning-based Adam Optimizer[C] //2022 6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022: 901-904.