基于稀疏平滑自蒸馏的差分隐私深度学习方法

赵登峰,薛大暄,赵素云,陈 红*
(中国人民大学信息学院,北京 100081)

摘 要: 为了减少深度学习中隐私泄露的风险,许多研究利用差分隐私技术来训练神经网络. 然而,这些隐私保护方法通常会导致模型性能显著下降. 为了在隐私保护与模型效用之间实现平衡,本文提出了一种基于稀疏平滑自蒸馏的差分隐私深度学习(Differentially Private learning with sparse and smooth Self-Distillation,DP3SD)方法,通过双温度缩放机制来增强隐私保护深度学习的效用. 具体而言,该方法设计了一种由稀疏分类损失和光滑蒸馏损失组成的双温度缩放损失函数. 通过将较低温度应用于分类损失,能够使学生模型的类别预测分布更加锐化,从而减少低概率类别的影响,这些类别通常可能是由噪声引起的. 相反,较高温度应用于蒸馏损失,能够平滑教师模型和学生模型的预测分布,从而在差分隐私约束下实现稳定和高效的知识迁移. 在差分隐私随机梯度下降的严格隐私保障下,本文提出的双重缩放机制能够减轻噪声带来的扰动,提升学生模型的泛化能力. 在三个公开数据集上的大量实验表明:本文提出的方法能够在确保严格数据隐私的同时,增强模型的可用性.

关键词: 深度学习;差分隐私;隐私保护;知识蒸馏;随机梯度下降

基金项目: 国家重点研发计划(No.2023YFB4503600);国家自然科学基金(No.U23A20299,No.U24B20144,No.62172424,No.62276270,No.62322214)

中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3310-09

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250133

Differentially Private with Sparse and Smooth Self-Distillation

ZHAO Deng-feng, XUE Da-xuan, ZHAO Su-yun, CHEN Hong*
(*School of Information*, *Renmin University of China*, *Beijing* 100081, *China*)

Abstract: To mitigate privacy leakage risks in deep learning, numerous studies utilize differential privacy techniques to train neural networks. However, substantial performance degradation is often unavoidable. To address the privacy-utility trade-off, we propose the differentially private learning with sparse and smooth self-distillation (DP3SD) method, which leverages dual temperature scaling to enhance the utility of privacy-preserving deep learning. Specifically, DP3SD proposes a dual scaling loss function composed of a sparse classification loss and a smooth distillation loss. By incorporating a lower temperature into the classification loss, the class prediction distribution of student model is sharpened, thereby reducing the influence of low-probability classes that are likely noise-induced. Conversely, applying a higher temperature to the distillation loss, the prediction distributions of both the teacher and student models are smoothed, thus promoting stable and efficient knowledge transfer under differential privacy constraints. This dual scaling mechanism, under strict privacy guarantees through differential privacy stochastic gradient descent, facilitates the student model in progressively enhancing its learning from the teacher model while simultaneously alleviating the perturbations caused by privacy constraints. By extensive experiments on three public datasets, we find that DP3SD can effectively improve model performance while ensuring rigorous data privacy.

Key words: deep learning; differential privacy; privacy protection; knowledge distillation; stochastic gradient descent

Foundation Item(s): National Key Research and Development Program of Chnia (No.2023YFB4503600); National Natural Science Foundation of China (No.U23A20299, No.U24B20144, No.62172424, No.62276270, No.62322214)

收稿日期:2025-02-21;录用日期:2025-06-24;责任编辑:覃怀银
*通讯作者:陈红

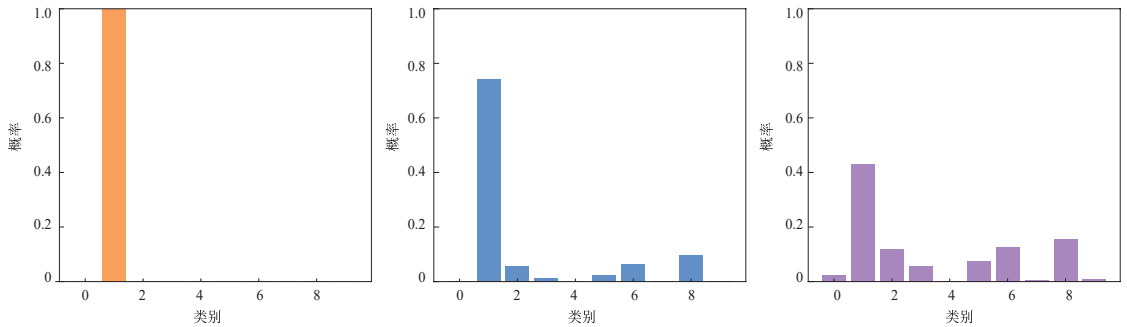
1 引言

在当前数据驱动的智能时代,深度学习已广泛应用于医疗诊断、金融风控、自动驾驶、智慧城市以及社交媒体内容推荐等关键领域,成为推动人工智能快速发展的核心技术. 然而,深度学习模型通常依赖大规模的敏感数据进行训练,近期研究表明[1,2]:这些模型可能无意中记忆并泄露训练数据,从而导致严重的隐私风险. 例如,在大型语言模型的研究中,有攻击者成功提取出用户的信用卡信息[3];在图像领域,也有攻击者通过反推模型参数重构用户的原始医疗数据[4]. 因此,如何在保障深度学习模型效能的同时,有效保护用户隐私,已成为深度学习研究与应用中亟须解决的重要问题.

差分隐私[5]在深度学习中广泛应用,因为理论上它能确保攻击者无法判断某个特定用户的数据是否用于模型训练. 此外,研究表明:差分隐私模型能够有效抵御多种攻击,包括成员推断攻击[6]、属性推断攻击[3]和数据提取攻击[4].

在训练深度学习模型时,为了确保隐私,差分隐私随机梯度下降(Differentially Private Stochastic Gradient Descent,DPSGD)[7]算法被广泛使用. 它通过两种关键机制保护隐私:一是对每个训练样本的梯度进行裁剪,从而减少单个样本对模型训练的影响;二是通过加入高斯噪声,掩盖样本的贡献,进一步防止隐私泄露. 尽管DPSGD能够有效提供隐私保障,但噪声的引入会降低模型性能. 为了在有限的隐私预算下优化模型性能,近年来的研究[8,9]提出了在训练过程中使用中间检查点的方法. 这些检查点代表了模型训练的中间状态,通过将中间检查点的参数或输出进行聚合,能够在不增加隐私预算的情况下,提升模型的泛化能力. 然而,这些方法未能充分利用中间检查点所包含的有价值的信息,与非隐私模型相比,仍然存在一定的性能差距.

自蒸馏[10]技术能够在不改变模型架构的情况下,利用模型自身输出提升泛化能力. 本文探索了自蒸馏在差分隐私保护模型中的应用潜力,特别值得注意的是,使用中间检查点作为教师模型不会带来额外的隐私消耗. 为了缩小隐私与非隐私模型之间的性能差距,本文利用中间检查点作为教师模型来进行差分隐私深度学习,提出了差分隐私稀疏平滑自蒸馏(Differentially Private with sparse and smooth Self-Distillation,DP3SD)方法. DP3SD通过双温度缩放构建了一个稀疏平滑自蒸馏框架,通过调节温度来实现稀疏性和平滑性,从而提升模型的性能. 具体来说,DP3SD结合低温度下的稀疏分类损失和高温度下的平滑蒸馏损失来进行优化. 高概率类别是指那些模型在训练过程中具有较高置信度的类别,即模型认为最有可能正确预测的类别. 低概率类别是指那些模型认为较不可能出现的类别,通常在模型输出中显示较低的概率. 如图1所示,通过将较低的温度应用于分类损失,可以增强高概率类别的预测分布,从而减少低概率类别的影响,这些低概率类别可能是由噪声引起的. 相反,将较高的温度应用于蒸馏损失,可以使教师和学生模型的预测分布变得平滑,从而在差分隐私约束下实现稳定且高效的知识迁移. 总的来说,DP3SD通过双温度缩放机制实现稀疏性和平滑性. 稀疏性忽略了低概率类别,从而减少了噪声的影响,增强了模型在隐私约束下的鲁棒性. 与此同时,平滑性更有效地捕捉了类别间的关系,利用知识蒸馏提高模型的泛化能力. 通过将稀疏性和平滑性的优势结合起来,DP3SD实现了模型效用与隐私保护之间的有效平衡.



(a) $\tau=0.3$　(b) $\tau=1.0$　(c) $\tau=2.0$

图1　在不同温度缩放系数τ取值下的概率分布可视化

本文的主要贡献如下:

(1)提出双温度缩放损失函数,包括由低温度控制的稀疏分类损失和由高温度引导的平滑蒸馏损失. 这个双温度缩放机制能够在深度学习的隐私保护和模型效用之间进行平衡.

(2)基于双温度缩放机制,构建了一个稀疏平滑自蒸馏框架. 利用中间检查点作为辅助教师模型,有效地指导模型的迭代,且不增加额外的隐私预算,增强了模

型的效用.

(3)在三个公开数据集上的实验结果表明,本文方法在相同隐私成本下,优于现有的三种隐私保护方法,具有更高的准确性.

2 基础知识与相关工作

2.1 差分隐私

差分隐私[11]是一个严格的数学框架,用于定义和量化数据隐私保护. 它确保个体数据的加入或删除不会显著改变任何分析结果,从而实现可量化隐私保护,防止个体身份的泄露. 具体而言,差分隐私通过控制查询输出结果的概率分布,保证无论某个个体是否参与数据集,输出的分布几乎保持一致,从而确保个体隐私不被泄露.

本文考虑差分隐私如下所述:

一个机制满足(ϵ,δ)差分隐私,当且仅当对于任意两个只相差一个元素的相邻数据集$D,D'\in D^n$,以及对所有$S\subseteq R$,以下不等式成立:

$$\Pr\left[\mathcal{A}(D)\in S\right]\leqslant\exp(\epsilon)\times\Pr\left[\mathcal{A}(D')\in S\right]+\delta \quad (1)$$

式中,ϵ为一个非负参数,用于控制算法的隐私损失,较小的ϵ提供更强的隐私保护,但可能会增大结果误差;参数δ理想情况下应小于数据集规模大小的倒数,以确保隐私泄露的概率可以忽略不计.

除了差分隐私,当前仍有多种隐私保护机制被广泛研究,以适应不同场景下的隐私需求. 例如,同态加密技术[12]允许在加密数据上直接计算,保障数据在计算过程中的隐私;联邦学习[13]通过分布式模型训练避免数据集中上传,从根本上减少隐私泄露的风险,并逐步结合差分隐私与安全多方计算等机制以提升安全性. 然而,这些方法也各自存在局限性:同态加密计算开销巨大,难以应用于大规模神经网络;联邦学习则在通信效率、系统异构性和跨设备隐私保障方面仍面临挑战. 相比之下,差分隐私具有可组合性和后处理不变性,在深度学习中更容易集成与分析,因此成为当前研究的主流.

2.2 差分隐私深度学习

为了在训练深度学习模型时保护隐私,最常用的算法是DPSGD[7]. DPSGD通过在梯度计算中加入噪声,修改了传统的随机梯度下降算法,从而保护参与训练的单个数据点的隐私. 在DPSGD中,梯度在添加噪声之前会被裁剪到预定义的范数阈值. 梯度裁剪确保了任何单一数据点对模型更新的影响是有限的,从而限制了梯度的敏感性. 在数学上,迭代t轮时的梯度g_t 裁剪方式为

$$g_t\leftarrow g_t\cdot\min\left(1,\frac{C}{\|g_t\|_2}\right) \quad (2)$$

为了实现差分隐私,噪声被添加到裁剪后的梯度中. 噪声通常来自均值为0、标准差与裁剪阈值及隐私参数成比例的高斯分布. 噪声梯度$\widetilde{g_t}$的计算方式为

$$\widetilde{g_t}=g_t+N\left(0,\sigma^2C^2I\right) \quad (3)$$

其中,σ 为噪声系数,用于控制模型效用和隐私之间的权衡.

该方法中的隐私损失是通过分析高斯机制来评估的,并结合子采样中的隐私放大效应以及多次迭代中的合成定理进行计算. 这种方法使得在训练过程中生成的所有中间检查点都可以公开发布. 为了获得更严格的隐私界限,通常采用如Rényi差分隐私(Rényi Differential Privacy,RDP)[14]或数值合成算法[15]等技术. 自适应裁剪[16]和基于生成对抗网络的差分隐私数据合成方法[17]促进了差分隐私深度学习的进一步发展.

2.3 知识蒸馏

知识蒸馏[18]是深度学习中的一种先进技术,用于将知识从一个更大、更复杂的模型(通常称为“教师”模型)转移到一个较小的模型(通常称为“学生”模型). 在知识蒸馏中,学生模型不仅从传统的独热标签中学习,还从教师模型生成的预测概率中学习. 教师模型生成的预测概率包含了关于类别之间相对差异的更多信息. 这种方法使学生模型能够模拟教师模型的行为,从而即使在模型规模较小的情况下,也能获得不错的性能.

知识蒸馏的目标是最小化教师模型和学生模型输出之间的Kullback-Leibler(KL)散度. 知识蒸馏的损失函数为

$$L_{\mathrm{KD}}=\mathrm{KL}\left(p^t\|p^s\right)=\sum_{j=1}^{C}p_j^t\ln\left(\frac{p_j^t}{p_j^s}\right) \quad (4)$$

式中,p^t和p^s分别为教师模型和学生模型的预测概率分布. 具体而言,p_j^t和p_j^s分别为教师模型和学生模型对样本属于第j类的预测概率分布.

与知识蒸馏不同,自蒸馏的教师模型和学生模型使用相同的神经网络架构. 一些研究[10,19]利用学生模型本身作为教师,利用其自身的知识来提升模型的性能. 与渐进式自蒸馏方法[20]相似,DP3SD方法将模型前一轮的参数作为新一轮训练开始时的教师模型. 这种优化方法能够动态地学习目标,从而提升模型性能.

3 方法

本文提出的DP3SD方法,如图2所示. 在DP3SD中,中间检查点作为教师模型提供知识,学生模型利用差分隐私随机梯度下降进行训练. DP3SD方法采用双温度缩放策略,通过调节温度参数,低温度有助于模型学习高置信度的类别,高温度则使蒸馏过程能够提供更丰富的知识. 通过将较低的温度应用于分类损失,可

以增强高概率类别的预测分布,从而减少低概率类别的影响,这些低概率类别可能是由噪声引起的. 相反,将较高的温度应用于蒸馏损失,可以使教师和学生模型的预测分布变得平滑,从而在差分隐私约束下实现稳定且高效的知识迁移.

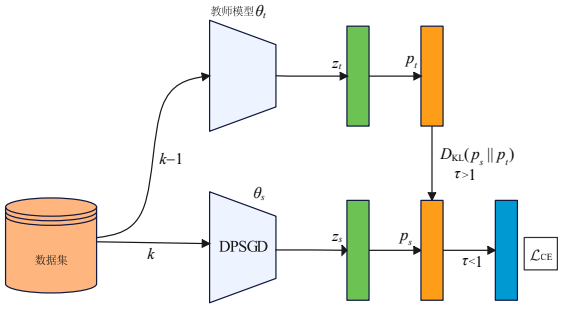


图2　DP3SD框架图

由于差分隐私深度学习中的隐私损失评估依赖于组合定理,允许公开发布中间训练检查点,本文在模型优化迭代过程中,结合来自第k–1次迭代时模型的历史知识来进行第k次迭代训练. 利用中间检查点作为辅助教师模型,能够有效地引导后续的模型迭代,提升模型效用的同时不会增加额外的隐私成本.

3.1　模型构建

假设有一个私有数据集D,由来自C个不同类别的标注实例组成. 数据集表示为$D=\left\{\left(x_i,y_i\right)\right\}_{i=1}^{N}$,其中$N$是数据集的大小,每个实例的标签属于集合$Y$,即$Y=\left\{y_i\right\}_{i=1}^{N}$,每个$y_i$属于集合$\{1,2,\cdots,C\}$.

对于一个具有C类的分类任务,给定模型的输出z,第i个类的概率p_i定义为

$$\widehat{p}_i=\mathrm{Softmax}\left(\frac{z_i}{\tau}\right)=\frac{\mathrm{e}^{z_i/\tau}}{\sum_{c=1}^{C}\mathrm{e}^{z_c/\tau}} \tag{5}$$

式中,τ为温度参数. 较低的τ值会导致更锐化的概率分布,而较高的τ值会生成更平滑的概率分布.

为了减轻DPSGD引入的噪声影响,本文设置较低的温度参数$\tau_s<1$来对学生模型的输出进行缩放. 较低的温度会使得模型的输出更加稀疏,关注高概率类别的预测结果,忽略可能是由噪声引起的低概率类别的预测结果. 稀疏分类损失的计算公式为

$$\mathcal{L}_{\mathrm{CE}}\left(p_s,y\right)=-\sum_{i=1}^{N}y_i\ln\left(\mathrm{Softmax}\left(\frac{z_s}{\tau_s}\right)\right) \tag{6}$$

式中,z_s为学生模型的输出;y为真实标签;p_s为通过应用带有温度缩放的softmax函数生成的稀疏概率分布. 这个损失函数鼓励模型关注高置信度的预测,从而减少低置信度预测带来的噪声影响.

除了稀疏分类损失外,为了促进从教师模型到学生模型的知识传递更加有效,本文引入了平滑蒸馏损失. 在蒸馏过程中,本文使用较高的温度$\tau_t>1$来对教师和学生模型的softmax输出进行平滑. 平滑蒸馏损失定义为教师模型和学生模型平滑输出之间的KL散度,计算式为

$$\mathcal{L}_{\mathrm{KL}}\left(q_t,p_t\right)=\tau_t^2\cdot\mathrm{KL}\left(\mathrm{Softmax}\left(\frac{z_t}{\tau_t}\right)\middle\|\mathrm{Softmax}\left(\frac{z_s}{\tau_t}\right)\right) \tag{7}$$

式中,q_t为教师模型的平滑输出;p_t为学生模型的平滑输出. 这种平滑蒸馏损失促使学生模型将其输出与教师模型的输出对齐,从而更有效地捕捉不同类别之间的关系. 教师模型和学生模型使用相同的温度,可以确保输出分布的一致性,从而提升知识迁移的效果.

DP3SD的总体损失函数是稀疏分类损失和平滑蒸馏损失的加权组合,定义为

$$\mathcal{L}=\alpha\cdot\mathcal{L}_{\mathrm{CE}}\left(p_s,y\right)+\left(1-\alpha\right)\cdot\mathcal{L}_{\mathrm{KL}}\left(q_t\|p_t\right) \tag{8}$$

式中,α为一个平衡参数,控制分类损失和蒸馏损失之间的权衡. 本质上,交叉熵损失保证了学生模型在训练过程中的基础完整性,而自蒸馏损失缓解了教师与学生之间的经验差距,确保学生能够捕捉到来自教师的有效知识,从而提升差分隐私深度学习模型的效用.

需要注意的是,传统知识蒸馏方法通常仅在蒸馏阶段引入统一温度参数,以调节教师模型输出的软标签平滑性. 而本文提出的双温度缩放机制,在差分隐私场景中实现了针对分类与蒸馏目标的差异化调控机制,有效提升模型性能与稳定性. 一方面,较低的温度参数τ_s应用于分类损失,有助于强化学生模型对高置信度类别的响应能力,从而抑制由隐私噪声引起的低概率预测干扰;另一方面,较高的温度参数τ_t用于蒸馏损失,可有效平滑教师与学生模型的输出分布,增强模型对类间结构关系的建模能力. 在差分隐私后处理不变性的保障下,该双温度机制无需额外隐私开销,即可实现对监督信号形态的自适应调控,从而提升模型在隐私约束下的训练稳定性与表达能力.

3.2　基于稀疏平滑自蒸馏的差分隐私方法

在DPSGD及其变体中,通常假设迭代训练过程的每一步都是公开的,这使得攻击者可以利用所有中间检查点进行潜在攻击. 同时这一假设有助于对整体机制进行必要的隐私预算评估. 然而,由于通常只有最终模型用于预测,这就引发了一个问题:是否可以利用中间模型来提高效用. 最近的研究[21,22]表明:利用中间检查点可以在不损害隐私保障的前提下,提高差分隐私训练模型的效用. 这些方法通过聚合检查点的参数或输出,以提高最终模型的准确性. 与之前的方法不同,本文主要关注差分隐私的后处理特性,同时在不使用公开数据的情况下,改进差分隐私训练算法. 本文提出

了一种替代方法,使用中间检查点作为辅助教师模型,从而提高预测准确性. DP3SD的方法采用自蒸馏框架,其中学生模型通过DPSGD进行训练来提供隐私保障. 具体来说,本文首先通过DPSGD训练学生模型,获得一个预训练模型. 然后,教师模型动态地从当前学生模型的前一次迭代结果中选取,从而实现在训练过程中提供额外的知识. 与标准的知识蒸馏方法不同,DP3SD方法使用动态更新的教师模型,而非静态教师模型. 随着训练的进行,教师模型通过学生模型的中间模型来不断更新. 在每次迭代中,教师模型生成软标签,这些标签用于指导学生模型的训练. 本文的稀疏分类损失函数鼓励学生模型优先关注其更有信心的预测,有效地减轻了低置信度预测带来的噪声影响. 平滑蒸馏损失确保教师模型的预测输出与学生模型的预测输出对齐,从而进行更有效的知识迁移. 这种动态变化的教师-学生模式避免了对公开数据的依赖. 同时利用训练过程中生成的中间检查点和差分隐私的后处理特性,本文的教师模型并没有增加额外的隐私消耗. 通过将稀疏性和平滑性的优势结合起来,DP3SD方法不仅帮助学生模型减少噪声的影响,还让它能够有效地从教师模型中学习知识,从而提升模型的性能. 算法1对该优化过程进行了全面的概述.

算法1 基于差分隐私的稀疏平滑自蒸馏(DP3SD)

输入:数据集D,学习率η,噪声比例σ,梯度裁剪阈值C,温度参数τ_t,τ_s,训练轮次K,批量大小B

输出:差分隐私模型参数θ_s

初始化 随机初始化学生模型参数θ_s

FOR k = 1 to K do

设置教师模型参数 $\theta_T \leftarrow \theta_S^{k-1}$

FOR 数据集D中的每个小批量B

计算学生模型的稀疏预测分布

$$p_S(x_i)=f(\theta_S,x_i,\tau_s),\ \forall x_i\in B$$

计算教师模型的平滑预测分布

$$p_T(x_i)=f(\theta_T,x_i,\tau_t),\ \forall x_i\in B$$

计算学生模型的平滑预测分布

$$p_S(x_i)=f(\theta_S,x_i,\tau_t),\ \forall x_i\in B$$

计算交叉熵损失 $\mathcal{L}_{\mathrm{CE}}(p_s,y)$

计算蒸馏损失 $\mathcal{L}_{\mathrm{KL}}(q_t,p_t)$

计算总损失 $\mathcal{L}=\alpha\mathcal{L}_{\mathrm{CE}}+(1-\alpha)\mathcal{L}_{\mathrm{KL}}$

裁剪梯度 $\widetilde{g}_i=\dfrac{\nabla_{\theta\mathcal{L}_{\mathrm{total}}}}{\max\left(1,\left|\left|\nabla_{\theta\mathcal{L}_{\mathrm{total}}}\right|\right|_2/C\right)},i\in B$

添加噪声并更新参数 $\theta_S^k\leftarrow\theta_S^k-\eta\left(\dfrac{1}{B}\sum_{i\in B}\widetilde{g}_i+N\left(0,\sigma^2C^2I\right)\right)$

END FOR

END FOR

RETURN θ_S^k

3.3 隐私分析

在DPSGD及其变体的隐私分析中,(ϵ,δ)隐私预算是基于高斯机制的分析得出的,该分析包含了通过子采样实现的隐私放大效应以及多次迭代中的组合效应. 需要注意的是,隐私成本是在训练过程中生成的所有中间模型上累积的. 这意味着每个中间模型都会贡献到整体隐私预算中,无论是否只发布最终模型. 本文提出方法的隐私保证源自DPSGD的差分隐私属性. 由于学生模型是通过DPSGD进行训练的,基于差分隐私的后处理特性,使用中间模型作为动态教师不会增加额外的隐私成本,因此整个训练过程遵循(ϵ,δ)-差分隐私框架.

定理1 存在常数c_1和c_2,使得在给定采样概率$q=\dfrac{B}{N}$和迭代次数T的情况下,当噪声尺度σ满足以下条件时,算法1在任意$\epsilon\leqslant c_1q^2T$和$\delta>0$的情况下实现$(\epsilon,\delta)$差分隐私:

$$\sigma\geqslant c_2\frac{q\sqrt{T\ln(1/\delta)}}{\epsilon} \tag{9}$$

由于DPSGD训练得到的模型参数已经满足差分隐私保证,根据差分隐私的后处理性质,本文提出的DP3SD方法的最终输出同样满足差分隐私要求. 此外,近期的研究[23,24]探讨了在训练过程中不发布中间模型的隐私优势. 这些研究表明:发布DPSGD训练过程中的中间模型可能会浪费部分隐私成本,从而导致隐私与准确性之间的权衡不理想. 本文的自蒸馏方法的有效性支持了这一假设,表明可能存在更高效利用隐私预算的方式.

4 实验

为了验证本文提出的DP3SD方法的有效性,本文在三个图像分类数据集上进行了全面的实验:混合国家标准与技术研究院数据库(Mixed National Institute of Standards and Technology,MNIST)[25]、时尚混合国家标准与技术研究院数据库(Fashion Mixed National Institute of Standards and Technology,FMNIST)[26]和CIFAR-10(Canadian Institute For Advanced Research-10)[27]. 本文重点评估了模型的效用和隐私保护性能,并将DP3SD方法与三种最新的技术进行了对比:DPSGD[7]、DPEMA(Differentially Private Exponential Moving Average)[22]和差分隐私知识蒸馏(Differentially Private Knowledge Distillation,DPKD)[28]. 由于本文的问题设定中不使用额外的公开数据集,对于需要使用公开数据集的教师集成的隐私聚合(Private Aggregation of Teacher Ensembles,PATE)[8]等方法,不进行比较. 同样,关注特征工程[29]或替代模型结构[30]的方法也未包括在本文的对比范围内. 本文实验的主要目标是评估DP3SD在保证差分隐

私的同时提升模型准确性的效果,以验证其作为隐私保护深度学习方法的优越性. 本文的实验基于Pytorch集成的Opacus仓库实现.

4.1 实验设置

MNIST和FMNIST是包含10个类别的数据集,每个数据集包含60 000张训练图像和10 000张测试图像. 这两个数据集的图像均为28 × 28的灰度图,其中MNIST专注于手写数字,而FMNIST则聚焦于时尚物品. CIFAR-10包含60 000张32 × 32的彩色图像,同样分为10个类别,其中50 000张用于训练,10 000张用于测试. 其类别包括飞机、汽车和动物等.

本文提出的DP3SD方法作为一种基于模型输出的损失函数调控机制,不依赖具体网络结构设计,可广泛适用于各类深度学习分类模型. 本文采用了Opacus库中的网络架构,对于MNIST和FMNIST数据集,使用的模型由2个卷积层和2个全连接层组成. 对于CIFAR-10数据集,本文采用包含4个卷积层和1个全连接层的模型.

文中B表示批量大小,C表示裁剪阈值,η表示学习率,τ表示温度参数. 根据研究表明,较小的裁剪阈值通常效果最佳,因此本文在所有数据集中均将C设置为0.1. 对于MNIST数据集,参数设置为B=1 200,η=0.8,τ_s=0.1,τ_t=5,α=0.1. 在非隐私训练中,模型经过60轮训练后达到了0.992的准确率. 对于FMNIST数据集,参数设置为B=1 600,η=3,τ_s=0.3,τ_t=5,α=0.3. 模型经过60轮训练后达到了0.898的准确率. 对于CIFAR-10数据集,参数设置为B=1 000,η=3,τ_s=0. 1,τ_t=5,α=0.3. 模型经过100轮训练后达到了0.825的准确率.

DPSGD算法[7]是将差分隐私引入深度学习的一种基础方法,通过对梯度进行扰动来保障隐私. DPEMA算法[22]则通过聚合DPSGD的中间检查点参数,并利用聚合后的参数进行推断. DPKD算法[28]通过知识蒸馏技术实现差分隐私保护. 它采用两步过程:首先,训练一个具有差分隐私保证的教师模型;然后,将教师模型中的知识蒸馏到学生模型中.

4.2 实验结果

为了展示模型准确性与隐私之间的权衡,表1显示了模型准确性随隐私成本变化的情况. 如表1所示,DP3SD方法在相同的隐私预算下始终表现出更高的准确性. 例如,在CIFAR-10数据集上,当隐私水平 $\epsilon=3$ 时,本文提出的DP3SD方法达到了63.34%的准确率. 相比之下,DPSGD的准确率为59.99%,下降了3.35个百分点. DPKD为56.19%,降低了7.15个百分点,而DPEMA为61.62%,减少了1.72个百分点.

这些结果突出了DP3SD框架的主要优势. 通过将自蒸馏与差分隐私结合,DP3SD利用历史模型检查点作为教师,实现了更高效的知识迁移. 该方法能够在提供严格隐私保障的同时,保留更多有用的信息,从而在相同的隐私预算下提升模型的准确性. 与传统知识蒸馏方法需要单独的教师模型并消耗额外隐私预算不同,DP3SD在无需增加隐私开销的情况下提高了模型的可用性.

表1　在不同隐私预算下,不同数据集上的准确率对比　单位:%

数据集	方法	$\epsilon=1$	$\epsilon=2$	$\epsilon=3$
MNIST	DPSGD	95.51	96.46	97.01
	DPKD	94.31	95.59	96.88
	DPEMA	96.63	97.51	97.93
	DP3SD	**97.21**	**97.95**	**98.42**
FMNIST	DPSGD	81.29	85.28	86.13
	DPKD	79.58	84.66	85.85
	DPEMA	82.54	86.06	86.91
	DP3SD	**83.22**	**86.53**	**87.68**
CIFAR-10	DPSGD	48.47	55.05	59.99
	DPKD	46.91	52.58	56.19
	DPEMA	51.68	58.51	61.62
	DP3SD	**52.36**	**59.87**	**63.34**

注:加粗字体为最佳结果.

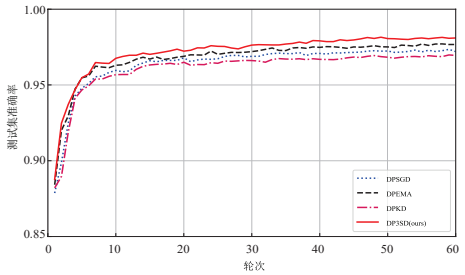
为了更直观地展示模型在训练过程中的性能表现,实验绘制了测试集准确率随训练轮次变化的曲线,如图3所示. 实验结果表明:DP3SD不仅在准确率上优于DPSGD、DPEMA和DPKD,还能够更快地达到较高的准确率. 将中间检查点用作辅助教师模型的设计,有效地引导了学习过程,使模型能够更快收敛,在较少的训练轮次内达到较高的准确率.

接下来,本文分析DP3SD方法相较于其他方法实现更高准确率的原因. DP3SD通过利用中间检查点作为辅助教师模型提供的额外知识,能够更有效地引导学习过程,从而优于DPSGD. 此外,DP3SD通过其双温度机制引入稀疏性,使模型能够专注于高概率类别,同时忽略可能是噪声引起的低概率类比,从而提升了模型的效用. 与DPKD不同,DP3SD方法无需在私有数据集上预训练教师模型,避免了额外的隐私成本. 相较于DPEMA,DP3SD在整个训练过程中自适应地从中间检查点中学习,不仅提升了收敛速度,还增强了模型的效用.

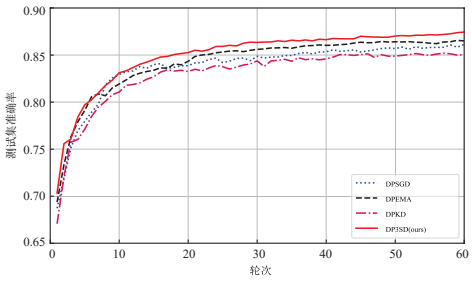
4.3 参数影响

本节重点评估DP3SD算法中各参数对其性能的影响. 这些参数包括平衡系数α,噪声系数σ,教师温度参数τ_t,学生温度参数τ_s. 在CIFAR-10数据集上展示DP3SD方法的实验结果.

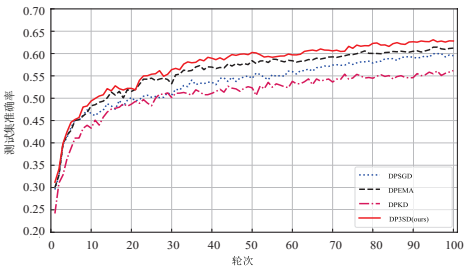
如图4(a)所示,平衡系数α在0.1~0.6时保持相对稳定,并在0.3时达到峰值. 较低的平衡系数会导致性能不佳,而较高的平衡系数也会导致准确率下降,这表明精心调节平衡系数对于实现模型的最佳性能非常重要.



(*a*) Mnist数据集



(*b*) Fashion Mnist数据集



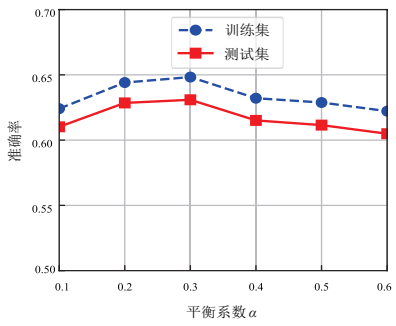
(*c*) CIFAR-10 数据集

图3 测试集准确率随训练轮次的变化趋势

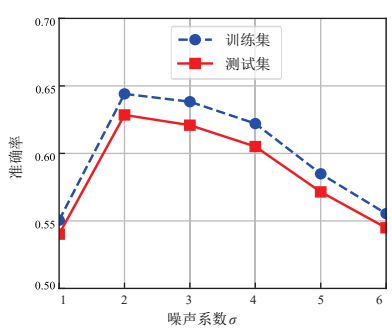
如图4(*b*)所示,噪声规模σ的选择对模型准确率具有显著影响. 噪声系数σ决定了每一步更新时添加的高斯噪声量. 较小的σ会降低噪声的影响,但同时也意味着训练步数减少,使得模型收敛更加困难;而较大的σ允许进行更多训练步数,但过大的噪声可能会损害模型的整体性能.

温度参数τ_t影响教师模型在自蒸馏过程中输出的平滑性. 当τ_t值在2~7时进行了实验,如图4(*c*)所示. 较高的τ值会产生更平滑的输出分布,有助于在差分隐私约束下实现更高效的知识迁移. 当$\tau_t=5$时,模型表现最佳,此时在过度平滑处和欠平滑之间达到了良好的平衡.

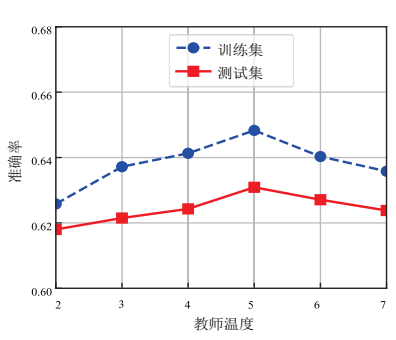
温度参数τ_s是控制模型输出分布稀疏性的重要超参数. 本文在实验中尝试了τ_s取值范围为0.01~0.4,如图4(*d*)所示. 较低的温度会增加输出的稀疏性,突出高概率类别,并减少由噪声引起的低概率预测的影响. 实验结果表明:当$\tau_s=0.1$时,模型性能最佳,能够在稀疏性和稳定性之间实现良好的平衡.



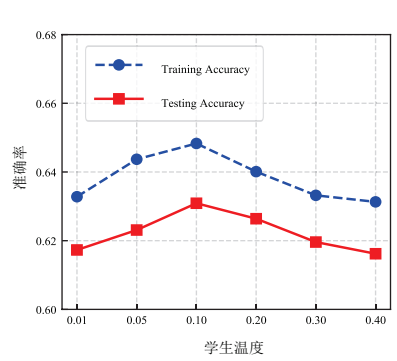
(*a*) 当平衡系数α变化时



(*b*) 当噪声系数σ变化时



(*c*) 当教师温度变化时



(*d*) 当学生温度变化时

图4 当一个参数变化、其他参数固定在参考值时的CIFAR-10数据集准确率

4.4 消融实验

为了更深入地理解本文提出的DP3SD方法中各个

组件的有效性,本文在FashionMnist数据集上进行了消融实验,系统分析了稀疏分类损失和平滑蒸馏损失对模型性能的具体影响. 稀疏组件通过在交叉熵损失中引入较低温度$\tau_s<1$来锐化学生模型的输出概率分布,从而增强对高置信度类别的关注,抑制由差分隐私噪声引起的低概率扰动. 这一设计有助于提升模型的鲁棒性与稳定性. 平滑组件则在蒸馏损失中引入较高温度$\tau_t>1$,使教师与学生模型的输出分布更加平滑,从而凸显类别间的相对关系,提升学生模型对类间结构信息的建模能力,增强泛化性能.

为了评估这些组件的重要性,本文系统地去除了损失函数中的稀疏或平滑损失,衡量模型性能的变化,结果如表2所示. 消融实验结果表明:稀疏性和平滑性均对提升性能具有积极作用,且二者协同效果最优. 其中,去除平滑性所造成的性能下降更为明显,表明其在自蒸馏中起到了更关键的知识迁移作用. 总体而言,消融实验结果表明:在差分隐私约束下,稀疏和平滑组件对提升模型性能都至关重要.

表2　损失函数不同组件的影响

稀疏性	平滑性	准确率
√	√	0.869 3
×	√	0.864 2
√	×	0.862 6
×	×	0.861 3

5　结论

本文提出了一种融合差分隐私与双温度缩放自蒸馏技术的隐私保护框架,旨在实现模型效用与隐私保护的平衡. 该框架采用双重机制:一方面,学生模型通过差分隐私技术为敏感数据提供严格的理论隐私保障,并利用较低温度参数促进预测分布的稀疏化,有效抑制低概率类别的扰动,从而提升模型的鲁棒性;另一方面,通过中间检查点作为教师模型进行自蒸馏,采用较高温度参数来捕捉类别间的潜在关联,降低了差分隐私噪声对模型性能的影响. 大量实验结果表明:本文提出的方法能够使学生模型在保持高效用的同时实现强隐私保护. 未来的研究可以探索其在大模型方面的潜力,持续提升其综合性能.

参考文献

[1] SONG C Z, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 587-601.

[2] WANG L, THAKKAR O, MATHEWS R. Unintended memorization in large ASR models, and how to mitigate it[C]// ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 4655-4659.

[3] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[C]//Proceedings of the 30rd USENIX Security Symposium. Berkeley: USENIX Association, 2021: 2633-2650

[4] WANG L J, WANG J J, WAN J, et al. Property existence inference against generative models[C]//Proceedings of the 33rd USENIX Security Symposium. Berkeley: USENIX Association, 2024: 2423-2440.

[5] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography. Berlin: Springer, 2006: 265-284.

[6] LI Z, WU Y T, CHEN Y H, et al. Membership inference attacks against large vision-language models[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. New York: ACM, 2025: 98645-98674.

[7] 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. New York: ACM, 2016: 308-318.

[8] TANG X Y, PANDA A, SEHWAG V, et al. Differentially private image classification by learning priors from random processes[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: ACM, 2023: 35855-35877.

[9] NASRSAEED MAHLOUJIFAR M, MAHLOUJIFAR S, TANG X Y, et al. Effectively using public data in privacy preserving machine learning[C]//Proceedings of the 40th International Conference on Machine Learning. Cambridge: PMLR, 2023: 25718-25732.

[10] FURLANELLO T, LIPTON Z, TSCHANNEN M, et al. Born again neural networks[C]//Proceedings of Machine Learning Research. Cambridge: PMLR, 2018: 1607-1616.

[11] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[12] GENTRY C, HALEVI S. Implementing gentry's fully-homomorphic encryption scheme[C]//Advances in Cryptology-EUROCRYPT 2011. Berlin: Springer, 2011: 129-148.

[13] 刘艺璇, 陈红, 刘宇涵, 等. 联邦学习中的隐私保护技术[J]. 软件学报, 2022, 33(3): 1057-1092.
LIU Y X, CHEN H, LIU Y H, et al. Privacy-preserving techniques in federated learning[J]. Journal of Software, 2022, 33(3): 1057-1092. (in Chinese)

[14] MIRONOV I. Rényi differential privacy[C]//2017 IEEE 30th Computer Security Foundations Symposium. Piscataway: IEEE, 2017: 263-275.

[15] GOPI S, LEE Y T, WUTSCHITZ L. Numerical composition of differential privacy[C]//Proceedings of the 35th In-

ternational Conference on Neural Information Processing Systems. New York: ACM, 2021: 11631-11642.
[16] 方晨, 郭渊博, 王娜, 等. 基于生成对抗网络的差分隐私数据发布方法[J]. 电子学报, 2020, 48(10): 1983-1992.
FANG C, GUO Y B, WANG N, et al. Differential private data publishing method based on generative adversarial network[J]. Acta Electronica Sinica, 2020, 48(10): 1983-1992. (in Chinese)
[17] 康海燕, 王骁识. 基于数据特征相关性和自适应差分隐私的深度学习方法研究[J]. 电子学报, 2024, 52(6): 1963-1976.
KANG H Y, WANG X S. Research on the deep learning method based on data feature relevance and adaptive differential privacy[J]. Acta Electronica Sinica, 2024, 52(6): 1963-1976. (in Chinese)
[18] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2025-02-20]. https://arXiv.org/abs/1503.02531.
[19] 刘松, 罗杨宇, 许佳培, 等. 基于轻量自蒸馏的低成本联邦学习[J]. 电子学报, 2025, 53(1): 259-269.
LIU S, LUO Y Y, XU J P, et al. Low-cost federated learning based on lightweight self-distillation[J]. Acta Electronica Sinica, 2025, 53(1): 259-269. (in Chinese)
[20] KIM K, JI B, YOON D, et al. Self-knowledge distillation with progressive refinement of targets[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 6547-6556.
[21] SHEJWALKAR V, GANESH A, MATHEWS R, et al. Recycling scraps: Improving private learning by leveraging intermediate checkpoints[EB/OL]. (2024-09-17)[2025-02-20]. https://arxiv.org/abs/2210.01864.
[22] DE S, BERRADA L, HAYES J, et al. Unlocking high-accuracy differentially private image classification through scale[EB/OL]. (2022-06-16)[2025-02-20]. https://arXiv.org/abs/2204.13650.
[23] ALTSCHULER J M, TALWAR K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss[C]//Proceedings of the 36th Conference on Neural Information Processing Systems. San Diego: NeurIPS, 2022: 3788-3800.
[24] BOENISCH F, MÜHL C, DZIEDZIC A, et al. Have it your way: Individualized privacy assignment for DP-SGD[C]//Proceedings of the 37th Conference on Neural Information Processing Systems. San Diego: NeurIPS, 2024: 8-12.
[25] DENG L. The MNIST database of handwritten digit images for machine learning research [best of the web] [J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.
[26] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[EB/OL]. (2017-08-25)[2025-02-20]. https://arXiv.org/abs/1708.07747.
[27] KRIZHEVSKY A. Convolutional deep belief networks on CIFAR-10[EB/OL]. (2010-12-21)[2025-02-20]. https://www.cs.utoronto.ca/~kriz/conv-cifar10-aug2010.pdf.
[28] FLEMINGS J, ANNAVARAM M. Differentially private knowledge distillation via synthetic text generation[C]//Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 12957-12968.
[29] TRAMÈR F, BONEH D. Differentially private learning needs better features (or much more data)[EB/OL]. (2021-02-18)[2025-02-20]. https://arXiv.org/abs/2011.11660.
[30] PAPERNOT N, THAKURTA A, SONG S, et al. Tempered sigmoid activations for deep learning with differential privacy[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(10): 9312-9321.

作者简介

赵登峰 男,1992年6月出生于河南省驻马店市. 现为中国人民大学信息学院博士研究生. 主要研究方向为深度学习、隐私保护.
E-mail: zhaodengfenglnu@163.com

赵素云 女,1979年9月出生于河北省石家庄市. 现为中国人民大学信息学院教授、博士生导师. 主要研究方向为人工智能、机器学习.
E-mail: zhaosuyun@ruc.edu.cn

薛大暄 男,1994年11月出生于陕西省西安市. 现为中国人民大学信息学院博士研究生. 主要研究方向为深度学习、隐私保护.
E-mail: 2021000907@ruc.edu.cn

陈 红 女,1965年5月出生于江西省上饶市. 现为中国人民大学信息学院教授、博士生导师. 主要研究方向为大数据、隐私保护.
E-mail: chong@ruc.edu.cn