

差分隐私 GAN 梯度裁剪阈值的自适应选取方法

郭鹏^{1,2}, 钟尚平^{1,2}, 陈开志^{1,2}, 程航^{1,2}

(1. 福州大学数学与计算机科学学院, 福建 福州 350116;
2. 网络系统信息安全福建省高校重点实验室, 福建 福州 350116)

摘要: 提出自适应选取差分隐私 GAN 梯度裁剪阈值的方法。该方法假设可以接触到与隐私数据同分布的小部分公开数据, 通过从公开数据中随机选取一批数据, 设置裁剪阈值为这批数据的平均梯度范数, 迭代上述操作直到网络聚合。在 Mnist 和 Cifar10 数据集上对所提方法进行了实验验证, 结果表明, 在合理隐私预算下与差分隐私辅助分类 GAN 相比, 卷积神经网络(CNN)分类器准确率提高 1%~4%, 而评估分数(inception scores)提升 0.6~1.2。

关键词: 生成对抗网络; 差分隐私保护; 梯度裁剪阈值; 自适应选取

中图分类号: TP309

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018041

Adaptive selection method of differential privacy GAN gradient clipping thresholds

GUO Peng^{1,2}, ZHONG Shangping^{1,2}, CHEN Kaizhi^{1,2}, CHENG Hang^{1,2}

1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
2. Network System Information Security Fujian Provincial University Key Laboratory, Fuzhou 350116, China

Abstract: A method of adaptive selection of differential privacy GAN gradient clipping threshold was proposed. The method assumes that a small portion of public data that is identically distributed with the private data can be accessed, a batch of data is randomly selected from the public data, a clipping threshold is set as an average gradient norm of the batch of data, and the above operations are iterated until the network converges. The method was verified on the Mnist and Cifar10 data sets. The results show that under a reasonable privacy budget, the accuracy of CNN classifiers is improved by 1%~4% compared with the differential privacy auxiliary classifier GAN, inception scores increased by 0.6~1.2.

Key words: GAN, differential privacy protection, gradient clipping thresholds, adaptive selection

1 引言

近年来, 随着计算机硬件和软件技术的快速发展, 大量关于个人语义丰富的数据(如图像、

文本、音频)正在被收集。尽管分析和利用这些数据带来了巨大的商业价值(如有针对性的广告和个性化推荐), 但是, 政府、组织部门都已经认识到在这种实践中保护个人隐私的重要性^[1]。

收稿日期: 2018-03-15; 修回日期: 2018-04-02

通信作者: 郭鹏, n160320070@fzu.edu.cn

基金项目: 福建省自然科学基金资助项目(No.2017J01502); 福州大学博士科研基金资助项目(No.XRC-17015)

Foundation Items: The Natural Science Foundation of Fujian Province (No.2017J01502), The Scientific Research Foundation of Fuzhou University (No.XRC-17015)

数据的隐私保护主要解决两方面问题^[2]：数据集不能泄露个人的隐私，以及使数据的可利用性和隐私性达到平衡。Mohassel 等^[3]提出数据隐私保护的双服务器协议，数据管理员结合安全双方计算（2PC, two-party computation）与同态加密技术利用数据集训练多种模型，但计算消耗较大，复杂度较高。Sweeney 等^[4]提出 k -anonymity 的目的是保证公开数据中包含的个人信息至少 $k-1$ 条不能通过其他个人信息确定出来，但是目前有多种根据背景知识针对匿名保护的攻击，并且有较高成功率； l -diversity^[5]、 t -closeness^[6]等方法是针对背景知识攻击以及对 k -anonymity 的改进方法。尽管以上方法一定程度保护数据隐私，但是计算复杂，不容易部署实施，属性缺失造成数据可利用性降低，或者因为背景知识容易被攻击等问题对上述方法提出挑战。

Dwork^[7]提出的差分隐私具有优雅的理论证明，自定义的“隐私预算”来调整数据的实用性和隐私性，差分隐私与机器学习的结合是隐私保护的热点。Shokri 等^[1]提出分布训练，使用互斥数据集，对每一个分布的模型参数添加噪声处理后，上传到中心服务器处，其他分布者再从中心处下载平均化处理后的参数，更新自己的模型参数，但 Abadi^[8]和 Papernot^[9]等认为这种方法会产生很高的隐私损失；Papernot 等^[9]应用独立的教师模型集成器和半监督知识迁移训练的学生模型对 Mnist 和 Svhn 数据集实现最好的分类准确率，但是实施这种方法的前提是确保对手不能接触到教师集成器和聚合结果，并且允许学生模型对教师集成器的查询次数尽可能少，因为查询会增加隐私损失；Abadi 等^[8]提出使用差分隐私随机梯度下降，对深度学习的优化过程实施隐私保护，在保证较低差分隐私预算同时不要求分布式训练，但是这种方法只是发布固定任务的模型，第三方无法根据实际任务训练自己的模型。

因此，为了保护数据集的个人隐私，同时平衡数据的实用性和隐私性而满足第三方的需求，已有研究者将差分隐私和生成模型结合，实现差分隐私保护的生成模型，第三方可以依靠生成模型无限制生成满足源数据统计特性的“虚假”数

据。Gergely 等^[10]提出了新颖的隐私保护生成模型，将 k 个生成神经网络混合一起学习训练数据的分布，构建生成器模型。先使用差分隐私核 k -means 将数据分成 k 类，使用单个生成神经网络（如限制玻尔兹曼机或者变分自编码器）学习每一类训练数据分布，学习过程中利用了 Abadi 等^[8]的差分隐私优化过程方法实验差分隐私保护，但是 VAE 是通过直接计算生成图像和原始图像的均方误差学习，使生成的图像模糊；Beaulieu-Jones 等^[11]提出差分隐私辅助分类生成对抗网络（DPACGAN, differential privacy auxiliary classifier generative adversarial network）生成医疗临床数据。这是第一篇实现对生成对抗网络差分隐私保护的文章。DPACGAN 和 Gergely 等^[10]方法相比具有很多优点。首先，GAN 确保可扩展性，更适用于现实生活中复杂结构的数据；其次，这种方案是比较容易实施的，即不需要在互斥数据集上分布式训练模型后再聚合；最后，DPACGAN 借鉴了 Abadi^[8]的隐私保护结构，即通过设定梯度裁剪阈值，限制每个样例的敏感度，然后对裁剪处理后的梯度添加合适的噪声，再来更新对应的网络参数。DPACGAN 具有完整的理论证明，可以根据标签生成需要的图像。但是，DPACGAN 直接固定梯度裁剪阈值 C ，这会影响优化器性能，原因为：1) 不同层的梯度相差很大；2) 梯度值是持续改变的。Zhang 等^[12]提出梯度裁剪自适应选取方法，先将权重等参数进行归类操作，在训练前和训练过程中监测梯度值的变化，再对每一类设置梯度阈值。

本文借鉴 Zhang 等^[12]思想，以 DPACGAN 框架为基础，假设除了可以接触到隐私数据 D_{pri} 来训练模型，还可以接触到小部分的公开数据集 D_{pub} 。在每一步训练中，笔者随机从 D_{pub} 抽取一批样例，设置训练模型的梯度裁剪阈值为这批数据的平均梯度范数，迭代上述操作直到网络聚合。在 Mnist 和 Cifar10 数据集上对本文方法进行了实验验证，结果表明，合理隐私预算下，和 DPACGAN 相比，使用生成图像训练得到的卷积神经网络（CNN）分类器准确率提高 1%~4%，评估分数（Inception Scores）提升 0.6~1.2。

本文的主要贡献如下。

1) 发现原始 DPACGAN 固定梯度阈值的弊端, 提出结合 C 自适应选取的方法。直观地说, 对每一个输入, 梯度值的 L_2 范数应该近似等于裁剪阈值, 这保证了输入对梯度的贡献最大并且相关误差最小。

2) 验证了改进后的 DPACGAN 在不同复杂度数据集上的实用性, 在合理的隐私预算下, CNN 分类器分类准确率提高, 生成的图像更清晰和多样性更好。

2 相关技术概述

2.1 差分隐私理论基础

本文方法使用的隐私模型是差分隐私^[7,13,14], 差分隐私被认为是最强的隐私保护之一。

2.1.1 差分隐私定义

定义 1 给定 2 个至多相差一条记录的数据集 d 和 d' (即 $|d \Delta d'| \leq 1$), 对于一个设定的随机算法 A , 其取值范围为 $Range(A)$, 若算法 A 在数据集 d 和 d' 上的任意输出结果 $S(S \in range(A))$ 满足式(1), 则称算法 A 满足 (ϵ, δ) -差分隐私。

$$\Pr[A(d) \in S] \leq e^\epsilon \times \Pr[A(d') \in S] + \delta \quad (1)$$

其中, ϵ 是隐私保护参数, 表示隐私保护的程度, ϵ 越小意味着对于真实输出的扰动越大, 隐私保护程度也越高。 δ 表示严格的差分隐私下隐私泄露的概率。

2.1.2 差分隐私实现方法

1) 函数的敏感度

给定函数 f , 通过对 f 输出添加随机噪声实现差分隐私, 噪声的大小取决于 f 输出的敏感度。如果 f 输出是向量值, 它的敏感度大小定义为

$$\Delta f = \max_{d, d'} \|f(d) - f(d')\| \quad (2)$$

表示单个数据对 f 输出的最大影响, 量化评估最坏情况下隐藏单个数据存在要添加的不确定性。

2) 高斯机制

如果 f 的敏感度使用 L_2 范式定义, 高斯机制常被用来随机化 f 的输出。高斯机制 (式(3)) 是给 f 的真实输出添加高斯噪声^[14]。

$$A(d) = f(d) + N(0, (\Delta f \sigma)^2 I) \quad (3)$$

其中, 高斯噪声是均值为 0、协方差为 $(\Delta f \sigma)^2 I$ 的

高斯分布, I 是单位矩阵。

3) 属性

在差分隐私对抗网络中利用差分隐私的以下特性。

① 后处理性。对差分隐私机制下输出的任何计算不会增加隐私损失。

② 序列化组合原理。差分隐私机制的序列化组成依然满足差分隐私保护。

2.1.3 差分隐私优化算法

本文的隐私保护方法和差分隐私深度学习^[8]的技术很相似, 通过在优化程序中添加噪声实现。算法 1 说明了差分隐私深度学习中通过最小化经验损失函数 $L(\theta)$ 训练关于参数 θ 模型的基本方法。在优化程序的每一步, 计算随机样例的梯度 $g(x)$, 裁剪每个梯度 L_2 范数, 计算平均值, 添加噪声, 如步骤 5)~步骤 7), 再使用平均梯度更新对应参数。这里对梯度的裁剪是为了限制单个数据对整体的影响, 也是方便敏感度的计算。添加噪声是实现了对梯度输出的扰动, 实现保护。

算法 1 差分隐私优化算法

输入 训练样例 $X = \{x_1, x_2, \dots, x_n\}$, 损失函数

$$L(\theta) = \frac{1}{N} \sum L(\theta, x_i), \text{ 噪声 } \sigma, \text{ 学习率 } \alpha, \text{ 组大小}$$

m , 梯度裁剪阈值 C , 训练轮数 T

输出 θ , 并且使用隐私统计方法计算全局隐私损失

1) 初始化 θ

2) for $t \in T$ do

3) 随机选取 m 个样例

4) for $x_i \in m_x$ do

5) $g(x_i) \leftarrow \nabla_{\theta} L(\theta, x_i)$ // 计算梯度值

6) $g(x_i) \leftarrow \frac{g(x_i)}{\max\left(1, \frac{\|g(x_i)\|}{C}\right)}$ // 梯度裁剪

7) $g(x_i) \leftarrow g(x_i) + N(0, \sigma^2 C^2 I)$ // 添加噪声

8) $\bar{g} = \frac{1}{m} \sum g(x_i)$

9) $\theta = \theta - \alpha \bar{g}$ // 更新参数

10) End (for)

11) End (for)

12) return θ

2.2 梯度裁剪阈值的自适应选取

差分隐私随机梯度下降将梯度裁剪阈值 C 作为输入, 没有给出解释如何计算该值。另外, C 的选择对优化器的性能影响较大。如果 C 太小, 会导致梯度的过度截断, 造成聚合缓慢。相反, 如果 C 太大, 较多不必要的噪声将被添加到梯度更新, 会导致生成图像质量不好。由于不同层权重, 偏置、梯度相差较大, 并且训练中梯度是持续改变的, 很难找到接近最优的固定设置 C 。直观地说, 对每个输入的样例 x , $\|g(x)\|_2 \approx C$, 这样保证了输入 x 对平均梯度的贡献具有最小相关误差的最大保护^[10]。

因此, 与其对整体训练集的 C 做固定输入, 不如自适应地选取每批的 C 。本文假设训练模型时, 除了可以接触到隐私数据 D_{pri} 还可以接触到小部分的公开数据 D_{pub} , 在每一次训练时, 随机从 D_{pub} 选取一批样例, 设置隐私数据的裁剪阈值为这一批训练样例梯度范式的平均值。这种自适应的方法确保具有较小误差的较快聚合。

2.3 生成对抗网络

如图 1 所示, 生成对抗网络^[15]同时训练 2 个模型: 生成模型 G 是学习将潜在分布 p_z 映射到训

练数据的真实分布 p_{data} , 同时判别器训练过程是计算输入样例来自真实训练数据的概率, 生成器的训练过程是最大化判别器犯错误的概率, 即判别器误以为样例为真实训练样例而不是生成器生成的“虚假”样例, 目标函数为

$$\min_G \max_D E_{X_{\text{real}} \sim p_{\text{data}}} [\log D(X_{\text{real}})] + E_{X_{\text{fake}} \sim p_{\text{data}}} [\log(1 - D(G(X_{\text{fake}})))] \quad (4)$$

X_{real} 表示从真实训练数据取样, X_{fake} 表示从“虚假”数据取样。

GAN 被用在各种无监督和半监督的学习任务。其中, 一个是将训练好的判别器作为特征提取器, 用在各种设置中; 另一个工作是针对潜在变量 X_{fake} , 要么使用正则化使 X_{fake} 语义有意义, 要么直接在潜在空间提取信息。尽管 GAN 理论简明, 但是原始 GAN 在训练中不稳定并且效率不高。

如图 2 所示, ACGAN^[16]是 GAN 的变形。ACGAN 除了输入对应噪声 z , 还有标签 $c \sim p_c$ 用来生成每个样例。G 使用两者生成图像 $X_{\text{fake}} = G(c, z)$ 。判别器判别输入图像的真假源概率分布和类别标签的分布。目标函数有式(5)和式(6)两部分。

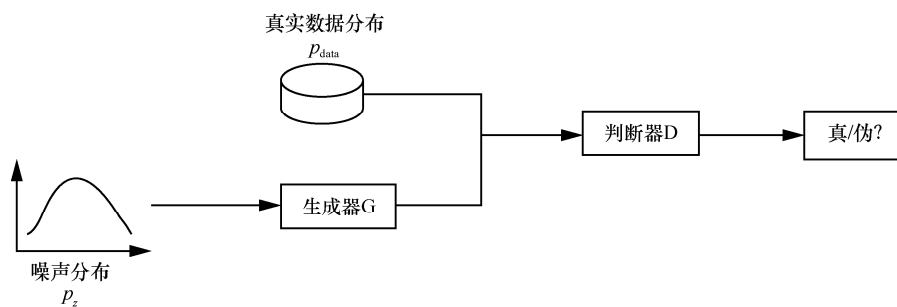


图1 生成对抗网络结构

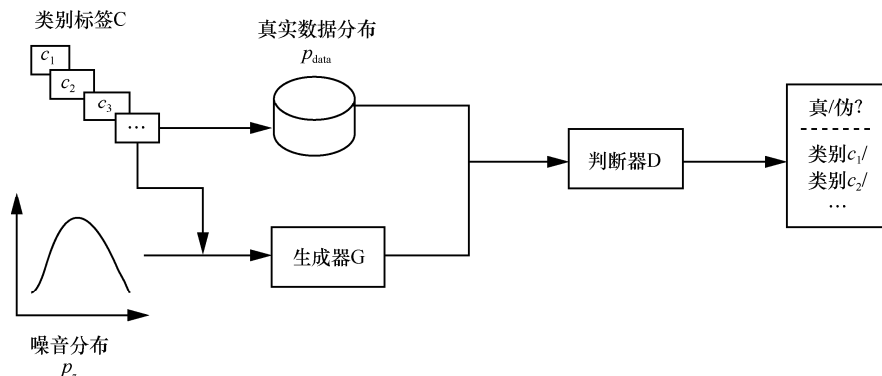


图2 辅助分类生成对抗网络

$$L_s = E[\log P(S = \text{real} | X_{\text{real}})] + E[\log P(S = \text{fake} | X_{\text{fake}})] \quad (5)$$

$$L_c = E[\log P(C = \text{real} | X_{\text{real}})] + E[\log P(C = \text{fake} | X_{\text{fake}})] \quad (6)$$

其中, L_s 表示输入数据的真实来源, L_c 表示输入数据正确的标签。 D 被训练最大化 $L_s + L_c$, 同时 G 最大化 $L_c + L_s$ 。ACGAN 和标准 GAN 相比, 产生极好的效果并且训练更稳定。早期实验表明, 增加训练集类别数量会降低模型输出的质量。ACGAN 的模型结构允许将训练集根据类别划分多个子集, 使用每个子集训练生成器和分类器。

2.4 差分隐私生成对抗网络

DPACGAN 是^[11]在优化程序中添加随机噪声实现对训练数据的隐私保护。本文尝试仅对判别器 D 添加随机扰动, 合理性如下: 首先, 在对抗网络中, 只有 D 可直接接触训练真实数据, 这样足够控制训练 D 的隐私损失; 其次, 和生成器 G 比较, D 仅需要较简单的结构和小部分参数, 这样比较容易估计隐私损失。但是, 通过对 D 实施隐私保护, 生成器的参数依然是差分隐私保护。这是因为差分隐私的后处理性, 也就是说, 对差分隐私输出的任何计算不会增加隐私损失。这里的计算是指对生成器参数的计算, 输出指判别器输出的差分隐私保护参数, 因此训练后的生成器用来生成数据是安全的。

2.5 隐私分析

差分隐私生成对抗网络的关键组成部分是计算训练期间的逆累积隐私损失, 即隐私统计 A , 包含 2 个部分: 时刻统计^[8]和子集采样^[8]。

时刻统计: Abadi 等提出了时刻统计——一个隐私统计的方法, 该方法比组合原理^[17]对隐私损失的评估更严谨。随机变量 Z 表示隐私损失 (式(7))。

$$Z(o, M, d, d') = \log \frac{\Pr[M(d) = o]}{\Pr[M(d') = o]} \quad (7)$$

其中, d, d' 表示临近数据集, M 是随机机制算法, $o \in R$ 表示输出。隐私损失随机变量的时刻生成函数日志为

$$\alpha_M(\lambda, d, d') = \log E_{o \sim M(d)} [\exp(\lambda Z(o, M, d, d'))] \quad (8)$$

为了满足差分隐私, 需要遍历所有可能的 d, d' , 得到 α_M 。

$$\alpha_M \triangleq \max_{d, d'} (\lambda, d, d') \quad (9)$$

使用马尔可夫不等式, 可以证明任何 $\varepsilon > 0$, 对于 $\delta = \min_{\lambda} (\alpha_M - \lambda \varepsilon)$, M 满足 $(\varepsilon, \delta) - DP$ 。

子集采样: 在每次 D 的迭代训练中, 从真实数据集采样一批数据。由隐私应用定理可知, 每次采样程序关于整个数据集实现 $(O(q\varepsilon, q\delta) - DP)$, q 表示每批的采样概率 $q = \frac{m}{n}$, m 表示批大小, n 表

示训练集大小, $\sigma = \frac{\sqrt{2 \log \left(\frac{1.25}{\delta} \right)}}{\varepsilon}$, $\varepsilon \in (0, 1)$ 。

如果噪声 σ 和梯度裁剪阈值 C 选择合适, 使用时刻统计^[7]可以证明算法 1 是 $(O(q\varepsilon\sqrt{t}, \delta) - DP)$, t 表示主循环中总的迭代轮数。

3 梯度裁剪阈值自适应选取和 DPACGAN 的结合

DPACGAN 因其更适用于生活中复杂数据结构, 更简单的实施方案, 更能保证数据利用和数据隐私之间达到平衡的优点而引起数据管理者的青睐。在差分隐私层面提供隐私保护控制的 GAN 可以生成和原始真实数据同分布的“虚假”数据, 在允许第三方利用训练数据实现不同机器学习任务的同时, 保护数据的隐私性, 满足数据分析师的需求, 引导长远的研究。

然而, 笔者发现在隐私保护的方法中, 直接固定裁剪阈值 C 会造成生成数据的可利用性降低。因为 GAN 训练中梯度是不断变化的, 每层相差较大。而如果 C 太小, 会导致梯度的过度裁剪, 造成聚合缓慢。相反, 如果 C 太大, 较多不必要的噪声将被添加到梯度更新, 会导致生成图像质量不好, 很难找到满足全局需求的最优的 C 作为输入。因此, 本文将 DPACGAN 与 2.2 节中方法结合, 保证在每一轮迭代中自适应地选取 C 。

3.1 自适应选取梯度阈值 DPACGAN

算法 2 列出了采用自适应选取梯度裁剪阈值的

DPACGAN 的结构。本文首先从 D_{pub} 选取一批数据, 计算输入样例的梯度, 然后使用该批的梯度平均值作为隐私训练的梯度裁剪阈值 (步骤 4)~步骤 5)), 再从真实数据取样, 计算梯度 (步骤 6)~步骤 7)), 对梯度裁剪和添加噪声, 更新判别器参数 (步骤 8)~步骤 12))。迭代这样的操作, 直到网络聚合或者隐私预算耗尽 (步骤 15)~步骤 16))。

算法 2 自适应选取阈值 DPACGAN 算法

输入: 数据集 $X=\{x_1, x_2, \dots, x_n\}$; D_{pub} 公共数据集, D_{pri} 隐私数据集, m 隐私数据集批大小, m_{pub} 公开数据集的批大小; D 判别器, G 生成器; $(\alpha, \beta_1, \beta_2)$ 是 Adam 的参数; C 裁剪阈值; σ 噪声大小; (ϵ_0, δ_0) -总隐私预算。

输出: G , 差分隐私生成器

```

1)  $(w, \theta) \leftarrow$  使用  $D_{\text{pub}}$  训练 ACGAN
2) While  $\theta$  没有聚合 do
3) while  $D$  do
4)    //从  $D_{\text{pub}}$  取样,  $\{x_i\}_{i=1}^{m_{\text{pub}}}$  类别标签
5)     $C \leftarrow \{g(x_i)\}_{i=1}^{m_{\text{pub}}} \leftarrow \nabla_{\theta} L(\theta, x_i)$  计算梯度, 求平均值得梯度裁剪阈值  $C$ 
6)    从  $D_{\text{pri}}$  取样  $\{x_i\}_{i=1}^m$ , 类别标签
7)     $\{g(x_i)\}_{i=1}^m \leftarrow \nabla_{\theta} L(\theta, x_i)$  //计算梯度
8)    for  $i=1, \dots, m$  do
9)        $g(x_i) \leftarrow \frac{g(x_i)}{\max(1, \|g(x_i)\|_2 / C)}$  //裁剪
10)     $g(x_i) \leftarrow g(x_i) + N(0, \sigma^2 C^2 I)$  //添加扰动噪声, 更新隐私统计
11)    //求批次平均梯度, 更新判别器参数
12) End (while)
13) //采样噪声, 类别标签, 更新生成器参数  $\theta$ 
14) //计算隐私算法 A 关于  $\epsilon_0$  的  $\delta$ 
15) If  $\delta > \delta_0$  结束训练
16) End while
17) return  $G$ 

```

3.2 算法性能分析

在算法 2 每次判别器的训练中, 本文增加了对 D_{pub} 的随机取样并且计算梯度值, 每次选取 m 个数据, 进行计算, 时间复杂度为 $O(m)$, 然后, 从 D_{pri}

选取 n 个数据计算梯度, 则时间复杂度 $O(n)$, 故一次迭代总复杂度为 $O(m+n)$ 。之后的每次迭代相同, 假设为 N 次, 则判别器层总的复杂度为 $NO(m+n)$ 。

4 实验

4.1 实验设置

本节评估改进后 DPACGAN 的性能。所有实验均是在 keras+tensorflow+gpu 平台上完成, python 版本为 3.5.2。此外使用了 2 块 NVIDIA GTX 1080Ti 显卡实现硬件加速。

本文实验使用 2 种标准数据集。

1) Mnist, 由 70 000 张 10 类手写数字图像组成, 包括 60 000 张训练数据集和标签, 10 000 张测试数据集和标签。每张图像大小 28×28 。

2) Cifar10, 由 60 000 张 10 类图像组成, 每类 6 000 张。包括 50 000 张训练数据集和标签, 10 000 张测试数据集和标签。每张图像 32×32 。

本文分别将 Mnist 和 Cifar10 数据集的训练集按 4:96 比例分为公共可利用数据集 D_{pub} 和隐私数据集 D_{pri} 。针对 Mnist 数据集的改进后 DPACGAN 超参数设置如图 3 所示, 针对 Cifar10 数据集的改进后 DPACGAN 超参数设置如图 4 所示。在判别器的优化中, 本文使用 NoisyAdam() 优化器, 实现隐私保护。

4.2 实验结果及分析

本节对生成的图像质量进行评估。笔者主要设计 2 个实验, 回答关于改进后 DPACGAN 的 3 个关键问题。第一, 在隐私约束条件下, 改进后 DPACGAN 是否可以生成视觉效果较好的图像? 第二, 生成的图像是否保持足够的能力实现数据分析任务? 第三, 生成的图像是否具有足够好的质量和多样性?

4.2.1 改进后 DPACGAN 生成的图像

本节主要是展示一些使用改进后方法生成的图像数据。图 5 和图 6 分别使用 Mnist 和 Cifar10 数据集训练的生成模型生成的图像数据, 可以看出, 改进后 DPACGAN 可以生成视觉上清晰的图像。

4.2.2 Mnist 数据集和 Cifar10 数据集的分类任务

笔者使用 Mnist 训练集在 $\delta = 10^{-5}$, $\epsilon = \left[\frac{1}{2}, 2, 4, 8 \right]$ 下训练 4 个不同隐私预算的生成模型, 再使用每一个模型生成数据, 对每一类数

Opt	units	kernel	strides	Feature maps	BN	Dropout	NonLinearity
生成器部分							
全连接层	128×7×7	—	—	—	—	—	Relu
Conv2DTranspose	—	5	1	256	是	—	Relu
UpSampling2D	—					—	—
Conv2DTranspose	—	5	1	128	是	—	Relu
UpSampling2D	—					—	—
Conv2DTranspose	—	2	1	1	是	—	Tanh
生成器优化器	Adam ($\alpha = 0.000\ 2, \beta_1 = 0.5, \beta_2 = 0.999$)						
判别器部分							
Cov2D	—	3	1	32	是	0.3	LeakyRelu
Cov2D	—	3	1	64	是	0.3	LeakyRelu
Cov2D	—	3	1	128	是	0.3	LeakyRelu
Cov2D	—	3	1	256	是	0.3	LeakyRelu
输出	1	—	—	—	—	—	Sigmoid
输出	10	—	—	—	—	—	Softmax
判别器优化器	NoisyAdam ($\alpha = 0.000\ 2, \beta_1 = 0.5, \beta_2 = 0.999$)						
批大小	100						

图 3 Mnist 数据集生成模型超参数设置

Opt	units	kernel	strides	Feature maps	BN	Dropout	NonLinearity
生成器部分							
全连接层	384×4×4	—	—	—	—	—	Relu
Conv2Dtranspose	—	5	2	192	是	—	Relu
Conv2Dtranspose	—	5	2	96	是	—	Relu
Conv2DTranspose	—	5	2	3	是	—	Tanh
生成器优化器	Adam ($\alpha = 0.000\ 1, \beta_1 = 0.5, \beta_2 = 0.999$)						
判别器部分							
Cov2D	—	3	2	16		0.5	LeakyRelu
Cov2D	—	3	1	32	是	0.5	LeakyRelu
Cov2D	—	3	2	64	是	0.5	LeakyRelu
Cov2D	—	3	1	128	是	0.5	LeakyRelu
Cov2D	—	3	2	256	是	0.5	LeakyRelu
Cov2D	—	3	1	512	是	0.5	LeakyRelu
输出	1	—	—	—	—	—	Sigmoid
输出	10	—	—	—	—	—	Softmax
判别器优化器	NoisyAdam ($\alpha = 0.000\ 1, \beta_1 = 0.5, \beta_2 = 0.999$)						
批大小	100						

图 4 Cifar10 数据集生成模型超参数设置

图 5 Mnist 生成模型生成图像 ($\epsilon=4$, $\delta=10^{-5}$)图 6 Cifar10 生成模型生成图像 ($\epsilon=8$, $\delta=10^{-5}$)

字随机挑选 3 000 张组成生成图像训练集, 则共得到 4 组训练集。每一组训练集训练得到一个 CNN 分类器, 再使用每一个训练好的分类器对 Mnist 测试集测试计算准确率。然后, 重复这样的操作 50 次, 计算平均准确率。作为对比, 使用 DPACGAN 设置在 $\delta=10^{-5}$, $\varepsilon=\left[\frac{1}{2}, 2, 4, 8\right]$ 下训

练生成模型, 同样操作计算平均准确率; 还使用真实 Mnist 训练集训练分类器, 应用测试集计算分类准确率; 使用无隐私保护 ACGAN 生成的数据训练分类器, 应用测试集计算准确率。实验结果如图 7 所示, 横坐标表示隐私预算大小, 纵坐标表示 CNN 的分类准确率, 每一部分前两列分别表示在固定隐私预算下, 改进前 DPACGAN 生成数据分类器准确率和改进后 DPACGAN 生成数据分类器准确率。虽然 ACGAN 数据和原始数据没有使用固定隐私预算, 但是作为对比, 分别放在每部分的后两列, 这样很容易得到对比效果。

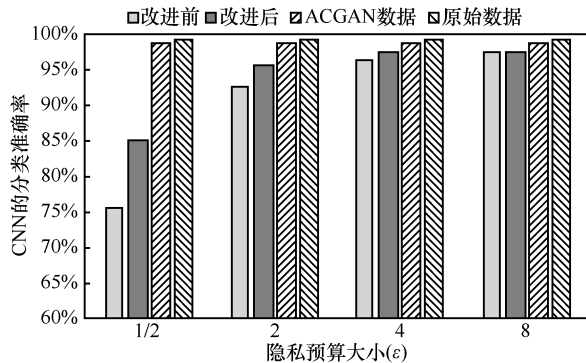


图7 Mnist 数据 CNN 分类准确率

笔者使用 Cifar10 数据集执行和 Mnist 数据集相同的操作。实验结果如图 8 所示, 每一部分前两列分别表示在固定隐私预算下, 改进前 DPACGAN 生成数据训练的分类器准确率和改进后 DPACGAN 生成数据训练的分类器准确率。虽然 ACGAN 数据和原始数据没有使用固定隐私预算, 但是作为对比, 分别放在每部分的后两列, 这样很容易得到对比效果。和 Mnist 数据集比较, Cifar10 数据集为彩色图像, 并且比较复杂, 通过自适应选取切割阈值发现, 生成数据的准确率提升更加明显, 这是因为改进后方法在固定隐私预算下, 自适应选取梯度裁

剪阈值, 保证了输入样例对梯度的贡献具有最小相关误差的最大保护。

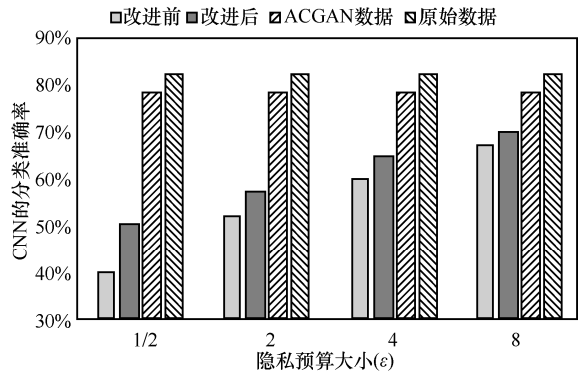


图8 Cifar10 数据 CNN 分类准确率

图 7 和图 8 表明分类准确率会随着隐私预算增大而提升, 这正是因为 ε 越小隐私保护程度越高, 训练中增加噪声越多, 对优化器参数更新造成干扰。为了调节可用性和隐私性, 应该适当增大隐私预算。并且, 笔者发现采用自适应选取裁剪阈值的方法后, 同隐私预算下 CNN 分类器分类效果有很大提升, 分类准确率接近无隐私保护情况下 ACGAN 生成的图像训练得到分类器的准确率。

4.2.3 Mnist 数据集和 Cifar10 数据集的统计特性

Salimans 等^[16]提出使用 Inception Scores 来评价 GAN 生成数据的质量。生成器 G 的 Inception Scores 定义为

$$s(G) = \exp(E_{x \sim G(z)} KL(\Pr(y|x) \| \Pr(y))) \quad (10)$$

这里, x 是 G 生成的样例; $\Pr(y|x)$ 是提前训练的预测 x 标签 y 的分类器所属的条件分布, 如果 x 是和真实样例相似的, 则期望 $\Pr(y|x)$ 的熵是小的; $\Pr(y) = \int \Pr(y|x = G(z)) dz$ 是 y 的边缘分布, 如果 G 能够生成多样性样例, 则期望 $\Pr(y)$ 的熵是大的。因此, 通过测量 2 个分布 KL 的差异, $s(G)$ 可以评估生成数据的质量和多样性。

表 1 总结了 Mnist 和 Cifar10 数据集分别训练的 DPACGAN, 改进后 DPACGAN 的生成数据 Inception Scores, 并且作为对比, 本文给出原始数据, ACGAN 生成数据^[16]的 Inception Scores。不难发现, 改进后 DPACGAN 生成的图像比改进前的分数提高, 更接近原始数据和 ACGAN 生成数据的分数。

表 1 真实数据和合成数据的 Inception Scores

数据集	类型	数量	(ε, δ)	分数
Mnist	真实数据	60 000	—	9.96 ± 0.03
	ACGAN 生成数据	60 000	—	6.59 ± 0.03
	DPACGAN 生成数据	60 000	$(4, 10^{-5})$	9.0 ± 0.03
	本文方法生成数据	60 000	$(4, 10^{-5})$	7.76 ± 0.03
Cifar10	真实数据	50 000	—	11.24 ± 0.12
	ACGAN 生成数据	50 000	—	8.25 ± 0.07
	DPACGAN 生成数据	50 000	$(8, 10^{-5})$	6.68 ± 0.06
	本文方法生成数据	50 000	$(8, 10^{-5})$	7.25 ± 0.07

4.2.4 分析总结

由 CNN 的分类准确率和 Inception Scores 的评分 2 个标准, 可以发现改进后的 DPACGAN 性能提升, 不仅更适用于数据分析任务, 而且生成的图像质量和多样性更好。这是因为改进后自适应选取的梯度裁剪阈值更接近每个样例输入时引起的梯度变化, 不会导致梯度的过小截断, 也不会导致梯度的过度放大。在差分隐私优化过程中, 可知梯度裁剪阈值大小即为敏感度的上界, 而添加噪声^[14], 如下式。

$$\sigma \geq \sqrt{2 \ln \left(\frac{1.25}{\delta} \right) \frac{\Delta f}{\varepsilon}} \quad (11)$$

可知噪声大小和敏感度相关。当自适应选取梯度阈值时, 梯度不会因为过小而导致训练缓慢, 隐私损失计算较大, 训练迭代轮数减少, 生成图像可利用性降低; 也不会因为过大而导致敏感度过大, 噪声添加过多, 以及训练参数的更新偏离最优值, 生成图像可利用性降低。

5 结束语

本文结合 DPACGAN 和梯度裁剪阈值自适应选取的优点, 不再将梯度裁剪阈值作为输入。因为生成对抗网络不同层的梯度相差很大, 训练时, 梯度值持续改变, 通过自适应选取阈值既可以防止阈值过小导致网络聚合缓慢, 又可以避免阈值过大导致敏感度增高, 生成图像质量可利用性低。本文通过实验评估, 改进后的 DPACGAN 比改进前 DPACGAN 生成的图像视觉上更清晰, 用于数据分析任务时性能更优,

图像质量和多样性更好。

关于本文工作还有很多需要长远研究的内容。例如, 本文实验都是聚焦于生成图像数据, 然而对其他语义丰富数据类型(如语言模型任务)是值得深入研究的。

参考文献:

- [1] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//The 53rd IEEE Annual Allerton Conference on Communication, Control, and Computing. 2015: 909-910.
- [2] ZHU T, LI G, ZHOU W, et al. Differentially private data publishing and analysis: a survey[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(8): 1619-1638.
- [3] MOHASSEL P, ZHANG Y. SecureML: a system for scalable privacy-preserving machine learning[C]//The 38th IEEE Symposium on Security and Privacy. 2017: 19-38.
- [4] SWEENEY L. K-Anonymity: a model for protecting privacy[J]. International Journal of UncertainTy Fuzziness and Knowledge Based Systems, 2002, 10(5): 557-570.
- [5] MACHANAVAJHALA A, GEHRKE J, KIFER D, et al. l-diversity: privacy beyond k-anonymity[C]//The 22nd IEEE International Conference on Data Engineering. Atlanta Georgia. 2006: 24-24.
- [6] LI N, LI T, VENKATASUBRAMANIAN S, et al. T-closeness: privacy beyond k-anonymity and l-diversity[C]//The 23rd IEEE International Conference on Data Engineering. 2007: 106-115.
- [7] DWORK C. Differential privacy[M]//Encyclopedia of Cryptography and Security. Springer, 2011: 338-340.
- [8] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//The ACM SIGSAC Conference on Computer and Communications Security. 2016: 308-318.
- [9] PAPERNOT N, ABADI M, ERLINGSSON U, et al. Semi-supervised knowledge transfer for deep learning from private training

- data[C]//The 5th International Conference on Learning Representations. 2017
- [10] ACS G, MELIS L, CASTELLUCCIA C, et al. Differentially private mixture of generative neural networks[C]//The 17th IEEE International Conference on Data Mining. 2017:715-720.
- [11] BEAULIEU-JONES B K, WU Z S, WILLIAMS C, et al. Privacy-preserving generative deep neural networks support clinical data sharing[R]. 2017.
- [12] ZHANG X, JI S, WANG T. Differentially private releasing via Deep Generative Model[EB/OL]. arXiv preprint arXiv: 1801.01594v1. 2018.
- [13] DWORK C. Differential privacy: a survey of results[C]// International Conference on Theory and Applications of Models of Computation. Springer, Berlin, Heidelberg. 2008: 1-19.
- [14] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.
- [15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.
- [16] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs[C]//International Conference on Machine Learning. 2017: 2642-2651.
- [17] KAIROUZ P, OH S, VISWANATH P, et al. The composition theorem for differential privacy[J]. IEEE Transactions on Information Theory, 2013, 63(6): 4037-4049.
- [18] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Advances in Neural Information Processing Systems. 2016: 2234-2242.

[作者简介]



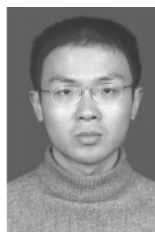
郭鹏（1991-），男，河南郸城人，福州大学硕士生，主要研究方向为信息安全、机器学习。



钟尚平（1969-），男，福建武平人，福州大学教授，主要研究方向为信息安全，机器学习、智能图像分析。



陈开志（1983-），男，福建福州人，福州大学讲师，主要研究方向为信息安全、图像安全。



程航（1979-），男，福建连江人，福州大学副教授，主要研究方向为多媒体安全、加密域图像检索、图像处理。