从记忆的角度论数据增强的隐私效应

匿名 CVPR 提交

论文编号 2529

抽象的

机器学习带来了严重的隐私问题,因为已经证明学习的模型可以揭示敏感信息
关于他们的训练数据的信息。许多工作研究了广泛采用的数据增强(DA) 和对抗训练 (AT) 技术（在本文中称为数据增强)对机器学习模型隐私泄露的影响。这种隐私影响通常通过成员推理攻击 (MIA) 来衡量,其目的是确定特定示例是否属于训练集。我们建议从称为记忆的新角度来研究隐私。通过记忆的镜头,我们发现之前部署的 MIA 会产生误导性的结果,因为与隐私风险较低的样本相比,它们不太可能将隐私风险较高的样本识别为成员。为了解决这个问题,我们部署了一种最近的攻击,可以捕获单个样本的记忆度以进行评估。通过广泛的实验,我们揭示了关于机器学习模型的三个基本属性之间联系的重要发现,包括隐私、泛化差距和对抗鲁棒性。我们证明,与现有结果不同,泛化差距与隐私泄露没有高度相关。此外,更强的对抗鲁棒性并不一定意味着该模型更容易受到隐私攻击。

一、简介

多项研究表明[3, 4, 34, 41]机器学习模型,尤其是深度神经网络(DNN) 会引发严重的隐私问题,因为它们往往会记住有关训练数据的敏感信息。为了量化评估机器学习模型揭示其训练数据的隐私泄漏,一种被广泛使用的基本方法是所谓的成员推理[32]。也就是说,如果可以访问目标模型,对手的目标是确定特定数据点是否用于训练该目标模型（作为成员)或不用于（作为非成员）。这样的会员资格信息可以揭示有关个人的非常敏感的信息,例如健康状况[28] ,并作为更强大类型的隐私攻击的基础[4]。

多项研究表明,成员推理攻击 (MIA) 的攻击成功率与泛化差距高度相关,即训练和测试精度之间的差异[21,31,32,35,39]。当应用不同的数据增强方法（包括数据增强（DA)和对抗训练（AT）)时,也会观察到这种相关性。 [35]表明应用 AT 会使模型更容易受到 MIA 的影响,他们得出的结论是,一个主要原因是应用 AT 后的泛化差距比标准训练更大。另一方面,DA 方法被广泛认为可有效减少隐私泄露[30,32,39] ,因为它们通常有助于避免过度拟合。然而,作为一种特殊的 DA 方法,标签平滑[36]被认为会增加隐私泄漏,同时减少泛化差距[14、20]。

然而,上述作品中显示的结果可能会产生误导,因为部署的用于测量隐私泄漏的 MIA 具有以下局限性:1)在一些作品[2,14,29]中受到批评,以前的 MIA 通常具有相当大的高误报率（FPR） ,即许多非会员被错误地识别为会员。

然而,一个好的攻击应该在低 FPR 区域获得有意义的攻击率,因为它对于计算机安全等实际应用更现实[22, 24]。如[2]所示的例子,如果一个总体准确率为 50.05% 的攻击能够可靠地识别出只有 0.1% 的成员而没有任何误报,即 FPR=0,并通过随机猜测以 50% 的准确率判断其余样本,则它与以 50.05% 的概率猜测任何样本的另一种攻击相比,给模型带来的风险要大得多。在这种情况下,后者具有较高的 FPR。 2) 我们发现以前的 MIA 与单个数据点的隐私风险不一致,尽管它们可能具有很高的总体成功率（参见第4.3 节）。具体来说,与隐私风险低的样本相比,他们在识别隐私风险高的训练样本作为成员方面更加困难,这与隐私风险高的样本的直觉不一致

风险应该更容易识别。

我们建议通过一种称为记忆的新视角来解决上述限制[10, 11]。如果模型的输出对这个单独的数据点非常敏感，则称该数据点已被记忆，例如，学习模型对该特定数据点的预测置信度可能非常低，除非它出现在训练集中[10].记忆的概念从根本上捕获了差分隐私（DP） [8, 9]框架下的隐私风险，这被认为是一个强隐私定义。根据经验，我们发现最近一种称为似然比攻击 (LiRA) [2]的攻击在反映记忆程度方面是有效的，正如我们在第 1 节中展示的那样。 4.3.

与传统的 MIA 相比，LiRA 在低 FPR 区域也表现出更好的性能[2]。

因此，我们采用 LiRA 来重新研究 DA 和 AT 的隐私影响。通过广泛的实验，我们揭示了几个重要的发现（详见第6节），这些发现促使社区重新思考机器学习模型的三个重要属性之间的关系，包括隐私泄露、泛化差距和对抗性鲁棒性。主要发现包括：

- 与之前的研究[17,21,31,32,35,39]表明泛化差距和隐私泄漏高度相关不同，我们的结果表明相关性要弱得多。
- 应用AT可以增加训练样本的记忆度，从而导致更多的隐私泄露。此外，它表明更强的对抗鲁棒性并不一定会以隐私泄露为代价。

据我们所知，这是第一次通过记忆的镜头对 DA 和 AT 进行系统评估。

二、相关工作

数据增强和隐私在[30,32]中根据经验观察到，DA 在减轻 MIA 方面是有效的。[20]进一步表明，很难使用 DA 来实现针对 MIA 的实质性缓解效果，同时实现更好的泛化差距。此外，标签平滑显示能够同时增加隐私泄漏和测试准确性[14、20]。至于 AT，研究其隐私效果也很重要，因为它被认为是实现对抗鲁棒性的最有效方法之一，这是安全界的另一个重要问题。 [35]使用各种 AT 方法进行系统调查，发现所有这些方法都可以使模型更容易受到 MIA 的影响。然而，这些研究中使用的 MIA 在从记忆的角度反映个体样本的隐私风险方面存在局限性。此外，他们都没有在低 FPR 区域下按指标报告结果。

增强信息改善隐私攻击多项研究表明[5, 18, 40] ，利用增强数据的信息将有助于提高攻击成功率。它们可以分为两种类型：增强未知攻击和增强感知攻击。前者假设对手不知道增强数据，只是简单地使用随机增强来探测模型。已经表明，通过多次查询模型，使用高斯噪声生成的随机增强数据，可以提高攻击成功率[18]。后者假设了一个更强的场景，其中训练中使用的特定增强数据为对手所知。 [5]表明，利用增强数据的知识可以显着提高攻击成功率。此外， [40]表明，增强感知攻击可以在经过一些数据增强训练的模型上获得比没有增强的模型更高的成功率。

三、预赛

我们考虑通常 su 下的前馈 DNN

受监管的设置。假设我们有一个训练集Dtr = {(x, y)|(x, y) ∈ X × Y}，其中 x 是特征向量（图像），y 是对应的标签。我们将具有参数 θ 的 DNN 模型表示为 fθ : X → Y。在训练期间，将数据增强 T : X × Y → P 应用于训练数据以提高其多样性，其中 P 是定义的所有概率度量的集合在幂集 2 X×Y上，模型的最优参数 θ 拟合为：

$$\theta^{*} = \underset{\theta}{\text{参数最小值}} \underset{(x,y)\in Dtr}{E(x,y)} T(x,y)\,[L(f\theta(x), y)], \quad (1)$$

其中 L(·, ·) 是损失函数。

3.1.数据增强

由于DA和AT都涉及将某些示例添加到训练集中以增强性能的过程，因此它们被称为数据增强技术。

在本文中，我们研究了八种流行的 DA 方法和两种 AT 方法，如下所述。

DA技术1)随机裁剪
和翻转[33]：通过从训练集中的原始特征随机裁剪和水平翻转补丁来采样新特征。

2)标签平滑[36]：通过将概率统一分配给其他类，用软连续标签替换硬标签。因此，aug (n－1)ε的概率

标记的标签是 p i = 1 − for i = y 并且它是 p i = for i = y，其中 $\frac{\varepsilon}{n}$ ∈ (0, 1) 并且 n 表示类的数量。

3) DisturbLabel [38]：将一部分ground truth (GT) labels改为不正确的labels，即y≈ = y +

CVPR #2529

(1 −)yf ,其中 是从 {0, 1} 中随机采样的， yf ∈ {1, 2, ..., n} \ {y} 表示错误的标签。

4) Gaussian Augmentation [6]:为每个特征添加高斯噪声。新特征 x≈ = x + ε,其中 ε N (0, σ2 I)。

5) Cutout [7]:从每个特征中屏蔽掉一个大小为 M × M 的随机正方形区域。

6) Mixup [42]:将两个特征x0、 x1按随机比率 γ 混合并创建一个新特征 x≈ = γx0 + (1 − γ)x1。相应的标签是 y~ = γy0 + (1 − γ)y1。

7)抖动[23]:随机改变每幅图像的亮度、对比度、饱和度和色调。

8) Distillation [16]:用原始训练集训练一个辅助 DNN ^f,并在训练目标 DNN 时使用辅助 DNN 的软输出和温度 T 作为训练特征的 GT 标签。温度 T 决定了软标签的平整度。

AT 技术1) PGD-
AT [26] :使用 PGD 攻击基于原始特征生成对抗样本xadv ,并在训练的每次迭代中用对抗样本替换原始特征,即 x~ = xadv。

2)TRADES [43]:也使用 PGD 攻击生成对抗样本xadv 。它与 PGD-AT 的不同之处在于它的损失函数由两部分组成： L(fθ(x), y) + L(fθ(x), fθ(xadv))/λ。第一个组成部分与标准训练的损失相同,而第二个组成部分鼓励模型平等对待 x 和xadv 。这两个分量由 λ 加权。在等式的框架内。 (1)、变换 T(x, y) 的离散分布可以表示为 Pr(x, y) = 和 Pr(x, fθ(xadv)) =

$$\frac{\lambda}{\lambda+1}$$

1 λ+1 。

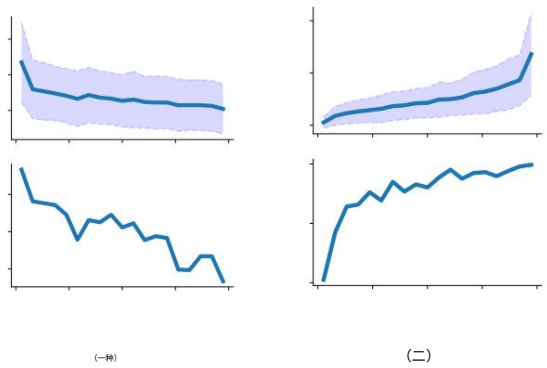
3) AWP [37]:使用正则化通过双扰动机制明确地平坦化 PGD-AT 的减肥景观。

4)TRADES-AWP [37]:将AWP的正则化机制纳入TRADES方法。

3.2.成员推理攻击

MIA 的目标是确定特定数据样本是否用于训练特定模型。由于其简单性,MIA 已成为研究最广泛的隐私攻击之一。许多现有的MIA 方法[19、25、31、32、39]能够通过利用机器学习模型通常与用于或不用于训练的数据表现不同这一事实来实现高攻击准确性。

例如,模型通常对训练数据比对测试数据更有信心。因此,通过为某些特征（例如损失、置信度分数、熵等)设置阈值,攻击可以在区分成员和非成员方面实现高精度。有关 MIA 的全面概述,我们建议读者参阅[17]。



(一种) (二)

图 1. 使用 (a) 在 CIFAR-100 上训练的目标模型的记忆得分方面的特征得分（顶部)和 TPR（底部）
MaxPreCA 和 (b) LiRA。为了进行比较,左上面板的特征得分被缩放到与右上面板相同的比例。

在本文中,我们假设一个黑盒设置,其中对手只能查询访问给定样本的目标模型的输出。大多数 MIA 遵循[32]来训练许多所谓的影子模型,这些模型与目标模型的训练方式类似,以模仿其行为。

不失一般性,这里我们假设输出是预测置信度,并且影子模型使用与目标模型相同的数据增强方法进行训练。

4. MIA 和记忆的一致性

在下文中,我们首先解释如何定义记忆,然后研究 MIA 和记忆的一致性。

4.1.背诵

如果数据点对模型的行为有很大影响,则称该数据点被模型记住了。为了确保这种影响完全是由这个特定样本引起的,通常需要使用留一法设置。

也就是说,除了这个特定的样本,所有其他设置都完全相同。例如,[10]定义了一个记忆分数,它衡量模型记忆了多少关于单个数据样本标签的信息。具体来说,给定训练集Dtr和学习算法 A,对于任意样本 (x, y) ∈ Dtr,其记忆分数定义为：

内存（A， Dtr，(x,y)）

$$= \Pr_{f_\theta \sim A(Dtr)}[f_\theta(x) = y] - \Pr_{f_\theta \sim A(Dtr\setminus(x,y))}[f_\theta(x) = y], \quad (2)$$

其中Dtr \(x, y)表示去除样本(x, y)后的数据集Dtr 。这个定义被证明是有效的,因为它能够在包括 CIFAR 在内的各种数据集上分配具有高记忆分数的非典型示例或异常值以及具有低记忆分数的典型或简单样本

100 和 ImageNet 数据集[11]。这完全符合直觉,即非典型示例或离群值是10,具有更高的隐私风险,因为模型在训练集内或训练集外时的行为会大不相同。

4.2. TraditionalMIAs 获得低一致性

记忆分数

大多数传统的 MIA 利用模型对训练数据过于自信的事实[15]。我们以广泛采用的基于最大预测置信度的攻击[31] （本文称为MaxPreCA)为例,其主要思想是将最大预测置信度高于给定阈值的样本分类为成员,否则为非成员。为了研究这种攻击是否可以捕获单个数据点的隐私风险,在图1(a) 中,我们展示了它的特征得分,即最大预测置信度（上图)和真阳性率（TPR)（下图）) 与记忆分数[11]。

我们首先根据记忆分数将所有样本分成 20 个 bin。顶部面板显示了不同 bin 中样本的平均特征得分（蓝色实线)及其标准差（紫色阴影） 。底部面板显示通过使用所有样本的最佳阈值计算的 TPR。很明显,记忆分数越高,被正确识别为成员的可能性就越小。因此,攻击结果与记忆分数的一致性较低。

4.3. LiRA 获得与记忆分数的高度一致性

LiRA 与传统 MIA 的主要区别

是 LiRA 考虑了模型在单个数据点上的预测在训练集中或训练集外时的分布。它需要训练多个阴影模型,使得对于每个样本 (x, y),一半模型将其包含在训练集中,另一半模型则不包含,分别表示为 IN 和 OUT 模型。将fθ (x)y表示为模型对标签 y 的 x 的置信度。将使用 IN 和 OUT 模型计算的样本 (x, y) 的缩放置信度集分别表示为Qin和Qout：

$$\begin{aligned} Qin &= \{ \quad (f\theta(x)y) : (x, y) \in Dtr\} \\ Qout &= \{ \quad (f\theta(x)y) : (x, y) \in / Dtr\}, \end{aligned} \tag{3}$$

其中 (p) = log 1−p 。 Qin和Qout用于拟合两个高斯分布,分别表示为 IN 分布 N µin， σ2 out 。给定一个任意样本,执行标准似然比检验以确定它更有可能属于哪个分布,其中似然比为：

$$\Lambda = \frac{p \quad (f\theta(x)y) \mid N\ \mu in, \sigma 2\ p}{(\quad (f\theta(x)y) \mid N\ (\mu out, \sigma 2\ out\ ))}。 \tag{4}$$

如果 Λ 高于阈值,样本将被归类为成员。与传统的 MIA 相比,LiRA 在低 FPR 区域显示有效[2]。至于记忆,在图1（b)中,我们展示了特征得分,即 Λ（上图)和真阳性率（TPR)（下图)与使用 LiRA的记忆得分[11] 。显然,我们可以看到样本的记忆分数越高,它被正确检测为成员的可能性就越大,这是一个合理的结果。因此,在下文中,我们将使用 LiRA 作为工具来研究不同的 DA 和 AT 方法如何影响隐私。

5. 评估数据增强的隐私效果

换货

我们现在开始评估 DA 和 AT 的隐私影响。通过实验,我们旨在研究隐私、对抗鲁棒性和泛化差距之间的关系,因为它们都是机器学习模型的关键属性。

5.1.实验设置

我们使用了 32 个 3080 GPU 来执行实验。实验代码由 Pytorch [27]实现并与论文一起提交。

数据集在[2, 12, 20] 之后,我们使用 CIFAR-10 和 CIFAR-100 数据集[1]进行 MIA 评估。 CIFAR 10 和 CIFAR-100 都包含 60,000 张分辨率为 32×32 的自然图像,分别来自 10 和 100 个类别。

MIA 设置我们使用 LiRA 来测量不同 DA 和 AT 方法的隐私泄露年龄。对于每个数据集上的每个数据增强,我们使用从 60,000 个数据点中随机选择的大约 30,000 个数据点训练了 128 个模型。对于每个单独的数据点,我们保证有 64 个 IN 模型和 64 个 OUT 模型。所有模组

除了数据增强策略外,els 使用相同的训练方法。然后我们随机选择一个训练好的模型作为目标模型,剩下的 127 个模型作为影子模型。之后,我们对所有的 60,000 个数据点进行了评估。为了执行 MIA,我们在每个数据点上查询目标模型十次,包括原始图像、四个移位（±4 像素)变体及其翻转版本。对于每种 DA 和 AT 方法,我们对十个随机选择的目标模型重复评估,然后报告每个指标的均值和标准差。

超参数我们在所有实验中都使用了 ResNet-18 [13] 。由于计算资源的限制,我们没有在其他架构上进行相同的实验。每个模型都通过随机梯度下降进行了优化,初始学习率为 0.1,动量为 0.9,用于单个 GPU 上的 100 个时期。在第 75 和第 90 个时期使用了将学习率缩放 0.1 的多步衰减。批量大小设置为 256。

CVPR #2529

方法	Training Acc	Test Acc	TPR @ 0.1% FPR	TPR @ 0.001% FPR	Log-scale AUC	MIA Balanced Acc
根据	100.0 ± 0.0	92.8 ± 0.2	8.20 ± 0.45	2.45 ± 0.93	0.815 ± 0.007	63.34 ± 0.26
光滑的	100.0 ± 0.0	92.9 ± 0.3	5.22 ± 0.66	0.14 ± 0.07	0.734 ± 0.012	62.28 ± 0.86
打扰标签	99.9 ± 0.0	92.7 ± 0.3	5.88 ± 0.83	0.70 ± 0.45	0.775 ± 0.013	61.69 ± 0.24
噪音	100.0 ± 0.0	92.6 ± 0.2	8.33 ± 0.26	2.79 ± 0.68	0.819 ± 0.004	63.56 ± 0.24
剪下	100.0 ± 0.0	93.1 ± 0.4	7.71 ± 0.39	2.48 ± 1.03	0.811 ± 0.010	63.23 ± 0.26
混合	99.7 ± 0.1	93.0 ± 0.2	5.17 ± 0.51	1.31 ± 0.40	0.779 ± 0.008	60.05 ± 0.53
抖动	100.0 ± 0.0	92.7 ± 0.2	8.24 ± 0.35	2.97 ± 0.76	0.819 ± 0.004	63.41 ± 0.31
蒸馏	99.9 ± 0.0	93.2 ± 0.2	7.04 ± 0.33	2.19 ± 0.70	0.805 ± 0.005	61.57 ± 0.39
PGD-AT	99.2 ± 0.1	82.2 ± 0.2	23.78 ± 0.89	10.52 ± 2.30	0.897 ± 0.005	78.82 ± 0.37
贸易	96.2 ± 0.2	80.0 ± 0.4	17.88 ± 1.56	8.14 ± 1.12	0.881 ± 0.006	77.21 ± 0.65
AWP	93.2 ± 2.0	82.6 ± 0.9	10.58 ± 3.48	3.06 ± 1.81	0.828 ± 0.045	72.13 ± 3.76
交易-AWP	91.9 ± 0.5	80.5 ± 0.2	12.43 ± 0.89	3.48 ± 1.36	0.848 ± 0.006	74.86 ± 0.80

表 1. 不同数据增强对 CIFAR-10 的攻击成功率。第 2 列和第 3 列分别显示每种方法的训练和测试精度。第 4-7 列显示了四个指标来评估隐私泄露的程度。我们强调不同 DA 和 AT 方法的 MIA 成功率高于 Base。

方法	Training Acc	Test Acc	TPR @ 0.1% FPR	TPR @ 0.001% FPR	Log-scale AUC	MIA Balanced Acc
根据	100.0 ± 0.0	70.3 ± 0.3	34.17 ± 1.05	17.24 ± 2.93	0.922 ± 0.002	83.17 ± 0.24
光滑的	100.0 ± 0.0	72.2 ± 0.4	39.21 ± 1.25	19.88 ± 3.86	0.932 ± 0.004	86.35 ± 0.22
打扰标签	98.0 ± 0.2	69.9 ± 0.3	19.53 ± 0.64	6.54 ± 2.58	0.879 ± 0.007	76.65 ± 0.28
噪音	100.0 ± 0.0	69.7 ± 0.3	33.83 ± 0.91	18.31 ± 2.78	0.923 ± 0.003	83.26 ± 0.13
剪下	100.0 ± 0.0	70.3 ± 0.3	34.71 ± 1.58	17.25 ± 5.02	0.923 ± 0.005	83.53 ± 0.22
混合	99.7 ± 0.1	71.2 ± 0.4	32.73 ± 1.13	19.18 ± 2.48	0.922 ± 0.003	82.39 ± 0.50
抖动	100.0 ± 0.0	70.3 ± 0.3	34.19 ± 0.90	18.37 ± 3.62	0.924 ± 0.003	83.35 ± 0.17
蒸馏	99.8 ± 0.0	72.6 ± 0.3	28.70 ± 0.83	14.58 ± 2.29	0.911 ± 0.002	79.46 ± 0.14
PGD-AT	99.5 ± 0.0	51.3 ± 0.3	68.63 ± 0.88	47.85 ± 4.26	0.972 ± 0.001	93.62 ± 0.10
贸易	98.0 ± 0.3	49.0 ± 0.5	60.23 ± 0.87	37.45 ± 5.15	0.963 ± 0.002	92.19 ± 0.23
AWP	85.3 ± 0.8	54.4 ± 0.3	39.92 ± 2.40	17.34 ± 4.22	0.931 ± 0.006	88.10 ± 0.38
交易-AWP	95.9 ± 0.6	51.3 ± 0.4	57.51 ± 2.02	35.57 ± 3.73	0.960 ± 0.002	91.92 ± 0.36

表 2. 不同数据增强对 CIFAR-100 的攻击成功率。使用与 Tab 中相同的约定。 1.

每个 DA 方法的超参数设置为通过搜索实现相对较高的测试精度（详见补充材料）。除非另有说明,否则对于所有 AT 方法,我们将无限范数下的最大扰动 ε 设置为 8。我们将步长设置为 ε/8,迭代步数设置为 10。另外,遵循默认值在每种方法的设置中,正则化参数 λ 对于 TRADES 设置为 1/6,扰动强度 γ 对于 TRADES-AWP 设置为 1e。

$^{-2}$ 适用于 AWP 和 5e $^{-3}$

5.2.评估结果

标签。图1和图2 分别显示了 CIFAR-10 和 CIFAR 100 上多次查询的训练和测试精度以及 MIA 结果。我们用 Base 方法表示 Random Cropping 和 Flip ping。与之前的大多数研究[20、35、40]不同,我们评估了从标签平滑到蒸馏的七种 DA 方法的隐私泄漏以及四种 AT 方法（第3 节)与 Base 相结合。这是一个实用的设置,因为 Random Cropping and Flipping 现在已经成为计算机视觉领域的默认设置,并且通常会显着提高测试精度。

在[29]中受到批评的是,测试精度低的模型对于评估隐私泄漏实际上没有用。请参阅补充材料中不含 Base 的不同 DA 方法的结果,其中 Base 的测试精度确实超过其他 DA 方法,在 CIFAR-10 上至少高出 8.4%,在 CIFAR-100 上高出 12.6%。除非另有说明,否则所有 DA 和 AT 方法也默认使用 Base。

对于 LiRA,我们使用四个指标评估了所有模型:TPR @ 0.1% FPR、TPR @ 0.001% FPR、曲线下对数刻度面积 (AUC) 和平衡精度。 ± 后的数字表示标准偏差。如前所述,在低 FPR 区域下使用 TPR 度量来评估隐私泄漏更为合理

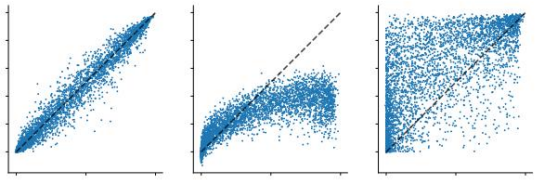


图 2. 使用 Jitter（上）、Disturblabel（中)和 PGD-AT 模型（下)与 Base 模型对比 5,000 个随机选择的样本的记忆分数。

年龄。然而，一方面，我们根据经验发现 TPR @ 0.001% FPR 的结果不稳定，因为它们的标准差相对较大。另一方面，0.001% FPR 可能过于严格，因为在[11]中已经表明CIFAR-10 和 CIFAR-100 都包含相当多的非常相似的硬样本对。当这些样本的对应物在成员集中时，这些样本将不可避免地被错误分类为成员，从而导致一些假阳性案例并导致 FPR 超过 0.001% 的容差。因此，我们主要使用TPR@0.1%FPR作为下面分析的攻击成功率。

6.分析

下面我们首先验证我们的攻击结果的有效性，然后分析隐私、泛化差距和对抗鲁棒性之间的关系。

6.1.攻击结果和记忆度

验证攻击结果是否如表所示。图1和图2确实反映了记忆程度，在图2中，我们比较了使用与[11]中相同的方法计算的 5,000 个随机选择样本的记忆分数，用于 CIFAR-100 的三种情况：Jitter、Disturblabel 和 PGD AT与基地。其中攻击成功率分别与Base相近、低于、高于 Base。显然，记忆分数的相应变化与攻击成功率是一致的。 Base 和 Jitter 样本的记忆分数相似（位于对角线附近），这解释了针对这两种方法的相似攻击成功率。 Disturlabel 的很多样本的记忆分数低于 Base，尤其是 Base 记忆分数高的样本。因此，Disturblabel 减少隐私泄漏的原因之一是它可以降低许多非典型样本的记忆分数。PGD-AT 的记忆分数普遍高于 Base（分数分布在对角线的上部）。因此，AT 导致更高隐私泄漏的一个主要原因是它记忆了许多非 AT 模型没有记忆的训练样本。

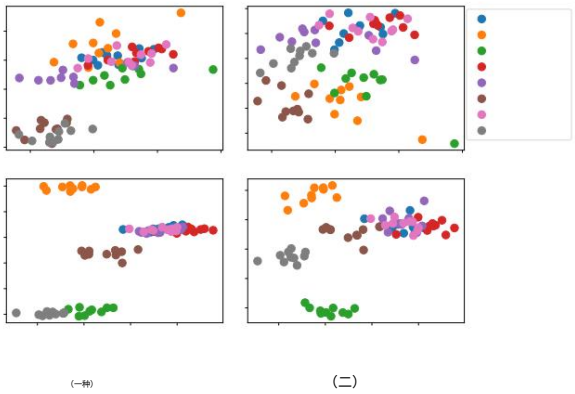


图 3. 分别使用 (a) MaxPreCA 和 (b) LiRA 在 CIFAR-10 和 CIFAR-100 上使用不同 DA 模型的攻击成功率与训练测试差距。 r 代表皮尔逊相关系数。

6.2.隐私和泛化差距

缩小泛化差距并不一定会降低 MIA 的脆弱性。许多先前的研究表明，MIAs 的攻击成功率与泛化差距，即过度拟合的程度高度相关[21, 31, 32, 35, 39]。为了从记忆的角度验证这种高相关性是否仍然正确，在图3中，我们使用 MaxPreCA 和 LiRA 证明了针对 CIFAR 10 和 CIFAR-100 数据集的所有 DA 模型的训练测试精度差距方面的攻击成功率，分别。很容易观察到，与传统的攻击结果相比，我们的结果表现出更分散的分布。我们还计算了每个图的 Pearson 相关系数 r。如图所示，我们结果的 Pearson 相关系数 r 明显低于传统结果，例如，对于 CIFAR-10，我们的 r 仅为 0.221，远低于使用传统攻击的 0.708。因此，从记忆的角度来看，泛化差距和隐私泄露的相关性似乎低于之前的结果。

很容易理解为什么许多传统攻击结果对泛化差距敏感，因为它们的成功率在很大程度上取决于模型对训练和测试样本的行为有多大不同。我们注意到记忆和过度拟合之间存在区别：记忆只是过度拟合所必需的但不是充分的[10]，即记忆一些训练样本并不总是导致过度拟合。事实上，在[10,11]中已经从理论上证明和经验上验证了记忆某些长尾样本将有助于减少泛化差距。因此，传统攻击可能会低估非过度拟合模型的隐私泄漏，从而使相关系数不必要地高。

当我们测量

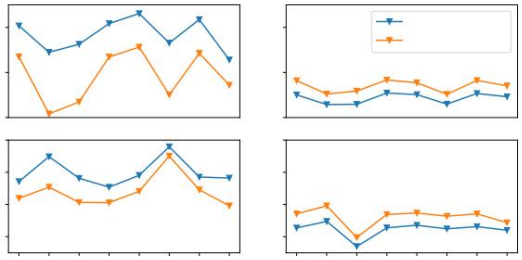


图 4. 两种情况下单个查询和多个查询的攻击成功率：增强未知（左）和增强感知（右）。我们分别在 CIFAR-10 和 CIFAR-100 数据集上评估了不同的 DA 方法。图中的 None 代表没有任何 DA 训练的模型。

数据集	方法	对抗TPR@ 加速器	0.1% 固定利率
CIFAR-10	PGD-AT	38.8 ± 0.4	23.78 ± 0.89
	贸易	45.2 ± 0.3	17.88 ± 1.56
	AWP	45.9 ± 0.1	10.58 ± 3.48
	交易-AWP	48.8 ± 0.2	12.43 ± 0.89
CIFAR-100	PGD-AT	16.9 ± 0.1	68.63 ± 0.88
	贸易	19.7 ± 0.4	60.23 ± 0.87
	AWP	23.9 ± 0.1	39.92 ± 2.40
	交易-AWP	23.3 ± 0.2	57.51 ± 2.02

表 3. 不同 AT 模型在 CIFAR-10 和 CIFAR-100 上的对抗样本（Adversarial Acc）和隐私泄露的准确率。使用 PGD 以 ε = 8 和 20 个迭代步骤评估准确度。

从记忆角度看隐私泄露，隐私泄露的根本原因。我们注意到，尽管[39]也指出过度拟合并不是导致隐私攻击脆弱性的唯一原因，但他们并没有明确指出其他因素是什么，他们的攻击结果与我们的相比仍然表现出更高的相关性。

数据增强不是 MIA 的有效防御所必需的。通过检查 DA 模型的攻击结果，隐私效果在不同的 DA 方法中存在显着差异。例如，Disturblabel 和 Disturblabel 被证明可以有效降低隐私攻击的脆弱性。 Mixup、Cutout、Jitter和Gaussian noise方法似乎对攻击成功率影响不大。

主要原因是应用 DA 并不总是会降低训练样本的记忆分数，例如图2 中所示的 Jitter 模型。此外，在所有 DA 方法中，标签平滑引起了很多关注，因为它已在[14,20]应用标签平滑会使模型更容易受到 MIA 的影响。为了验证这一点，我们

使用传统攻击 MaxPreCA [31] 计算平衡精度。如图 3的左图所示，与 Base 相比，Label Smoothing 确实提高了两个数据集的攻击精度。然而，从记忆的角度来看，它并没有表现出同样的趋势。

通过检查图3的右图，我们注意到标签平滑在不同数据集上表现出不一致的行为。在 CIFAR-100 上，隐私泄露高于 Base，而在 CIAFR-10 上，隐私泄露较低。

我们推测标签平滑的隐私效果可能取决于数据集的复杂性。因此，标签平滑会持续放大隐私泄漏的说法是不正确的。总的来说，我们得出的结论是，很难就 DA 是否有助于减轻隐私攻击给出一般性的断言。我们提醒您，在依赖 DA 作为防御 MIA 的技术时应格外注意。

如果增强方法已知，则多个查询只能增强攻击。如第二节所述。 1，先前的研究表明，使用增强数据进行多次查询会提高攻击成功率。

为了研究这一点，我们使用 Base 方法为每个数据点生成的十个增强对应物查询目标模型。然后，我们将所有在 Base 上训练的 DA 模型作为增强感知案例。然后通过在不使用 Base 的情况下以 DA 模型为目标来评估增强未知案例。如图4 所示，多个查询确实有助于提高增强感知案例的攻击成功率，而对于增强未知案例，它们会产生相反的效果，即降低攻击成功率。

6.3.隐私和对抗鲁棒性

应用 AT 会使模型记住更多的训练样本，从而导致更多的隐私泄露。如选项卡所示。如图 1和2 所示，与非 AT 模型相比，应用 AT 显着增加了隐私泄露。例如，对于 CIFAR-100 上的 TRADES，TPR @ 0.1% FPR 从 34.17% 增加到 60.23%。一个原因是应用 AT 会迫使模型拟合每个训练样本周围ℓ_∞球中发现的所有对抗样本，这通常会增加每个样本对训练模型的影响，从而导致更高的隐私风险。为了可视化应用 AT 的效果，在图5中，我们在 CIFAR-100 中选择了三个具有不同记忆分数的示例，并绘制了它们相应的归一化置信度分布，这些归一化置信度由使用 Base 的 IN 和 OUT 模型以及所有四种类型的 AT 模型评估.显然，如果特定样本的 IN 和 OUT 分布更加分离，则意味着该样本处于更高的隐私风险中。我们可以看到，对于隐私风险较低的样本（例如 Raccoon 和 Train），应用 AT 会使分布更加可分离。请注意，有一个机器人

CVPR
#2529

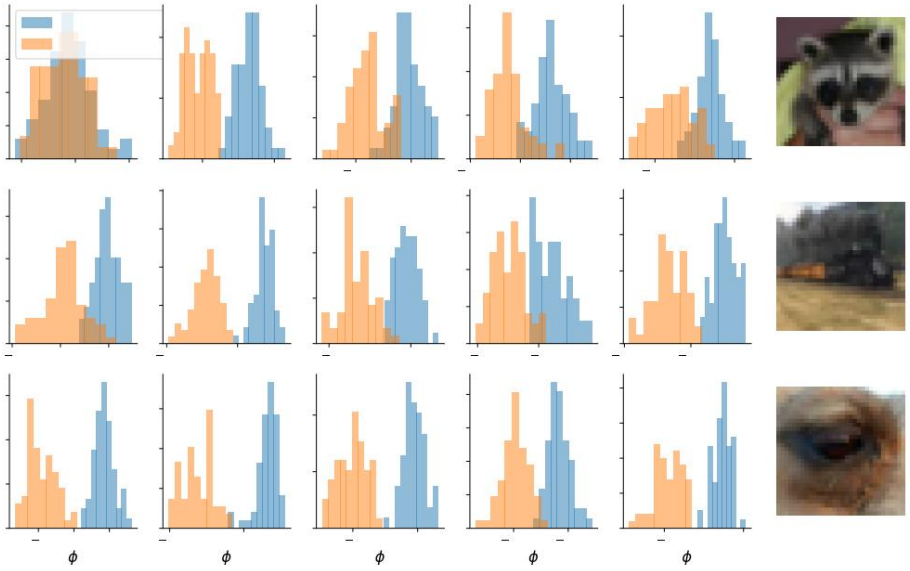


图 5. 使用 Base 和四个 AT 模型具有不同记忆分数的三个样本的归一化置信度 分布。
每个原始对应一个样本。

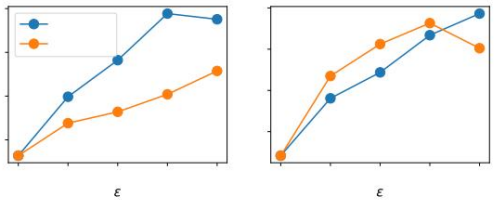


图 6. PGD-AT 和 TRADES 在 CIFAR-10 和 CIFAR-100 数据集上不同 ε 下的攻击成功率。

这种影响的缺点:对于已经处于高隐私风险的样本（例如,具有高记忆分数的 Camel）,应用 AT 不会产生太大差异,因为所有模型的分布都非常分离。总的来说,我们得出结论,AT 导致更高隐私泄漏的一个主要原因是它记住了许多非 AT 模型没有记住的训练样本。

更好的对抗鲁棒性并不一定会使模型更容易受到隐私攻击。为了进一步研究对抗鲁棒性和隐私泄露之间的关系,在表中。 3我们在 CIFAR-10 和 CIFAR-100 上使用不同的 AT 方法比较了对抗性鲁棒性和攻击成功率。我们可以清楚地看到,与 TARDES 和 PGD-AT 相比,AWP 和 TRADES-AWP 都获得了更高的对抗精度,但攻击成功率较低。因此,提高对抗性鲁棒性并不一定会以隐私为代价。除了使用不同 AT 方法的攻击结果外,看看攻击结果随着参数的变化而变化。由于 ε 是 AT 的关键参数,在图6中,我们使用 PGD-AT 和 TRADES 模型在 CIFAR-10 和 CIFAR-100 上比较了不同 ε 的攻击结果。我们可以看到总体上攻击成功率随着 ε 的增加而增加（至少对于 ε < 8）。此外,我们还观察到类似的瓶颈效应（参见补充材料中的图A1 ）。

因此,对于记忆分数较低的样本,增加 ε 会增加隐私风险,而对于记忆分数较高的样本,则几乎没有太大区别。

七、结论

在本文中,我们通过一个新的视角,即记忆程度,重新研究了将数据增强和对抗训练应用于机器学习模型的隐私效果。这种重新调查是非常必要的,因为我们发现在以前的研究中部署的用于测量隐私泄漏的攻击产生了误导性结果:与具有高隐私风险的训练样本相比,具有低隐私风险的训练样本更容易被识别为成员。通过系统的评估,我们揭示了一些发现与之前的结果相冲突,例如,泛化差距和隐私泄露显示出比之前的结果更不相关,标签平滑并不总是放大隐私泄露。此外,我们还表明,提高对抗鲁棒性（通过对抗训练)并不一定会使模型更容易受到隐私攻击。我们的结果要求从记忆的角度对机器学习模型的隐私进行更多调查。

CVPR #2529

参考

[1] Krizhevsky Alex、Hinton Geoffrey 等。从微小图像中学习多层特征。 2009. 4 [2] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer.来自第一原则的成员推理攻击。在 IEEE 安全和隐私研讨会 (SP),第 1897–1914 页。 IEEE, 2022. 1, 2, 4 [3] Nicholas Carlini、Chang Liu、Ulfar Erlingsson、Jernej Kos 和 Dawn Song。秘密分享者:评估和测试神经网络中的无意识记忆。在 USENIX 中

安全研讨会 (USENIX),第 267–284 页,2019年。1 [4] Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts，
Tom B. Brown、Dawn Song、Ulfar Erlingsson、Alina Oprea 和 Colin Raffel。从大型语言模型中提取训练数据。在 USENIX 安全研讨会 (USENIX),第 2633–2650 页,2021年。1

[5] Christopher A. Choquette-Choo、Florian Tramer、Nicholas Carlini 和 Nicolas Papernot。推理攻击中的仅标签成员资格。在诠释。会议。马赫。学。 (ICML),第 1964-1974 页。 PMLR,2021。2 [6] Jeremy M. Cohen、Elan Rosenfeld 和 J. Zico Kolter。通过随机平滑证明了对抗性的鲁棒性。在诠释。会议。马赫。学。 (ICML),第 1310–1320 页,2019年。3 [7] Terrance Devries 和 Graham W. TaylorW。改进了带切口的卷积神经网络的正则化。

arXiv 预印本 arXiv:1708.04552, 2017. 3

[8] 辛西娅·德沃克。差分隐私。在 ICALP 中,第 1-12 页，
2006. 2

[9] Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam D. Smith。在私人数据分析中将噪声校准为灵敏度。 J. 私人。机密性,7(3):17–51, 2016. 2 [10] Vitaly Feldman。学习需要记忆吗?一个关于长尾巴的小故事。在年度 ACM SIGACT 计算理论研讨会 (STOC) 中,第954-959 页， 2020年。2、3、6

[11] 维塔利·费尔德曼和张驰远.神经网络记住什么以及为什么:通过影响估计发现长尾。进阶神经信息。过程。系统。 (NeurIPS), 33:2881–2891, 2020. 2, 4, 6

[12] Ganesh Del Grosso、Hamid Jalalzai、Georg Pichler、Catus cia Palamidessi 和 Pablo Piantanida。利用对抗样本来量化成员信息泄漏。

IEEE 会议。电脑。可见。模式识别。 (CVPR), 2022. 4 [13] 何开明,张翔宇,任少卿，孙健。
用于图像识别的深度残差学习。在 IEEE 会议上。
电脑。可见。模式识别。 (CVPR),第 770–778 页,2016 年。4

[14] Dominik Hintersdorf、Lukas Struppek 和 Kristian Kersting。信任或不信任成员推理攻击的预测分数。 arXiv 预印本 arXiv:2111.09076, 2022. 1, 2, 7

[15] Dominik Hintersdorf、Lukas Struppek 和 Kristian Kersting。信任或不信任成员推理攻击的预测分数。摘自 IJCAI 编辑 Luc De Raedt,第 3043–3049 页,2022年。4

[16] Geoffrey E. Hinton、Oriol Vinyals 和 Jeffrey Dean。在神经网络中提取知识。 arXiv 预印本 arXiv:1503.02531, 2015. 3 [17] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang.对机器学习的成员推理攻击:一项调查。 ACM 计算调查 (CSUR),2021. 2, 3 [18] Bargav Jayaraman、Lingxiao Wang、Katherine Knipmeyer、Quanquan Gu 和 David Evans。重新审视现实假设下的成员推理。隐私增强技术论文集, 2021(2), 2021. 2 [19] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard:通过对抗性示例防御黑盒成员推理攻击。在 ACM SIGSAC 计算机和通信安全会议 (CCS),第 259–274 页,2019年。3 [20] Yigitcan Kaya 和 Tudor Dumitras。数据增强何时有助于成员推理攻击?在诠释。

会议。马赫。学。 (ICML),第 5345–5355页， 2021年。1、2、4、5、7

[21] 克拉斯雷诺和马特弗雷德里克森。被盗的记忆:利用老化的模型记忆进行校准的 {White-Box} 成员推理。在 USENIX 安全研讨会 (USENIX),第1605–1622 页,2020。1、2、6 [22] Jeremy Z. Kolter 和 Marcus A. Maloof。学习在野外检测和分类恶意可执行文件。机器学习研究杂志,7(12)， 2006。1 [23] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E. Hinton。

使用深度卷积神经网络进行 Imagenet 分类。进阶神经信息。过程。系统。 (NeurIPS), 25, 2012. 3

[24] Aleksandar Lazarevic、Levent Ert”oz、Vipin Kumar、Aysel Ozgur 和 Jaideep Srivastava。网络入侵检测中异常检测方案的比较研究。

SIAM 数据挖掘国际会议论文集,第 25-36 页,2003 年。1

[25] 龙云辉,Vincent Bindschaedler,王磊,卜迪岳,王晓峰,唐海旭,Carl A. Gunter,陈凯。

理解对广义学习模型的成员推理。 arXiv 预印本 arXiv:1802.04889, 2018. 3 [26] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向抵抗对抗性攻击的深度学习模型。在诠释。会议。学。

代表。 (ICLR), 2018. 3 [27] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zem ing Lin、Natalia Gimelshein、Luca Antiga、Alban Desmai son、Andreas Kopf、Edward Z . Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。 Py torch:一种命令式的高性能深度学习库。在副词中。神经信息。过程。系统。 (NeurIPS),第 8024–8035 页,2019。4 [28] Amir Hossein Poorjam、Yordan P. Raykov、Reham Badawy、Jesper Rindom Jensen、Mads Graesboll Christensen 和 Max A. Little。使用无限隐马尔可夫模型对远程帕金森病监测中的录音质量进行控制。在过程中。诠释。会议。 Acoust., Speech, Signal Process., pp. 805-809, 2019. 1

[29] Shahbaz Rezaei 和 Xin Liu.关于成员推理攻击的难度。在 IEEE 会议上。电脑。可见。模式识别。 (CVPR),第 7892–7900 页,2021.1、5 [30] Alexandre Sablayrolles、Matthijs Douze、Cordelia Schmid、Yann Ollivier 和 Herve Jegou。白盒与黑盒：成员推理的贝叶斯最优策略。在诠释。 , ,

会议。马赫。学。 (ICML),第 5558–5567 页,2019.1、2 [31] Ahmed Salem、Yang Zhang、Mathias Humbert、Pascal Berrang、Mario Fritz 和 Michael Backes。 Ml-leaks：模型和数据独立的成员推理攻击和机器学习模型的防御。网络和分布式系统安全研讨会,2019 . 1, 2, 3, 4, 6, 7

[32] Reza Shokri、Marco Stronati、宋从政和 Vitaly Shmatikov。针对机器学习模型的成员推理攻击。在 IEEE 安全与隐私 (SP) 研讨会,第 3–18 页， 2017年。1、2、3、6

[33] 凯伦·西蒙尼安和安德鲁·齐瑟曼。用于大规模图像识别的非常深的卷积网络。 arXiv 预印本 arXiv:1409.1556, 2014.2 [34] Congzheng Song、Thomas Ristenpart 和 Vitaly Shmatikov。

记住太多的机器学习模型。在 ACM SIGSAC 计算机和通信安全会议 (CCS),第 587–601 页,2017 年。1 [35] Liwei Song、Reza Shokri 和 Prateek Mittal。保护机器学习模型免受对抗性示例的隐私风险。 ACM SIGSAC 计算机和通信安全会议 (CCS),第 241–257 页,2019年,第1、2、5、6页

[36] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jonathon Shlens 和 Zbigniew Wojna。重新思考计算机视觉的初始架构。在 IEEE 会议上。

电脑。可见。模式识别。 (CVPR), pages 2818–2826, 2016. 1, 2 [37] Dongxian Wu, Shu-Tao Xia, and Yisen Wang.对抗性权重扰动有助于稳健的泛化。在副词中。神经信息。过程。系统。 (NeurIPS), 2020. 3 [38] Lingxi Xie, Jingdong Wang, Zhen Wei, Meng Wang, and Qi Tian. Disturlabel：在损失层上正则化 cnn。

在 IEEE 会议上。电脑。可见。模式识别。 (CVPR),第 4753–4762 页,2016 年。2 [39] Samuel Yeom、Irene Giacomelli、Matt Fredrikson 和 Somesh Jha。机器学习中的隐私风险：分析与过度拟合的联系。在 IEEE 计算机安全基金会研讨会 (CSF),第268–282 页， 2018年。1、2、3、6、7

[40] 大禹,张会帅,陈伟,尹建,刘铁雁。数据增强如何影响机器学习中的隐私?在 AAAI,第 35 卷,第 10746–10753 页,2021年。2、5

[41] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。理解深度学习（仍然)需要重新思考泛化。公社。 ACM, 64(3):107–115, 2021. 1 [42] Hongyi Zhang、Moustapha Cisse、Yann N. Dauphin 和 David Lopez-Paz。混合：超越经验风险最小化。在诠释。会议。学。代表。 (ICLR), 2018. 3 [43] 张红阳、于耀东、焦建涛、邢鹏、Laurent El Ghaoui 和 Michael I. Jordan。理论上

鲁棒性和准确性之间的原则权衡。在诠释。
会议。马赫。学。 (ICML),第 7472–7482 页,2019年。3

方法	Training Acc	Test Acc	TPR @ 0.1% FPR	TPR @ 0.001% FPR	Log-scale AUC	MIA Balanced Acc
没有任何	100.0 ± 0.0	82.9 ± 0.5	20.35 ± 4.31	9.44 ± 3.24	0.885 ± 0.013	76.25 ± 2.18
无+平滑	100.0 ± 0.0	83.7 ± 0.5	14.48 ± 3.03	2.37 ± 1.79	0.839 ± 0.024	72.91 ± 1.60
无 + 干扰标签	100.0 ± 0.0	84.1 ± 0.6	16.26 ± 1.03	3.43 ± 2.48	0.853 ± 0.016	72.34 ± 0.89
无 + 噪音	100.0 ± 0.0	82.4 ± 0.7	20.84 ± 3.69	8.59 ± 2.88	0.886 ± 0.011	76.90 ± 2.52
无 + 镂空	100.0 ± 0.0	84.0 ± 0.6	23.07 ± 0.80	10.53 ± 1.85	0.894 ± 0.004	77.42 ± 0.37
无+混合	100.0 ± 0.0	83.7 ± 0.4	16.53 ± 1.42	4.84 ± 1.58	0.867 ± 0.006	76.53 ± 0.99
无 + 抖动	100.0 ± 0.0	82.0 ± 1.3	21.73 ± 5.85	8.77 ± 3.64	0.886 ± 0.020	77.63 ± 2.61
无 + 蒸馏	100.0 ± 0.0	84.4 ± 0.4	12.79 ± 1.81	5.16 ± 1.18	0.851 ± 0.010	69.25 ± 0.99

表 A1。在 CIFAR-10 上不使用 Base 的不同 DA 方法的攻击成功率。第 2 列和第 3 列分别显示每种方法的训练和测试精度。第 4-7 列显示了四个指标来评估隐私泄露的程度。我们强调不同 DA 方法的 MIA 成功率大于 None 的成功率。

方法	Training Acc	Test Acc	TPR @ 0.1% FPR	TPR @ 0.001% FPR	Log-scale AUC	MIA Balanced Acc
没有任何	100.0 ± 0.0	54.0 ± 0.9	54.26 ± 10.72	30.68 ± 11.03	0.954 ± 0.014	92.96 ± 2.28
无+平滑	100.0 ± 0.0	53.7 ± 2.0	69.71 ± 3.51	41.81 ± 9.55	0.972 ± 0.004	96.76 ± 0.33
无 + 干扰标签	100.0 ± 0.0	55.5 ± 0.5	56.30 ± 1.22	37.37 ± 4.58	0.959 ± 0.003	90.96 ± 0.23
无 + 噪音	100.0 ± 0.0	53.5 ± 1.2	50.75 ± 8.61	28.61 ± 10.75	0.950 ± 0.011	91.79 ± 1.70
无 + 镂空	100.0 ± 0.0	54.1 ± 0.8	58.14 ± 5.60	36.64 ± 7.94	0.961 ± 0.007	92.87 ± 1.04
无+混合	100.0 ± 0.0	49.3 ± 0.7	75.88 ± 1.09	51.04 ± 6.80	0.978 ± 0.002	96.13 ± 0.06
无 + 抖动	100.0 ± 0.0	53.1 ± 0.9	57.05 ± 11.22	30.21 ± 14.71	0.955 ± 0.016	93.47 ± 2.18
无 + 蒸馏	100.0 ± 0.0	57.7 ± 1.7	56.45 ± 3.95	35.61 ± 7.31	0.959 ± 0.006	90.39 ± 0.88

表 A2。在 CIFAR-100 上不使用 Base 的不同 DA 方法的攻击成功率。使用与 Tab 中相同的约定。 A1.

A. 每个数据增强的超参数

如论文所述,每个 DA 方法的超参数被设置为通过尝试各种值来获得相对较高的测试精度。在这里,我们报告了在为每种 DA 方法训练 128 个影子模型时我们尝试的值和使用的最终值。

随机裁剪和翻转首先,分辨率为 32×32 的图像在每一端用 4 个像素的零填充。然后随机裁剪出分辨率为 36 × 36 的填充图像,形成分辨率为 32 × 32 的输入。最后,输入被随机水平翻转。除非另有说明,否则所有其他 DA 和 AT 方法也默认使用此方法。

标签平滑我们尝试了 ε 包括 0.01、0.05、0.1、0.2、0.3、0.4、0.5、0.6、0.7 和 0.8。最后,我们在 CIFAR-10 上选择了 0.2,在 CIFAR-100 上选择了 0.3。

Disturlabel我们尝试了 ε,包括 0.01、0.05、0.1、0.2、0.3、0.4、0.425、0.45、0.5、0.525、0.55、0.575 和 0.6。最后,我们在 CIFAR-10 上选择 0.05,在 CIFAR-100 上选择 0.3。

高斯增强我们尝试了 σ,包括 0.025、0.01、0.05、0.075、0.1、0.125、0.15、0.175、0.2、0.225、0.25、0.275、0.3、0.325 和 0.35。最后,我们在两个数据集上选择 σ 为 0.01。

Cutout我们尝试了 M,包括 4、8、12、16 和 20。最后,我们在两个数据集上都选择 M 为 8。

Mixup中使用的 Mixup γ 从 beta 分布 γ Be(α, α) 中采样。我们尝试了 α,包括 0.5、0.1、0.25、1、2、4、8、16、32、64、128 和 256。最后,我们在两个数据集上都选择 α 为 0.5。

抖动我们直接使用了 Pytorch 中的 ColorJitter 函数。我们尝试了亮度、对比度、饱和度和色调对应的参数,包括0.05、0.1、0.2、0.15、0.25、0.3、0.35、0.4、0.45和0.5。最后,我们在两个数据集上都选择了 0.05。

蒸馏我们尝试了包括 1、2、3、5 和 10 的 T。最后,我们在两个数据集上都选择了 T 为 3。

B. 不使用 Base 的模型的成员推理攻击结果

表A1和A2展示了所有未使用 Base 训练的 DA 模型的训练和测试精度以及 MIA 结果。这里使用单个查询是因为它获得了比多个查询更高的攻击成功率,如论文中的图4所示。这里 None 代表没有任何 DA(只有原始图像数据)训练的模型。未使用 Base 训练的模型的测试精度远低于使用 Base 训练的模型。

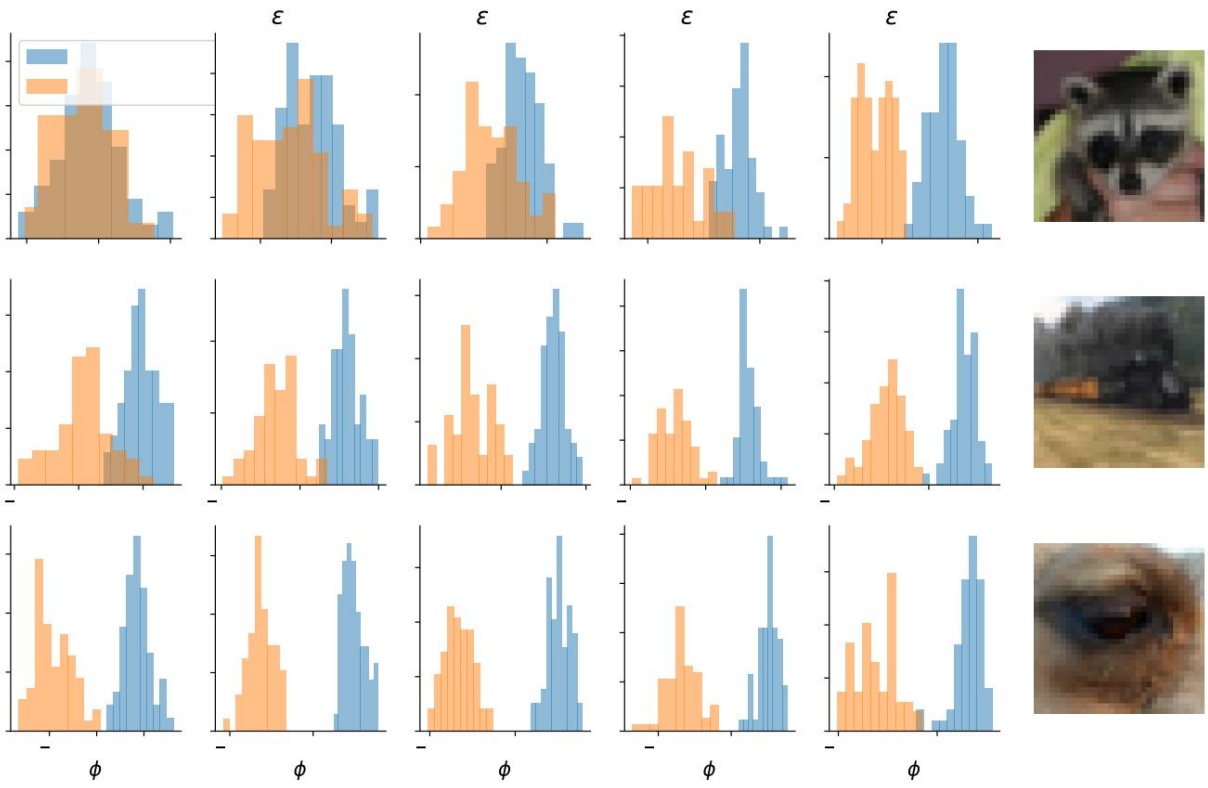


图 A1。使用 Base 和 PGD-AT 在四种不同 ε 下具有不同记忆分数的三个样本的归一化置信度 分布。每个原始对应一个样本。