文章编号:1671-4229(2024)05-0076-09

基于扩散模型和混合样本的成员推理防御方法

余良钊[a],李啸宇[b],周灿彬[b*]
(广州大学 a. 网络空间安全学院, b. 计算机科学与网络工程学院, 广东 广州 510006)

摘 要: 在深度学习中,成员推理攻击是指对于任意给定的样本,推断其是否属于目标模型的训练数据集。由于训练数据集中包含隐私信息,对成员推理攻击进行防御属于重要的隐私保护问题。文章首先介绍了成员推理攻击的定义和造成攻击的原因;接着详细地介绍了现有的防御算法;最后提出防御方案,并描述防御方案的防御途径。与当下最先进的成员推理攻击防御措施相比,该方案在保护成员隐私和保护模型效用之间提供了最优越的权衡。文章对所使用的方案做了详细的解释,帮助研究者更好地理解成员推理攻击及其防御,为进一步降低训练数据集的隐私风险、更好地取得模型效用和隐私安全之间的平衡性,提供了一定的参考。

关键词: 成员推理攻击; 防御技术; 隐私保护; 机器学习
中图分类号: TP181;TP309 **文献标志码:** A

Defense method for membership inference attacks based on diffusion model and mixed samples

YU Liang-zhao[a], *LI Xiao-yu*[b], *ZHOU Can-bin*[b*]
(a. Cyberspace Institute of Advanced Technology,
b. School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China)

Abstract: Membership inference attacks in deep learning refer to inferring whether a given sample belongs to the training dataset of a target model. Due to the presence of privacy-sensitive information in the training dataset, defending against membership inference attacks is crucial for privacy protection. This paper begins by defining membership inference attacks and elucidating the underlying reasons causing such attacks. Subsequently, existing defense algorithms are comprehensively reviewed. Finally, a novel defense mechanism is proposed, delineating the defensive approach adopted in this paper. Compared to state-of-the-art defenses against membership inference attacks, this method offers superior trade-offs between preserving member privacy and maintaining model utility. Detailed explanations of the employed techniques are provided to facilitate a better understanding of membership inference attacks and their defenses, thereby furnishing valuable insights for mitigating privacy risks in training datasets and striking a balance between model utility and privacy security.

Key words: membership inference attack; defense technology; privacy protection; machine learning

收稿日期: 2024-02-22; **修回日期:** 2024-04-26
基金项目: 国家自然科学基金资助项目(U1936116)
作者简介: 余良钊(1999—),男,硕士研究生. E-mail:2112106067@e.gzhu.edu.cn
* 通信作者. E-mail:864235575@qq.com
引文格式: 余良钊, 李啸宇, 周灿彬. 基于扩散模型和混合样本的成员推理防御方法[J]. 广州大学学报(自然科学版),2024,23(5):76-84.

机器学习在近些年在诸多领域有飞速发展,而支撑它发展的数据有许多是敏感的。在许多敏感领域(如医疗,金融)的机器学习模型部署中存在相当多的隐私问题,这些隐私问题给个人带来了严重的隐私风险。近期的工作表明,机器学习模型常因为记住了敏感的训练数据存在严重的隐私风险,攻击者可以从模型中提取得到个体的详细信息。在成员推理攻击中,攻击者试图知晓是否使用了特定的数据样本来训练目标受害者模型(意即特定的数据样本是否在目标受害者模型的训练集中)。这项攻击广泛存在于各个敏感数据领域中(如图像、医疗数据、交易记录),即使服务提供方仅提供 API 查询接口或者只提供部分的输出预测,攻击者仍然能获取到私人信息。

成员推理攻击对隐私的威胁促使学界在防御成员推理攻击上已经取得一些成果。有两个方向的防御措施[1]:①提供可证明隐私的防御措施,主要使用差分隐私技术为输入数据提供可证明的隐私保护,然而差分隐私技术已经被证明会显著降低诸多机器学习模型的性能;②提供经验性成员隐私的防御措施,隐私保护通过实际的成员推理攻击来进行经验评估。本文的方法属于第二种。

本文的主要贡献包括以下两个方面:

(1)基于模型的损失值分布与模型是否易受成员推理攻击之间存在的强烈联系[2],提出了一种基于扩散模型和混合样本的成员推理攻击防御方法,其中,扩散模型用于生成与训练集同分布的生成数据集,混合样本用于在训练过程中提升模型的泛化性能。

(2)在两个基准数据集上对两种网络架构进行了实验测试,验证了本文方法的有效性。实验结果表明,与当下最先进的成员推理攻击防御措施相比,本文的方法在保护成员隐私和保护模型效用之间提供了最优越的权衡。

1　相关工作

1.1　成员推理攻击

成员推理攻击是一种针对机器学习模型的隐私攻击,在黑盒成员推理攻击中,敌手仅能访问到全体目标数据集 D_{target} 以及目标模型的置信向量 p_{target}[3]。一个完整的成员推理攻击分为 3 个阶段:

(1)准备阶段。敌手在准备阶段将构建与目标数据集同分布的阴影数据集 D_{shadow},同时使用 D_{shadow} 训练出和目标模型同架构的阴影模型。

(2)训练攻击模型阶段。敌手首先确定最能反映出样本成员信息的特征,如模型的预测向量[4]、模型的损失值分布[5]、梯度范数[6],等等;然后,通过阴影数据集和阴影模型提取这些特征的共同模式,常用的方法有通过全连接神经网络、SVM 等来提取。

(3)攻击阶段。给定任意样本,通过访问目标模型,得到该样本的成员信息特征,将此成员信息特征作为攻击模型的输入,由攻击模型判断此样本是否在目标数据集中。

在传统的机器学习中进行成员推理攻击的准确率相当高,可对用户信息安全造成严重的威胁。

1.2　成员推理防御

目前针对成员推理攻击的防御机制主要可分为 3 类[7]:①缓解模型过拟合;②采用对抗性训练来混淆攻击者;③提供严格的差分隐私机制以保障隐私。常用的方法有以下 5 种:

(1)早停法(early stopping):早停法的核心目标在于在模型训练过程中避免过度拟合,从而提高模型的泛化能力。通过及时停止训练,该方法可以防止模型在训练数据上过度拟合,避免模型在测试集上的性能下降。

(2)标签平滑(label smoothing)[8]:标签平滑是一种用于改善深度学习模型训练和泛化性能的技术。其基本思想是在训练过程中,将原始的硬标签(通常为 one-hot 编码)转换为软标签(概率分布),并对其进行平滑处理。这个过程可以通过将原始的标签分布进行平滑化,将标签分布中的极端概率值降低,同时提高其他类别的概率值来实现。

(3)知识蒸馏(knowledge distillation)[9]:知识蒸馏是一种用于提高深度神经网络模型性能和压缩模型大小的技术。其基本思想源自师生网络的概念,即通过将一个复杂的大型模型(教师模型)的知识转移到一个简化的小型模型(学生模型)中,从而实现模型的精简和性能的提升。

(4)采用对抗性训练来混淆攻击者(如 Jia 等[10]和 Nasr 等[11]提出 Adversarial Regularization):此类方法都是建立在对抗性训练的理念之

上的,通常依赖于对攻击模型的强烈假设,使得它们的有效性高度依赖于假设攻击者和真正攻击者之间的相似性,换句话说,如果对攻击模型的假设偏差过大,则防御方法效果极差。

(5)提供严格的差分隐私机制以保障隐私[12]:差分隐私机制提供严格的隐私保障,但是会不可避免地牺牲受保护模型的性能[13]。同时差分隐私机制也会增加额外的计算开销。

1.3 扩散模型

一个扩散模型(diffusion model)由两个参数化马尔可夫链组成,并使用变分推断以在有限时间生成与原始数据分布一致的样本[14]。前向链的作用是扰动数据,它根据预先设计的噪声进度向数据逐渐加入高斯噪声,直到数据的分布趋于简单先验分布,即标准高斯分布。反向链从给定的先验开始并使用参数化的高斯转换核,学习逐步恢复原数据分布[15]。用 $x_0 \sim q(x_0)$ 表示原始数据及其分布,则前向链的分布是可由式(1)和式(2)表达:

$$q(x_1,\cdots,x_T \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \quad (1)$$

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)。 \quad (2)$$

逆向过程的概率分布函数由式(3)表达:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))。 \quad (3)$$

可以将变分下界作为损失函数进行学习:

$$\mathbb{E}[-\log p_\theta(x_0)] \leqslant \mathbb{E}_q[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}] = E_q[-\log p(x_T) - \sum_{t \geqslant 1} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})}]。 \quad (4)$$

运用扩散模型,可以完成图像生成、图像超分辨率和图像编辑等任务。

2 本文方法

本节将详细介绍本文的防御方案。现有的防御方法都是直接使用目标数据集,然而直接使用目标数据集一定会造成隐私上的泄露,本文方法提出通过减少对目标训练集的直接使用,将抹平成员样本与非成员样本在目标模型上的行为差异,从而减少成员推理攻击造成的隐私泄露风险。此外,还使用混合样本技术对目标模型进行训练,以进一步提升其泛化能力。本文方法主要包括生成数据和混合训练两个阶段。

2.1 生成数据

在生成数据阶段,本文首先使用 D_{target} 训练出扩散模型 DM。扩散模型 DM 的作用是生成大量的、高质量的与 D_{target} 来自同一数据分布的图像。使用生成图像可以有效丰富原始数据的多样性,提高目标模型的鲁棒性。对于所有的生成图像,如果不加挑选地将其加入到模型训练,最终会使模型的性能衰减。为了减少生成图像对模型性能的影响,本文额外训练了一个教师模型 $\theta_{teacher}$,用于评估生成的图像。教师模型 $\theta_{teacher}$ 是使用 D_{target} 训练得到的高精度分类器。所有的生成图像都将经过教师模型的筛选与标注。具体流程首先是将所有的生成图像都作为教师模型的输入,将教师模型的输出 $P_{teacher}$(置信向量)作为每个图像对应的软标签,同时根据软标签的最大置信分数(置信向量所有维度中值最大的那一维度的值)与事先设定的阈值比较,只保存高于阈值的图像及其软标签,最终得到生成数据集 D_{gen}。图 1 展示了生成数据集样本配对的过程。图 2 展示了教师模型筛选生成样本的流程。

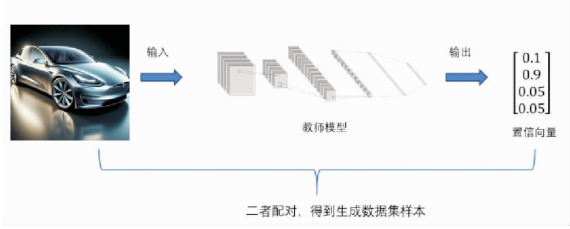


图 1 生成数据集样本配对示意图

Fig. 1 Schematic diagram of generating dataset sample pairings

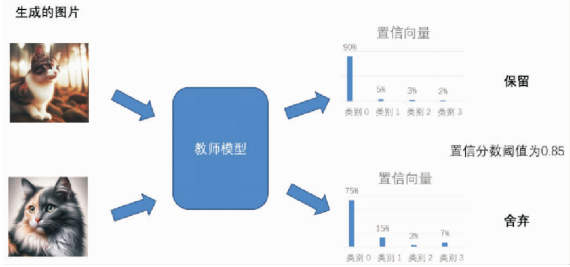


图 2 筛选样本流程图

Fig. 2 Process flowchart for sample filtering

综上所述,生成数据阶段的整体流程如图 3 所示。

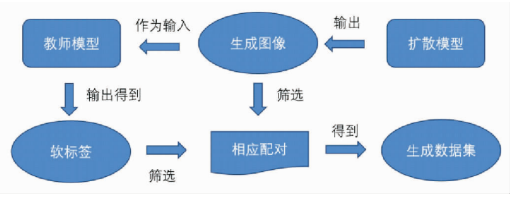


图 3　制作生成数据集流程

Fig. 3　Process for Generating Synthetic Datasets

2.2　混合训练

在得到生成数据集 D_{gen} 之后,将开始混合训练阶段。过去的研究表明,模型在训练样本(成员样本)上与在未训练样本(非成员样本)上的差异表现与模型是否易受成员推理攻击之间存在强烈联系,因此,本文提出不使用全部目标数据集直接训练目标模型。在目标数据集中,对于未直接参与训练的样本,这些样本在目标模型上的表现将更接近于非成员样本,从而进一步混淆敌手。同时,通过使用 D_{gen} 混合训练目标模型,将提高目标模型的泛化能力,缓解目标模型的过拟合现象。

具体的混合训练阶段首先需要从 D_{target} 中随机抽样出一个子集 D_{sub},得到 D_{sub} 后,将 D_{sub} 与 D_{gen} 组合成为一个数据集,按照混合样本技术的训练方法训练目标模型。混合样本技术是一种用于数据增强的技术,其核心思想是在训练过程中将两个不同的样本的特征和标签进行线性混合,生成一个新的样本[16]。混合样本的过程如式(5)所示:

$$\begin{aligned} x &= \lambda x_i + (1-\lambda)x_j, \\ y &= \lambda y_i + (1-\lambda)\, y_j, \end{aligned} \tag{5}$$

其中,x_i 和 x_j 是两个原始的样本,对应的标签分别为 y_i 和 y_j,λ 是一个服从 Beta 分布的随机变量,通常取值在 0 ~ 1 之间。通过这种方式,模型在训练过程中不仅能够学习到单个样本的特征,还能够学习到样本之间的关系,从而提高了泛化能力。混合训练的全部流程如图 4 所示。

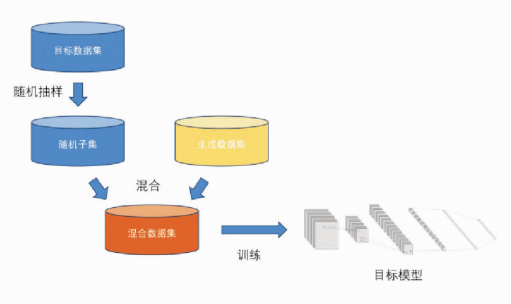


图 4　混合训练流程图

Fig. 4　Mixed Training Process

3　实验评估

本节将首先介绍数据集和实验设置,然后评估本文方法应对多种成员推理攻击的性能。

3.1　数据集和实验设置

3.1.1　数据集说明

本文采用图像识别领域中的两个基准数据集,即 CIFAR10 和 CIFAR100 来进行实验评估。

CIFAR10 数据集包含来自 10 个不同类别的 60 000 张 32 × 32 像素彩色图片,每个类别有 6 000 张图片。这些类别包括飞机、汽车、鸟类等。CIFAR10 通常用于评估图像分类算法的性能。

CIFAR100 数据集包含了来自 100 个不同类别的相同尺寸和数量的图像。每个类别包含 600 张图片,其中包括细粒度的类别,如各种动物、植物、交通工具等。CIFAR100 的数据集结构与 CIFAR10 类似,但类别数量更多,类别之间的区分更加困难,因此,对于评估模型的鲁棒性和泛化能力更有挑战性。

3.1.2　实验设置说明

数据集划分配置。同现有研究保持一致,本文实验方案将 CIFAR10 数据集随机等分为 4 个子集,每个子集之间无交集。分别作为目标模型训练集、目标模型测试集、影子模型训练集和影子模型测试集。每个子集有 15 000 个样本。因为 CIFAR100 数据集与 CIFAR10 数据集具有相同数量的图像,所以 CIFAR100 数据集划分设置与 CIFAR10 数据集划分设置一致。

生成数据集配置。本文使用了两种扩散模型。对于 CIFAR10 数据集而言,本文使用 15 000 张 CIFAR10 目标模型训练集来训练无条件引导的扩散模型。对于 CIFAR100 数据集而言,因为其具有更多的类别,所以本文使用 15 000 张 CIFAR100 目标模型训练集来训练带类别条件引导的扩散模型。这两种扩散模型使用了相同的网络架构,即 UNet。UNet 是一种对称的神经网络框架,分为编码器(encoder)与解码器(decoder),编码器进行卷积操作,而解码器进行上采样操作。对于无条件引导的扩散模型的超参数设置如下:T 即扩散步数,设置为 1 000;$\beta_t \in [10^{-4}, 0.02]$;Dropout 率为 0.15;学习率为 10^{-4};训练迭代次数为 70 000。对于带类别条件引导的扩散模型的超参数设置如

下:T 即扩散步数,设置为 500;$\beta_t \in [10^{-4}, 0.028]$;Dropout 率为 0.15;学习率为 10^{-4};训练迭代次数为 50 000。此外,用于控制类别条件对生成模型影响程度的参数 w 设置为 1.8。

目标模型配置。目标模型使用了两种不同的网络架构,分别是 ResNet18 与 VGG11。阴影模型与教师模型皆使用了相同的网络架构。默认情况下,设置训练迭代次数为 200、学习速率为 0.01。

成员推理攻击配置。为了验证本文防御策略的有效性,在实验方案中设置了多种成员推理攻击:

(1)Salem 等[4]所提出的神经网络攻击模型黑盒攻击(记为 NN Attack)。

(2)Sablayrolles 等[2]所提出的基于损失值阈值的成员推理攻击(记为 Loss Attack)。

(3)Song 等[17]所提出的基于熵阈值和修改熵阈值的成员推理攻击(记为 Entropy attack 和 M-Entropy Attack)。

对比防御方法配置。本文选取了多种防御方法进行对比,所有的防御方法将分别应用至目标模型上。在这些对比方法中,混合样本技术、标签平滑、早停法、Dropout 属于机器学习训练过程中常用的正则化方法,旨在缓解模型的过拟合现象。另外,超参数的选择参考现有研究的设置。

以下是对各防御方法的超参数配置说明:

(1)混合样本技术(Mixup):对于 Beta 分布的 α 参数设置为 0.1、0.2、0.3、0.4 和 0.5。

(2)标签平滑(label-smoothing):在标签平滑中,将原始的离散标签转换为概率分布,ε_{ls} 用于控制其他非正确标签类别的总概率值,$1-\varepsilon_{ls}$ 即为正确标签类别的概率值。ε_{ls} 的取值为 0.1、0.2、0.3、0.4 和 0.5。

(3)早停法:所选取的迭代次数为 5、25、50、75、100 和 125。

(4)Relaxloss:Relaxloss 是 Chen 等[7]提出的一种松弛损失的防御方法,区别于传统训练方法使用梯度下降算法将损失值收敛到 0,Relaxloss 允许损失值在 $(0,\alpha]$ 区间内,具有更容易的学习目标,从而缩小了泛化差距,减少了隐私泄漏。α 的取值为 0.1、0.2、0.3、0.4、0.5、1.0、1.5 和 2.0。α 越大,则具有更强的隐私保护性。

(5)Dropout:Dropout 是一种正则化技术,用于减少神经网络的过拟合现象。在模型的训练过程中,Dropout 随机地将一部分神经元的输出设置为零,从而在每次迭代中随机地丢弃一部分神经元的信息。本文中设置了不同的丢弃概率,分别为 0.1、0.2、0.3、0.4 和 0.5。

评价指标说明。本文从两个方面来评估防御措施的优劣:①模型效用方面,通过目标模型的测试集精确度来衡量;②隐私保护方面,与以往的研究一致,采用攻击模型的 AUC 值(ROC 曲线下面积)来评估。ROC 曲线是一种以假阳率(False Positive Rate, FPR)为横轴,真阳率(True Positive Rate, TPR)为纵轴的曲线。一个完美的防御措施对应的 AUC 值为 0.5。本文期望理想的防御方法是在完美保护隐私的同时做到模型效用不受损失的影响,即在当 AUC 值为 0.5 的同时,保证模型效用相比基线不受损失。

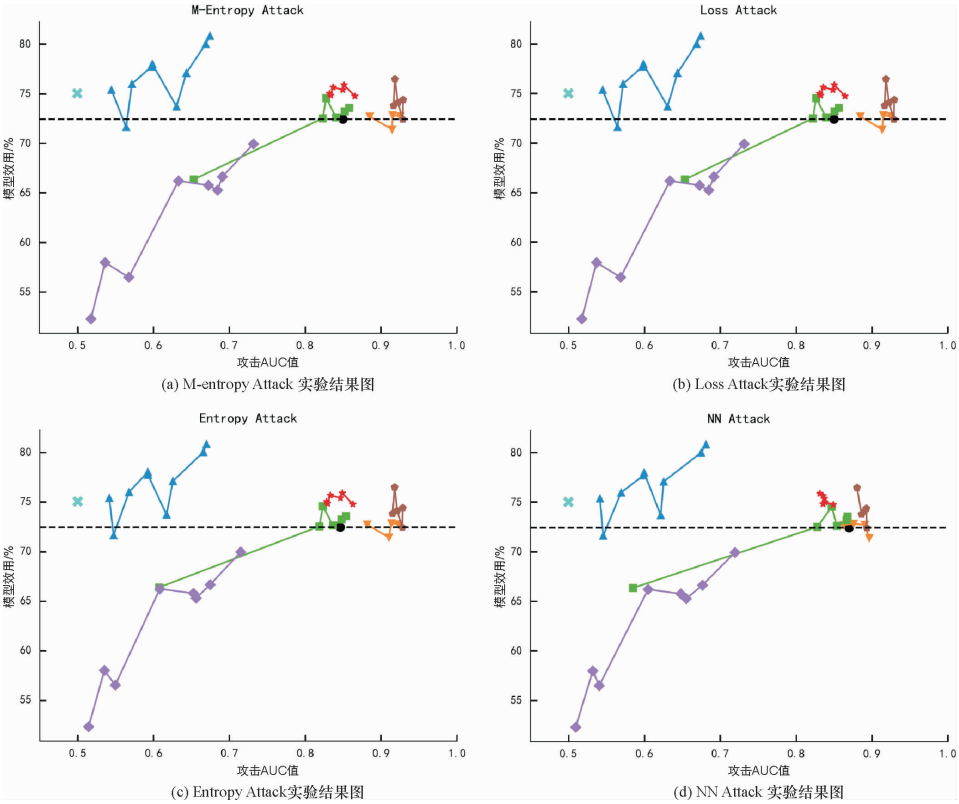
3.2 实验结果说明

本文在 CIFAR10 和 CIFAR100 数据集上对本文方案及对比方案进行攻击实验评估。首先,使用影子模型训练集训练出同架构的影子模型;接着,通过对影子模型最能反映出样本成员信息的特征进行采集,训练出攻击模型;然后,在目标模型训练集上以常规训练方式训练出目标模型作为基线,使用上述对比防御方法以及本文方法分别训练出各种受防御的目标模型;最后,使用目标模型训练集、目标模型测试集和攻击模型对上述模型进行攻击并记录实验结果。

对实验结果(图 5 ~ 图 8)中的符号进行说明:×代表理想防御措施,即在完全防御成员推理攻击的情况下拥有最高的模型性能;●代表未经任何防御的目标模型;⬟代表使用了标签平滑的目标模型;✱代表使用了 Dropout 的目标模型;◆代表使用了 Ralexloss 的目标模型;▼代表使用了混合样本技术的目标模型;■代表使用了早停法的目标模型;▲代表本文的防御方案。图中横轴表示攻击的 AUC 值,其值越接近 1 则代表目标模型的隐私泄露风险越大。纵轴表示目标模型的效用,其值越大则代表模型的可用性越好。平行于横轴的黑色虚线表示未经防御的目标模型效用基线。

在 CIFAR10 数据集上,ResNet18 的实验结果如图 5 所示。

在 CIFAR100 数据集上,ResNet18 的实验结果如图 6 所示。

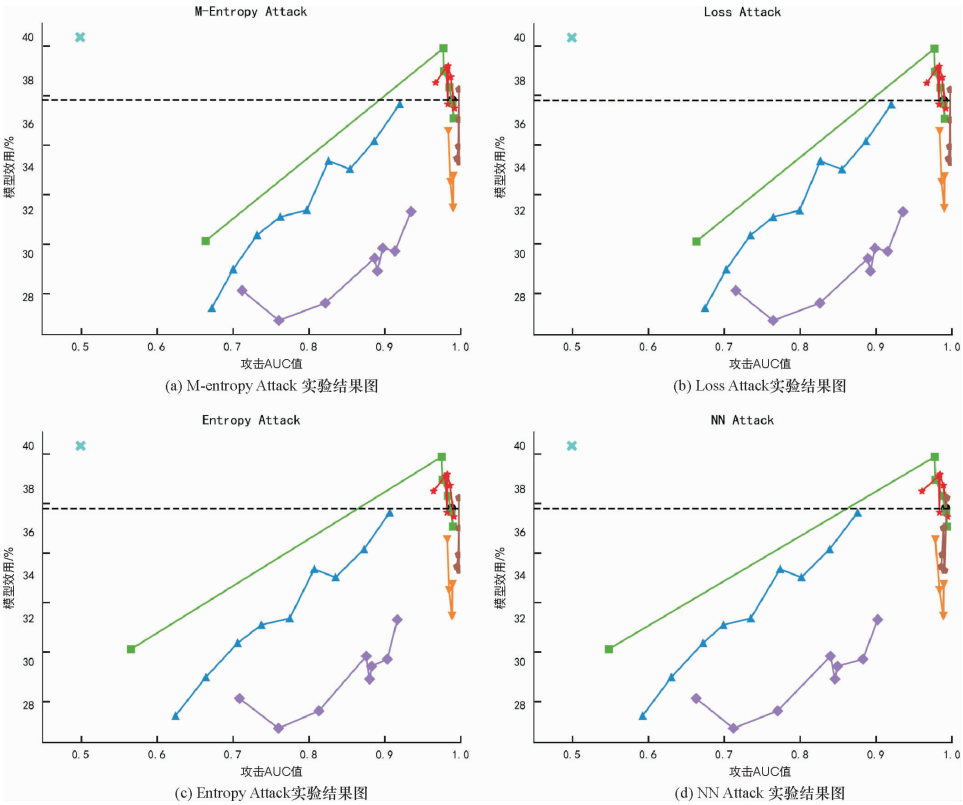


(a) M-entropy Attack 实验结果图　(b) Loss Attack实验结果图

(c) Entropy Attack实验结果图　(d) NN Attack 实验结果图

图 5　ResNet18(CIFAR10 数据集)各攻击实验结果图

Fig. 5　ResNet18 (CIFAR10 dataset) results of various attack experiments



(a) M-entropy Attack 实验结果图　(b) Loss Attack实验结果图

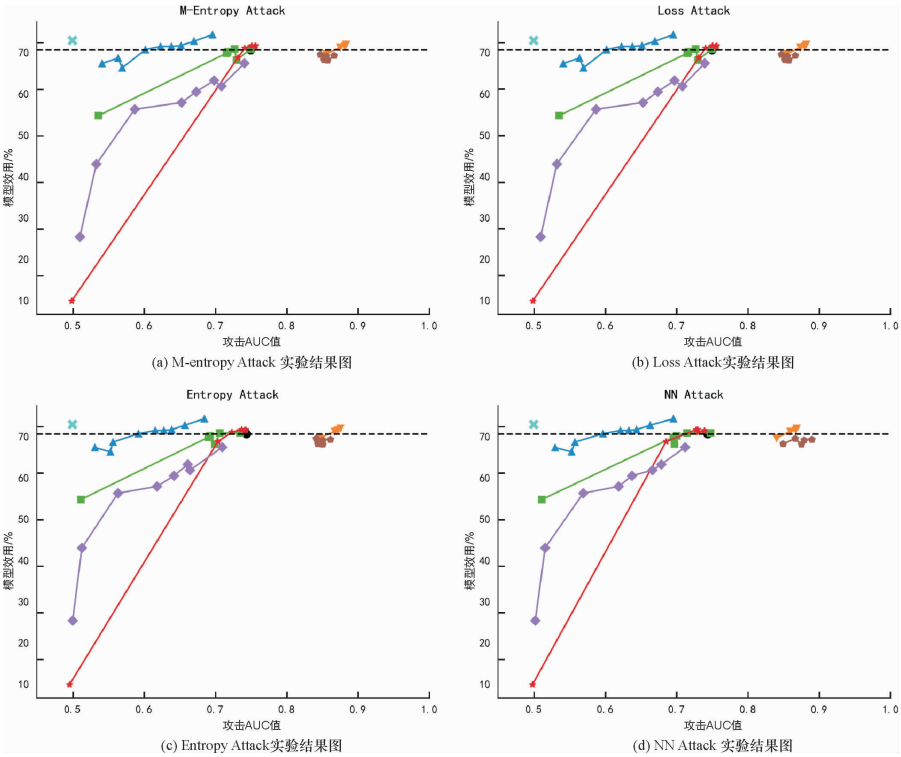
(c) Entropy Attack实验结果图　(d) NN Attack 实验结果图

图 6　ResNet18(CIFAR100 数据集)各攻击实验结果图

Fig. 6　ResNet18 (CIFAR100 dataset) results of various attack experiments

在 CIFAR10 数据集上,VGG 的实验结果如图 7 所示。

在 CIFAR100 数据集上,VGG 的实验结果如图 8 所示。

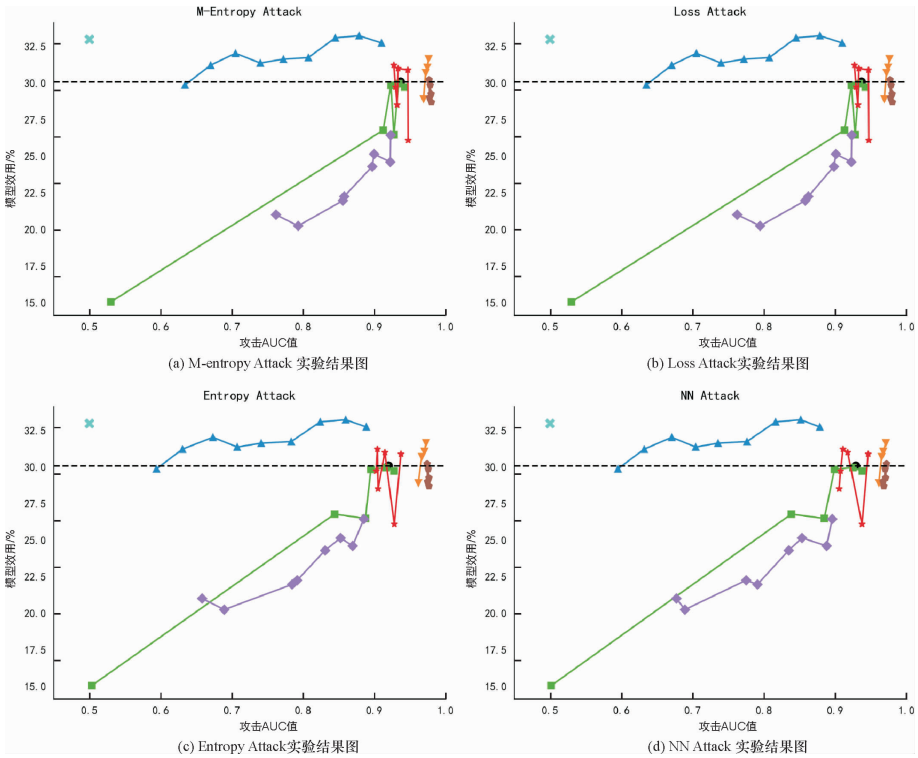


(a) M-entropy Attack 实验结果图 (b) Loss Attack实验结果图

(c) Entropy Attack实验结果图 (d) NN Attack 实验结果图

图 7 VGG11(CIFAR10 数据集)各攻击实验结果图

Fig. 7 VGG11 (CIFAR10 dataset) results of various attack experiments



(a) M-entropy Attack 实验结果图 (b) Loss Attack实验结果图

(c) Entropy Attack实验结果图 (d) NN Attack 实验结果图

图 8 VGG11(CIFAR100 数据集)各攻击实验结果图

Fig. 8 VGG11 (CIFAR100 dataset) results of various attack experiments

从图 5 ~ 图 8 中可以观察到,在 CIFAR10 数据集的不同网络架构上,本文方法在防御 Loss attack、Entropy attack 和 M-Entropy attack 方面均领先于所有对比的现有防御方法。在对比的现有方法中,除了 Relaxloss 和早停法在防范这些攻击时有一定效果,其余方法皆无法很好地防御此类攻击。同时,Relaxloss 和早停法并不能完全保证模型效用,与之相反的,本文方法在防御有效的情况下仍保证了模型效用。

在防御 NN attack 方面,早停法和 Relaxloss 能接近完美防御,但问题仍是无法保证模型效用。而本文方法依旧表现出色,在提升模型效用的同时还保护了隐私安全。以 ResNet18 网络为例,当攻击者 AUC 值被降低到 0.6 以下时,只有本文方法做到了效用提升;以 VGG11 网络为例,当攻击者 AUC 值被降低到 0.6 以下时,只有本文方法保护模型效用不受损失,其他方法都出现了不同程度的效用损失;在 CIFAR100 数据集上可以得到相同的结论。这展现出本文方法在多种网络架构上通用的优势,这意味着其更具灵活性和适用性,从而为对抗成员推理攻击提供了更可靠的解决方案。

表 1 和表 2 展示了 CIFAR10 数据集以及 CIFAR100 数据集上各个模型的泛化间隙(generalization gap),其中包括应用各种防御框架的目标模型,和未经任何防御框架的目标模型。

表 1　CIFAR10 数据集下 ResNet18 的泛化差距

Table 1　Generalization Gap of ResNet18 on the CIFAR10 Dataset　　%

防御方法	普通训练	混合样本	标签平滑	Dropout	Relaxloss	早停法	本文方案
训练集精度	100.0	100.00	100.00	100.00	97.80	100.00	82.48
测试集精度	72.4	72.66	76.43	74.87	69.89	73.47	75.95
泛化差距	27.6	27.34	23.57	25.13	27.91	26.53	6.53

表 2　CIFAR100 数据集下 VGG11 的泛化差距

Table 2　Generalization Gap of VGG11 on the CIFAR100 Dataset　　%

防御方法	普通训练	混合样本	标签平滑	Dropout	Relaxloss	早停法	本文方案
训练集精度	100.0	99.97	99.99	100.00	97.31	99.99	84.01
测试集精度	30.1	28.94	29.89	36.40	26.64	36.99	33.11
泛化差距	69.9	71.03	70.10	63.59	70.85	63.00	50.90

从表中可以观察到,对所有数据集而言,无论哪种网络模型,本文方法都最大程度地减少了泛化间隙。本文方法是从两个方面减小泛化间隙的:

(1)因为在蒸馏训练时仅使用了部分目标数据集,所以受防御目标模型在全体目标数据集上的训练精度下降。

(2)因为在混合训练时额外使用了生成数据集,并且这些生成数据集是经过教师模型挑选的高质量样本。此外,混合训练可以帮助目标模型提升泛化能力,所以应用本文方法的目标模型在测试数据集上的测试精度也有一定幅度的提升。

4　总　结

本文对成员推理防御方法进行了全面探索,首先,提出了基于扩散模型和混合样本的成员推理防御方法,其中,扩散模型用于生成与目标数据集同分布的高质量样本,混合样本技术用于缓解模型的过拟合现象;其次,有效降低了泛化差距,保证了模型在成员数据集上的与在非成员数据集上的表现近似一致;最后,为了进一步混淆敌手,本文方案选择只使用目标数据集的一个随机子集。在未来的工作中,将探索完全不直接使用目标数据集情况下的成员推理防御,在这种情况下,如何保证模型的可用性(模型的分类准确率)将是一个新的挑战。

参考文献:

[1] Tang X Y, Mahloujifar S, Song L W, et al. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture[EB/OL]. (2021-10-15)[2023-10-21]. 2021: arXiv:2110.08324. http://arxiv.org/abs/2110.08324.

[2] Sablayrolles A, Douze M, Ollivier Y, et al. White-box vs black-box: Bayes optimal strategies for membership inference [EB/OL]. (2019-06-09)[2023-11-27]. 2019:arXiv.1908.11229. http://arxiv.org/abs/1908.11229.

[3] Shokri R, Stronati M, Song C Z, et al. 2017 IEEE symposium on security and privacy (SP), May 22-26, 2017[C]. Piscataway: IEEE, 2017.

[4] Salem A, Zhang Y, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models[EB/OL]. (2019-02-14)[2023-11-30]. 2018:arXiv.1806.01246. http://arxiv.org/abs/1806.01246.

[5] Yeom S, Giacomelli I, Fredrikson M, et al. 2018 IEEE 31st Computer Security Foundations Symposium, July 09-12, 2018 [C]. Piscataway: IEEE, 2018.

[6] Rezaei S, Liu X. On the difficulty of membership inference attacks[EB/OL](2020-05-27)[2023-12-05]. 2020:arXiv:2005.13702. http://arxiv.org/abs/2005.13702.

[7] Chen D, Yu N, Fritz M. RelaxLoss: Defending membership inference attacks without losing utility[EB/OL]. (2022-04-25)[2023-12-10]. 2022:arXiv:2207.05801. http://arxiv.org/abs/2207.05801.

[8] Kaya Y, Hong S, Dumitras T. On the effectiveness of regularization against membership inference attacks[EB/OL]. (2020-06-09)[2023-12-12]. 2020:arXiv:2006.05336. http://arxiv.org/abs/2006.05336.

[9] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2023-12-12]. 2015: arXiv:1503.02531. http://arxiv.org/abs/1503.02531.

[10] Jia J Y, Salem A, Backes M, et al. Memguard: Defending against black-box membership inference attacks via adversarial examples[EB/OL]. (2019-09-23)[2023-12-12]. 2019:arXiv:1909.10594. http://arxiv.org/abs/1909.10594.

[11] Nasr M, Shokri R, Houmansadr A. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, October 15, 2018[C]. New York: ACM, 2018.

[12] Dwork C. International Conference on Theory and Applications of Models of Computation, April 25-29, 2008[C]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[13] Kaya Y, Dumitras T. When does data augmentation help with membership inference attacks? [EB/OL]. (2021-07-01)[2024-11-27]. http://proceedings.mlr.press/v139/kaya21a/kaya21a.pdf.

[14] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[EB/OL]. (2021-05-11)[2023-12-12]. 2021:arXiv:2105.05233. http://arxiv.org/abs/2105.05233.

[15] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[EB/OL]. (2020-06-19)[2023-12-13]. 2020:arXiv: 2006.11239. http://arxiv.org/abs/2006.11239.

[16] Zhang H Y, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization[EB/OL]. (2017-10-25)[2023-12-15]. 2017:arXiv:1710.09412. http://arxiv.org/abs/1710.09412.

[17] Song L W, Mittal P. Systematic evaluation of privacy risks of machine learning models[EB/OL]. (2020-03-24)[2023-12-20]. 2020:arXiv:2003.10595. http://arxiv.org/abs/2003.10595.

【责任编辑:孙向荣】