计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(1): 243−254, 2026

面向扩散模型的时序成员推理攻击方法

高志鹏[1] 张 祎[1] 尤玮婧[2] 柴 泽[1] 杨 杨[1] 芮兰兰[1]
[1](网络与交换技术全国重点实验室(北京邮电大学) 北京 100876)
[2](福建师范大学计算机与网络空间安全学院 福州 350117)
(gaozhipeng@bupt.edu.cn)

Temporal Membership Inference Attack Method for Diffusion Models

Gao Zhipeng[1], Zhang Yi[1], You Weijing[2], Chai Ze[1], Yang Yang[1], and Rui Lanlan[1]
[1](*State Key Laboratory of Networking and Switching Technology* (*Beijing University of Posts and Telecommunications*), *Beijing* 100876)
[2](*College of Computer and Cyber Security*, *Fujian Normal University*, *Fuzhou* 350117)

Abstract Diffusion models have gained significant attention in recent years due to their potential in various generation tasks, including image and text generation. However, the widespread use of these models has also raised concerns regarding data privacy, particularly the vulnerability of these models to membership inference attacks (MIA). These attacks aim to determine whether a specific data point is part of the model's training set, posing significant risks to privacy. We provide an overview of the latest developments in privacy protection for diffusion models, with a specific focus on MIAs and their challenges. Existing MIA methods often struggle with balancing the effectiveness of short-term and long-term attacks, and their applicability to diffusion models has not been thoroughly explored. To address these issues, we propose a novel temporal membership inference attack method designed to enhance the attack success rate (ASR) for both short-term and long-term attacks. The proposed method leverages gradient information from noise during short-term attacks and temporal noise patterns to bolster the effectiveness of long-term attacks. Experimental results demonstrate that our method improves the ASR by approximately 5% for short-term attacks and 1% for long-term attacks compared with conventional approaches on common diffusion models. This work contributes to the ongoing efforts to understand and mitigate privacy risks in diffusion model applications.

Key words membership inference attack (MIA); diffusion model; gradient; privacy breach; machine learning

摘 要 扩散模型是一种用于描述信息传播或影响传递过程的数学模型,近年来被越来越多地应用在计算机视觉和自然语言处理等领域的生成式任务中,针对扩散模型的数据隐私攻击与保护也随之得到了广泛关注。成员推理攻击(membership inference attack, MIA)是一种针对机器学习模型的经典攻击,然而,现有的MIA方法实施的深度和广度不足,尤其是在短时攻击和长时攻击的效果平衡上存在挑战。提出了一种新的面向扩散模型的时序成员推理攻击方法(temporal membership inference attack method for diffusion models, TMIA-DM),通过噪声梯度信息保证短时攻击的攻击成功率(attack success rate, ASR),同时采用时序噪声信息来提高长时攻击效果。实验表明,提出的方法在常见扩散模型上短时攻击的ASR提升约5

收稿日期: 2024-08-13; **修回日期**: 2025-04-29
基金项目: 国家自然科学基金项目(62202102, 62372050); 北京市自然科学基金项目(L244010, 4232029); 福建省自然科学基金项目(2024J08162); CCF-绿盟科技"鲲鹏"科研基金项目(CCF-NSFOCUS2024004)
This work was supported by the National Natural Science Foundation of China (62202102,62372050), the Beijing Natural Science Foundation (L244010, 4232029), the Fujian Provincial Natural Science Foundation of China (2024J08162), and the CCF-NSFOCUS "Kunpeng" Research Fund (CCF-NSFOCUS2024004).
通信作者: 尤玮婧(youweijing@fjnu.edu.cn)

个百分点,长时攻击的 ASR 提高约 1 个百分点。

关键词 成员推理攻击;扩散模型;梯度;隐私泄露;机器学习

中图法分类号 TP309

DOI: 10.7544/issn1000-1239.202440687 **CSTR:** 32373.14.issn1000-1239.202440687

扩散模型[1-2](diffusion model, DM)作为一类强大的生成模型,依托生成式模型[3]的发展助力,在多个领域如计算机视觉[4]、自然语言处理[5]、多模态学习[6]、时间序列分析[7]及对抗学习[8]中取得了显著成果。与传统的生成对抗性网络相比,扩散模型采用了一种迭代的生成过程[9],通过逐步增加噪声的方式生成逼真的图像或数据样本。这种迭代的生成方式使得扩散模型在生成复杂、高质量数据时具有很强的稳定性和可控性,也使得扩散模型的架构易受攻击。扩散模型面临的攻击技术主要包括参数反推攻击、梯度泄露攻击和黑盒攻击等多种方式。例如,扩散模型在训练过程中可能通过梯度更新的方式传播训练数据的特征信息。攻击者可以利用这些梯度信息推断出训练数据的分布特征,进而可能推断出训练数据的成员身份或敏感信息。再如,扩散模型的生成过程中可能会通过噪声注入或层间传递的方式传播隐含的信息。攻击者可以通过分析生成过程中的噪声注入模式或层间传递规律,推断出模型的内部结构和参数特征,从而进行隐私侵犯攻击。在敏感信息泄露方面,成员推理攻击能够快速推理模型是否对特定数据点进行了记忆,近年来逐渐流行。

成员推断攻击(membership inference attack, MIA)是一种针对机器学习模型的特定攻击[10],它专门针对模型是否“记住”了训练集中的个别数据点进行推断。MIA 攻击通常通过分析模型输出的概率分布来推断出模型是否对特定数据点进行了记忆[11]。攻击者可能会构建一个针对性的数据集,其中包含目标数据点以及一些无关数据点。然后,攻击者通过向模型输入这些数据集,并观察模型输出的概率分布变化来推断出目标数据点是否被模型记住。同时,MIA 攻击也可以通过分析模型的梯度信息来进行推断。攻击者可能会利用模型的反向传播机制获取模型在某个数据点上的梯度信息,然后根据这些梯度信息推断出该数据点是否被用于模型的训练。另外,攻击者还可以利用模型的输出结果和标签信息来进行推断。通过比较模型对目标数据点的输出结果和真实标签[12],攻击者可以推断出模型对目标数据点的学习程度,从而判断是否存在 MIA 的可能性。MIA 攻击对机器学习模型的隐私和安全构成重大威胁,因为它可以直接暴露模型对训练数据的记忆和学习情况,从而导致隐私泄露和模型安全性问题[13]。

在 MIA 方式中,黑盒攻击(black-box attack)和白盒攻击(white-box attack)是常见的攻击方法,它们分别通过不同的途径来对扩散模型进行攻击[14]。黑盒攻击是一种基于模型输出的攻击方式,攻击者通过观察模型的输出结果来推断模型的内部结构和特性。例如,攻击者可以通过向模型输入不同的噪声样本,并观察模型对这些样本的输出结果来逐步推断出模型的参数设置和生成过程。黑盒攻击的难度在于攻击者无法直接获取模型的内部信息,需要通过观察模型的输出来进行推断。相比之下,白盒攻击则是一种更直接的攻击方式,攻击者可以获取模型的内部结构和参数设置。例如,攻击者可能通过恶意访问或者模型泄露等方式获取模型的详细信息,然后利用这些信息来进行攻击。

灰盒攻击(gray-box attack)假设攻击者对目标模型有部分了解,通常包括模型的结构、部分权重参数或部分输入输出数据,但并不掌握模型的完整信息。由于黑盒攻击训练影子模型的方式会产生较大计算开销,而白盒攻击对模型内部参数细节的需求不适用于大部分开源扩散模型,因此灰盒攻击更适用于针对开源场景的高效攻击。在灰盒攻击中,攻击者能够根据已知的模型结构信息和部分参数,精确设计针对性的攻击策略,缩小攻击空间,从而提高攻击的效率和精度。例如,对于联邦学习或边缘计算等实际应用场景,在对系统进行 MIA 的过程中,攻击者可以掌握模型的架构细节、推断过程,或对部分输入输出数据进行监测,而不需要完整的权重或全部训练数据。在此有限信息共享场景下,灰盒攻击可以帮助分析模型的安全性,系统设计者可以在确保模型功能和性能的同时,减少潜在的隐私泄露风险。

根据 Chang 等人[15]的实验结果,在针对扩散模型的成员推理过程中,当扩散步数较少时,被攻击样本的梯度大幅度变化,随着扩散步数的增加,梯度逐渐平滑,表明梯度信息不适用于长时间的成员推理。根据 Duan 等人[16]的实验结果,尽管基于后验的噪声

相比基于随机噪声的攻击效果有所提升，但仅使用噪声信息进行推理时，短时的攻击成功率接近随机猜测，表明噪声信息不适用于短时的成员推理。为提高成员推理攻击在整个扩散过程中的准确度，本文提出一种基于扩散的时序 MIA 方法。该方法通过提取扩散噪声的时序信息和梯度信息，扩大成员集和非成员集间的差异，进而提升攻击成功率。本文方法采用基于查询的灰盒攻击方式，仅获取扩散模型训练过程中的部分图像，直接针对目标模型进行攻击。这种攻击方式无需重复训练多个影子分类器，能够节省推理时间，同时，由于流行的开源扩散模型提供了访问中间步输出的接口，灰盒攻击能够获得足够信息来进行成员推理。具体来说，本文方法首先获取扩散模型生成过程中某些步的图像，将基于时序迭代的噪声加入前向传播推理过程和反向传播推理过程。其次，在 2 个攻击过程中累计损失梯度信息和时序噪声信息作为推理依据，采用二分类阈值判断方式判别成员样本。最后，本文进行了基于不同数量级的数据集实验和评估，验证了本文方法的有效性和稳健性，并展示了它在传统 MIA 技术上的优越性。

本文的主要贡献包括 3 个方面：

1）针对扩散模型，提出了一种面向扩散模型的时序成员推理攻击方法（temporal membership inference attack method for diffusion models, TMIA-DM），通过逐步计算时序噪声及其梯度信息提取成员集特征；

2）采用基于查询的灰盒攻击方式获取噪声及梯度特征序列，结合基于阈值的二分类法判断成员样本；

3）应用不同数据集的实验结果表明，我们提出的 TMIA-DM 可以在短时和长时情况有效鉴别扩散模型的成员数据样本。

1 相关工作

本节介绍 MIA 的现有研究，将其分为基于影子模型的 MIA 和基于非影子模型的 MIA 这 2 类，并简要介绍每一类的方法。

基于影子的攻击主要学习成员和非成员的预测概率分布的区别[17]，攻击质量取决于影子模型与预测目标之间的相似性。影子模型与目标模型相似性越高，攻击效果就越好。现有的基于影子的攻击方法主要侧重于通过减少攻击所需的知识依赖降低计算消耗。在准备攻击阶段，首先使用影子数据集训练影子模型，再使用影子模型和影子数据集一起训练攻击模型，完成攻击准备。在实施攻击阶段，将目标样本输入影子模型，随后将影子模型的输出作为攻击模型的输入进行攻击。最后，根据攻击模型的输出来判断该样本是否为模型成员。Reza 等人[18] 提出基于深度学习分类任务的 MIA，使用目标模型判别成员和非成员样本输出的预测向量，使用影子模型和分类模型作为攻击模型来学习目标模型的输出，从而判别目标样本。这是一个多影子模型，目标模型数据集和多个影子数据集的分布相同。训练影子模型时，每个影子模型从对应的影子数据集中分别抽取部分数据作为训练集和测试集；训练攻击模型时，判断样本是否来自影子数据集，如果来自影子数据集，则输出标记为成员，如果判断是来自非影子数据集，则输出标记为非成员。由于影子数据集和目标数据集具有相同分布，因此对于相同的样本产生的输出相似，影子模型因此可以产生“模仿”目标模型的效果。然而，多影子攻击不可避免地会造成较大的资源开销，包括内存占用和较长运行时长，这是由于攻击者不仅需要使用多个影子模型，还需要为每个影子模型准备对应的影子数据集。同时，Reza 等人[18] 使用的影子数据集与目标数据集具有相同分布这一条件，对 MIA 在现实世界的扩展产生限制。

为解决上述问题，Ahmed 等人[19] 提出了 3 种优化方式。第 1 种方式使用 1 个影子模型，仍然采用和目标模型相同分布的数据集和模型结构。第 2 种方式采用黑盒攻击，既无法获取目标训练集的数据分布也不知道目标模型的结构，因此采用不同的影子数据集构造多个影子模型，再使用这些影子模型对目标样本进行攻击。第 2 种方式的实验结果表明，在这些影子模型中，采用与目标模型的数据集相同分布的数据进行训练的影子模型，其攻击效果更好。此外，使用 Purchase-10 数据集训练的影子模型对 News 数据集训练的目标模型进行攻击，其攻击成功率较同分布数据集训练的影子模型更高。这一实验结果佐证了利用来自不同领域的数据集训练的影子模型，当攻击生效时，不同的数据集遵循相似的决策边界，可以有效地实施相应的攻击。Liu 等人[20] 使用 GAN 生成和训练集相似分布的方法改进得到影子模型的方式。Stacey 等人[21] 扩展了适用于不同模型的单影子攻击框架。Alexandre 等人[22] 改进影子模型获取信息的方式，使用损失函数代替模型参数进行学习推理，对齐黑盒和白盒攻击的效果。

基于非影子的攻击中，攻击者无法学习预测概率分布等关于模型的详细信息，仅能获取如目标模

型的预测标签等信息，依靠较少的信息实现攻击。现有的基于非影子的攻击方法主要侧重于通过对稀疏信息进行更详细的分析来改进攻击技术，包括基于预测标签 MIA 和基于对比策略的 MIA。

一方面，基于预测标签的 MIA 通常使用阈值作为决策边界判定值，通过比较阈值与数据点到决策边界的距离，如果距离小于阈值则为非成员样本，否则为成员样本。Jia 等人[23]认为训练集中的成员样本具有更高鲁棒性，在攻击过程中非成员样本更容易受到影响而靠近决策边界。Li 等人[24]向目标模型的数据中加入扰动噪声增强其鲁棒性，扩大成员样本与非成员样本的区别。Duan 等人[16]将扰动噪声优化为基于时序的噪声，优化长时攻击过程的准确性。另一方面，基于对比策略的 MIA 可以参考的判断条件有预测后验概率、预测熵、预测损失等，简单决策边界可以快速产生预测结果。Ahmed 等人[19]同时使用后验统计量和预测熵，认为非成员样本的预测结果熵值更高，Samuel 等人[25]结合基于预测标签的策略和对比策略，提出比较阈值 τ 和预测损失，如果预测损失小于阈值则判定为成员样本，与 Ahmed 等人[19]计算后验概率不同，Samuel 等人[25]计算 L2 范数距离。同时，也有方法通过在样本中添加噪声扰动生成对抗样本，再应用对比策略。但是，决策边界维度较高时，后验概率和 L2 范数均无法有效判别样本，Hui 等人[26]提出 BLINDMI 方法，在再生核希尔伯特空间中利用差分比较计算 2 个核空间的距离，差分比较越接近则非成员样本和目标模型越接近。

基于影子模型的攻击存在的问题有：1)影子模型的训练过程需要与目标模型相似的攻击数据和训练参数，多影子模型不可避免地使用额外计算资源；2)获取攻击数据集涉及数据隐私，这些都限制了影子模型的扩展。因此，一些研究提出基于预测标签的 MIA 和基于对比的 MIA，但是基于非影子模型的攻击也存在一些问题：1)决策边界较简单，简单二分类难以适用高维度决策边界场景；2)白盒攻击需要的模型信息较为细致，难以适用于现实场景。随着深度学习的广泛应用，基于生成模型和基于联邦学习的 MIA 也取得一些进展。

在基于生成模型和基于联邦学习的 MIA 方面，Jamie 等人[27]观察到，在 GAN 中鉴别器和生成器在博弈过程中传递的信息可能会导致白盒成员攻击信息泄露，据此针对生成模型的 MIA[28]提出了黑盒 MIA 的蒙特卡洛攻击，以及通过利用重建损失项对 VAE 的重建攻击。Chen 等人[8]提出了针对 GANs 的 MIA 分类法，并提出了一种针对仅具有生成器访问权限攻击的基于优化的方法。Hu 等人[29]扩展 Benjamin 等人[28]的 MC 攻击法，使用 MC 分数划分成员样本和非成员样本。Duan 等人[16]通过在样本中加入噪声扰动生成对抗样本，采用迭代方法来获取特定时间的确定性输出噪声，但这需要更多的查询，从而导致攻击时间变长。随着模型规模的增加，攻击所需时间也随之增加，为了降低时间开销，受 Hu 等人[29]和 Duan 等人[16]的启发，Kong 等人[30]提出了一种近端初始化攻击方法，利用扩散模型在时间点 t=0 时的输出作为噪声。近端初始化攻击是一种基于查询的 MIA，仅依赖于推理结果，不仅可以适用于离散时间扩散模型，还可以适用于连续时间扩散模型。在联邦学习中，客户端易受到用户级 MIA 攻击，客户端向中央服务器上传数据时也会产生数据隐私泄露问题，Hu 等人[31]提出利用源推理攻击窃取中央服务器的成员样本信息。

现有的研究通过探索训练集中数据样本的成员特征，对已有辨别成员信息的方法进行改进。对于影子模型，需要考虑影子数据集的获取及隐私问题。对于非影子模型，现有基于简单决策边界判断成员样本的方法较为简单。本文方法摒弃影子模型带来的开销，使用基于查询的灰盒攻击方法，结合时序噪声及其梯度信息推理，提升成员推理攻击成功率。

2 背景知识

本节形式化定义成员推理过程和扩散过程，并对其进行简要概述。

2.1 成员推理

成员推理(membership inference, MI)是一种隐私推理方法，攻击者通过成员推理确定某个特定的数据点是否在成员集中。本文遵循 Alexandre 等人[22]的假设，将数据集 D 分为成员集 D_{m}和非成员集 D_{n}，$D=D_{\mathrm{m}}\cup D_{\mathrm{n}}$，$D_{\mathrm{m}}\cap D_{\mathrm{n}}=\varnothing$。模型 $f_{\boldsymbol{\theta}}$在成员集 D_{m}上进行训练，$\boldsymbol{\theta}$是参数化的权重。二分类标签 c 表示数据点 $\boldsymbol{x}$ 是否为训练集成员，如果 $\boldsymbol{x}$ 在训练集中，则 c=1；否则 c=0。攻击者不了解数据 $\boldsymbol{x}_i$在 D_{m}还是 D_{n}中，可以通过攻击模型 $f_{\boldsymbol{\theta}}$得到 $\boldsymbol{x}_i$在 D_{m}的概率 $p(c_i=1|\boldsymbol{\theta},\boldsymbol{x}_i)$，再使用二分类方法判断 $\boldsymbol{x}_i$是否在 D_{m}中。通过上述过程，攻击者利用目标模型对成员数据和非成员数据的行为差异进行成员推理，从而判断某个特定的数据点是否在训练集中。

2.2 扩散模型

扩散模型接受输入图像、扩散参数 β_{T}和时间序

列$T=\{t_1,t_2,\cdots,t_K\}$等参数，进行前向传播和反向传播训练。输入图像$\boldsymbol{x}_0$代表初始时刻的数据，扩散参数β_{T}用于控制扩散速率。在前向传播实现高斯扩散过程的模拟阶段，$\boldsymbol{x}_0$经过一系列扩散加噪步骤，每一步都按照高斯分布进行数据变换，直到达到目标时间点t_K，过程中时间点t的图像$\boldsymbol{x}_t$为

$$\boldsymbol{x}_t=\sqrt{\alpha_t}\boldsymbol{x}_0+\sqrt{1-\alpha_t}\boldsymbol{\varepsilon}_t\text{。} \tag{1}$$

噪声$\boldsymbol{\varepsilon}$由UNet模块计算得到。UNet通过编码器-解码器结构在不同尺度下提取图像特征和预测噪声，并在每个时间点调节生成样本的质量。在高斯扩散的前向传播阶段，UNet接收输入图像$\boldsymbol{x}_0$，逐步加入噪声，并在每个时间点t生成噪声预测值。UNet通过编码器模块提取多层次特征，并在解码器阶段利用上采样模块还原分辨率，实现特征重建。从实现角度，UNet的子模块Swish激活函数能够维持模型梯度稳定性和非线性表达能力。其次，时间嵌入(time embedding)模块根据输入的时间点t、模型维度和目标维度，生成包含位置信息编码的时间嵌入矩阵，为UNet提供时间嵌入信息，使得UNet能够“感知”时间点t的变化。此外，UNet还包括注意力机制模块(attnblock)和残差模块(resblock)，attnblock通过注意力权重突出高优先级特征，resblock通过残差连接提升深层特征提取能力。经过UNet的处理，扩散模型在每个时间点产生噪声来预测中间值，生成不断逼近目标数据分布的输出结果。

扩散模型训练过程需要计算每个时间点的损失函数，并根据损失函数的反向传播来调整模型参数，以最小化损失并逼近目标数据分布。

$$\boldsymbol{L}_t(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\varepsilon}_t}\left[\left\|\boldsymbol{\varepsilon}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)-\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|_2^2\right]\text{。} \tag{2}$$

扩散模型使用上述损失函数来衡量模型输出与目标数据之间的差异，通过反向传播算法来计算损失函数关于模型参数的梯度，并根据梯度更新模型参数实现训练过程的迭代优化。

3 面向扩散模型的时序成员推理攻击方法

在本节中，本文首先定义成员推理攻击的威胁模型；其次描述TMIA-DM在扩散模型中获取噪声信息的流程；最后描述TMIA-DM的攻击过程。

3.1 威胁模型

本文提出的威胁模型的相关概念和定义如下。

1)攻击者知识范围

①灰盒访问权限。攻击者可以访问目标扩散模型的输入输出对，能够对模型进行多次查询，并获取模型中间步的输出。攻击者无法直接访问模型的权重参数、完整网络结构或训练数据集。

②时序观测。攻击者可以跟踪模型在不同步长下的响应输出，观测输入样本在扩散过程中逐步去噪的中间状态。通过分析这些输出的变化趋势，攻击者推测样本是否属于训练数据集。

2)攻击目标

识别目标模型的训练数据集中的成员样本。攻击者通过在时序推理过程中分析模型的输出变化来捕捉成员样本在输出稳定性和变化特征上的差异，以此推断出成员关系。

3)攻击方法

①时序噪声信息分析。攻击者通过记录时序噪声对输入样本的扰动程度，并分析模型对这些扰动的响应，来比较模型输出在时间序列中的变化。此方法允许攻击者捕捉样本在模型推理过程中表现出的特定特征以推断其成员身份。

②梯度信息与距离度量。攻击者利用噪声的梯度信息，结合特定的距离范数优化攻击效果。通过分析模型对噪声扰动的敏感性及其梯度变化趋势，攻击者推断输入样本与训练集之间的潜在关系。

4)威胁假设

①攻击者假设模型对成员样本在扩散去噪过程中的输出特征存在相对稳定的表现，且成员样本和非成员样本在噪声注入后的输出变化上有微小但可检测的差异。

②假设扩散模型逐步去噪生成输出，攻击者利用该模型的中间状态输出、噪声信息及其时序梯度来区分成员样本和非成员样本。

5)隐私风险

该攻击揭示扩散模型在灰盒环境下的隐私漏洞。在不完全了解模型内部细节的情况下，攻击者能够通过模型去噪过程中的输出波动，推断出输入样本的成员性。

3.2 时序成员推理方法

TMIA-DM架构如图1所示，核心思想是利用扩散过程来建模数据的生成和变化，通过逐渐添加噪声并逆向消除噪声的方式，实现对数据分布的建模和采样。此外本文在短时攻击过程中加入了时序迭代优化的梯度特征信息并在长时攻击过程中使用时序噪声信息，以在扩散模型整体生成过程中提高攻击成功率。

TMIA-DM特征获取算法通过前向和后向计算，

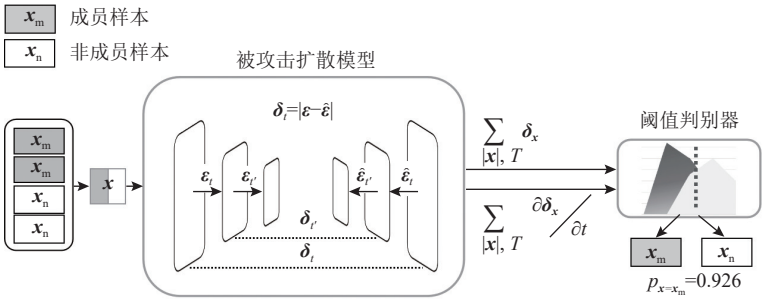


Fig. 1 Flow chart of TMIA-DM

图 1 TMIA-DM 工作流程图

逐步更新和评估样本的状态和损失,并最终通过时序噪声及其梯度信息计算来评估样本是否为成员,从而实现 MIA。

在正向特征计算过程中,算法通过不断更新样本的状态来模拟扩散过程。在每一个时间点 t,算法利用上一个时间点 $\boldsymbol{x}_{t-1}$ 的状态来计算当前时间点 $\boldsymbol{x}_t$ 的状态。然后,通过当前时间点的状态 $\boldsymbol{x}_t$ 计算损失 $\boldsymbol{L}_t$,这些损失将用于后续的反向特征计算。在反向特征计算过程中,算法从时间点 t 开始倒序遍历到时间步长 0。在每一个时间点 t,算法利用当前时间点 $\boldsymbol{x}_t$ 的状态来计算上一个时间点 $\boldsymbol{x}_{t-1}$ 的状态。然后,通过当前时间点的状态计算损失 $\hat{\boldsymbol{L}}_t$,得到损失差 $\boldsymbol{\delta}_t=|\boldsymbol{L}_t-\hat{\boldsymbol{L}}_t|$。最后,利用损失计算梯度 $\boldsymbol{G}_t=\dfrac{\partial\boldsymbol{\delta}_{\boldsymbol{x}}}{\partial t}$,算法将所有时间点 t 的梯度信息和噪声信息汇总得到最终的特征向量,这个特征向量用于评估待推理样本 $\boldsymbol{x}$ 是否为成员样本。

算法 1. TMIA-DM 特征获取算法。

输入:目标模型噪声 $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}$,时间序列 T,待推理样本 $\boldsymbol{x}$;

输出:梯度向量 $\boldsymbol{G}$ 和噪声向量 $\boldsymbol{L}$ 结合的向量。

① 初始化 $\boldsymbol{\varepsilon}_0$;

② for t in T

③ 根据 $\boldsymbol{x}_{t-1}$ 计算 $\boldsymbol{x}_t$;

④ 根据 $\boldsymbol{x}_t$ 计算 $\boldsymbol{L}_t$;

⑤ 根据噪声信息 $\boldsymbol{L}_t$ 计算梯度信息 $\boldsymbol{G}_t$;

⑥ end for

⑦ for reverse(t) in T

⑧ 根据生成的 $\boldsymbol{x}_t$ 逆向计算 $\boldsymbol{x}_{t-1}$;

⑨ 根据 $\boldsymbol{x}_t$ 计算噪声向量 $\boldsymbol{L}'_t$;

⑩ 根据噪声信息 $\boldsymbol{L}'_t$ 计算梯度信息 $\boldsymbol{G}'_t$;

⑪ end for

⑫ $\Delta\boldsymbol{L}=|\boldsymbol{L}_t-\boldsymbol{L}'_t|,\Delta\boldsymbol{G}=|\boldsymbol{G}_t-\boldsymbol{G}'_t|$。

3.3 攻击流程

TMIA-DM 推理攻击可以分为 5 个步骤:第 1 步使用成员数据集 D_{m} 训练目标扩散模型 M_{Diff}。第 2 步从训练集中采样得到成员数据 $\{(\boldsymbol{x}_i,y_i)|\boldsymbol{x}_i\in D_{\mathrm{m}},c_i=1\}$,再从测试集中采样得到非成员数据 $\{(\boldsymbol{x}_j,y_j)|\boldsymbol{x}_j\in D_{\mathrm{n}},c_j=0\}$,其中 y_i 是数据样本 $\boldsymbol{x}_i$ 的标签。第 3 步对于每个数据点 $\boldsymbol{x}_i$,使用目标模型 M_{Diff} 提取特征 p 的预测时序噪声及梯度信息,随后将张量展平并沿着最后 1 个维度求和,该张量包含了目标模型对数据点的内部处理信息,可能会泄露数据点是否在训练集中。第 4 步将成员和非成员的数据点及其对应的特征 p 作为输入,训练攻击模型 M_{Attack}。攻击模型的目标是预测 c: $M_{\mathrm{Attack}}=Train(\{(p_i,1)|\boldsymbol{x}_i\in D_{\mathrm{m}}\}\cup\{(p_j,0)|\boldsymbol{x}_j\in D_{\mathrm{n}}\})$。第 5 步执行成员推理,对于一个新的数据点 $\boldsymbol{x}$,使用目标模型 M_{Diff} 提取特征 $p=f(x)$,使用攻击模型 M_{Attack} 预测 p 对应的成员标签 $\hat{c}=M_{\mathrm{Attack}}(p)$,如果 $\hat{c}=1$,则 $\boldsymbol{x}$ 被预测为成员数据;否则 $\boldsymbol{x}$ 被预测为非成员数据。攻击模型 M_{Attack} 是一个二分类模型,其输出通常是一个概率值 $p_{M_{\mathrm{Attack}}}$,表示数据点 $\boldsymbol{x}$ 为成员数据的概率。攻击模型使用了特征向量的比较方式,对生成数据的特征进行比较,通过计算生成数据之间的 L4 范数距离,判别数据样本之间的差异和相似度。与欧氏距离、曼哈顿距离或闵可夫斯基距离计算方法相比,TMIA-DM 能够简练地表达成员样本和非成员样本特征区域之间的差异。在同时考虑时序噪声和噪声梯度信息的情况下,能够统一预测图像差异,便于推理成员集和非成员集。

4 实 验

本节中,为了验证 TMIA-DM 是否优于针对扩散模型的最新 MIA 攻击,采用对比实验来评估 TMIA-DM 的性能,并通过消融实验讨论范数距离的级数对 TMIA-DM 攻击效果的影响。

4.1 数据集和模型配置

本文在实验中采用 CIFAR-10 数据集,并从中随机抽取 50% 的数据作为训练集。同时,使用了

CIFAR-10 余下的数据以及 CIFAR-100 和 TINY-IN 这 2 个数据集作为测试集，本文对这 2 个数据集也进行 50% 的数据抽样作为测试集数据。由于这 3 个数据集本身已有对应于图片的文本标识，无需使用生成模型对其进行标记。本文还讨论了使用 pokemon 数据集作为成员样本、使用 pokemon, laion, coco 数据集分别作为非成员样本的攻击情况。pokemon 数据集包含了近 800 种口袋妖怪的图片和妖怪名称，laion 数据集包含 10 亿图像和文本对，coco 数据集包含超过 20 万张标注的日常场景图像，是一个被广泛使用的对象检测、分割和图像字幕数据集。超参数设置如表 1 所示。

Table 1 Hyperparameter Setting

表 1 超参数设置

超参数	值
学习率α	0.001
攻击时间 T	20，40，60，80，100
噪声增加速率β	$\beta_1=0.000\,1$，$\beta_T=0.02$

本文在训练被攻击的扩散模型时，采用默认的超参数进行从头训练。对于攻击模型，选用 FCNN 作为主干网络，设置学习率为 0.001，并分别在不同攻击时间点 t 下评估其攻击效果。

4.2 评价指标及对比方法

在评估模型攻击效果时，本文采用 MIA[3] 中所使用的核心指标，包括 ROC 曲线下面积（area under the ROC curve, AUC）以及假阳性率为 0.01% 时的真阳性率（true positive rate at 0.01% false positive rate, TPR@0.01%FPR）。其中，TPR@0.01%FPR 指标表示在极低误报率（0.01%）下模型检测到真实阳性样本的比例，即模型在严苛条件下识别目标的能力。在 ROC 曲线中，该点的表现能够直观反映模型在低误报环境下的实际检测效果。

为验证 TMIA-DM 的有效性，本文将其与 3 种基准模型 MIDM[32], SecMI[16], PIA[30] 进行比较。根据 Duan 等人[16] 的实验结果，传统攻击方法在扩散模型上的效果较普通，因此本实验主要关注基于 MIA 方法。在 MIA 方面，本文选取 SecMI, MIDM, PIA 作为对比模型。以下简要介绍这些基准模型的相关特性：

1）SecM[16]。利用逐步增加的噪声水平来生成反向攻击。通过计算每个步骤的输入值，对中间结果进行去噪处理，通过迭代中间结果获取去噪后的输入值。

2）PIA[30]。通过逆向生成噪声，实现了一种反向推导攻击，根据反向传播的噪声生成攻击。利用中间结果和条件获取去噪后的攻击。

3）MIDM[32]。采用随机噪声生成攻击的方法，在每个步骤中生成随机噪声作为攻击。利用中间结果和条件，获取去噪后的攻击结果。

4.3 实验分析

为验证 TMIA-DM 的有效性，本文在 3 个实际数据集 CIFAR-10, CIFAR-100, TINY-IN 上进行实验，比较 4 个模型的性能。为确保公平性和结果的可靠性，对所有对比方法进行超参数调优。对比方法涉及多个关键参数，主要包括攻击间隔 N 和攻击时间序列 T，具体参数如表 1 所示。在表 2~4 中，本文分别设置攻击时间点为 t=20, t=60, t=100。此外，为更直观地展示改进模型的性能提升，图 2 展示了在 CIFAR-10 数据集上 4 种攻击模型随时间变化的性能表现。表 2~4 包含 4 种攻击方法（MIDM, SecMI, PIA, TMIA-DM）在不同数据集（CIFAR-10, CIFAR-100, TINY-IN）和不同攻击时间点（t=20, t=60, t=100）下的性能指标，主要包括 AUC 和 TPR@0.01%FPR 等。

Table 2 Comparison of Attack Effectiveness of Four Methods on Three Datasets when t=20

表 2 t=20 时4种方法在3个数据集上的攻击效果对比 %

方法	CIFAR-10		CIFAR-100		TINY-IN	
	AUC	T/FPR	AUC	T/FPR	AUC	T/FPR
MIDM	56.54	4.96	57.60	4.61	53.36	6.20
SecMI	49.81	12.73	50.37	15.23	47.34	15.64
PIA	49.24	15.55	50.63	16.99	46.54	15.17
TMIA-DM（本文）	62.45	19.59	59.16	20.58	57.35	20.70

注：T/FPR 是 TPR@0.01%FPR 的简写。

Table 3 Comparison of Attack Effectiveness of Four Methods on Three Datasets when t=60

表 3 t=60 时4种方法在3个数据集上的攻击效果对比 %

方法	CIFAR-10		CIFAR-100		TINY-IN	
	AUC	T/FPR	AUC	T/FPR	AUC	T/FPR
MIDM	71.90	3.02	72.38	2.98	70.22	5.29
SecMI	86.85	10.90	86.34	12.85	83.24	11.17
PIA	76.43	6.17	77.25	6.85	76.69	6.99
TMIA-DM（本文）	87.03	16.38	87.16	16.11	80.15	15.57

注：T/FPR 是 TPR@0.01%FPR 的简写。

MIDM 模型采用随机噪声作为攻击手段，通过引入随机性扰动降低模型鲁棒性。在较短攻击时间

Table 4 Comparison of Attack Effectiveness of Four Methods on Three Datasets when t=100

表 4 t=100 时4种方法在3个数据集上的攻击效果对比 %

方法	CIFAR-10		CIFAR-100		TINY-IN	
	AUC	T/FPR	AUC	T/FPR	AUC	T/FPR
MIDM	80.35	1.49	82.84	1.45	78.32	1.37
SecMI	89.54	1.29	89.48	1.38	88.84	1.32
PIA	87.80	1.03	87.01	1.15	87.57	1.08
TMIA-DM（本文）	90.48	1.16	90.34	1.09	87.50	1.12

注：T/FPR 是 TPR@0.01%FPR 的简写。

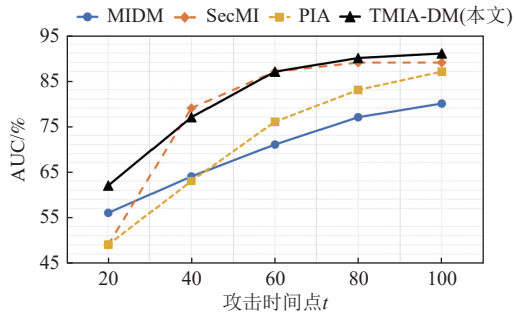


Fig. 2 The trend of AUC over t under four attack methods

图 2 4 种攻击方式下 AUC 随 t 的变化趋势

内，MIDM 在 CIFAR-10 和 CIFAR-100 数据集上表现较优，可能是由于随机噪声在短时间内能引入显著干扰，导致模型判断产生偏差，从而提升 AUC。然而，TINY-IN 数据集的类别数量比 CIFAR-10 和 CIFAR-100 大 20 倍和 2 倍，随机噪声对多样性高的数据集干扰较小，因此随机噪声对 TINY-IN 的影响小于对 CIFAR 数据集的影响。随着攻击时间增加，MIDM 性能稍有提升，后期则开始下降，可能是随机噪声在长时间内的高随机性导致攻击效果减弱。

SecMI 和 PIA 方法在所有数据集上的性能相对较低，尤其是在 TPR@0.01%FPR 方面。SecMI 基于成员推理的攻击方法，利用中间样本的生成和应用，通过调整模型输入使得模型输出更易于推断是否为成员，其技术点主要在于中间样本的生成和选择，以及在推理过程中的判断和优化。SecMI 在较短时间可能仍然处于对抗样本积累阶段，因此 TPR@0.01%FPR 性能不如使用随机噪声的 MIDM。SecMI 在 CIFAR-10 和 CIFAR-100 数据集上表现较好，但在 TINY-IN 数据集上稍逊。PIA 同样需要依赖中间样本的生成和选择，此时性能也较低。随着时间积累，对抗样本的积累相比于 t=20 时更加完善，因此 SecMI 的 TPR@0.01%FPR 指数迅速上升至 10% 以上。与 PIA 相比，SecMI 的噪声生成是基于后验的，而 PIA 的噪声仅与初始噪声相关，因此，在同样的时间内，SecMI 能够更加迅速地提升噪声干扰性能。在攻击时间 t=100 时，SecMI 和 PIA 方法在所有数据集上的性能进一步提升。

TMIA-DM 方法的短时攻击指标优于 MIDM。基于时序噪声的梯度信息，TMIA-DM 可以在短时间内对模型产生较大偏差引导。TMIA-DM 的 TPR@0.01% FPR 结果可以达到 16.38%，相对于 MIDM 模型、SecMI 模型和 PIA 模型，TPR@0.01%FPR 效果分别增长了 13.36 个百分点、5.48 个百分点、10.21 个百分点。随着时间增加，TMIA-DM 利用时序噪声差异进行判断，表现相对稳定，即使是长时间的攻击也依然拥有最佳 AUC 结果，其 TPR@0.01%FPR 表现与其他 3 个攻击模型效果相近，是综合最优模型。进一步分析不同攻击时间下的表现（t=20, t=60, t=100），TMIA-DM 在各时间段的 AUC 均保持在较高水平，这表明 TMIA-DM 在不同攻击时长下均能维持较强的攻击效果和稳定性。较高的 AUC 值说明 TMIA-DM 生成的对抗样本能够有效扰乱模型的判断，使得模型在这些样本上的预测与原始标签产生较大偏差。在低假阳性率（0.01%FPR）下，TMIA-DM 的真阳性率显著提升，表明其对抗样本生成策略具有较高的攻击成功率。较高的 TPR@0.01%FPR 反映了 TMIA-DM 在模型决策上产生了显著影响，使模型更易将对抗样本误判为特定类别。综上所述，TMIA-DM 在 AUC 和 TPR@0.01%FPR 上的优异表现显示了其在攻击效果和稳定性上的显著优势，具有较强的综合性能。

随着时间的增加，4 种攻击方法的 AUC 均呈现出提升的趋势，其中 SecMI 和 TMIA-DM 攻击在 AUC 上的表现相对较好，提升速度较快且达到了较高的水平；而 MIDM 和 PIA 攻击在 AUC 上的提升相对较慢，且最终达到的水平较低。MIDM 采用的是一种较为简单的攻击策略，导致在 AUC 上的提升速度较慢，而且随时间变化的趋势相对平稳。SecMI 采用了更加复杂和高效的攻击策略，可能包含了更多的攻击技巧和特征，因此在 AUC 上的提升速度较快，且趋势相对稳定。PIA 使用了一种中等复杂度的攻击策略，在 AUC 上的提升速度介于 MIDM 和 SecMI 之间，并且随着时间的增加有一定的提升。TMIA-DM 采用了一种高度有效且复杂的攻击策略，其结合时序迭代和梯度度量，因此在 AUC 上的提升速度最快，且趋势呈现稳定增长。TMIA-DM 作为最具优势的攻击方法，兼顾了长时间攻击与短时间攻击的效果，在攻击深度和广度上都有较好的覆盖，能够更全

面地攻击模型，因此提升速度最快。

此外，实验部分还采用了在 pokemon 数据集上进行微调的扩散模型，并使用加入时序噪声及其梯度信息的 TMIA-DM 进行攻击。在此场景中，TPR 值在大于 95% 的情况下都能够达到 1，显示出良好的攻击效果，如图 3 所示。

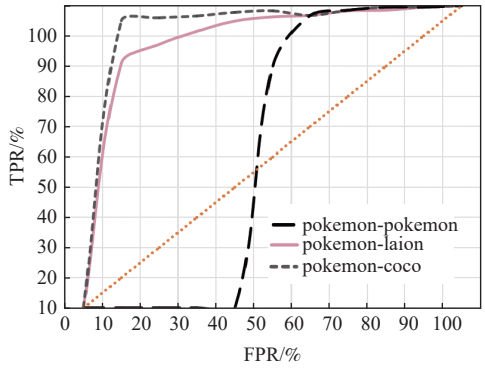


Fig. 3 AUC attack effect maps based on pokemon member samples on three different datasets

图 3 基于 pokemon 成员样本在 3 种不同数据集上的 AUC 攻击效果图

此处主要分析 TMIA-DM 的攻击效果，其 AUC 曲线如图 3 所示，展示了在 pokemon 数据集作为成员样本以及 pokemon, laion, coco 数据集分别作为非成员样本时，TMIA-DM 方法在不同假阳性率(FPR)下的真阳性率(TPR)变化情况。以 pokemon-pokemon 的结果为例，攻击方法的性能随着 FPR 增加而发生变化。随着 FPR 的上升，TPR 也呈上升趋势，显示出 TMIA-DM 方法在较高假阳性率下仍能维持较高的攻击成功率。从图 3 中可以观察到，当假阳性率在 40% 以下时，TPR 的增长比较平缓；而当假阳性率超过 40% 时，TPR 的增长速度明显加快，这可能与 TMIA-DM 方法在更高 FPR 下展现出更强攻击性能有关。特别地，当 FPR 接近 70% 时，TPR 呈现急剧上升趋势，这一拐点可能是 TMIA-DM 方法实现显著攻击效果的阈值点。当 FPR 接近 100% 时，TPR 达到 100%，即实现最高攻击效果。这表明在高 FPR 条件下，TMIA-DM 方法能够准确攻击目标模型。

表 5 和表 6 分别使用准确率(ACC)和 AUC 指标展示不同成员样本与非成员样本组合的攻击效果。当 pokemon 数据集被当作成员样本时，TMIA-DM 能较好地还原其特征，推理结果会更准确，因此 AUC 较高。在此基础上，使用 coco 数据集作为非成员样本时，由于 2 个数据集的分布差异，TMIA-DM 能够以较高成功率区分二者。与使用 pokemon 数据作为成员样本相反，当 coco 数据集作为成员样本时，TMIA-DM 在该数据集上没有经过训练，对其特征的识别能力较弱，难以生成与该数据集一致的特征。在此基础上，使用 pokemon 数据集作为非成员样本时，由于 TMIA-DM 的推理偏向于识别 pokemon 数据集的特征，而 2 个数据集的数据特征差异较大，包括图像的内容、风格、分辨率等，扩散模型的数据集样本特征的潜在空间分布可能不一致。面对 coco 数据集成员样本，TMIA-DM 难以准确还原，从而导致 AUC 极低，接近反向区分的效果。这也意味着 TMIA-DM 在识别 coco 数据集的成员样本时几乎失败，因此 AUC 接近 3.34%。

Table 5 ACC Attack Effect of Combinations of Different Member Samples and Non-Member Samples

表 5 不同成员样本与非成员样本组合的 ACC 攻击效果 %

数据集	非成员样本数据集		
	pokemon	laion	coco
pokemon	50.00	14.36	14.30
laion	85.69	50.00	59.93
coco	85.69	50.05	50.00

Table 6 AUC Attack Effect of Combinations of Different Member Samples and Non-Member Samples

表 6 不同成员样本与非成员样本组合的 AUC 攻击效果 %

数据集	非成员样本数据集		
	pokemon	laion	coco
pokemon	90.94	65.11	96.65
laion	34.88	49.96	56.39
coco	3.34	43.56	50.00

根据表 5 和表 6，当使用同一数据集并采用随机抽样的方式得到成员样本和非成员样本时，得到 ACC 和 AUC 值为 50%，表明 TMIA-DM 在成员推理攻击任务中的表现接近随机猜测。一方面，成员样本和非成员样本本身的分布非常接近，导致 TMIA-DM 在训练时学习到的潜在空间没有明显区分这 2 类样本，在攻击过程中，TMIA-DM 可能无法有效地通过噪声或梯度进行区分。另一方面，攻击方法本身对该数据集的特征不够敏感，或者 TMIA-DM 无法有效地捕捉成员和非成员样本之间的微小差异，那么即使样本来自同一数据集，攻击效果仍然可能较差。最后，pokemon 数据集本身有多种不同的图像类型和风格，即使所有样本都来自同一数据集，成员和非成员之间的差异可能被 TMIA-DM 处理得比较模糊。

总体来说，TMIA-DM 在 AUC 和 TPR@0.01%FPR

等指标上的表现说明其作为一种综合攻击方法,在攻击效果稳定性、对模型决策的影响力等方面具有一定优势,适用于对抗性攻击任务的需求。

4.4 消融实验

本研究进一步探索时序噪声及其梯度信息在成员推理攻击中的贡献,通过消融实验分别检验了二者对长时和短时攻击效果的影响,结果如表 7 所示。首先,在仅使用时序噪声和不使用梯度信息进行推断时,攻击的长时 AUC 略低于完整的 TMIA-DM 方法,短时(t=20)成功率相比 TMIA-DM 低约 4.57 个百分点。这一结果表明梯度信息在短时内能够提升攻击成功率,而在长时攻击中其提升效果较小。其次,在仅使用随机噪声的梯度信息和不使用时序噪声时,攻击的 AUC 值在各时间点均低于 TMIA-DM 方法,在 t=100 时,AUC 仅为 80.33%,较 TMIA-DM 降低了 9.33 个百分点。与此同时,去除时序噪声计算步骤后,攻击效率得到了约 50% 的显著提升。这一现象说明时序噪声虽然在攻击性能上具有重要作用,但在计算效率和性能之间需要权衡。

Table 7 AUC Results of Ablation Experiment on Temporal Noise and Gradient Information

表 7 时序噪声及其梯度信息的消融实验 AUC 结果 %

实验方法	t=20	t=60	t=100
TMIA-DM	62.45	87.52	89.63
TMIA-DM(不使用梯度信息)	57.88	86.58	88.41
TMIA-DM(不使用时序噪声)	57.21	71.92	80.33

此外,本文还讨论了距离计算过程中范数级数 L 的最优选择,具体结果如图 4 所示。在攻击时间 t=100 且查询次数为 10 的情况下,选择 L=4 的范数可在短时间内达到 88% 的攻击成功率,同时其长时间攻击效果略优于其他级数,成功率达到 90%。基于此,本文最终采用 L4 范数来进行推理攻击。

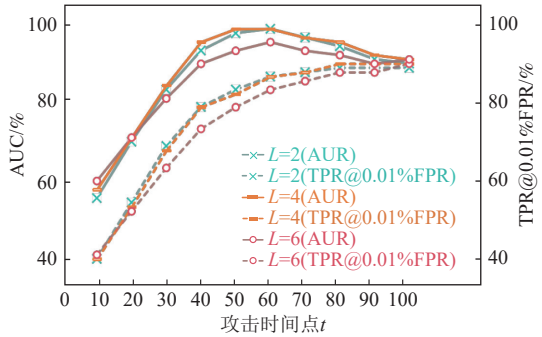


Fig. 4 Ablation study of series of distance norm

图 4 距离范数级数的消融实验

5 总 结

本文针对扩散模型在成员推理攻击方面存在的局限性,提出了一种基于时序的成员推理攻击方法 TMIA-DM。通过对攻击在长时与短时区段的增强,TMIA-DM 方法能够有效地攻击扩散模型,并提高攻击的准确性和成功率。首先,使用基于噪声梯度信息的 TMIA-DM 攻击方法,保证短时攻击的有效性;其次,将时序噪声加入 TMIA-DM 扩散模型和攻击方法,利用扩散模型的特性提高攻击的有效性和准确性;最后通过实验验证了 TMIA-DM 方法在不同数据集上的攻击效果,结果表明 TMIA-DM 相较于传统方法具有更高的攻击成功率和更广泛的适用性。

尽管 TMIA-DM 在各种模型和数据集上显示出出色的攻击性能,但在物理世界中应用时仍存在一些局限性。扩散模型通常作为黑盒 API 服务提供,攻击者访问中间结果的可能性较小,而 MIA 需要访问这些结果来训练扩散模型和攻击模型;TMIA-DM 仅在公共数据上进行评估,尚未调查在特殊数据集如人脸数据集上的有效性和敏感性。为了减少本文可能造成的滥用,所有实验都是在公共数据集和通用模型架构上进行的。

未来可以进一步探究如何增强攻击模型的鲁棒性,使其能够应对更复杂的防御策略和模型结构;可以通过引入对抗性评估指标,对攻击方法的效果进行更全面的评估,包括对抗性鲁棒性和攻击的隐蔽性;可以结合隐私保护技术,探索如何在保证攻击效果的前提下,降低对用户隐私的侵犯程度。同时,针对多模态数据和模型的攻击研究也是未来的发展方向,可以探索跨模态攻击的方法和技术。

作者贡献声明: 高志鹏和张祎提出算法思路、完成实验并撰写论文;尤玮婧对算法思路提出改进意见;高志鹏、尤玮婧、柴泽、杨杨和芮兰兰提出指导意见并修改论文。

参 考 文 献

[1] Jonathan H, Ajay J, Pieter A. Denoising diffusion probabilistic models[C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 6840–6851

[2] Song Yang , Stefano E. Generative modeling by estimating gradients of the data distribution[C/OL]//Proc of the 33rd Advances in Neural

Information Processing Systems. Cambridge, MA: MIT, 2019[2025-04-21]. https://proceedings.neurips.cc/paper_files/paper/2019/hash/30-01ef257407d5a371a96dcd947c7d93-Abstract.html

[3] Song Yang, Jascha S, Diederik P, et al. Score-based generative modeling through stochastic differential equations[J]. arXiv preprint, arXiv: 2011.13456, 2020

[4] Jonathan H, Tim S, Alexey G, et al. Video diffusion models[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022, 8633–8646

[5] Alec R, Jong W, Chris H, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748–8763

[6] Aditya R, Mikhail P, Gabriel G, et al. Zero-shot text-to-image generation[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8821–8831

[7] Kong Zhifeng, Wei Ping, Huang Jiaji, et al. Diffwave: A versatile diffusion model for audio synthesis[J]. arXiv preprint, arXiv: 2009.09761, 2020

[8] Chen Dingfan, Yu Ning, Zhang Yang, et al. Gan-leaks: A taxonomy of membership inference attacks against generative models[C]//Proc of the 2020 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2020: 343–362

[9] Tim S, Jonathan H. Progressive distillation for fast sampling of diffusion models[J]. arXiv preprint, arXiv: 2202.00512, 2022

[10] Wei Lifei, Chen Congcong, Zhang Lei, et al. Security issues and privacy preserving in machine learning[J]. Journal of Computer Research and Development, 2020, 57(10): 2066−2085 (in Chinese)
(魏立斐, 陈聪聪, 张蕾, 等. 机器学习的安全问题及隐私保护[J]. 计算机研究与发展, 2020, 57(10): 2066−2085)

[11] Li Minghui, Jiang Peipei, Wang Qian, et al. Adversarial Attacks and defenses for deep learning models[J]. Journal of Computer Research and Development, 2021, 58(5): 909−926 (in Chinese)
(李明慧, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J]. 计算机研究与发展, 2021, 58(5): 909−926)

[12] Li Zituo, Sun Jianbin, Yang Kewei, et al. A review of adversarial robustness evaluation for image classification[J]. Journal of Computer Research and Development, 2022, 59(10): 2164−2189 (in Chinese)
(李自拓, 孙建彬, 杨克巍, 等. 面向图像分类的对抗鲁棒性评估综述[J]. 计算机研究与发展, 2022, 59(10): 2164−2189)

[13] Jonathan H, Tim S. Classifier free diffusion guidance[J]. arXiv preprint, arXiv: 2207.12598, 2022

[14] Qin Zhen, Zhuang Tianming, Zhu Guosong, et al. Survey of security attack and defense strategies for artificial intelligence model[J]. Journal of Computer Research and Development, 2024, 61(10): 2627−2648 (in Chinese)
(秦臻, 庄添铭, 朱国淞, 等. 面向人工智能模型的安全攻击和防御策略综述[J]. 计算机研究与发展, 2024, 61(10): 2627−2648)

[15] Chang Hongyan, Shamsabadi A S, Katevas K, et al. Context-aware membership inference attacks against pre-trained large language models[J]. arXiv preprint, arXiv: 2409.13745, 2024

[16] Duan Jinhao, Kong Fei, Wang Shiqi, et al. Are diffusion models vulnerable to membership inference attacks?[C]//Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 8717–8730

[17] Zhou Chunyi, Chen Dawei, Wang Shang, et al. Research and challenge of distributed deep learning privacy and security attack[J]. Journal of Computer Research and Development, 2021, 58(5): 927−943 (in Chinese)
(周纯毅, 陈大卫, 王尚, 等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. 计算机研究与发展, 2021, 58(5): 927−943)

[18] Reza S, Marco S, Song Congzheng, et al. Membership inference attacks against machine learning models[C]//Proc of the 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3–18

[19] Ahmed S, Zhang Yang, Mathias H, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models[J]. arXiv preprint, arXiv: 1806.01246, 2018

[20] Liu Gaoyang, Wang Chen, Peng Kai, et al. Socinf: Membership inference attacks on social media health data with machine learning[J]. IEEE Transactions on Computational Social Systems, 2019, 6(5): 907−921

[21] Stacey T, Liu Ling, Mehmet E G, et al. Demystifying membership inference attacks in machine learning as a service[J]. IEEE Transactions on Services Computing, 2019, 14(6): 2073−2089

[22] Alexandre S, Matthijs D, Cordelia S, et al. White-box vs black-box: Bayes optimal strategies for membership inference[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 5558–5567

[23] Jia Jinyuan, Ahmed S, Michael B, et al. Memguard: Defending against black-box membership inference attacks via adversarial examples[C]//Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 259–274

[24] Li Zheng, Zhang Yang. Membership leakage in label-only exposures[C]//Proc of the 2021 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 880–895

[25] Samuel Y, Irene G, Matt F, et al. Privacy risk in machine learning: Analyzing the connection to overfitting[C]//Proc of the 31st IEEE Computer Security Foundations Symp (CSF). Piscataway, NJ: IEEE, 2018: 268–282

[26] Hui Bo, Yang Yuchen, Yuan Haolin, et al. Practical blind membership inference attack via differential comparisons[J]. arXiv preprint, arXiv: 2101.01341, 2021

[27] Jamie H, Luca M, George D, et al. Membership inference attacks against generative models[J]. arXiv preprint, arXiv: 1705.07663, 2017

[28] Benjamin H, Martin H, Daniel B. Monte Carlo and reconstruction membership inference attacks against generative mode[J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(4): 232−249

[29] Hu Hailong, Pang Jun. Membership inference attacks against GANs by leveraging over-representation regions[C]//Proc of the 2021 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 2387–2389

[30] Kong Fei, Duan Jinhao, Ma RuiPeng, et al. An efficient membership inference attack for the diffusion model by proximal initialization[J]. arXiv preprint, arXiv: 2305.18355, 2023

[31] Hu Hongsheng, Zoran S, Sun Lichao, et al. Source inference attacks in federated learning[C]//Proc of the 2021 IEEE Int Conf on Data Mining (ICDM). Piscataway, NJ: IEEE, 2021: 1102–1107

[32] Matsumoto T, Miura T, Yanai N. Membership inference attacks against diffusion models[C]//Proc of the 2023 IEEE Security and Privacy Workshops (SPW). Piscataway, NJ: IEEE, 2023: 77−83



Gao Zhipeng, born in 1980. PhD, professor, PhD supervisor. Director and distinguished member of CCF. His main research interests include block chain, edge intelligence, and federated learning.

高志鹏, 1980 年生。博士，教授，博士生导师。CCF 理事、杰出会员。主要研究方向为区块链、边缘智能、联邦学习。



Zhang Yi, born in 2001. Master. Student member of CCF. Her main research interest includes artificial intelligence security.

张　祎, 2001 年生。硕士。CCF 学生会员。主要研究方向为人工智能安全。



You Weijing, born in 1994. PhD, associate professor, master supervisor. Member of CCF. Her main research interests include AI security, ownership validation on data elements, and cryptographic engineering security.

尤玮婧, 1994 年生。博士，副教授，硕士生导师。CCF 会员。主要研究方向为 AI 安全、数据要素确权、密码工程安全。



Chai Ze, born in 1998. PhD. Student member of CCF. His main research interests include federated learning and block chain.

柴　泽, 1998 年生。博士。CCF 学生会员。主要研究方向为联邦学习、区块链。



Yang Yang, born in 1981. PhD, associate professor, PhD supervisor. Senior member of CCF. Her main research interests include data mining and artificial intelligence.

杨　杨, 1981 年生。博士，副教授，博士生导师。CCF 高级会员。主要研究方向为数据挖掘、人工智能。



Rui Lanlan, born in 1979. PhD, associate professor, PhD supervisor. Her main research interests include network management, mobile network, and edge computing.

芮兰兰, 1979 年生。博士，副教授，博士生导师。主要研究方向为网络管理、移动网络、边缘计算。