中国科学 : 信息科学 2025 年 第 55 卷 第 11 期: 2759–2779
SCIENTIA SINICA Informationis
人工智能应用中的数据安全专刊 · 论文

DMS-MIA: 面向 TEE 保护机器学习模型的磁盘重放与多指标序列成员推理攻击

董一凡[1,2], 冯伟[1], 李昊[1*], 张敏[1], 秦宇[1], 冯登国[1]

1. 中国科学院软件研究所, 北京 100190
2. 中国科学院大学计算机科学与技术学院, 北京 100049
* 通信作者. E-mail: lihao@iscas.ac.cn

收稿日期: 2025–05–30; 修回日期: 2025–08–01; 接受日期: 2025–09–10; 网络出版日期: 2025–11–07

国家重点研发计划 (批准号: 2022YFB4501500, 2022YFB4501503) 资助项目

摘要 随着机器学习即服务 (machine learning as a service, MLaaS) 的广泛应用, 其数据隐私问题日益凸显. 可信执行环境 (trusted execution environment, TEE) 通过硬件隔离为 MLaaS 提供了强大的保护, 尤其在防止模型参数泄露和训练数据泄露方面具有显著优势. 这种隔离机制使得主流的成员推理攻击 (membership inference attack, MIA) 方法面临挑战: 由于攻击者难以直接访问模型内部参数和状态, 且受到查询频率的严格限制, 现有 MIA 方法在 TEE 下效果不佳. 然而, 本文发现并利用了 TEE 保护下 MLaaS 存在的一个新型攻击面, 提出了一种名为 DMS-MIA (disk replay-based multi-metric sequence membership inference attack) 的成员推理攻击方法. 该方法的核心思想在于, 攻击者利用对宿主机及虚拟化管理程序的控制权, 对 TEE 保护下机密虚拟机的加密磁盘进行周期性快照与重放, 获取 MLaaS 模型训练过程中的中间状态的输出. 随后, DMS-MIA 从这些历史输出中构建多维度指标的时间序列, 并采用 Mamba 模型作为攻击模型, 放大序列中的成员关系信号以识别样本的成员身份. 实验结果表明, 在 3 个图像数据集及 2 个非图像数据集上, DMS-MIA 在诸如 TPR @ 0.1% FPR 和 AUC 等多项关键指标上取得了显著效果.

关键词 成员推理攻击, 可信执行环境, 机器学习, 磁盘重放, 多指标序列

1 引言

随着大数据与人工智能的快速发展, 机器学习即服务 (machine learning as a service, MLaaS) 已广泛应用于人们的日常生活. 尽管 MLaaS 为用户带来了极大的便利, 但其在隐私保护方面也面临严峻挑战. 当数据所有者将训练数据上传至服务器进行模型训练时, 攻击者或恶意服务提供方可能直接获取模型参数甚至训练数据, 造成严重的隐私泄露风险. 为应对这一问题, 可信执行环境 (trusted execution environment, TEE), 如 Intel SGX [1,2], AMD SEV [3,4] 和 Intel TDX [5], 提供了基于硬件的安全机制.

引用格式: 董一凡, 冯伟, 李昊, 等. DMS-MIA: 面向 TEE 保护机器学习模型的磁盘重放与多指标序列成员推理攻击. 中国科学: 信息科学, 2025, 55: 2759–2779, doi: 10.1360/SSI-2025-0248
Dong Y F, Feng W, Li H, et al. DMS-MIA: disk replay-based and multi-metric sequence membership inference attack on TEE-protected machine learning models. Sci Sin Inform, 2025, 55: 2759–2779, doi: 10.1360/SSI-2025-0248

TEE 能够在处理器内部构建一个由硬件强制隔离的安全执行环境, 确保模型训练或推理等关键任务在此环境中运行时, 其代码和数据免受包括操作系统在内的高权限软件的访问或篡改. 这种硬件级隔离机制有效抵御了针对模型和数据的外部攻击, 显著降低了模型参数与训练数据被直接窃取的风险. 因此, 在 TEE 保护下进行模型训练与推理, 已成为实现 MLaaS 等远程模型训练与部署场景中的主流安全方案之一 [6~10].

尽管 TEE 能有效防止包括服务提供方在内的攻击者对模型参数和训练数据的直接访问, 但它并不能完全消除所有类型的隐私风险. 即使在 TEE 保护下, 模型仍可能面临来自间接隐私推理攻击的威胁 [11~17]. 其中, 成员推理攻击 (membership inference attacks, MIAs) [15,16,18~26] 是当前最受关注的一类攻击形式. 该类攻击通过分析模型的输入与输出行为, 推断某一特定样本是否属于其训练集. 若该样本曾用于训练, 则称为 “成员样本”; 否则为 “非成员样本”. 一旦攻击成功, 可能直接侵犯数据所有者的隐私. 例如, 当一个模型的训练数据中包含特定疾病的医疗记录时, 攻击者可能通过成员推理攻击判断某人数据是否参与训练, 从而推断其潜在的健康状况. 因此, 成员推理攻击不仅威胁模型训练数据的安全性, 还可能直接侵害个人隐私, 给数据提供者带来严重的隐私泄露风险.

早期的黑盒成员推理攻击主要基于目标模型对输入样本的最终预测输出行为进行推断, 如后验概率分布、预测置信度或损失值等度量指标 [15,19~22,27,28]. 由于成员样本来源于训练数据集, 模型通常对其具有更低的预测损失或更高的正确类别置信度. 为进一步提升攻击效果, 一些研究提出在白盒场景下利用模型对成员和非成员样本在梯度或中间计算结果上的差异来增强攻击性能 [16,24]. 近期研究表明, 模型在训练过程中对成员与非成员样本的损失变化趋势也存在可区分的差异. 例如, Trajectory-MIA [23] 和 SeqMIA [29] 发现, 成员样本的损失下降速度普遍快于非成员样本, 这一特性可用于增强成员身份识别能力, 进而加剧隐私泄露风险. 值得庆幸的是, TEE 的隔离机制与内存加密机制有效阻止了攻击者获取模型参数及其训练过程中的明文信息, 从而在一定程度上防止了白盒场景下的成员关系隐私泄露.

然而, 这并不意味着 TEE 能够天然地降低成员推理攻击的效果. 目前尚缺乏针对 TEE 保护下机器学习模型成员关系隐私泄露的有效分析工作, 导致其真实隐私泄露风险尚未被系统评估. 鉴于此, 本文聚焦于 TEE 下的成员推理攻击问题, 旨在揭示其潜在的隐私威胁. 具体而言, 我们发现 TEE 场景中存在一种由加密的模型中间状态带来的新型攻击面, 并据此提出了一种新的成员推理攻击方法——DMS-MIA (disk replay-based multi-metric sequence membership inference attack). 该方法的核心思想是, 攻击者利用其对虚拟化管理程序的控制权限, 在模型训练期间系统性地对 TEE 保护的机密虚拟机 (confidential virtual machine, CVM) 的加密磁盘执行快照与重放操作. 尽管 TEE 提供了强大的运行时内存保护, 但该手段仍使攻击者能够访问模型在多个历史时间点的中间状态, 并收集其预测输出序列. 通过整合这些历史输出构建多指标序列, 攻击者可以从多个指标维度提取信息, 从而增强对成员身份的辨识能力. 更进一步地, 为高效捕捉这些序列中的复杂时序依赖关系, 我们采用 Mamba [30] 架构来构建攻击模型, 并在 3 个图像数据集与 2 个非图像数据集上验证了 DMS-MIA 的有效性. 本文的主要贡献如下.

(1) 首次揭示了一种针对 TEE 保护下机器学习模型的新型攻击面. 攻击者通过控制宿主机及其虚拟化管理程序, 周期性地对 TEE 加密磁盘执行快照与重放操作, 在不破坏 TEE 运行时内存保护机制的前提下, 获取目标模型在不同训练阶段的预测输出, 从而拓展了成员推理攻击的信息来源.

(2) 构建了一个多维时序特征分析框架, 通过融合损失、熵等多个指标的序列, 捕获更加丰富且具有区分性的成员关系动态信号. 考虑到多指标序列中可能存在的复杂长程时序依赖性, 本文采用 Mamba 模型对这些序列模式进行有效建模, 以实现对成员身份的高精度识别.

(3) 在多种类型的数据集上系统评估了 DMS-MIA 的有效性, 并通过消融实验深入分析了若干攻击效果影响因素, 充分验证了新型攻击面的实际威胁性. 此外, 还评估了主流防御方法在抵御 DMS-MIA 时的效果, 其结果不仅为理解该攻击的内在机制提供了更深视角, 也为 TEE 下的成员推理攻击

研究提供了实证支持与理论拓展.

本文余下内容的结构安排如下. 第 2 节概述相关工作与预备知识. 第 3 节重点阐述所提出的成员推理攻击方法, 包括威胁模型和攻击流程等. 第 4 节展示该方法的实验评估, 包括实验设置、结果对比与分析. 第 5 节讨论相关防御措施. 第 6 节总结全文.

2　相关工作及预备知识

2.1　成员推理攻击

成员推理攻击 [15] 是针对机器学习模型的常见隐私攻击之一, 其攻击目标是判断某一数据样本是否被用于训练目标模型. 我们将用于训练目标模型的数据样本视为成员样本, 而其他样本视为非成员样本, 即考虑一个数据样本 x 是否属于目标机器学习模型 f_θ 的训练集 D_{train}. 攻击者试图构建一个判别器 A, 该判别器接收目标模型 f_θ 和数据样本 x 作为输入, 并输出一个二元决策结果, 如式 (1) 所示:

$$A(f_\theta, x) \to \{1, 0\}, \tag{1}$$

其中, 1 表示 $x \in D_{\text{train}}$, 而 0 表示 $x \notin D_{\text{train}}$.

近年来, 成员推理攻击已在不同的攻击者知识背景下成功实施, 包括白盒攻击 [16,24]、黑盒攻击 [15,19,21,27,28] 和仅标签攻击 [25,26] 等.

具体而言, 早期的代表性工作是 Shokri 等 [15] 和 Salem 等 [19] 提出的影子训练 (shadow training) 技术. 该技术的核心思想是通过训练影子模型来模拟目标模型的预测行为, 并基于影子模型的成员和非成员样本来训练攻击模型. 另一类重要的攻击方法是基于度量的攻击 (metric-based attack). 例如, Song 等 [21] 和 Yeom 等 [27] 提出直接将样本的损失值 (或其他度量指标) 与预定义阈值进行比较来推断成员关系. 通常情况下, 成员样本会比非成员样本表现出更低的损失 (或更优越的其他度量指标). 此外, 仅标签攻击 [25,26] 则专注于在仅能获取预测标签的极端受限条件下, 通过在模型输入中添加扰动来提取成员关系信号.

近年来, 研究者们开始关注如何降低成员推理攻击的高假阳性率 (false positive rate, FPR). Carlini 等 [20] 提出的 LiRA 通过基于似然比的统计检验方法, 在有效控制 FPR 方面取得了显著进展. Bertran 等 [31] 通过分位数回归提出了一种新的攻击方法, 表明在计算开销更低的情况下, 能够达到与 LiRA 相近的攻击效果. Liu 等 [23] 提出的 TrajectoryMIA 开创性地利用了在目标模型训练过程中产生的成员关系信号, 他们收集并分析了模型在不同训练阶段对样本的损失值序列, 认为这种动态的轨迹信息比单一的最终模型状态能更有效地揭示成员身份. 进一步地, Li 等 [29] 提出的 SeqMIA 通过融合多维度的度量指标构建特征序列, 并提取其中的时序依赖关系, 从而有效增强了训练轨迹中成员关系信号的可区分性.

2.2　成员关系度量指标

现有成员推理攻击的成功, 很大程度上归因于机器学习模型固有的过拟合特性. 过拟合是指模型在训练数据上表现得异常良好, 但在未见过的数据上却难以泛化, 导致其在训练集中的预测准确性远高于在其他数据集上的表现. 这种模型在训练集成员样本上表现更好的现象可以通过多种指标进行度量, 从而作为预测一个样本的成员关系的依据. 以下是一些现有成员推理攻击研究中常用的度量指标.

最大值 (max). 机器学习模型的预测输出通常为类别概率分布, 通常以最大概率值所对应的类别作为预测结果. 值得注意的是, 成员样本的预测最大值普遍高于非成员样本, 反映了模型对其训练数据更强的置信度. 这一特性已在成员推理攻击中被广泛利用, 如 Salem 等 [19] 和 Song 等 [21] 所提出的方法, 即通过分析最大值分布差异来识别成员身份.

损失 (Loss). 损失函数用于量化模型预测与真实标签之间的误差, 并作为优化目标指导模型参数更新. 由于模型对其训练数据具有更强的拟合能力, 成员样本的损失值通常显著低于非成员样本. 这一特性已被广泛用于成员推理攻击 [20, 22, 32, 33], 攻击者通过分析损失值即可推断某样本是否属于训练集, 从而引发隐私泄露问题.

标准差 (SD). 标准差是衡量模型输出后验概率分布相对于其均值离散程度的一项指标, 用于量化预测概率的分散性. 相关研究 [19] 表明, 成员样本的标准差通常大于非成员样本. 这是因为模型对成员数据的预测置信度更高, 其概率分布倾向于集中于正确类别, 而其他类别的概率则显著较低, 从而导致概率分布的离散程度更大.

熵 (Entropy). 熵是机器学习领域用以量化模型预测不确定性的重要指标. 较低的熵值表明模型输出的预测概率集中于某一特定类别, 反映出模型对该预测具有较高的确定性与置信度; 反之, 较高的熵值则意味着预测概率分布较为分散, 模型预测的不确定性较大. 相关研究 [15,19,21] 指出, 成员样本的熵值通常显著低于非成员样本. 这一现象揭示了模型在处理其训练数据与未见过数据时, 在预测置信度上存在的差异, 从而为成员推理攻击提供了有效依据.

修正熵 (M-Entropy). 修正熵通过结合模型的预测概率与样本的真实标签, 能够更准确地反映模型在已知真实标签条件下的预测不确定性. 具体而言, 当模型的预测完全正确时, 修正熵为 0; 而当预测完全错误时, 修正熵则为无穷大. 这种定义方式使得修正熵不仅能够衡量模型预测的分散性, 还能敏感地捕捉预测准确性与置信度之间的交互关系, 从而为评估模型性能提供了更为全面的视角. 相关工作 [21] 使用该指标来区分成员和非成员样本, 取得了良好的效果.

2.3　TEE 保护机制

数据隐私保护. 在 TEE 保护下, 深度神经网络的执行过程呈现为黑盒模型, 这意味着模型的内部运行机制, 包括其架构细节、参数权重以及中间计算状态, 对于包括宿主机操作系统在内的高权限软件而言, 在运行时是完全不透明且不可篡改的 [7, 9]. 通过硬件强制的内存加密和精细的访问控制, TEE 将模型计算封装在一个隔离的执行环境中. 此类受保护的模型仅通过明确定义的接口向用户返回最终的预测输出, 而不会泄露额外的中间信息或内部计算细节. 这种设计使攻击者无法获取白盒场景中成员推理攻击所需的隐私信号, 限制了攻击者的能力.

模型保护. 在深度神经网络的安全研究中, 模型窃取攻击是一种典型威胁. 该攻击通过查询目标 DNN 并利用其响应来训练一个功能相似的学生模型. TEE 通过运行时内存加密防止攻击者直接读取或窃取模型参数, 从而有效抵御直接获取模型权重的攻击 [34~37]. 此外, TEE 还能限制对外部查询接口的访问频率, 如限制单位时间内的查询次数或总数. 这种策略显著增加了基于查询的模型窃取攻击难度, 因为大规模查询是实现精确复制所需的关键步骤. 具体而言, 攻击者需要通过大量查询获取足够的标签数据以训练学生模型, 而查询限制使得这一过程难以实现. 因此, TEE 显著降低了模型窃取攻击的成功率和实用性. 然而, 某些 TEE 实现存在安全缺陷. 例如, AMD SEV [3, 4] 由于设计和实现上的不足, 可能遭受密文侧信道攻击, 导致模型权重泄露, 进而增加模型窃取风险 [38]. 相比之下, Intel TDX [5] 等其他主流 TEE 技术依然能有效防御此类攻击, 确保模型安全性.

综上所述, TEE 保护使得机器学习模型仅以黑盒形式对外呈现, 同时严格限制了攻击者对模型的查询次数.

3　成员推理攻击

3.1　威胁模型

首先描述 TEE 中与成员推理攻击相关的保护机制, 并在此基础上界定攻击者的能力范围.

TEE 保护机制. 假设目标机器学习模型运行于基于 TEE 构建的机密虚拟机 (CVM) 中, 其安全保护机制主要包括以下方面.

- **运行时机密性与完整性:** CVM 依赖 TEE 硬件来保障其计算过程的机密性与完整性. 这意味着即使是拥有最高系统权限的宿主机软件 (如 Hypervisor 或主机操作系统内核), 也无法直接读取或恶意篡改 CVM 加密内存中的敏感数据, 包括模型参数、中间激活值以及其他运行时状态信息.
- **安全的加密机制:** TEE 提供了安全的加密算法. 这些算法被用于运行时内存的实时加密/解密, 以及可能的数据密封操作, 以保护持久化状态的机密性和完整性.
- **可信根与硬件安全:** TEE 硬件本身及其固件构成一个可信根, 并被认为是安全的, 无法被攻击者在物理层面或通过软件手段攻破.
- **启动时完整性验证:** TEE 支持对 CVM 的启动过程进行度量与验证, 从而确保加载到 CVM 中的镜像未被篡改. 该镜像通常包含操作系统内核、应用程序以及训练或部署的机器学习模型等内容.
- **加密的持久化存储:** 本文假设 CVM 的虚拟磁盘采用加密方式存储, 以实现持久化数据的机密性保护. 这种加密设计可有效阻止攻击者通过宿主机上的磁盘文件对 CVM 中的机器学习模型参数等敏感数据进行离线解析或篡改.

攻击模型. 攻击者的目标是针对部署于 TEE 保护下的机器学习模型发起成员推理攻击, 以推测某一特定样本是否属于其训练集. 具体而言, 攻击者的能力定义如下.

- **攻击者的知识:** 与现有成员推理攻击研究 [15, 18~23] 的主流假设相同, 假设攻击者知道目标模型训练集的分布, 并能从该分布中采样数据, 得到一个独立同分布的数据集 (本文记为影子数据集). 同时, 攻击者了解目标模型的基本信息, 包括网络架构类型和主要训练超参数, 如学习率和训练轮数以及模型参数检查点的保存间隔等.
- **CVM 控制能力:** 在 TEE 中, 攻击者可能掌握宿主机的全部权限, 成为恶意宿主机 (malicious host), 并对承载 CVM 的物理平台具有完全控制能力. 虽然攻击者无法直接读取或修改 CVM 内部受保护的内容, 但其可操控 Hypervisor、主机操作系统内核及核心虚拟化管理组件, 进而干预 CVM 的启动、加载与持久化存储操作.
- **模型访问能力:** 攻击者能够与运行在 TEE 中的目标模型进行有限交互, 但其行为严格受限于 TEE 所规定的黑盒访问模型, 即攻击者仅能通过 CVM 提供的标准查询接口提交输入样本, 并接收模型返回的预测输出, 无法获取任何中间状态或模型参数. 此外, 出于安全考虑, TEE 对查询接口施加了严格的访问控制策略, 限制了攻击者在单位时间内的查询次数, 防止其通过无限次查询积累更多的成员关系隐私信息.

3.2　攻击方法概述

本小节系统阐述我们的攻击方法 DMS-MIA. 攻击框架如图 1 所示, 该攻击方法利用了 TEE 场景中潜在的攻击面来增强成员关系信号. 具体而言, 该方法主要包含以下 4 个关键阶段.

(1) 影子模型构建. 根据第 3.1 小节所设定的威胁模型, 攻击者已知目标模型的训练数据分布, 因此可以从中采样生成影子数据集, 并据此训练影子模型. 由于影子模型与目标模型均基于独立同分布的数据集进行训练, 二者在知识表征上具有较高相似性, 预测行为亦趋于一致. 此外, 影子模型在训练过程中产生的多个中间状态, 也将用于模拟目标模型的训练演化过程.

(2) 基于磁盘重放的中间状态输出获取. 根据第 3.1 小节中对攻击者控制能力的设定, 攻击者可在目标模型训练过程中周期性地对 CVM 使用的磁盘进行快照, 并以加密形式持久化存储. 为获取模型在不同训练阶段对特定样本的预测输出变化情况, 攻击者可将此前保存的 CVM 磁盘快照重新加载至 TEE 中运行, 并对其中的历史模型版本发起查询. 通过这种方式, 攻击者能够基于磁盘重放机制, 获得目标模型对某一输入样本在训练过程中的中间状态输出.

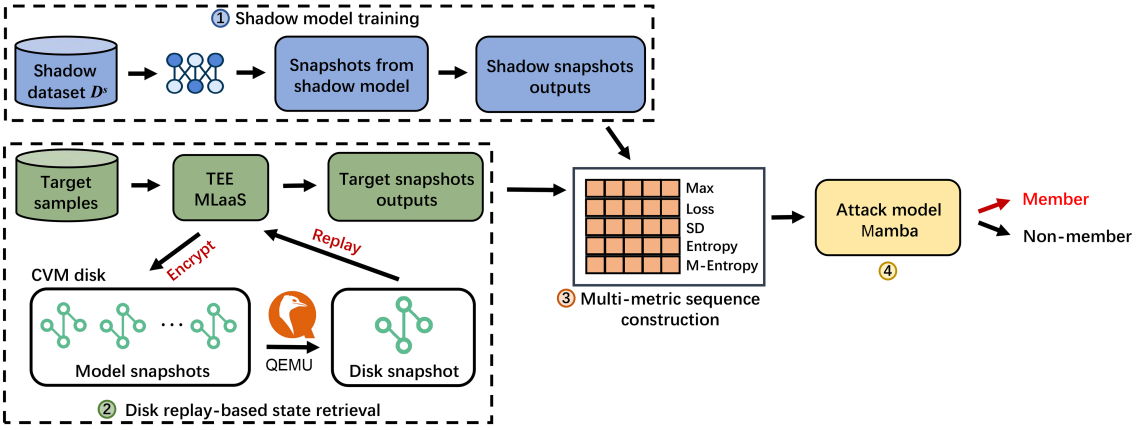

图 1　(网络版彩图) DMS-MIA: 基于磁盘重放的多指标序列成员推理攻击框架.

Figure 1　(Color online) DMS-MIA: disk replay-based multi-metric sequence member inference attack framework.

(3) 多指标序列构建. 在阶段 1 或 2 的基础上, 攻击者能够获取某一特定样本在影子模型或目标模型训练过程中对应的预测输出序列. 进一步地, 攻击者可从这些预测输出中提取第 2.2 小节所述的多种度量指标, 并依据训练阶段对每个指标的序列值进行排序, 从而构建多维度的指标序列, 实现对成员关系特征的有效建模.

(4) 攻击模型构建和成员推理. 攻击者利用影子数据集及其对应的影子模型所生成的多指标序列, 训练用于成员身份判别的攻击模型. 考虑到这些指标序列具有显著的时序依赖特性, 本文采用 Mamba 架构 [30] 作为攻击模型的基础结构. 该模型通过学习成员样本与非成员样本在多指标序列中表现出的特征差异, 最终实现对目标样本是否属于训练集的准确推断.

3.3　阶段 1: 影子模型构建

在影子模型构建阶段, 攻击者基于与目标模型相同的网络架构和训练超参数训练一个影子模型, 以模拟其训练过程. 具体地, 攻击者持有的影子数据集 D^s 被划分为两个互斥子集: 一部分用于模型训练, 记作成员数据集 D^s_{train}; 另一部分不参与训练, 作为非成员数据集 D^s_{test}. 由于影子数据集的划分和影子模型训练全过程均由攻击者控制, 攻击者能够准确标记 D^s 中每个样本的成员状态. D^s_{train} 被标记为成员, 表示其参与了影子模型的训练; D^s_{test} 被标记为非成员, 表示其未被用于训练. 此外, 攻击者还对影子模型训练过程中产生的多个中间状态进行快照并本地保存, 用于后续攻击分析. 一个样本在这些中间模型上所获得的预测输出序列, 构成了该样本的中间状态输出. 对于影子模型而言, 此类中间状态输出是攻击者可以直接获取的.

3.4　阶段 2: 基于磁盘重放的中间状态输出获取

与影子模型不同, 攻击者无法直接获取样本在目标模型上的中间状态输出. 这是由于目标模型运行于 TEE 保护之下, 其训练过程中的中间模型版本均存储于加密内存中, 对外不可见. 为克服这一限制, 攻击者可借助威胁模型中假定的 CVM 控制能力, 利用磁盘快照与重放技术, 从持久化存储中恢复目标模型的不同训练阶段, 从而获取样本在其上的预测输出序列, 即样本在目标模型训练过程中的中间状态输出.

具体地, 如算法 1 所示, 假设攻击者预设了 CVM 的快照时间间隔, 并由此获得一个快照时间点集合 T. 首先, 攻击者执行步骤 1, 启动 CVM 并开始模型训练. 在训练过程中, 执行步骤 2~6, 对 CVM 使用的磁盘进行周期性快照. 当对目标样本集合 Q 发起成员推理攻击时, 执行步骤 7~16, 依次加载各快照并查询模型, 以获得每个样本在不同训练阶段的预测输出, 进而构建其完整的中间状态输出序列.

算法 1 基于磁盘重放的中间状态输出获取算法.

输入: 初始加密虚拟磁盘文件 DS_0, 训练任务的快照时间点集合 $T = \{t_1, t_2, \ldots, t_k\}$, 目标查询样本集合 Q;
输出: 每个查询样本 q 在不同快照时间点 $t_1, t_2, \ldots, t_k$ 下对应的模型预测输出序列 $\{o_{1,q}, o_{2,q}, \ldots, o_{k,q}\}$;

使用加密虚拟磁盘 DS_0 启动 CVM 并正常执行训练任务;
for 每个快照时间点 $t_i \in T$ **do**
    暂停 CVM 运行;
    利用虚拟化管理接口执行磁盘快照操作, 生成加密虚拟磁盘快照文件 DS_i;
    恢复 CVM 运行;
end for
for 每个快照磁盘 DS_i **do**
    暂停当前运行的 CVM;
    将正在使用的虚拟磁盘替换为快照磁盘 DS_i;
    使用快照磁盘 DS_i 重新启动或恢复 CVM;
    for 每个查询样本 $q \in Q$ **do**
        通过 CVM 标准查询接口向模型输入样本 q;
        获取模型输出 $o_{i,q}$ 并存储;
    end for
end for
组成每个查询样本 q 的输出序列 $\{o_{1,q}, o_{2,q}, \ldots, o_{k,q}\}$.

需要指出的是, 目前对整个 CVM (包括运行时内存、寄存器状态等) 进行完整快照仍存在较大挑战. 因此, 本方法采用 QEMU [39] 提供的管理功能, 仅对 CVM 的虚拟磁盘进行快照与重放. 具体而言, 管理功能基于 QEMU Monitor[1)] 实现, 通过 stop 命令暂停 CVM 的执行, snapshot_blkdev 命令备份当前的磁盘快照, cont 命令恢复 CVM 的执行. 重复该过程可以备份多个磁盘快照 DS_i, 并使用 qemu-system-x86_64-drive file = DS_i 对中间状态进行重放. 由于模型训练是一个计算密集型过程, 通常训练框架会在训练过程中定期将模型参数备份至磁盘, 以防止因断电等因素导致训练中断. 根据威胁模型设定, 攻击者已知模型参数的备份时间间隔. 因此, 攻击者可根据该信息选择适当的快照时间点集合 T, 通过对加密磁盘的快照与重放, 实现对目标模型中间状态的有效恢复. 最后, 由于影子模型的训练过程完全由攻击者控制, 攻击者可以灵活调整其快照时机, 使得影子模型的中间状态与目标模型保持对齐. 这种对齐旨在确保攻击模型在影子模型的成员与非成员样本之间所学习到的中间状态输出差异, 能够与目标模型上的实际差异保持一致, 从而提升攻击模型的泛化能力和推理准确性.

3.5　阶段 3: 多指标序列构建

在阶段 1 的基础上, 攻击者获得了一个样本在影子模型训练过程中 k 个快照点上的输出后验概率向量 $o_1, o_2, \ldots, o_k$. 随后, 针对每个输出后验概率, 攻击者计算 m 种不同的度量指标, 如最大值 (Max)、损失 (Loss)、标准差 (SD)、熵 (Entropy) 等. 每种度量指标对应的 k 个取值按时间顺序排列, 形成一个单一的指标序列. 更进一步地, 所有 m 个指标序列被整合为一个 $m \times k$ 的矩阵, 称为该样本的多指标序列. 它刻画了该样本在影子模型训练过程中多个度量指标维度上的动态变化特征. 类似地, 攻击者对影子数据集 D_s 中的每一个样本重复上述过程, 生成相应的多指标序列, 并根据其真实成员状态 (成员或非成员) 进行标签标注, 作为后续攻击模型训练的输入样本. 由于各个指标序列彼此独立, 该构建方法具有良好的可扩展性, 可灵活纳入新的度量指标以适应不同的场景.

3.6　阶段 4: 攻击模型构建和成员推理

多指标序列蕴含了模型在训练过程中多个阶段的多种度量指标变化趋势, 但这些信息仍需通过高效的建模方法进行时序模式提取与特征分析. 为更有效地捕捉其中的时序依赖关系, DMS-MIA 采用

1) https://www.qemu.org/docs/master/system/monitor.html.

了一种高效的序列建模架构——Mamba [30]. 该模型能够高效处理具有时序特性的多维序列数据, 并具备强大的模式学习与表征能力. 具体而言, 攻击者利用影子模型生成并标注的成员与非成员样本的多指标序列集合, 作为训练数据输入至 Mamba 模型中. 在训练过程中, 采用二分类交叉熵损失函数对模型参数进行优化. 在完成训练后, 攻击者可将目标样本的多指标序列输入该攻击模型, 从而获得其成员身份的预测结果.

4 实验

本节在 3 个图像数据集和 2 个非图像数据集上评估 DMS-MIA 的性能, 并与现有攻击工作进行对比. 此外, 本节还通过消融实验探究了影响攻击性能的关键因素.

4.1 实验设置

4.1.1 TEE 配置

实验采用了 AMD SEV [3,4] 服务器来训练和部署目标模型. 该服务器搭载了 AMD EPYC 9745 128 核处理器. 在软件层面, Host 操作系统为 Ubuntu 24.04.2 LTS, 搭载 Linux 内核版本为 6.11.0+. 采用的 QEMU 版本为 9.1.50, 并结合内核自带的 KVM (kernel-based virtual machine) 模块, 作为底层的虚拟化管理程序来创建和管理 SEV-SNP CVM. CVM 内部配置为运行 Ubuntu 24.04 LTS 操作系统, 其内核版本与宿主机保持一致.

4.1.2 数据集

本文实验采用了 3 个图像数据集, 包括 CIFAR10 [40], CIFAR100 [40] 和 CINIC10 [41], 以及 2 个非图像数据集, 分别为 Purchase[2)]和 Location[3)]. 这些数据集被广泛应用于现有成员推理攻击 [20~23,29] 的性能评估中.

CIFAR10 和 CIFAR100 数据集. 这 2 个图像数据集均包含 6 万张 32×32 彩色图像. 其中, CIFAR10 含 10 个类别 (如飞机、汽车、鸟类等), 每类 6000 张图像, 类别平衡; CIFAR100 则包含 100 个更细粒度的类别.

CINIC10 数据集. 该数据集图像来源于 ImageNet [42] 和 CIFAR10, 总计 27 万张 32×32 像素的彩色图像, 旨在提供更具挑战性和泛化性的图像分类任务基线.

Purchase 数据集. 该数据集源于 Kaggle 的 Acquire Valued Shopper 挑战赛, 包含了 197324 条具有 600 个特征维度的购物记录. 与先前成员推理攻击研究 [15,19,29] 一致, 我们将其聚类为 100 个类别, 专门用于评估非图像分类模型的成员推理攻击性能.

Location 数据集. 该数据集源自 Foursquare [43], 包含大量用户的签到记录. 在 Shokri 等 [15] 的成员推理攻击研究中, 数据集被处理为包含 5010 个样本, 每个样本具有 446 维特征, 并分属于 30 个不同的类别.

参照现有成员推理攻击研究 [23,29], 将每个数据集划分为 4 个互不相交的部分, 以支持目标模型和影子模型的训练与评估: 目标模型的训练集 D^t_{train}、目标模型的测试集 D^t_{test}、影子模型的训练集 D^s_{train} 和影子模型的测试集 D^s_{test}. 每部分的数据量如表 1 所示.

4.1.3 模型

对于 CIFAR10, CIFAR100 和 CINIC10 图像数据集, 采用了 3 种广泛使用的卷积神经网络模型: ResNet-56 [44], VGG16 [45] 和 MobileNetV2 [46] . 对于 Location 和 Purchase 非图像数据集, 采用了一个

2) https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data.
3) https://sites.google.com/site/yangdingqi/home/foursquare-dataset.

表 1　目标模型与影子模型的数据集划分大小.

Table 1　Dataset partition sizes for target and shadow models.

Dataset	D^t_{train}	D^t_{test}	D^s_{train}	D^s_{test}
CIFAR10	10000	10000	10000	10000
CIFAR100	10000	10000	10000	10000
CINIC10	10000	10000	10000	10000
Purchase	20000	20000	20000	20000
Location	800	800	800	800

表 2　各目标模型的训练/测试准确率.

Table 2　Training/testing accuracy of each target model.

Target model	CIFAR10	CIFAR100	CINIC10	Purchase	Location
ResNet-56	0.981/0.661	0.993/0.249	0.938/0.455	–	–
VGG-16	1.000/0.753	1.000/0.292	0.999/0.564	–	–
MobileNetV2	0.983/0.672	0.882/0.182	0.956/0.451	–	–
MLPs	–	–	–	1.000/0.707	1.000/0.549

标准的 2 层多层感知机 (2-layer MLP). 在所有的实验中, 目标模型及其对应的影子模型均使用相同的架构, 并统一训练了 100 个周期. 各目标模型的性能如表 2 所示.

4.1.4　DMS-MIA 默认设置

在本研究中, DMS-MIA 攻击模型的输入由模型训练过程中多个快照的历史输出序列构成. 为了从这些序列中提取有效的成员关系信号, 我们在实验中使用 5 种常用的度量指标, 分别为最大值 (Max)、损失 (Loss)、标准差 (SD)、熵 (Entropy) 和修正熵 (M-Entropy), 其构建细节见第 2.2 小节.

4.1.5　基线方法

为验证 DMS-MIA 的有效性, 我们将其与以下成员推理攻击方法进行了对比. 这些基线方法涵盖了当前主流的攻击范式, 并包括了目前性能最先进的攻击技术.

首先, 是基于影子训练的经典方法 (shadow training) [15, 19], 其通过训练影子模型模拟目标模型的预测行为, 并利用影子模型的输出训练攻击模型以区分成员与非成员样本.

其次, 是基于度量指标的攻击方法 (metric-based attack) [21], 该方法仅依赖于目标模型的预测输出, 计算损失值、置信度或熵等度量指标, 作为成员身份判别的依据. 本文实验采用了其中的熵 (Entropy) 和修正熵 (M-Entropy) 两个度量指标来度量成员关系, 分别记为 MBA(Entropy) 和 MBA(M-Entropy).

此外, 我们还引入了近期通过知识蒸馏 (knowledge distillation) [47] 来模拟目标模型训练轨迹的先进方法: TrajectoryMIA [23] 和 SeqMIA [29], 它们试图用蒸馏技术模拟目标模型的训练过程, 并捕捉模型在训练过程中对成员样本的动态学习特征.

通过与上述多种类型的攻击方法进行比较, 可以更全面地展现 DMS-MIA 在 TEE 下所具有的优势.

4.1.6　评估指标

为了全面细致地评估本文提出的 DMS-MIA 方法以及各基线方法的性能, 采用了在成员推理攻击研究 [20, 22, 23, 29, 48] 中广泛使用的评估指标. 重点关注在极低假阳性率下的真阳性率 (TPR @ low FPR), 该指标反映了攻击者在严格控制误报的条件下识别真实成员的能力, 对评估成员推理攻击的实际威胁尤为重要. 此外, 平衡准确率 (balanced accuracy) 定义为成员推理攻击在一个成员与非成员样

表 3　在 3 个图像数据集上训练的 ResNet-56 模型的多种攻击效果评估.

Table 3　Performance of various attacks on ResNet-56 trained using three different image datasets.

MIA method	TPR @ 0.1% FPR (%)			Balanced accuracy			AUC		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.10	0.16	0.11	0.642	0.784	0.630	0.657	0.844	0.659
MBA(Entropy)	0.16	0.25	0.15	0.642	0.761	0.600	0.642	0.761	0.600
MBA(M-Entropy)	0.17	0.42	0.20	0.687	0.868	0.711	0.687	0.868	0.711
TrajectoryMIA	0.45	0.56	0.30	0.639	0.852	0.690	0.704	0.911	0.745
SeqMIA	3.93	20.53	5.80	0.712	0.897	0.776	0.803	0.965	0.863
DMS-MIA	**5.70**	**23.11**	**8.27**	**0.736**	**0.931**	**0.852**	**0.841**	**0.977**	**0.924**
MIA method	Precision			Recall			F1-score		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.581	0.694	0.603	0.897	0.936	0.746	0.705	0.797	0.667
MBA(Entropy)	0.609	0.712	0.594	0.792	0.878	0.635	0.688	0.786	0.614
MBA(M-Entropy)	0.631	0.807	0.661	**0.901**	0.968	0.866	0.742	0.880	0.750
TrajectoryMIA	0.630	0.824	0.691	0.653	0.891	0.681	0.641	0.856	0.686
SeqMIA	0.678	0.838	0.739	0.807	0.983	0.852	0.737	0.905	0.792
DMS-MIA	**0.725**	**0.888**	**0.793**	0.761	**0.986**	**0.954**	**0.743**	**0.934**	**0.866**

本数量均衡的数据集上作出正确预测的概率. ROC 曲线下面积 (area under the ROC curve, AUC) 则衡量攻击模型在所有可能阈值下的整体区分能力, 是评价攻击性能的综合性指标.

此外, 我们还引入了机器学习领域二分类任务中常用的评估指标. 精确率 (Precision) 定义为被预测为成员的样本中真实成员的比例, 反映了攻击结果的可信度; 召回率 (Recall) 表示所有真实成员样本中被正确识别的比例, 体现攻击的覆盖能力; F1 分数 (F1-score) 作为精确率与召回率的调和平均, 综合评价模型的整体性能.

4.2　攻击性能

本小节旨在全面评估本文所提 DMS-MIA 方法在不同数据集和模型架构下的攻击性能, 并与基线方法进行对比.

图像数据集. 表 3 展示了在 CIFAR10, CIFAR100 和 CINIC10 3 个图像数据集上, 针对 ResNet-56 目标模型, DMS-MIA 及其他主流成员推理攻击方法的性能对比. 更多实验结果见附录中的表 A1 和 A2. 本文表格中, 黑体数值表示最优结果, 带下划线数值表示次优结果.

从关键指标 TPR @ 0.1% FPR 来看, DMS-MIA 在 3 个数据集上均取得最优表现. 以 CIFAR100 为例, 其 TPR 达到 23.11%, 较次优方法 SeqMIA 提升约 12.57%. 这表明 DMS-MIA 在严格控制误报的前提下, 具备最优的识别成员样本的实际攻击能力. 而在 Balanced accuracy 和 AUC 指标上, DMS-MIA 同样表现最佳. 例如, 在 CINIC10 上, 其 Balanced accuracy 为 0.852, 较次优方法 SeqMIA 提升约 9.79%; AUC 达到 0.924, 相较次优提升约 7.07%, 这说明 DMS-MIA 对整个评估数据集的综合判别能力最优. 更进一步地, DMS-MIA 多数情况下可同时实现较高的 Precision 与 Recall, 从而获得最优的 F1-score (如 CINIC10 上达到 0.866). 尽管部分方法 (如 MBA(M-Entropy) 在 CIFAR10 上) 可能在召回率上接近甚至略高, 但其精确率偏低, 导致假阳性增加. 而 DMS-MIA 在保持高召回的同时实现了更高精度, 体现出攻击有效性与可靠性的良好平衡.

非图像数据集. 表 4 展示了 DMS-MIA 在非图像数据集上的成员推理性能. 实验表明, DMS-MIA 在 Location 和 Purchase 数据集上均优于现有方法. 以 TPR @ 0.1% FPR 为例, DMS-MIA 在 Location

表 4　在 2 个非图像数据集上训练的 MLPs 模型的多种攻击效果评估.

Table 4　Performance of various attacks on MLPs trained using two non-image datasets.

MIA method	TPR @ 0.1% FPR (%)		Balanced accuracy		AUC	
	Purchase	Location	Purchase	Location	Purchase	Location
Shadow training	0.11	0.13	0.881	0.961	0.856	0.952
MBA(Entropy)	0.42	1.37	0.876	0.943	0.876	0.943
MBA(M-Entropy)	0.45	1.43	0.884	0.944	0.884	0.944
TrajectoryMIA	0.37	8.54	0.798	0.954	0.853	0.982
SeqMIA	4.68	23.26	0.877	0.962	0.940	**0.992**
DMS-MIA	**5.04**	**25.49**	**0.890**	**0.966**	**0.949**	0.989

MIA method	Precision		Recall		F1-score	
	Purchase	Location	Purchase	Location	Purchase	Location
Shadow training	0.785	0.798	**0.999**	**0.999**	0.880	0.887
MBA(Entropy)	0.809	0.932	0.987	0.955	0.889	0.944
MBA(M-Entropy)	0.818	0.934	0.989	0.955	0.895	0.945
TrajectoryMIA	0.767	0.919	0.745	0.993	0.756	0.955
SeqMIA	0.810	0.936	0.984	0.989	0.889	0.963
DMS-MIA	**0.822**	**0.938**	0.996	0.998	**0.901**	**0.968**

数据集上达到 25.49%, 在 Purchase 数据集上为 5.04%, 高于所有基线方法. 同时, 在整体判别性能方面, DMS-MIA 在 2 个数据集上均获得最高 F1 分数, 展现出在精确率与召回率之间的良好平衡.

综上所述, DMS-MIA 在几乎所有实验场景和评估指标上均取得了最优性能. 我们认为, 这一优势主要源于 DMS-MIA 对 TEE 下模型训练与部署中新型攻击面的有效利用. 具体而言, 通过磁盘快照与重放技术, DMS-MIA 能够直接获取目标模型在训练过程中对目标样本的输出变化序列; 再借助 Mamba 模型对这些变化趋势进行建模, 从而更准确地区分成员与非成员样本. 相比之下, TrajectoryMIA 和 SeqMIA 虽也试图从训练过程中提取成员关系信号, 并借助知识蒸馏技术模拟目标模型的行为, 但这种间接方式不可避免地导致成员信号衰减, 限制了攻击效果. 更为关键的是, 知识蒸馏通常需要对目标模型进行大量查询, 在 TEE 保护机制下, 这类高频率访问行为难以实现, 进一步削弱了此类方法的实际可行性.

4.3　消融实验

本小节从目标模型过拟合程度、磁盘重放次数、不同模型架构与超参数、目标模型和影子模型训练过程相关性 4 个方面进一步展开探讨.

4.3.1　目标模型过拟合程度

现有研究表明 [15, 19, 20, 49], 模型过拟合是导致成员关系隐私泄露的关键因素之一. 因此, 本小节评估了过拟合对 DMS-MIA 攻击效果的影响. 具体而言, 我们通过调整目标模型训练数据集的规模来控制其过拟合程度, 并参照相关工作 [23, 29] 的做法, 使用训练集准确率与测试集准确率之间的差距作为过拟合程度的量化指标.

如表 5 所示, 随着训练数据集规模的减小, 目标模型的过拟合程度从 0.334 逐步上升至 0.435, DMS-MIA 的攻击性能也随之提升. 以 TPR @ 0.1% FPR 指标为例, 攻击成功率由 12.72% 增长至 19.64%, 提升了约 6.92 个百分点. 值得注意的是, 即便在过拟合程度仅为 0.334 的情况下, DMS-MIA 的攻击性能 (如 AUC 达到 0.936) 仍优于其他基线方法在更高过拟合水平 (0.435) 下的表现 (例如, 表

表 5　过拟合程度对 DMS-MIA 攻击性能的影响. 目标模型为 CINIC10 上训练的 VGG-16.

Table 5　Impact of overfitting level on DMS-MIA attack performance. The target model is VGG-16 trained on CINIC10.

	Training dataset size				
	30000	25000	20000	15000	10000
Overfitting level	0.334	0.356	0.383	0.403	0.435
TPR @ 0.1% FPR (%)	12.72	15.19	15.31	15.91	19.64
Balanced accuracy	0.853	0.863	0.879	0.878	0.879
AUC	0.936	0.943	0.952	0.953	0.953
Precision	0.832	0.843	0.862	0.869	0.837
Recall	0.885	0.891	0.901	0.890	0.940
F1-score	0.858	0.867	0.881	0.880	0.886

表 6　磁盘重放次数对 DMS-MIA 攻击性能的影响. 目标模型为 CIFAR10 上训练的 VGG-16.

Table 6　Impact of number of disk replays on DMS-MIA attack performance. The target model is VGG-16 trained on CIFAR10.

	Number of disk replays					
	10	30	50	70	90	100
TPR @ 0.1% FPR (%)	2.07	4.32	5.26	6.51	9.34	11.10
Balanced accuracy	0.785	0.791	0.809	0.818	0.830	0.822
AUC	0.874	0.882	0.899	0.907	0.916	0.916
Precision	0.736	0.762	0.805	0.819	0.800	0.831
Recall	0.889	0.846	0.816	0.818	0.878	0.809
F1-score	0.805	0.802	0.810	0.818	0.837	0.820

A1 中基线方法的 AUC 最高为 0.929). 我们将其归因于 DMS-MIA 在成员关系信息提取方面的更高效率.

4.3.2　磁盘重放次数

在现实场景中, 攻击者的磁盘重放次数可能受到计算资源或存储成本等因素的限制. 磁盘重放次数决定了 DMS-MIA 可利用的目标模型中间状态数量, 进而影响多指标序列长度与攻击性能. 同时, 过多的重放操作会降低攻击效率. 因此, 评估不同重放次数下的攻击表现, 对于理解 DMS-MIA 的效率及其实际威胁具有重要意义.

表 6 展示了在不同磁盘重放次数下 DMS-MIA 的性能. 为简化实验设置, 快照点按等时间间隔从训练过程中选取. 可以看出, 减少重放次数确实会影响攻击性能, 但 DMS-MIA 在重放次数较低时仍表现出较强攻击能力. 例如, 在仅进行 10 次磁盘重放的情况下, DMS-MIA 的 F1-score 达到 0.805, 仍优于所有基线方法. 这表明攻击者无须依赖高频的磁盘重放操作, 只需要捕获训练过程中的少数关键快照, 即可有效提取成员关系信号.

值得注意的是, 当重放次数从 90 次增加到 100 次时, 部分综合性指标 (如平衡准确率和 F1-score) 出现了轻微下降, 我们认为这是信号与噪声之间权衡的结果. 随着重放次数的增多, 攻击模型能从更长的轨迹中捕获到更多有效信号, 这一点可以从 TPR @ 0.1% FPR 持续上升得到验证, 但同时也可能引入了更多的随机噪声, 如训练后期模型收敛时的微小震荡. 攻击模型 (Mamba) 在学习过程中可能会对这些噪声产生一定程度的过拟合, 从而导致其在目标模型上的泛化能力在某些综合性指标上略有下降.

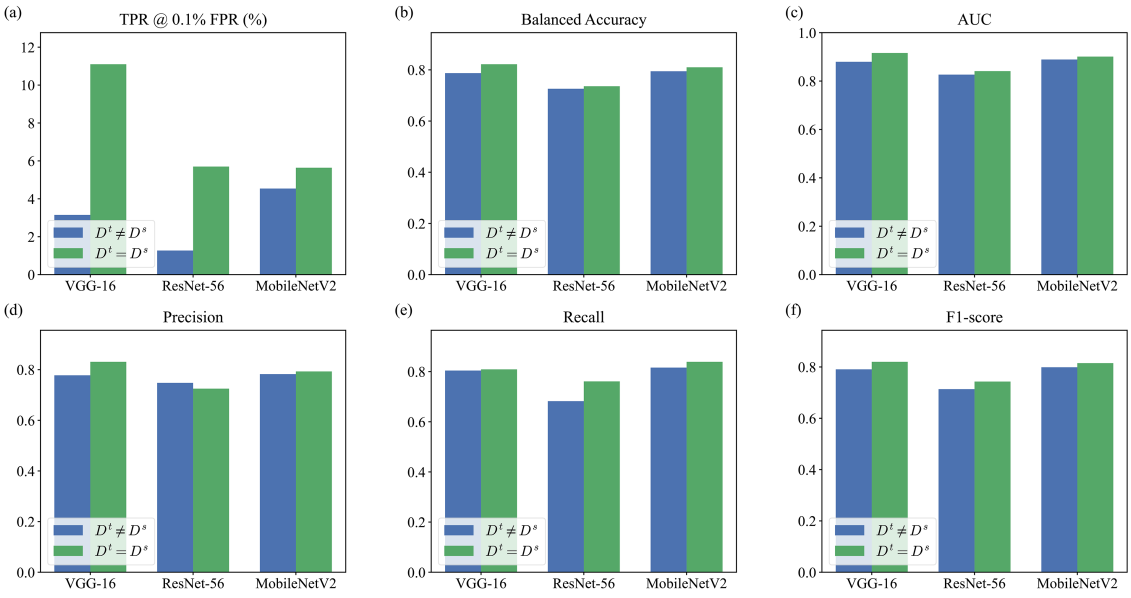

图 2　(网络版彩图) 独立同分布 (IID) 与非同分布 (Non-IID) 数据集对 DMS-MIA 攻击性能的影响. 目标模型均在 CIFAR10 上训练. (a) TPR @ 0.1% FPR (%); (b) Balanced accuracy; (c) AUC; (d) Precision; (e) Recall; (f) F1-score.

Figure 2　(Color online) Impact of IID vs. Non-IID datasets on DMS-MIA attack performance. Target models trained on CIFAR10. (a) TPR @ 0.1% FPR (%); (b) Balanced accuracy; (c) AUC; (d) Precision; (e) Recall; (f) F1-score.

4.3.3　非同分布数据集

依据第 3.1 小节威胁模型的设定, 攻击者能够获取与目标模型数据集 D^t 独立同分布的数据构建影子数据集 D^s. 为评估 DMS-MIA 在更具挑战性、也更符合实际的场景下的鲁棒性, 本小节进一步放宽该假设, 探讨目标数据集与影子数据集非同分布时的攻击性能表现.

具体而言, 设定目标数据集 D^t 为 CIFAR10. 在同分布场景下, 影子数据集 D^s 采用 CINIC10 数据集中的 CIFAR10 部分, 且与目标数据集样本不重叠. 而在非同分布场景下, D^s 则采用 CINIC10 中的 ImageNet 部分. 我们在 3 种不同模型架构上进行了对比实验, 结果如图 2 所示.

实验结果表明, 影子数据与目标数据分布的差异会削弱攻击性能. 例如, 对于 VGG-16 模型, 当影子数据从同分布切换为非同分布时, 其 TPR @ 0.1% FPR 从 11.10% 下降至 3.15%. 值得注意的是, 尽管处于非同分布这一更为严苛的攻击场景下, 但 DMS-MIA 仍展现出优越的攻击性能. 例如, VGG-16 模型在该场景下的 AUC 达到 0.879, 仍然优于基线方法在同分布设置下的最佳表现 0.869 (见表 A1). 这一结果表明, DMS-MIA 通过捕捉模型训练过程中的动态演化轨迹, 学得了更具泛化性的成员特征.

4.3.4　不同模型架构与超参数

本小节在第 3.1 小节所设定的威胁模型基础上进一步放宽假设条件, 不再要求攻击者完全掌握目标模型的架构与超参数. 我们转而探讨一种更具现实意义的场景: 攻击者在信息不完全的情况下, 可能采用与目标模型存在差异的架构或超参数, 在本地训练其影子模型. 我们在 CIFAR100 数据集上针对多种目标模型与影子模型的架构组合进行了实验, 结果如图 3 所示.

可以看出, DMS-MIA 的攻击性能通常在影子模型与目标模型架构一致时达到最优. 这是因为相同的模型架构和相近的训练配置有助于影子模型更准确地模拟目标模型的训练行为特征. 当影子模型架构与目标模型不一致时, DMS-MIA 的攻击性能通常会有所下降.

值得注意的是, 即使在目标模型与影子模型架构不匹配的情况下, DMS-MIA 在大多数情况下仍

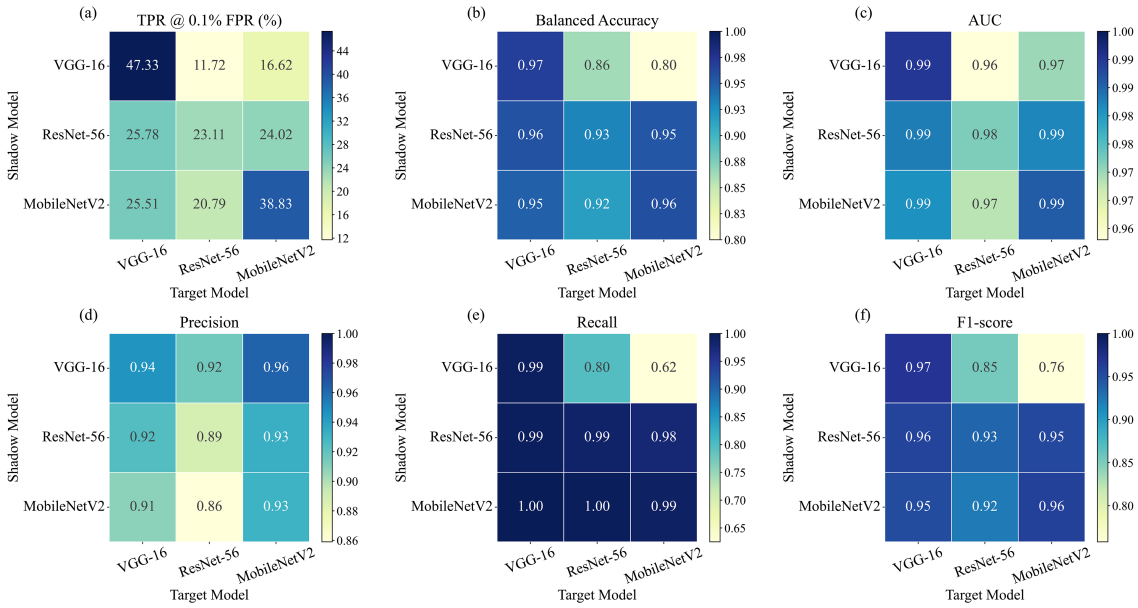

图 3　(网络版彩图) 不同影子模型与目标模型架构组合对 DMS-MIA 攻击性能的影响. 目标模型在 CIFAR100 上训练. (a) TPR @ 0.1% FPR (%); (b) Balanced accuracy; (c) AUC; (d) Precision; (e) Recall; (f) F1-score.

Figure 3　(Color online) Impact of different shadow and target model architecture combinations on DMS-MIA attack performance. Target models trained on CIFAR100. (a) TPR @ 0.1% FPR (%); (b) Balanced accuracy; (c) AUC; (d) Precision; (e) Recall; (f) F1-score.

优于现有基线方法. 例如, 表 3 显示, 在对 CIFAR100 上的 ResNet-56 目标模型进行攻击时, 当基线方法使用与其架构一致的模型作为影子模型时, 其最佳 TPR @ 0.1% FPR 为 20.53%; 而在图 3 中, DMS-MIA 使用 MobileNetV2 作为影子模型时, TPR 仍能达到 20.79%, 略高于基线方法的最佳表现.

4.3.5　攻击模型的选择

在 DMS-MIA 的设计中, 我们采用 Mamba 架构来构建攻击模型, 以捕捉多指标序列中的时序依赖关系. 为验证这一选择的合理性, 本小节将 Mamba 与两种主流的序列建模架构长短期记忆网络 (long short-term memory, LSTM) [50] 和 Transformer [51] 进行了性能对比. LSTM 是循环神经网络 (recurrent neural network, RNN) 的经典代表, 擅长处理时序数据; 而 Transformer 凭借其自注意力机制, 已成为序列处理领域的主流架构. 我们在 3 个图像数据集上对这 3 种攻击模型进行了评估, 结果如表 7 所示. 实验结果表明, 基于 Mamba 的攻击模型在绝大多数关键指标上均表现出最优性能. 在 TPR @ 0.1% FPR 指标上, Mamba 在 3 个数据集上均优于 LSTM 和 Transformer. 例如, 在 CIFAR10 上, Mamba 的 TPR 达到了 5.70%, 高于 LSTM 的 2.58% 和 Transformer 的 4.04%. 同样, 在 AUC 和平衡准确率等综合性指标上, Mamba 也达到最优.

我们认为 Mamba 的这种整体优势可以归因于其高效的状态空间模型 (state space model, SSM) 架构. 相比计算量随着序列长度平方增长、开销较大的 Transformer 和 Mamba 能以线性复杂度处理长序列, 效率更高. 而相较于 LSTM, Mamba 的选择性 SSM 机制能更有效地捕捉长程依赖关系, 并过滤掉序列中的无关信息, 这对于从多指标序列中精确提取成员关系信号至关重要. 因此, 实验结果和理论分析共同验证了 Mamba 作为 DMS-MIA 攻击模型的合理性.

4.3.6　目标模型和影子模型训练过程相关性

本小节对 DMS-MIA 攻击中影子模型与目标模型中间状态的相关性进行验证. 只有当两者在成员

表 7　不同攻击模型在 3 个图像数据集上的性能比较. 目标模型为 ResNet-56.

Table 7　Performance comparison of different attack models on three image datasets. The target model is ResNet-56.

Dataset	TPR @ 0.1% FPR (%)			Balanced accuracy			AUC		
	Mamba	LSTM	Transformer	Mamba	LSTM	Transformer	Mamba	LSTM	Transformer
CIFAR10	**5.70**	2.58	4.04	**0.736**	0.726	0.712	**0.841**	0.833	0.822
CIFAR100	**23.11**	22.86	12.25	**0.931**	0.927	0.929	**0.977**	0.974	0.973
CINIC10	**8.27**	4.54	4.86	**0.852**	0.849	0.849	**0.924**	0.922	0.922
Dataset	Precision			Recall			F1-score		
	Mamba	LSTM	Transformer	Mamba	LSTM	Transformer	Mamba	LSTM	Transformer
CIFAR10	**0.725**	**0.725**	0.703	**0.761**	0.727	0.736	**0.743**	0.726	0.719
CIFAR100	0.888	**0.892**	0.883	0.986	0.972	**0.989**	**0.934**	0.930	0.933
CINIC10	0.793	0.794	**0.800**	**0.954**	0.944	0.930	**0.866**	0.862	0.860

表 8　DMS-MIA 与现有成员推理攻击在信号利用上的对比.

Table 8　Comparison of signal utilization between DMS-MIA and existing membership inference attacks.

Attack	Model states utilized		Source of middle states
	Final state	Middle states	
[15, 17, 19∼21]	✓	–	–
SeqMIA, TrajectoryMIA	✓	✓	Knowledge distillation
DMS-MIA	✓	✓	Disk replay

关系上的表现具有相似性时, 攻击者基于影子模型中间状态构建的多指标序列所训练出的攻击模型, 才能有效迁移到对目标模型的攻击中. 值得强调的是, 尽管早期工作 [15, 17, 19∼21] 等已提出利用影子模型模拟目标模型, 但如表 8 所示, 这些方法通常仅利用模型的最终状态 (final state), 因此只要求影子模型在训练完成时具备与目标模型相似的预测行为. SeqMIA 和 TrajectoryMIA 虽然通过知识蒸馏模拟了目标模型的训练过程, 从而能够利用中间状态 (middle states), 但它们并未验证影子模型的中间状态能否真实反映目标模型的训练动态. 相比之下, 我们的 DMS-MIA 方法通过磁盘重放获取中间状态. 这提出了一个更强的要求: 影子模型需要在整个训练过程中, 对成员与非成员样本的学习效果差异与目标模型高度相似. 我们的攻击模型学习的是这种动态轨迹中的时序模式, 而非单一时间点的预测输出. 因此, 本小节旨在通过实验证实其合理性, 为 DMS-MIA 攻击的有效迁移提供坚实的理论依据.

为简化分析, 选取单一指标——损失值作为研究对象. 具体而言, 在模型训练的每个中间状态, 我们分别计算所有成员样本和非成员样本的损失均值, 相应记为 $\overline{L}_{\text{member}}$ 和 $\overline{L}_{\text{non-member}}$, 并将其差值定义为损失差距 (LossGap):

$$\text{LossGap} = \overline{L}_{\text{non-member}} - \overline{L}_{\text{member}}. \tag{2}$$

LossGap 反映了模型在某一特定中间状态下对成员与非成员样本预测行为的差异. 将多个中间状态下的 LossGap 连接起来, 即可形成 LossGap 曲线, 用于刻画模型在训练过程中对两类样本学习效果差异的变化趋势. 在此基础上, 我们分别为目标模型与影子模型构建了 LossGap 曲线, 如图 4 所示. 可以看出, 两者曲线高度吻合, 表明其在训练过程中对成员与非成员样本的学习差异变化趋势基本一致. 进一步地, 如表 9 所示, 计算了目标模型与影子模型 LossGap 曲线之间的相关系数. 结果显示, 在所有实验场景下, 相关系数均超过 0.98, 说明二者具有极高的相似性. 这表明, 使用影子模型模拟目标模型的训练过程是合理且有效的, 基于影子数据和影子模型训练得到的攻击模型, 也能够有效迁移到对目标模型的攻击中.

表 9　不同数据集下目标模型和影子模型训练过程损失差距的相关系数.

Table 9　Correlation coefficients of LossGap between target and shadow models across different datasets during training.

Dataset	CIFAR10	CIFAR100	CINIC10	Purchase	Location
Model	VGG	VGG	VGG	MLPs	MLPs
Correlation coefficient	0.988	0.999	0.994	0.999	0.999

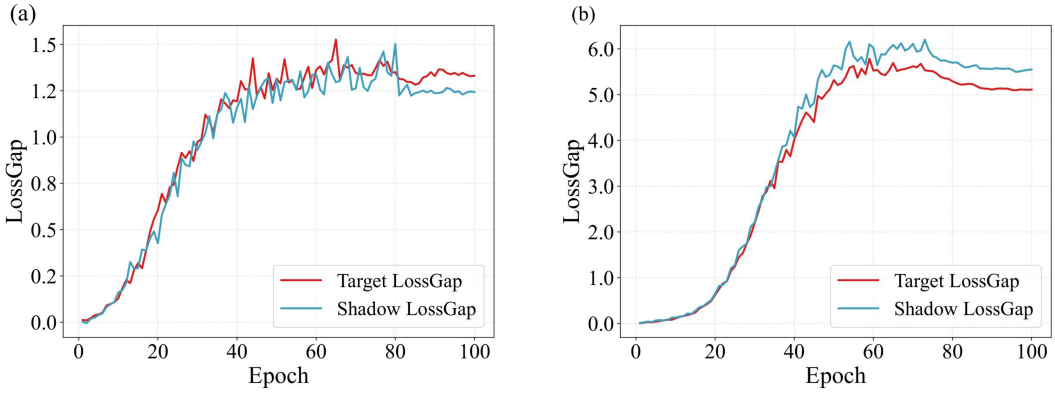

图 4　(网络版彩图) 目标模型与影子模型的损失差距曲线. (a) VGG-16 在 CIFAR10 上的损失差距曲线; (b) VGG-16 在 CIFAR100 上的损失差距曲线.

Figure 4　(Color online) LossGap curves of target and shadow models. (a) LossGap curve of VGG-16 on CIFAR10; (b) LossGap curve of VGG-16 on CIFAR100.

5　防御措施

本文提出的 DMS-MIA 攻击揭示了 TEE 保护下机器学习模型面临的新型隐私威胁, 对成员隐私构成严重挑战. 因此, 研究并部署相应的防御策略至关重要. 本节首先介绍两种主流的成员推理攻击防御方法, 随后通过实验评估它们在抵御 DMS-MIA 及其他基线攻击时的效果.

5.1　防御策略设计

为应对 DMS-MIA 所揭示的隐私威胁, 引入两种在成员推理攻击研究 [15, 23, 27, 29, 52] 中广泛使用的防御策略 L2 正则化与 MixupMMD [25], 并在本小节对其原理进行概述.

正则化. 正则化是一种在机器学习中广泛用于防止过拟合的经典技术. 通过在损失函数中增加一个惩罚项, 正则化能够限制模型参数的复杂度. 本研究采用了 L2 正则化 (权重衰减系数设为 0.0005), 其通过惩罚较大的权重值, 促使模型学习更平滑的决策边界, 从而降低其对训练数据中噪声和特定样本特征的敏感度. 这种机制可以提升模型的泛化能力, 从而有效削弱成员关系信号.

MixupMMD. MixupMMD 结合了数据增强方法 Mixup 与最大均值差异 (maximum mean discrepancy, MMD) 正则化. Mixup 通过对样本线性插值增强数据多样性, 平滑决策边界. MMD 则正则化隐藏层表示使不同类别在特征空间中分布更紧凑. MixupMMD 可提升模型泛化能力, 减少对训练数据的记忆, 有效抵御成员推理攻击.

5.2　防御效果评估

从表 10 可以看出, 在大部分情况下, 两种防御方法均能有效降低成员推理攻击的性能. 以 DMS-MIA 为例, 在应用 L2 正则化后, 其 TPR @ 0.1% FPR 从 8.27% 显著下降至 1.04%, AUC 从 0.924 降至 0.738. 这表明正则化通过抑制过拟合, 削弱了模型训练过程中的成员关系信号. 与 L2 正则化类似, MixupMMD 也有效地削弱了 DMS-MIA 的攻击能力. 实验结果显示, 在应用 MixupMMD 防

表 10　防御措施对成员推理攻击性能的影响. 目标模型为 CINIC10 上训练的 ResNet-56.

Table 10 Impact of defense measures on membership inference attack performance. The target model is ResNet-56 trained on CINIC10.

MIA method	TPR @ 0.1% FPR (%)			Balanced accuracy			AUC		
	No defense	L2	MixupMMD	No defense	L2	MixupMMD	No defense	L2	MixupMMD
Shadow training	0.11	0.06	0.10	0.630	0.514	0.622	0.659	0.515	0.658
MBA(Entropy)	0.15	0.11	0.30	0.600	0.504	0.503	0.600	0.504	0.503
MBA(M-Entropy)	0.20	0.14	0.15	0.711	0.572	0.605	0.711	0.572	0.605
TrajectoryMIA	0.30	0.20	0.32	0.690	0.539	0.615	0.745	0.563	0.677
SeqMIA	5.80	0.96	**4.86**	0.776	0.537	0.537	0.863	0.665	0.820
DMS-MIA	**8.27**	**1.04**	4.30	**0.852**	**0.644**	**0.843**	**0.924**	**0.738**	**0.916**

MIA method	Precision			Recall			F1-score		
	No defense	L2	MixupMMD	No defense	L2	MixupMMD	No defense	L2	MixupMMD
Shadow training	0.603	0.515	0.677	0.746	0.214	0.090	0.667	0.303	0.159
MBA(Entropy)	0.594	0.528	0.752	0.635	0.083	0.008	0.614	0.144	0.016
MBA(M-Entropy)	0.661	0.576	0.600	0.866	0.545	0.628	0.750	0.560	0.614
TrajectoryMIA	0.691	0.579	0.731	0.681	0.288	0.366	0.686	0.385	0.487
SeqMIA	0.739	**0.724**	**0.969**	0.852	0.119	0.076	0.792	0.204	0.140
DMS-MIA	**0.793**	0.675	0.811	**0.954**	**0.556**	**0.895**	**0.866**	**0.610**	**0.851**

御后, DMS-MIA 的 TPR @ 0.1% FPR 从 8.27% 下降至 4.30%, AUC 也从 0.924 降至 0.916. 此外, MixupMMD 对部分攻击方法的精确率有一定提升, 但同时其召回率出现了显著下降, 从综合评价指标 F1-score 来看, 这些攻击的性能仍然被削弱. 值得注意的是, 在部署了防御措施后, DMS-MIA 在多数关键指标上仍然是表现最强的攻击方法.

6　结论

本文针对 TEE 下机器学习模型的成员隐私泄露风险, 提出了一种名为 DMS-MIA 的新型成员推理攻击方法. 该方法的核心在于利用攻击者对宿主机及虚拟化管理程序的控制权, 对 TEE 保护下的 CVM 的加密磁盘进行周期性快照与重放, 从而在不破坏 TEE 运行时内存保护的前提下, 获取目标模型在不同历史快照时间点的预测输出序列. DMS-MIA 进一步从这些历史输出中构建融合损失、熵等多维度指标的时间序列, 并采用 Mamba 模型作为攻击分类器, 捕捉序列中的复杂时序关系以精确判别样本的成员身份. 在 3 个图像数据集和 2 个非图像数据集上的综合实验评估表明, DMS-MIA 在包括 TPR @ 0.1% FPR 和 AUC 在内的多项关键指标上均显著优于现有的基线攻击方法. 消融实验进一步揭示了模型过拟合程度、磁盘重放次数、数据分布、影子模型与目标模型配置差异、攻击模型架构以及目标模型与影子模型训练过程相关性对攻击性能的影响, 同时验证了所提方法的合理性并探索了如正则化等防御策略的效果. 本研究揭示了 TEE 中一种针对 CVM 的执行控制的独特攻击面, 并验证了由此带来的成员关系隐私的高泄露风险. 这一发现为 TEE 下的机器学习隐私保护研究提供了新的视角和挑战.

参考文献

1　Intel. Product brief, 3rd gen Intel Xeon scalable processor for IoT. 2023. https://www.intel.com/content/www/us/en/products/docs/processors/embedded/3rd-gen-xeon-scalable-iot-product-brief.html

2　Johnson S, Makaram R, Santoni A, et al. Supporting Intel SGX on multi-socket platforms. Intel Corp, 2021, 65

3 Kaplan D. Hardware VM isolation in the cloud. Queue, 2023, 21: 49–67
4 Sev-Snp A. Strengthening VM isolation with integrity protection and more. White Paper, 2020, 53: 1450–1465
5 Intel Corporation. Intel trust domain extensions. 2023. https://cdrdv2.intel.com/v1/dl/getContent/690419
6 Hu B, Wang Y, Cheng J, et al. Secure and efficient mobile DNN using trusted execution environments. In: Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security, 2023. 274–285
7 Lee T, Lin Z, Pushp S, et al. Occlumency: privacy-preserving remote deep-learning inference using SGX. In: Proceedings of the 25th Annual International Conference on Mobile Computing and Networking, 2019. 1–17
8 Li Y, Zeng D, Gu L, et al. Lasagna: accelerating secure deep learning inference in SGX-enabled edge cloud. In: Proceedings of the ACM Symposium on Cloud Computing, 2021. 533–545
9 Zhang Z, Gong C, Cai Y, et al. No privacy left outside: on the (in-) security of tee-shielded DNN partition for on-device ml. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), 2024. 3327–3345
10 Li D, Zhang Z, Yao M, et al. Teeslice: protecting sensitive neural network models in trusted execution environments when attackers have pre-trained models. ACM Trans Softw Eng Methodology, 2024, 34: 1–49
11 Al-Rubaie M, Chang J M. Privacy-preserving machine learning: Threats and solutions. IEEE Secur Privacy, 2019, 17: 49–58
12 Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015. 1322–1333
13 Yang Q, Liu Y, Chen T, et al. Federated machine learning: concept and applications. ACM Trans Intell Syst Tech, 2019, 10: 1–19
14 Ganju K, Wang Q, Yang W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018. 619–633
15 Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), 2017. 3–18
16 Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), 2019. 739–753
17 Liu Y, Wen R, He X, et al. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In: Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), 2022. 4525–4542
18 Long Y, Wang L, Bu D, et al. A pragmatic approach to membership inferences on machine learning models. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P), 2020. 521–534
19 Salem A, Zhang Y, Humbert M, et al. Ml-leaks: model and data independent membership inference attacks and defenses on machine learning models. ArXiv:1806.01246
20 Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles. In: Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP), 2022. 1897–1914
21 Song L, Mittal P. Systematic evaluation of privacy risks of machine learning models. In: Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), 2021. 2615–2632
22 Ye J, Maddi A, Murakonda S K, et al. Enhanced membership inference attacks against machine learning models. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, 2022. 3093–3106
23 Liu Y, Zhao Z, Backes M, et al. Membership inference attacks by exploiting loss trajectory. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, 2022. 2085–2098
24 Leino K, Fredrikson M. Stolen memories: leveraging model memorization for calibrated White-Box membership inference. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020. 1605–1622
25 Li Z, Zhang Y. Membership leakage in label-only exposures. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021. 880–895
26 Choquette-Choo C A, Tramer F, Carlini N, et al. Label-only membership inference attacks. In: Proceedings of the International Conference on Machine Learning, 2021. 1964–1974
27 Yeom S, Giacomelli I, Fredrikson M, et al. Privacy risk in machine learning: analyzing the connection to overfitting. In: Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF), 2018. 268–282
28 Zhang M, Ren Z, Wang Z, et al. Membership inference attacks against recommender systems. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021. 864–879
29 Li H, Li Z, Wu S, et al. Seqmia: sequential-metric based membership inference attack. In: Proceedings of 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024. 3496–3510

30 Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. ArXiv:2312.00752

31 Bertran M, Tang S, Roth A, et al. Scalable membership inference attacks via quantile regression. Adv Neural Inform Process Syst, 2023, 36: 314–330

32 Sablayrolles A, Douze M, Schmid C, et al. White-box vs black-box: Bayes optimal strategies for membership inference. In: Proceedings of the International Conference on Machine Learning, 2019. 5558–5567

33 Watson L, Guo C, Cormode G, et al. On the importance of difficulty calibration in membership inference attacks. ArXiv:2111.08440

34 Jagielski M, Carlini N, Berthelot D, et al. High accuracy and high fidelity extraction of neural networks. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020. 1345–1362

35 Chandrasekaran V, Chaudhuri K, Giacomelli I, et al. Exploring connections between active learning and model extraction. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), 2020. 1309–1326

36 Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction APIs. In: Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), 2016. 601–618

37 Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017. 506–519

38 Yuan Y, Liu Z, Deng S, et al. Hypertheft: thieving model weights from tee-shielded neural networks via ciphertext side channels. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024. 4346–4360

39 Bellard F. QEMU, a fast and portable dynamic translator. In: Proceedings of the Annual Conference on USENIX Annual Technical Conference, 2005

40 Krizhevsky A, Hinton G, et al. Dataset: learning multiple layers of features from tiny images. 2024. https://service.tib.eu/ldmservice/dataset/learning-multiple-layers-of-features-from-tiny-images

41 Darlow L N, Crowley E J, Antoniou A, et al. CINIC-10 is not Imagenet or CIFAR-10. ArXiv:1810.03505

42 Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255

43 Yang D, Zhang D, Qu B. Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Trans Intell Syst Tech, 2016, 7: 1–23

44 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

45 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv:1409.1556

46 Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4510–4520

47 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. ArXiv:1503.02531

48 Chang H, Edwards B, Paul A S, et al. Efficient privacy auditing in federated learning. In: Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24), 2024. 307–323

49 Hu H, Salcic Z, Sun L, et al. Membership inference attacks on machine learning: a survey. ACM Comput Sur (CSUR), 2022, 54: 1–37

50 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

51 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 6000–6010

52 Kaya Y, Hong S, Dumitras T. On the effectiveness of regularization against membership inference attacks. ArXiv:2006.05336

附录 A

表 A1　在 3 个图像数据集上训练的 VGG-16 模型的多种攻击效果评估.

Table A1　Performance of various attacks on VGG-16 trained using three different image datasets.

MIA method	TPR @ 0.1% FPR (%)			Balanced accuracy			AUC		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.14	0.82	0.17	0.738	0.946	0.764	0.748	0.962	0.805
MBA(Entropy)	0.20	1.18	0.27	0.736	0.937	0.775	0.736	0.937	0.775
MBA(M-Entropy)	0.21	1.31	0.30	0.749	0.944	0.815	0.749	0.944	0.815
TrajectoryMIA	0.26	3.81	1.09	0.647	0.910	0.740	0.699	0.953	0.817
SeqMIA	10.25	37.76	19.16	0.764	0.955	0.837	0.869	0.990	0.929
DMS-MIA	**11.10**	**47.33**	**19.64**	**0.822**	**0.968**	**0.879**	**0.916**	**0.995**	**0.953**
MIA method	Precision			Recall			F1-score		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.634	0.874	0.634	**0.999**	**0.999**	**0.989**	0.776	0.932	0.773
MBA(Entropy)	0.672	0.922	0.728	0.920	0.957	0.878	0.777	0.939	0.796
MBA(M-Entropy)	0.680	0.929	0.751	0.942	0.962	0.942	0.790	0.945	0.836
TrajectoryMIA	0.618	0.874	0.707	0.664	0.939	0.792	0.640	0.905	0.747
SeqMIA	0.706	0.930	0.785	0.906	0.985	0.927	0.793	0.957	0.850
DMS-MIA	**0.831**	**0.945**	**0.837**	0.809	0.993	0.940	**0.820**	**0.969**	**0.886**

表 A2　在 3 个图像数据集上训练的 MobileNetV2 模型的多种攻击效果评估.

Table A2　Performance of various attacks on MobileNetV2 trained using three different image datasets.

MIA method	TPR @ 0.1% FPR (%)			Balanced accuracy			AUC		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.22	0.48	0.18	0.657	0.691	0.669	0.685	0.758	0.718
MBA(Entropy)	0.17	0.20	0.19	0.646	0.675	0.667	0.646	0.675	0.667
MBA(M-Entropy)	0.18	0.43	0.26	0.688	0.835	0.754	0.688	0.835	0.754
TrajectoryMIA	0.16	0.33	0.44	0.662	0.837	0.761	0.729	0.891	0.835
SeqMIA	4.35	19.67	10.67	0.718	0.885	0.807	0.810	0.956	0.897
DMS-MIA	**5.64**	**38.83**	**14.65**	**0.810**	**0.959**	**0.903**	**0.901**	**0.991**	**0.963**
MIA method	Precision			Recall			F1-score		
	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10	CIFAR10	CIFAR100	CINIC10
Shadow training	0.598	0.637	0.620	**0.874**	0.807	0.814	0.710	0.712	0.704
MBA(Entropy)	0.623	0.672	0.654	0.740	0.684	0.707	0.678	0.678	0.680
MBA(M-Entropy)	0.649	0.810	0.726	0.819	0.876	0.817	0.724	0.842	0.769
TrajectoryMIA	0.659	0.820	0.754	0.640	0.860	0.750	0.649	0.840	0.752
SeqMIA	0.671	0.861	0.773	0.858	0.918	0.870	0.753	0.889	0.818
DMS-MIA	**0.793**	**0.931**	**0.880**	0.839	**0.991**	**0.933**	**0.815**	**0.960**	**0.906**

DMS-MIA: disk replay-based and multi-metric sequence membership inference attack on TEE-protected machine learning models

Yifan DONG[1,2], Wei FENG[1], Hao LI[1*], Min ZHANG[1], Yu QIN[1] & Dengguo FENG[1]

1. *Institute of Software, Chinese Academy of Sciences, Beijing 100190, China*
2. *School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China*
* Corresponding author. E-mail: lihao@iscas.ac.cn

Abstract With the widespread adoption of machine learning as a service (MLaaS), data privacy concerns have become increasingly prominent. Trusted execution environment (TEE) offers robust protection for MLaaS, particularly in preventing the leakage of model parameters and training data. The isolation mechanism provided by TEE presents a significant challenge to mainstream membership inference attack (MIA) methods: due to the difficulty in directly accessing the internal parameters and states of the model, as well as strict limitations on query frequency, existing MIA methods perform poorly under TEE protection. However, this paper identifies and exploits a novel attack surface inherent in TEE-protected MLaaS, proposing a new membership inference attack method called DMS-MIA (disk replay-based multi-metric sequence membership inference attack). The core idea of this method is that the attacker gains control over the host machine and virtualization management software, periodically snapshots and replays the encrypted disk of a TEE-protected confidential virtual machine, thereby obtaining intermediate state outputs during the training process of the MLaaS model. Subsequently, DMS-MIA constructs a multi-dimensional metric time series from these historical outputs and uses the Mamba model as an attack model to amplify membership signals within the series, thereby identifying the membership of the sample. Experimental results demonstrate that DMS-MIA achieves significant results in several key metrics, such as TPR @ 0.1% FPR and AUC, across three image datasets and two non-image datasets.

Keywords membership inference attack, trusted execution environment, machine learning, disk replay, multi-metric sequence