软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(10):4812−4826 [doi: 10.13328/j.cnki.jos.007310] [CSTR: 32375.14.jos.007310]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于视觉特征解耦的无数据依赖模型窃取攻击方法*

张锦弘[1], 刘仁阳[1], 韦廷楚[2], 董云云[3,4], 周 维[3,4]

[1](云南大学 信息学院, 云南 昆明 650500)

[2](云南大学 国际河流与生态安全研究院, 云南 昆明 650500)

[3](云南大学 软件学院, 云南 昆明 650500)

[4](云南大学 跨境网络空间安全工程研究中心, 云南 昆明 650500)

通信作者: 周维, E-mail: zwei@ynu.edu.cn

摘 要: 随着深度学习模型安全性和隐私性研究的不断深入, 研究者发现模型窃取攻击能够对神经网络产生极大的威胁. 典型的数据依赖模型窃取攻击可以利用一定比例的真实数据查询目标模型, 在本地训练一个替代模型, 从而达到目标模型窃取的目的. 2020 年以来, 一种新颖的无数据依赖模型窃取攻击方法被提出, 仅使用生成模型生成伪造的查询样本便能对深度神经网络开展窃取和攻击. 由于不依赖于真实数据, 无数据依赖模型窃取攻击具有更严重的破坏力. 然而, 目前的无数据依赖模型窃取攻击方法所构造查询样本的多样性和有效性不足, 存在模型窃取过程中查询次数大、攻击成功率较低的问题. 因此提出一种基于视觉特征解耦的无数据依赖模型窃取攻击方法 VFDA (vision feature decoupling-based model stealing attack), 该方法通过利用多解码器结构对无数据依赖模型窃取过程中生成的查询样本的视觉特征进行解耦与生成, 从而提高查询样本的多样性和模型窃取的有效性. 具体来说, VFDA 利用 3 个解码器分别生成查询样本的纹理信息、区域编码和平滑信息, 完成查询样本的视觉特征解耦. 其次, 为了使生成的查询样本更加符合真实样本的视觉特征, 通过限制纹理信息的稀疏性以及对生成的平滑信息进行滤波. VFDA 利用了神经网络的表征倾向依赖于图像纹理特征的性质, 能够生成类间多样性的查询样本, 从而有效提高了模型窃取的相似性以及攻击成功率. 此外, VFDA 对解耦生成的查询样本平滑信息添加了类内多样性损失, 使查询样本更加符合真实样本的分布. 通过与多个模型窃取攻击方法对比, VFDA 方法在模型窃取的相似性以及攻击的成功率上具有更好的表现. 特别在分辨率较高的 GTSRB 和 Tiny-ImageNet 数据集上, 相比于目前较好的 EBFA 方法, 在攻击成功率上 VFDA 方法平均提高了 3.86% 和 4.15%.

关键词: 模型窃取; 对抗样本; 迁移攻击; 生成模型; 模型隐私

中图法分类号: TP309

中文引用格式: 张锦弘, 刘仁阳, 韦廷楚, 董云云, 周维. 基于视觉特征解耦的无数据依赖模型窃取攻击方法. 软件学报, 2025, 36(10): 4812–4826. http://www.jos.org.cn/1000-9825/7310.htm

英文引用格式: Zhang JH, Liu RY, Wei TC, Dong YY, Zhou W. Data-free Model Stealing Attack Method Based on Visual Feature Decoupling. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4812–4826 (in Chinese). http://www.jos.org.cn/1000-9825/7310.htm

Data-free Model Stealing Attack Method Based on Visual Feature Decoupling

ZHANG Jin-Hong[1], LIU Ren-Yang[1], WEI Ting-Chu[2], DONG Yun-Yun[3,4], ZHOU Wei[3,4]

[1](School of Information Science and Engineering, Yunnan University, Kunming 650500, China)

[2](Institute of International Rivers and ECO-Security, Yunnan University, Kunming 650500, China)

* 基金项目: 国家自然科学基金 (62162067, 62101480); 云南省院士专家工作站项目 (202205AF150145)
收稿时间: 2023-06-27; 修改时间: 2024-01-11, 2024-07-18, 2024-10-07; 采用时间: 2024-10-31; jos 在线出版时间: 2025-03-12
CNKI 网络首发时间: 2025-03-13

[3](School of Software, Yunnan University, Kunming 650500, China)
[4](Engineering Research Center of Cyberspace, Yunnan University, Kunming 650500, China)

Abstract: With the continuous deepening of research on the security and privacy of deep learning models, researchers find that model stealing attacks pose a tremendous threat to neural networks. A typical data-dependent model stealing attack can use a certain percentage of real data to query the target model and train an alternative model locally to steal the target model. Since 2020, a novel data-free model stealing attack method has been proposed, which can steal and attack deep neural networks simply by using fake query examples generated by generative models. Since it does not rely on real data, the data-free model stealing attack can cause more serious damage. However, the diversity and effectiveness of the query examples constructed by the current data-free model stealing attack methods are insufficient, and there are problems of a large number of queries and a relatively low success rate of the attack during the model stealing process. Therefore, this study proposes a vision feature decoupling-based model stealing attack (VFDA), which decouples and generates the visual features of the query examples generated during the data-free model stealing process by using a multi-decoder structure, thus improving the diversity of query examples and the effectiveness of model stealing. Specifically, VFDA uses three decoders to respectively generate the texture information, region encoding, and smoothing information of query examples to complete the decoupling of visual features of query examples. Secondly, to make the generated query examples more consistent with the visual features of real examples, the sparsity of the texture information is limited and the generated smoothing information is filtered. VFDA exploits the property that the representational tendency of neural networks depends on the image texture features, and can generate query examples with inter-class diversity, thus effectively improving the similarity of model stealing and the success rate of the attack. In addition, VFDA adds intra-class diversity loss to the smoothed information of query samples generated through decoupling to make the query samples more consistent with real sample distribution. By comparing with multiple model stealing attack methods, the VFDA method proposed in this study has better performance in the similarity of model stealing and the success rate of the attack. In particular, on the GTSRB and Tiny-ImageNet datasets with high resolution, the attack success rate is respectively improved by 3.86% and 4.15% on average compared with the currently better EBFA method.

Key words: model stealing; adversarial examples; transfer attack; generative model; model privacy

随着深度学习在计算机视觉和自然语言处理等领域取得的跨越式发展, 人工智能的安全性、隐私性受到了越来越多的关注. 现有研究表明, 人工智能模型 (机器学习模型或深度学习模型) 很容易受到对抗攻击的影响[1−3], 即通过在干净样本中加入精心制作的微小扰动, 这些扰动能够使得人工智能模型产生错误的预测结果. 此外, 大模型的训练需要极其高昂的成本, 如训练数据的收集、标注人工成本、模型训练的硬件成本等. 模型研发和使用方迫切需要提升对这类高成本模型的防护能力, 以免模型泄露, 造成不必要的损失. 然而, 目前的研究表明, 部署在云端的深度学习模型存在被窃取和攻击的风险, 这类安全和隐私问题严重影响了深度学习的发展. 对深度学习模型进行窃取和攻击被称为模型窃取攻击, 其过程如图 1 所示.

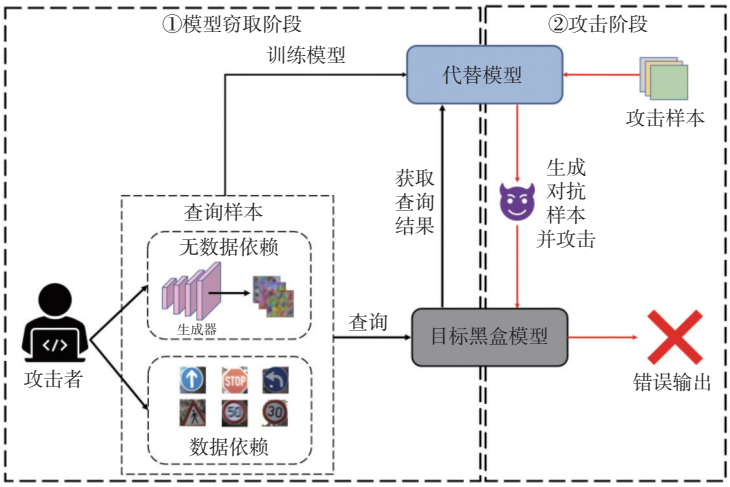


图 1　模型窃取攻击中模型窃取阶段及攻击阶段示意图

目前模型窃取攻击方法可以分为数据依赖的模型窃取攻击和无数据依赖模型窃取攻击. 数据依赖的模型窃取攻击通过利用真实数据对目标模型进行查询, 在黑盒场景下以“模型蒸馏[1–7]”的方式在本地训练一个代替模型, 从而完成模型的窃取. 在获得窃取到的模型之后, 便能通过攻击代替模型来构造对抗样本, 并利用对抗样本的可迁移性对目标模型发起迁移攻击. 这种攻击方法在成功攻击目标模型的同时, 能够窃取到一个具有相似功能的盗版模型, 对深度学习模型具有安全性和隐私性的双重威胁. 由于数据依赖的模型窃取攻击需要真实数据集作为查询样本, 这在现实场景中往往又难以达成, 因此数据依赖的方法适用性较差. 与需要真实数据的模型窃取攻击相对应, 近两年提出的无数据依赖模型窃取攻击更符合物理世界中的实际情况, 具有更大的应用潜力. 无数据依赖模型窃取攻击是一种非常具有挑战性的攻击方法, 因为它要求攻击者在未知目标模型内部结构的情况下, 使用伪造的查询样本在有限查询次数下探索目标模型的行为, 并窃取到一个相似的模型. 因此, 这些构造查询样本的质量成为影响代替模型训练效果的重要因素. 由于代替模型的训练数据为生成器生成的查询样本, 数据标签为查询目标黑盒模型得到的硬标签, 因此如何生成高质量的查询样本, 确保查询样本被黑盒模型分类输出的类别更加均衡是无数据依赖模型窃取攻击的核心问题. 即如何提高查询样本对于代替模型训练的有效性, 以及提高被目标模型预测输出的多样性？解决该问题需从神经网络分类决策的敏感性角度进行探索, 确定查询样本的哪些信息影响对神经网络的决策更为重要, 从而使查询样本更加有效、多样. Geirhos 等人[8]和 Wang 等人[9]指出, 神经网络的分类决策更多依赖样本的少量纹理, 而非平滑区域. 也就是说, 纹理信息较高程度地控制着神经网络的类间判别输出. 通过对无数据依赖模型窃取攻击方法调研, 本文认为现有无数据依赖的模型窃取攻击方法存在一些局限性:

(1) 现有的无数据依赖模型窃取攻击方法仅依赖一个生成器去生成查询样本, 生成查询样本对于代替模型训练的有效性不佳, 导致需要大量的查询次数, 增加攻击者暴露的风险.

(2) 目前的方法仅考虑了生成器生成查询样本的类间多样性, 却无顾及生成的查询样本对于目标黑盒模型的类内多样性, 导致代替模型和目标模型决策边界不相似, 攻击成功率不高.

(3) 目前的方法没有考虑从神经网络分类决策的敏感性角度提高查询样本的质量, 导致生成查询样本不能很好地探测到目标模型的决策边界, 攻击成功率不高.

针对上述方法存在的局限性, 本文提出了一种基于视觉特征解耦的无数据依赖模型窃取攻击方法 VFDA (vision feature decoupling-based model stealing attack), 该方法针对视觉特征中的纹理信息和平滑信息进行解耦, 使用不同的解码器针对性地生成查询样本信息, 以提高查询样本的多样性和有效性. 具体地说, VFDA 首先使用上采样生成器对高斯噪声进行采样, 并经过一个编码器编码, 得到生成查询样本的编码信息. 之后, 将编码信息分别输入 3 个解码器中. 3 个解码器在接收到编码信息后, 第 1 个解码器的目标是生成查询样本的纹理信息, 主要用于提高查询样本的类间多样性; 第 2 个解码器的目标是生成查询样本的各自信息的区域位置, 通过对第 2 个解码器的输出进行量化及取反操作得到查询样本的纹理区域掩码和平滑区域掩码; 第 3 个解码器的目标是生成查询样本的低频信息, 通过对解码器的输出进行滤波得到查询样本的平滑信息, 用于提高查询样本的类内多样性. 得到 3 个解码器的输出后, 我们对第 1 个解码器生成的纹理信息和纹理掩码, 以及第 3 个解码器生成的平滑信息和平滑掩码进行哈达玛积 (Hadamard product), 再将两个结果加和得到我们最终的查询样本. 此外, 我们提出了一种针对生成查询样本低频信息的类内多样性损失, 使 VFDA 方法生成的查询样本更加符合真实样本的分布. 最后, 我们将得到的查询样本分别送入代替模型和目标黑盒模型, 计算样本多样性损失来训练我们的查询样本生成模型和蒸馏损失来使代替模型学习目标模型.

本文的主要贡献如下:

(1) 提出了一种基于多解码器的无数据依赖模型窃取攻击的生成模型架构, 每个解码器分别用于生成查询样本的纹理信息、低频信息和各自信息的区域位置. 这种架构的设计可以帮助提高查询样本生成的多样性, 提高模型窃取的效率, 降低查询次数.

(2) 提出了一种基于代替模型 Logit 向量的类内多样性损失函数, 通过缩小该损失可以提高生成模型所生成的查询样本的类内多样性, 从而使查询样本更加符合真实样本的数据分布, 提高模型窃取的有效性.

(3) 与目前最先进的无数据依赖模型窃取攻击相比, 实验结果表明 VFDA 在查询样本的有效性方面具有优势,

并且在有限次数查询的情况下, 生成的对抗样本的攻击成功率有了明显的提高.

本文第 1 节介绍模型窃取攻击的相关方法和研究现状. 第 2 节对本文所提出的基于视觉特征解耦的无数据模型窃取攻击方法进行阐述. 第 3 节通过对比实验验证了所提模型的有效性. 最后总结全文.

1 相关工作

本文所提方法主要基于视觉特征解耦的无数据依赖模型窃取攻击方法, 下面就相关工作予以介绍.

1.1 对抗攻击

对抗攻击又称为对抗样本攻击, 是指向干净样本中添加精心设计的微小扰动, 从而导致被攻击目标模型错误输出的过程. 与传统的攻击不同, 对抗攻击不干扰模型的计算过程, 仅针对模型的输入进行微小扰动, 从而使模型通过正常的推理计算后输出错误的结果. 对抗样本的攻击过程可以简化为以下运算: 给定被攻击的目标模型 $F(\cdot)$, 被攻击的原始样本 $x_{\text{ori}} \in X_{\text{data}}$ 及其对应真实标签 $F(x_{\text{ori}}) = y_{\text{true}} \in Y$, 攻击者的目标是寻找到一个较小的对抗扰动 δ 添加到原始图像上, 从而使目标模型分类错误, 即 $F(x_{\text{ori}} + \delta) \neq y_{\text{true}}$. 同时, 为了限制扰动的不可察觉性, 扰动 δ 通常被攻击者使用 l_p 范数来约束扰动的大小, 即. 通常, 使用 l_0 范数作为约束的攻击被称为稀疏攻击, 旨在仅进行干净样本的局部来进行对抗扰动; 使用 l_2 范数作为约束的攻击被称为基于优化方法的攻击; 使用 l_∞ 范数作为约束的攻击被称为基于梯度的攻击. 整个攻击过程可以表示为:

$$F(x_{\text{ori}} + \delta) \neq y_{\text{true}} \text{ s.t. } \|\delta\|_p < \epsilon \tag{1}$$

对抗攻击的分类有许多种, 从攻击者能够获取到的被攻击模型信息程度来说, 攻击可以分为白盒攻击和黑盒攻击. 其中, 黑盒攻击中的迁移攻击因其在现实世界的适用性在近期受到广泛关注. 基于迁移的黑盒攻击是指利用对抗样本的可迁移性[10]来完成针对其他模型的黑盒攻击. 对抗样本的可迁移性指的是, 基于一个模型生成的对抗样本, 在一定程度上同样能够使其他模型输出错误的结果. Dong 等人[11]利用了深度学习中卷积神经网络的平移不变性, 对原始图像通过平移的数据增强方法来生成一组图像, 使用这一组图像共同基于一个白盒模型计算梯度并生成对抗扰动. 将基于这一组图像生成的对抗扰动添加到原始图像上, 并输入到黑盒模型中完成黑盒攻击. 基于这种方法生成的对抗样本减少了对固定的白盒模型过拟合的程度, 因此具有更强的可迁移性. Xie 等人[12]提出的 DIM 采用了类似的思想, 探索了多种数据增强方法对于提高对抗样本可迁移性的作用. 这种基于迁移的攻击方法能够在不需要查询目标模型的情况下完成黑盒攻击, 但是, 上述的传统迁移攻击方法完成攻击的前提是需要一个训练良好的白盒模型, 这就需要使用到目标模型的训练数据知识, 而通常攻击者是无法获取模型的训练数据的.

1.2 模型窃取攻击

为了解决在现实世界完成迁移攻击所需要本地白盒模型的问题, 近些年提出了模型窃取攻击, 即通过一系列样本查询目标模型得到返回结果, 并依据这些样本和返回的查询结果在本地训练一个白盒的代替模型, 来完成迁移攻击. 模型窃取攻击是一种针对机器学习模型的攻击方法, 其目的是通过向目标模型发起一系列查询来构建一个与目标模型具有相同行为模式的代替模型, 从而窥探目标模型的内部结构和参数.

在模型窃取攻击中, 攻击者会利用一组输入和相应的输出来发起查询, 以探测目标模型的行为模式. 攻击者通常会设计特定的查询策略, 以最小化查询次数并尽可能地获取有用的信息. 一旦攻击者构建出与目标模型具有相同行为模式的代替模型, 就可以利用对抗样本的可迁移性, 在代替模型上制作对抗样本来攻击目标模型. 模型窃取攻击是一种非常具有挑战性的攻击方法, 因为它要求攻击者在没有访问目标模型内部结构的情况下, 通过有限的查询来了解目标模型的行为. 模型窃取攻击与知识蒸馏具有一定的相似性, 但是模型窃取攻击无法像知识蒸馏一样以白盒的方式访问目标模型. 根据模型窃取是否对真实数据的依赖, 可以分为数据依赖的模型窃取攻击和无数据依赖的模型窃取攻击.

1.2.1 数据依赖的模型窃取攻击

Tramèr 等人[13]在 2016 年提出了一种针对机器学习中决策树、逻辑回归、支持向量机和简单深度学习模型

的模型窃取方法, 其按照模型输出的置信度与模型的输入结合建立求解方程, 尝试窃取到目标模型的参数. 然而, 这种方法仅能适用在目标模型给攻击者提供输出的置信度的情况, 极大限制了攻击者的模型窃取实施场景.

Papernot 等人[14]首先在代替模型上对原始输入使用基于雅可比矩阵 (Jacobi matrix) 进行数据增强并生成查询样本, 然后利用增强过的查询样本对目标模型窃取. 然而, 由于代替模型没有经过良好的训练, 在代替模型上使用雅可比矩阵进行数据增强而制作的查询样本在窃取过程中有效性较低.

Orekondy 等人[15]提出了 Knockoff 方法, 试图在一个巨大的数据集中找到有效的查询实例, 例如, ImageNet (ILSVRC 数据集的 1.2M 图像), 并在窃取过程中采取适应性策略. Knockoff 使用确定性损失来鼓励查询样本的多样性, 并使用多样性奖励来防止对单一标签的图像利用的退化情况. 但是, Knockoff 需要大量的查询数据, 并且无法适用于只有硬标签的场景.

Yang 等人[16]提出了 DSBF 方法, 该方法利用预先训练的生成对抗网络 (GAN) 来生成查询样本. 该方法采用类平衡和鲁棒筛选策略来提高查询样本的类间多样性问题.

He 等人[17]提出了一种新颖的利用扩散模型进行模型窃取攻击的方法, 该方法生成的查询样本更加真实, 视觉语义信息更强. 然而该方法依赖基于广泛收集数据训练的扩散模型, 无法在零知识的情况下对黑盒模型进行攻击.

1.2.2 无数据依赖的模型窃取攻击

Zhou 等人[18]在 2020 年提出了 DAST 方法, 这是第一个无数据依赖的模型窃取攻击方法. DAST 使用了多个生成器分别生成不同标签的查询样本, 来生成具有不同标签的合成查询数据集. 然而, DAST 的生成模型结构复杂、参数量较大且随着分类类别总数线性增长, 导致模型窃取效率极低, 查询次数大.

Truong 等人[19]提出了 DFME 方法, 该方法中代替模型的目标是缩小代替模型输出和目标模型输出之间的差异, 而生成器的目标是扩大两个模型输出的差异, 以对抗博弈的思想更新生成器的参数和代替模型的参数. 然而, 在这种对抗博弈中, 仅凭模型输出的差异来计算损失并不能如同典型生成对抗网络中的鉴别器一样给予生成器足够的指导, 导致生成的查询样本质量较差, 模型窃取的有效性不足.

Wang 等人[20]从生成器的中间层出发, 通过嵌入不同的条件信息, 并通过扩大不同样本之间的余弦距离来控制生成样本的类间多样性, 然而在黑盒的情况下目标模型无法为生成器提供足够的指导, 且在样本层面扩大样本距离并不能保证类别的多样性, 致模型窃取的效率较低.

Kariyappa 等人[21]提出了 MAZE 模型窃取方法, 该方法利用零阶梯度估计来指导生成器的训练, 然而该方法不能应用在目标模型仅能提供硬标签输出的场景.

Yu 等人[22]提出了 Fe-DaST 方法, 该方法在 DAST 方法的基础上, 压缩了生成器的大小, 并引入伪标签来提供信息熵, 从而提高查询样本的类间多样性.

Zhang 等人[23]在 2022 年提出的 EBFA 方法使用了伪标签, 通过缩小生成样本输入到目标模型得到的结果和伪标签之间的信息熵来使生成器生成具有类间多样性的生成样本. EBFA 提出的伪标签信息熵损失旨在提升了查询样本的多样性, 然而单一生成器的架构不能足够拟合伪标签计算的多样性损失, 无法充分实现标签带来的多样性提升.

在模型窃取攻击中, 攻击者需要构造请求目标模型的数据样本以获取对应的标签, 并用于代替模型的训练. 因此, 这些构造的样本的质量成为影响代替模型训练效果的重要因素. 在完成模型窃取后, 由于提取到的本地模型与目标模型高度相似, 攻击者能够基于提取到的本地模型来生成对抗样本, 从而利用对抗样本的可迁移性来完成对黑盒目标模型的对抗攻击. 由于在相似的模型中对抗样本的可迁移性更高, 可以看出, 由于模型窃取攻击的研究重点为如何使本地模型与目标模型决策边界更加相似, 从而提高对抗样本的迁移攻击成功率. 然而, 目前的无数据工作不能保证查询样本的类间多样性, 而且查询的有效性不足, 需要对目标模型进行大量的查询, 这大大增加了攻击者暴露的风险. 因此, 本文提出了一种新颖的无数据依赖模型窃取策略, 相比与上述无数据依赖的模型窃取攻击方法, 在查询样本的生成模型架构和查询样本的构造方法上提出了新的策略. 该策略基于神经网络的分类决策更依赖于样本的纹理信息启发[8,9], 从视觉特征解耦的角度生成查询样本. 并且提出了一种基于代替模型输出的 Logit

向量的类间多样性损失从而提高查询样本的多样性和模型窃取的有效性.

2 VFDA 基于视觉特征解耦的无数据依赖模型窃取攻击方法

在本文中, 我们提出的无数据依赖模型窃取攻击方法包含 3 个部分: 1) 使用生成模型生成查询样本. 2) 使用查询样本对目标黑盒模型进行模型窃取. 3) 基于提取到的代替模型生成对抗样本并利用对抗样本的可迁移性攻击目标黑盒模型. 本文所提出的基于视觉特征解耦的无数据依赖模型窃取攻击方法如图 2 所示.

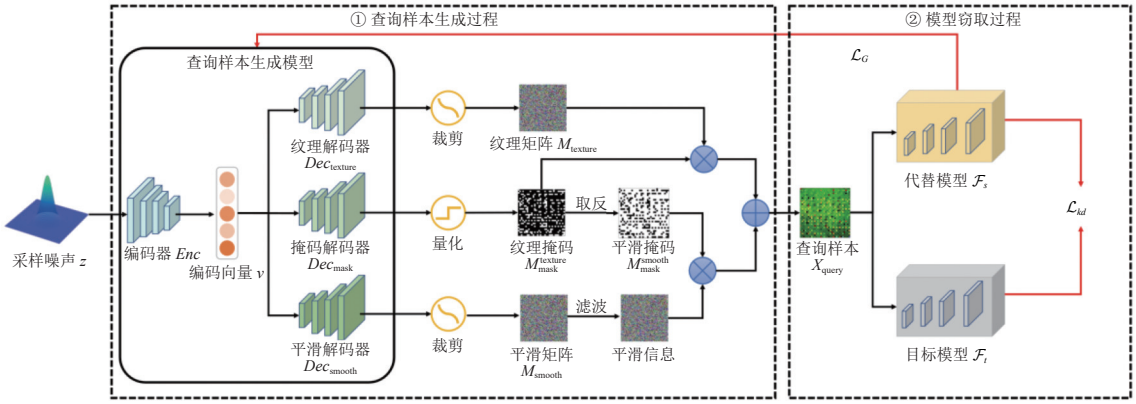


图 2 基于视觉特征解耦的模型窃取攻击方法总体框架图

2.1 问题描述

在 VFDA 无数据依赖模型窃取攻击方法的模型窃取阶段中, 有两个步骤: 1) 查询样本的生成, 2) 使用查询样本作为输入进行模型窃取. 对于步骤 1), 我们使用一个编码器 $Enc(\cdot)$ 对采样噪声 $z \in \mathcal{N}(0,1)$ 进行编码, 并将编码后得到的编码向量 $\mathcal{V}$ 送入 3 个不同的解码器 $Dec_{\text{texture}}(\cdot)$、$Dec_{\text{mask}}(\cdot)$ 和 $Dec_{\text{smooth}}(\cdot)$, 分别生成查询样本的纹理信息、位置掩码和平滑信息. 之后, 对 3 个解码器的输出进行融合来构造查询样本 X_{query}. 在步骤 2) 中, 我们将步骤 1) 中生成的 X_{query} 输入到代替模型 $\mathcal{F}_s(\cdot)$ 和目标模型 $\mathcal{F}_t(\cdot)$ 中, 使它们的输出差值最小, 从而使替代模型学习目标模型的行为模式, 提高两者的相似性. 模型窃取阶段的目标为:

$$\min_{\theta} D\left(\mathcal{F}_s^{\theta}(X_{\text{query}}), \mathcal{F}_t(X_{\text{query}})\right) \tag{2}$$

其中, θ 为代替模型的参数. D 为距离损失, 在本文中, 对于目标模型仅输出硬标签的情况, 我们使用交叉熵损失, 在目标模型能够输出 Logit 向量的情况, 我们使用 2 范数距离损失.

模型窃取阶段即为代替模型的训练阶段, 其训练样本为查询样本, 标签为查询目标模型返回的标签. 由此可见, 为了使代替模型更好地完成训练并模仿目标模型的决策输出, 查询样本的质量是关键因素. 即: 生成器生成的查询样本需要具有类间多样性和类内多样性. 最优查询样本的目标为:

$$min\left\{\left|\sum_{i=0}^{C}\mathbf{1}_{[\mathcal{F}_t(X_{\text{query}})=i]}-\frac{len(X_{\text{query}})}{C}\right|-\sum_{i}^{len(X_{\text{query}})}\sum_{j}^{len(X_{\text{query}})}\mathbf{1}_{[\mathcal{F}_t(X_{\text{query}}^{i})=\mathcal{F}_t(X_{\text{query}}^{j})]}*\left|X_{\text{query}}^{i}-X_{\text{query}}^{j}\right|\right\} \tag{3}$$

其中, $len(X_{\text{query}})$ 为查询样本的总数量, C 为目标模型分类类别数量, $\mathbf{1}$ 为指示函数, $|\cdot|$ 为绝对值计算. 在上式中, 第 1 项为查询样本类间多样性的优化目标, 即查询样本被目标模型分类的每一类别的样本分布数量应该是均衡的. 第 2 项为类内多样性的优化目标, 即同属于一个类别的查询样本之间应当具有一定的差异性. 因此, 最小化上述的目标, 生成的查询样本便能兼顾类间多样性和类内多样性, 更加符合真实样本的分布, 有利于代替模型的训练, 提高模型窃取的有效性.

因此, 在无数据模型窃取过程中, 生成器与替代模型的训练可视为一个最大最小化的对抗博弈过程. 生成器的优化目标是通过最大化查询样本在替代模型和目标模型预测输出的熵来提升查询样本的有效性, 而替代模型的训

练目标则是通过最小化其与目标模型输出的熵差异来逼近目标模型的行为. 因此, 模型窃取过程中生成器和代替模型的总体目标为:

$$\min_{F_s^\theta}\max_{G} E_{x\sim G}\left[H\left(\mathcal{F}_s^\theta(x)\right)-H\left(\mathcal{F}_t(x)\right)\right] \tag{4}$$

其中, H 为熵计算公式, 在图像分类任务中为交叉熵公式.

为了完成模型窃取之后的攻击, 经过模型窃取阶段后, 接下来是攻击阶段. 在模型窃取阶段中, VFDA 方法已经获取到一个与目标模型 $\mathcal{F}_t(\cdot)$ 相似的代替模型 $\mathcal{F}_s(\cdot)$, 因此接下来可以利用对抗样本的可迁移性进行攻击. 在攻击阶段中, 我们基于代替模型 $\mathcal{F}_s(\cdot)$ 来生成对抗样本 X_{adv}, 并期待生成的对抗样本能够最大程度地同样攻击成功目标模型. 攻击阶段的目标如下:

$$\max\mathcal{F}_t(X_{\text{adv}})\neq Y_{\text{true}},\ \text{s.t.}\ \|X_{\text{adv}}-X_{\text{ori}}\|_p\leqslant\epsilon \tag{5}$$

其中, X_{ori} 为对抗样本对应的原始样本, Y_{true} 为原始样本对应的真实标签, $\|\cdot\|_p$ 为计算对抗样本与原始样本差异的 p 范数, 即对抗样本扰动量的约束.ϵ 为对抗样本最大扰动量的限制.

因此, 基于上述两个阶段, $\mathcal{F}_s(\cdot)$ 可以以无数据的方式模仿 $\mathcal{F}_t(\cdot)$ 之后, 我们通过攻击 $\mathcal{F}_s(\cdot)$ 来产生对抗样本, 然后利用这些对抗样本的可迁移性来成功攻击. 值得注意的是, 代替模型 $\mathcal{F}_s(\cdot)$ 与目标模型 $\mathcal{F}_t(\cdot)$ 越相似, 即模型窃取的有效性越高, 对抗样本的可迁移性也就越高, 迁移攻击的成功率也就越高.

2.2 视觉特征解耦的查询样本生成

如上所述, 由于我们使用生成的查询样本作为代替模型的训练数据集, 而不是使用真实数据, X_{query} 应该具有与真实数据相似的类间多样性, 同时兼具提高模型窃取有效性的类内多样性. 关于这一点, VFDA 首先使用一个编码器 $Enc(\cdot)$, 以高斯噪声作为输入并输出与查询样本维度相等的编码向量 $\mathcal{V}$. 然后, VFDA 将编码向量分别输入 3 个解码器并分别生成查询样本的纹理信息、平滑信息和纹理信息的位置信息, 这个过程如图 2 的第 1 部分所示.

由于神经网络分类的类间决策依赖于图像的纹理信息, 我们使用解码器 $Dec_{\text{texture}}(\cdot)$ 和 $Dec_{\text{mask}}(\cdot)$ 来生成查询样本的纹理信息 X_{texture}. 其中, X_{texture} 由 $Dec_{\text{texture}}(\cdot)$ 生成的纹理矩阵 $M_{\text{texture}}=Dec_{\text{texture}}(\mathcal{V})$, 以及 $Dec_{\text{mask}}(\cdot)$ 生成的纹理信息所在区域的掩码矩阵 M_{mask} 计算组成. 纹理掩码矩阵 $M_{\text{mask}}^{\text{texture}}$ 的计算方式为:

$$M_{\text{mask}}^{\text{texture}}=\begin{cases}1, & \text{if } n \text{ in } M_{\text{mask}}>0.5\\ 0, & \text{if } n \text{ in } M_{\text{mask}}<0.5\end{cases} \tag{6}$$

其中, n 为 $Dec_{\text{mask}}(\mathcal{V})$ 输出矩阵中一个元素的值. VFDA 方法生成查询样本的纹理信息的过程为:

$$X_{\text{texture}}=M_{\text{texture}}\times M_{\text{mask}}^{\text{texture}},\quad \text{s.t.}\ \left\|M_{\text{mask}}^{\text{texture}}\right\|_0<\delta \tag{7}$$

其中, $\times$ 为矩阵的哈达玛积, $\|\cdot\|_0$ 为求矩阵的 0 范数, δ 为控制纹理稀疏性的超参数.

因此, 在 VFDA 方法中查询样本的纹理信息部分是由纹理矩阵以及纹理掩码矩阵共 $Dec_{\text{smooth}}(\cdot)$ 同组成的, 纹理矩阵负责生成纹理信息的像素值的大小, 纹理掩码矩 $\mathcal{V}$ 阵负责生成纹理信息的所在位置, 并同时考虑了纹理信息应具有一定的稀疏性.

其次, 为了生成更加具有类内多样性的样本, VFDA 还需利用解码器 $Dec_{\text{smooth}}(\cdot)$ 来生成查询样本的平滑信息 X_{smooth}.X_{smooth} 同样包含平滑矩阵 M_{smooth} 和平滑掩码 $M_{\text{mask}}^{\text{smooth}}$. 平滑矩阵是由解码器 $Dec_{\text{smooth}}(\cdot)$ 接收编码向量作为输入, 并生成查询样本的平滑矩阵 M_{smooth}, 之后, 我们对平滑矩阵进行双边滤波, 以消除其中包含的纹理信息.M_{smooth} 同样需要位置编码, 我们对 $Dec_{\text{mask}}(\cdot)$ 解码器生成的纹理信息位置编码 $M_{\text{mask}}^{\text{texture}}$ 进行取反, 得到平滑信息的位置编码 $M_{\text{mask}}^{\text{smooth}}$. 即, $M_{\text{mask}}^{\text{smooth}}=\sim M_{\text{mask}}^{\text{texture}}$. VFDA 方法生成查询样本平滑信息 X_{smooth} 的过程为:

$$X_{\text{smooth}}=G(M_{\text{smooth}})\times M_{\text{mask}}^{\text{smooth}},\quad \text{where } M_{\text{mask}}^{\text{smooth}}=\sim M_{\text{mask}}^{\text{texture}} \tag{8}$$

其中, 符号~表示按位取反操作, $G(\cdot)$ 为双边滤波函数.

得到由 3 个解码器生成构造的纹理信息 X_{texture} 和平滑信息 X_{smooth} 后, 由于纹理信息和平滑信息的位置编码互补, 因此我们将其直接相加便能得到我们的查询样本 X_{query}. 即:

$$X_{\text{query}}=X_{\text{texture}}+X_{\text{smooth}} \tag{9}$$

因此, 经过以上步骤, 我们使用了不同的解码器分别生成查询样本的纹理信息、位置编码和平滑信息, 达到所生成查询样本的视觉特征解耦.

2.3 通过查询样本进行模型窃取及攻击

在 VFDA 生成查询样本之后, 便可以使用查询样本 X_{query} 查询目标黑盒模型并根据输出来训练本地的代替模型 $\mathcal{F}_s(\cdot)$, 完成模型窃取, 如图 2 的第 2 部分所示. 在本文中, 我们考虑了模型窃取过程中的两种情况: 软标签 (soft label) 场景和硬标签 (hard label) 场景. 在软标签场景中, 我们使用查询样本 X_{query} 查询目标黑盒模型并获得黑盒模型输出的逻辑向量, 而在硬标签场景中, 我们只能获得黑盒模型输出的硬标签. 在这一过程中, VFDA 的目标为:

$$\min_{\theta}\begin{cases} D(\mathcal{F}_s^{\theta}(X_{\text{query}}),\mathcal{F}_t(X_{\text{query}})), & \text{soft label} \\ -\log(\mathcal{F}_s^{\theta}(X_{\text{query}})*\mathbf{1}_{\mathcal{F}_t(X_{\text{query}})=\text{argmax}\mathcal{F}_s(X_{\text{query}})}, & \text{hard label} \end{cases} \tag{10}$$

其中, $\mathbf{1}$ 为指示函数, 当黑盒模型输出的硬标签等于代替模型输出的硬标签时为 1, 反之为 0.

完成模型窃取后, 在攻击阶段使用常见的攻击算法来评估模型窃取的有效性, 在本文中, 我们评估了 3 个经典攻击算法 (FGSM[2]、BIM[24] 和 PGD[25]) 在模型窃取后的迁移攻击成功率.

2.4 目标函数

本节将用公式表示 VFDA 方法进行模型窃取任务中的上述目标. 在查询样本的生成中, 编码器 $Enc(\cdot)$ 和 3 个解码器 $Dec_{\text{texture}}(\cdot)$、$Dec_{\text{mask}}(\cdot)$ 和 $Dec_{\text{smooth}}(\cdot)$ 的损失函数包含类间多样性损失 $\mathcal{L}_{\text{inter-diverse}}$、类内多样性损失 $\mathcal{L}_{\text{intra-diverse}}$ 和纹理稀疏损失 $\mathcal{L}_{\text{sparse}}$. 为了生成类间多样性样本, 我们使用类间信息熵来指导生成器编码器和解码器. 也就是说, VFDA 随机设置一批目标标签 y_{target}, 并减少其与代替模型以 X_{query} 作为输入的计算输出之间的熵, 类间损失函数为:

$$\mathcal{L}_{\text{inter-diverse}} = -\sum_{i=1}^{B} y_{\text{target}}\log[\mathcal{F}_s(X_{\text{query}})] \tag{11}$$

其中, B 是批次大小. 类间多样性损失函数的目标为扩大生成样本对目标模型预测的信息熵, 从而使生成模型的训练满足与代替模型对抗博弈的过程.

其次, 为了使我们的查询样本具有更好的类内多样性, 即同属于一个类别的样本之间同样应该具有不同的视觉特征, 我们对生成平滑信息的解码器进行了类内多样性损失的约束. 类内多样性损失 $\mathcal{L}_{\text{intra-diverse}}$ 的损失函数为:

$$\mathcal{L}_{\text{intra-diverse}} = \log\sum_{i}^{B}\sum_{j}^{B}\mathbf{1}_{[i\neq j]}\text{Logit}\left[\mathcal{F}_s(X^i_{\text{query}})\right]\cdot\text{Logit}\left[\mathcal{F}_s(X^j_{\text{query}})\right] \quad \text{where} \quad \mathcal{F}_s(X^i_{\text{query}}) = \mathcal{F}_s(X^j_{\text{query}}) \tag{12}$$

其中, B 是批次大小, Logit 为未经过 Softmax 的代替模型输出. 通过扩大点积计算同一类别不同样本之间的 Logit 向量的距离, 便能扩大同一类别样本之间 Logit 向量的差异, 从而提高查询样本的类内多样性. 当同一类别样本之间具有类内多样性时, 样本之间 Logit 向量的差异更大. 不具有类内多样性时, 样本之间 Logit 向量的差异更小. 类内多样性损失函数的目标为扩大属于同一个类别的查询样本之间的 KL 散度, 从而进一步提高样本的多样性.

为了使查询样本更符合真实样本的视觉特征, 我们对解码器 $Dec_{\text{mask}}(\cdot)$ 生成查询样本的纹理区域稀疏性进行了约束, 使查询样本的纹理区域和平滑区域处于一个合理的范围. 纹理稀疏性约束的损失函数为:

$$\mathcal{L}_{\text{sparse}} = \left|\|\text{clip}(Dec_{\text{mask}}(\mathcal{V}))\|_0 - \delta\right| \tag{13}$$

其中, $\|$ 为计算绝对值, clip 表示对 $Dec_{\text{mask}}(\mathcal{V})$ 的输出进行取整, 输出为仅包含 0 或 1 的矩阵, $\|\cdot\|_0$ 表示矩阵的 0 范数, δ 为控制纹理稀疏度的超参数. 在本文中, 对于 $H\times W\times C$ 大小的图像, 我们设置 δ 的值为 $(H\times W)/10$.

总的来说, 生成模型的整体损失函数为:

$$\mathcal{L}_G = \alpha\times\mathcal{L}_{\text{inter-diverse}} + \beta\times\mathcal{L}_{\text{intra-diverse}} + \gamma\times\mathcal{L}_{\text{sparse}} \tag{14}$$

其中, α,β 和 γ 为控制损失平衡的超参数.

通过 3 个解码器生成我们的查询样本后, 下一阶段是将它们输入代替模型 $\mathcal{F}_s(\cdot)$ 和目标模型 $\mathcal{F}_t(\cdot)$, 并使它们的输出差值最小. 模型窃取的损失函数 $\mathcal{L}_{kd}$ 为:

$$\mathcal{L}_{kd} = \sum_{i=1}^{B} d(\mathcal{F}_s(X_{\text{query}}),\mathcal{F}_t(X_{\text{query}})) \tag{15}$$

其中距离函数 $d(\cdot)$ 在硬标签情况下为交叉熵 (CrossEntropy) 损失, 在软标签情况下为均方误差损失 (mean square error) 损失.

3 实验分析

3.1 实验数据

在 VFDA 的模型窃取攻击实验中, 本文评估了 VFDA、基于真实数据的模型窃取攻击和无数据依赖模型窃取攻击的效果. 针对图像分类模型, 本文研究了在 CIFAR-10[26] 、CIFAR-100[26] 、GTSRB[27] 和 Tiny-ImageNet[28] 标准数据集上的模型窃取和攻击效果.

3.2 评价指标及实验细节

我们利用 3 种经典的攻击方法来评估 VFDA 和对比方法的模型窃取效果, 其中包括 FGSM[2] 、BIM[24] 和 PGD[25] 来产生对抗样本. 在所有数据集上, 我们设定扰动预算 $\epsilon = 8/255$, 对于 BIM 和 PGD 方法, 我们设置步长 $\alpha = 2/255$. 迭代次数为 10 次. 在攻击成功率的评估过程中, 我们仅采用能够成功攻击本地模型的对抗样本去评估在黑盒模型上的攻击成功率. 攻击成功率 (attack success rate, ASR) 的计算方法为:

$$ASR = \frac{1}{N}\sum_{i=1}^{N}[\mathcal{F}_t(X_i^{\text{adv}}) \neq Y_i] \text{ where } \mathcal{F}_s(X_i^{\text{adv}}) \neq Y_i \text{ and } \mathcal{F}_s(X_i^{\text{ori}}) = \mathcal{F}_t(X_i^{\text{ori}}) = Y_i \tag{16}$$

其中, N 为评估过程样本总数, 在本文中, 评估样本数量为随机挑选的 1 000 张图片, 并攻击成功率取重复 10 次实验的平均值, X_i^{adv} 为对抗样本, X_i^{ori} 为原始样本, Y_i 为原始样本对应的真实标签, $\mathcal{F}_t$ 为黑盒模型, $\mathcal{F}_s$ 为本地代替模型. 在本文中, 所有攻击成功率实验结果均以百分比 (%) 形式呈现.

在 VFDA 和所有对比方法的查询次数限制上, 对于任何一个数据集我们限制 20 万次/张的查询, 批次大小为 256. 在超参数上, 为了控制每个损失之间数量级的一致性, 我们设置 α 为 1, β 为 0.000 1, γ 为 0.5. 对于生成模型, 我们使用 Adam 优化器, 学习率为 0.001, 动量衰减率为 [0.5, 0.999]. 对于代替模型, 我们使用 SGD 优化器, 学习率为 0.01, 动量衰减率为 0.9.

在代替模型的选择上, VFDA 与所有对比方法均使用 ResNet18[29] 作为代替模型, 目标黑盒模型为 ResNet50[29].

3.3 模型窃取攻击实验结果

3.3.1 模型窃取攻击成功率评估

为了评估 VFDA 模型窃取攻击的攻击能力, 我们首先在较低分辨率的数据集 CIFAR-10 和 CIFAR-100 中与 MAZE[21] 、Fe-DaST[22] 、Del[20] 和 EBFA[23] 方法进行对比实验. Soft label 是指目标模型返回结果为预测概率向量的情况下, 使用均方误差损失训练本地模型. Hard label 是指目标模型返回结果为硬标签的情况下, 使用交叉熵损失训练本地模型. 我们同时评估了使用 FGSM、BIM、PGD 这 3 种常见的攻击方法来进行攻击成功率的评估. 实验结果如后文表 1 所示. 可以看出, 在 CIFAR-10 和 CIFAR-100 数据集上, VFDA 在多数情况下取得了最高的攻击成功率, 尤其是在仅能查询硬标签情况下, VFDA 方法在攻击成功率上更具优势.

为了进一步评估 VFDA 的模型窃取攻击效果, 我们在分辨率更大的数据集 GTSRB 和 Tiny-ImageNet 上评估了 VFDA 与对比方法的模型窃取攻击成功率, 结果如后文表 2 所示. 可以看出, 在较大分辨率的数据集上, 我们提出的 VFDA 方法具有更加优异的攻击成功率表现. 从实验结果来看, VFDA 利用 FGSM、PGD、BIM 攻击方法生成对抗样本的攻击成功率均在 80% 以上.

3.3.2 模型窃取迁移攻击评估

此外, 我们在 Microsoft Azure 上进行了攻击在线模型的实验, 在不知道模型内部参数和结构的情况下, 仅通过模型输出的硬标签来窃取目标模型. 实验结果如表 3 所示. 其中, Valina 是在查询在线模型 100 个 Epoch 的常规模型窃取攻击的结果. Transfer 考虑了更加实际的情况, 是指在本地基于 Fashion 数据集的预窃取模型及生成样本,

对在线模型进行 (1~10) 个 Epoch 查询下窃取的结果, 两个数据集属于不同类型且数据集之间无交叉. 所有的对抗样本生成方法统一为 BIM. ACC 是指窃取到的代替模型的识别准确率, ASR 是指攻击成功率. 可以看出, 在模型窃取迁移攻击的实验中, 前 6 个 Epoch 的查询中 VFDA 在识别准确率和攻击成功率上均具有明显的优势. 在更加实际的对在线模型进行迁移窃取场景中, VFDA 具有更高的效率.

表 1　CIFAR-10 和 CIFAR-100 数据集上模型窃取攻击成功率 (%)

Datasets	Methods	Soft label			Hard label		
		FGSM	BIM	PGD	FGSM	BIM	PGD
CIFAR-10	Del	26.38	31.53	31.47	25.33	30.45	30.34
	MAZE	53.26	59.11	52.48	—	—	—
	Fe-DaST	63.99	64.16	64.36	65.98	63.76	68.25
	EBFA	83.89	87.68	**89.11**	**86.13**	87.02	84.32
	VFDA	**85.34**	**93.19**	85.41	83.66	**90.06**	**86.73**
CIFAR-100	Del	31.64	36.63	37.44	30.8	35.63	36.15
	MAZE	47.37	50.27	46.36	—	—	—
	Fe-DaST	65.81	65.77	67.74	66.33	65.91	67.94
	EBFA	83.69	**94.53**	**94.14**	78.61	85.31	81.21
	VFDA	**84.58**	92.64	90.19	**81.43**	**90.51**	**86.61**

表 2　GTSRB 和 Tiny-ImageNet 数据集上模型窃取攻击成功率 (%)

Datasets	Methods	Soft label			Hard label		
		FGSM	BIM	PGD	FGSM	BIM	PGD
GTSRB	Del	31.43	36.26	33.87	30.8	34.14	32.45
	MAZE	46.58	54.75	46.84	—	—	—
	Fe-DaST	67.55	74.3	72.95	67.7	70.86	68.87
	EBFA	75.19	79.94	80.16	78.61	80.93	**89.3**
	VFDA	**84.22**	**87.39**	**85.98**	**82.04**	**85.73**	81.93
Tiny ImageNet	Del	34.28	38.49	36.72	28.31	32.54	29.73
	MAZE	—	—	—	—	—	—
	Fe-DaST	72.83	76.5	75.45	70.82	73.16	72.83
	EBFA	80.26	85.32	78.29	78.29	81.12	78.23
	VFDA	**85.02**	**85.66**	**80.62**	**83.73**	**89.21**	**83.14**

表 3　Microsoft Azure 在线模型窃取攻击实验结果 (%)

Method	Metrics	Valina	Transfer									
			1	2	3	4	5	6	7	8	9	10
EBFA	ACC	90.62	**9.96**	16.31	24.61	32.42	39.75	56.54	64.84	75.59	81.64	**87.11**
	ASR	97.34	**0.89**	1.38	5.13	12.18	26.57	51.18	**75.39**	85.4	91.74	96.06
VFDA	ACC	**91.99**	**9.96**	**22.46**	**33.69**	**45.51**	**59.28**	**67.58**	**65.14**	**79.69**	**83.59**	86.04
	ASR	**97.46**	0.59	**9.72**	**16.99**	**28.8**	**44.38**	**62.44**	74.83	**88.67**	**93.79**	**96.55**

3.3.3　模型窃取攻击效率评估

为了证明我们所提出的 VFDA 方法利用 3 个解码器生成的查询样本在模型窃取阶段中的有效性, 并证明 VFDA 方法的高效, 我们对比了 VFDA 方法与目前最先进的无数据的模型窃取攻击方法 EBFA 在模型窃取过程中的模型识别准确率 (ACC) 和攻击成功率 (ASR), 如图 3 所示, 其中, 红色是我们的方法在模型窃取过程中的折线图, 蓝色是 EBFA 方法在模型窃取过程中的折线图. 纵坐标为窃取查询的次数, 每次表示一个 BatchSize 的查询. 图 3(a) 表示在模型窃取过程中, 使用真实数据集的测试集对代替模型进行测试得到的识别准确率, 这表示着代替模型窃取目标黑盒模型行为模式的有效性. 图 3(b) 是在模型窃取阶段 VFDA 和 EBFA 的攻击成功率对比. 可以看

出, 受益于 VFDA 的 3 个解码器结构, 即使解码器还未收敛时, 通过纹理信息和平滑信息在不同区域位置的计算得到查询样本, 在查询初期就能够大大提高查询样本的有效性. 同时, 受益于查询样本有效性带来的代替模型识别准确率的提高, VFDA 的攻击成功率在查询初期也能够快速提升.

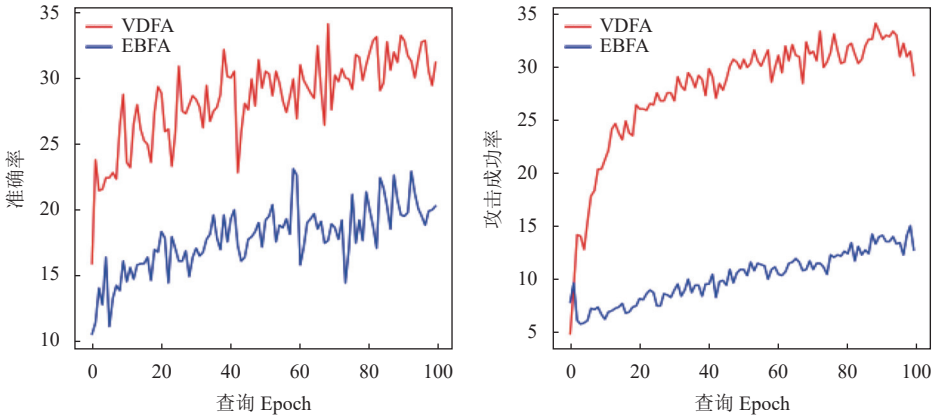


图 3　在 CIFAR-10 数据集上的模型窃取过程中模型收敛速度对比

3.3.4　样本生成及训练时间开销评估

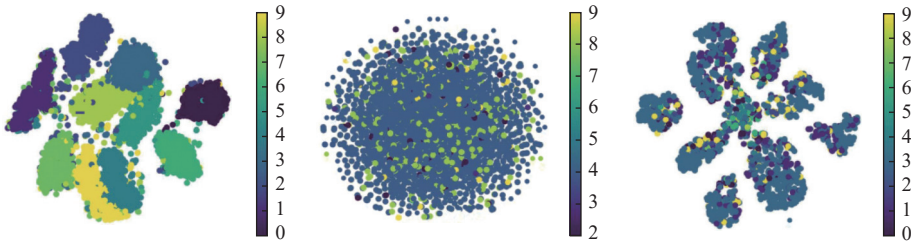
本文提出的 VFDA 方法的查询样本生成过程使用了 1 个编码器和 3 个解码器的结构, 相较于基线方法生成模型结构更复杂、待优化的参数量更多. 为评估 VFDA 方法的生成模型和基线方法效率之间的差异, 本文在 CIFAR10 数据集中对比不同方法之间的效率. 本项实验利用 VFDA 方法和 EBFA 方法生成 100 个 batch 的样本, batch 大小为 256, 实验环境为 TeslaA100×2. 计算时间包括生成器生成样本的前向传播过程、损失计算过程、反向传播过程和参数优化过程的时间消耗. 实验结果如表 4 所示. 可以看出, 使用 3 个解码器结构的 VFDA 方法在生成模型的训练过程中需要更长的时间, 但也在可接受的范围之内, 相比于整个模型窃取的总时间, 仅增加了少量的额外时间开销.

表 4　VFDA 方法模型训练耗时对比 (s)

Methods	生成模型训练每batch耗时			100个batch训练总耗时
	最小值	平均值	最大值	
EBFA	0.0753	0.0917	0.1440	1181.5027
VFDA	0.214	0.3129	0.4536	1250.0196

3.4　模型窃取查询样本对比分析

如前文所述, 在模型窃取任务中, 查询样本的生成质量是影响模型窃取效果和攻击成功率至关重要的因素. 因此, 我们对比了真实数据、EBFA 方法与我们提出的 VFDA 方法生成的查询样本经过主成分分析 (principal component analysis, PCA) 后画出的 T-SNE 图, 如图 4 所示. 我们使用相同的高斯分布样本作为生成器的输入, 共生成了 10 批查询样本, 批次大小为 256. 图中每种颜色代表一个类别, 可以看出, VFDA 生成的查询样本与真实数据相似, 具有更多的类间多样性. 对比 EBFA 和 VFDA, 可以看出 VFDA 生成的数据在特征空间中分布广泛, 分类差异比较明显, 而 EBFA 生成的数据类间差距比较小, 集中在特征空间的一部分, 不利于代替模型的训练. VFDA 方法能够在低维保持成群状的同时, 同一类别能够分布在不同的群落. 这意味着我们的方法生成的查询样本, 即使是同一个类别的查询样本, 其依然保持着足够的类内多样性, 这对于代替模型学习目标模型的决策边界是有帮助的. 这些也进一步验证了生成的查询样本数据分布对代替模型训练的重要性.



(a) 真实数据集 (b) EBFA (c) VFDA

图 4 在 CIFAR-10 数据集上, 原始数据、EBFA 生成的查询样本和 VFDA 生成的查询样本的 T-SNE 图

3.5 消融实验

3.5.1 不同生成器架构对模型窃取的影响

为进一步验证三解码器架构在视觉特征解耦方面的优势, 我们对比了在 CIFAR-10 数据集上三解码器和单生成器在生成查询样本时的性能表现, 特别是前 100 个批次的代替模型训练中的交叉熵和识别准确率, 结果如图 5 所示. 其中, 三解码器相比单生成器编码器部分完全相同, 仅在解码器部分具有额外的两个解码器. 正如前文公式 (4) 所述, 在模型窃取过程中, 生成查询样本的生成器与代替模型的训练过程构成了一种对抗博弈. 生成模型的目标是尽可能增大代替模型与目标模型输出之间的交叉熵, 而代替模型的训练目标则是尽量减小该交叉熵. 从图 5 可以看出, 在模型窃取初期, 三解码器能够生成具有更大交叉熵的查询样本, 从而更有效地为代替模型提供训练数据. 虽然在生成样本的交叉熵均值上, 三解码器和单生成器在后期表现接近, 但三解码器在代替模型训练中的优势更为显著.

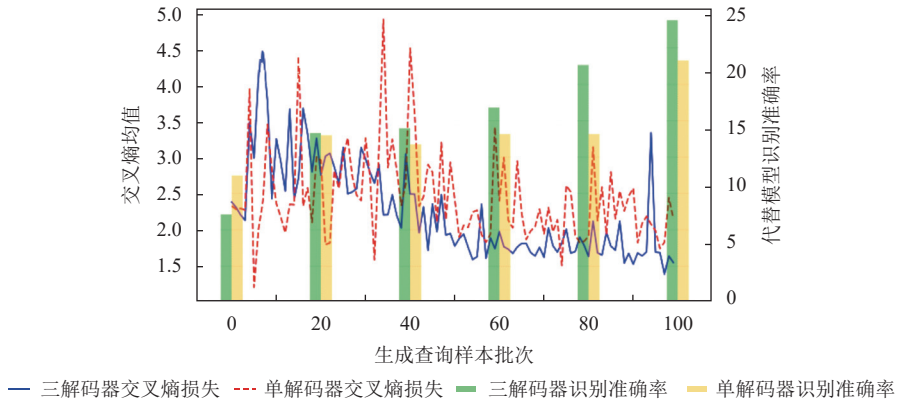


图 5 不同生成器架构在模型窃取过程中对代替模型训练交叉熵和识别准确率的影响

3.5.2 类间多样性损失和类内多样性损失对查询样本信息熵和 KL 散度的分析

为了进一步分析类间多样性损失和类内多样性损失对查询样本交叉熵的影响, 我们对比了在有类间多样性损失和无类间多样性损失的情况下, 不同类别查询样本的信息熵. 结果如图 6 所示. 可以看出, 在引入类间多样性损失时, 不同类别的查询样本对目标模型产生了更高的信息熵, 从而为代替模型的训练提供了更加有效的监督信息. 为了探究纹理解码器和平滑解码器对于信息熵的影响, 我们对比了在仅保留纹理解码器生成的查询样本和仅保留平滑解码器生成的查询样本的信息熵. 其中, 纹理解码器的输出仅保留了稀疏的纹理掩码后的结果, 图像其余大部分区域为 0 像素填充, 平滑解码器同样保留了平滑掩码后的结果, 极少部分稀疏纹理区域用 0 像素填充. 可以看出, 即使仅保留少量纹理区域, 多数情况下查询样本具有更高的信息熵. 此外, 我们还计算了在有无类内多样性损失情况下, 各类别生成 256 张查询样本, 每个类别内部每张样本 KL 散度的均值, 结果如图 7 所示. 结果表明, 类内

多样性损失使得每个类别的查询样本具有更高的 KL 散度, 从而提高了查询样本的多样性, 进一步说明了类内多样性损失有助于生成更具差异化的查询样本, 有效增强了代替模型学习目标模型决策的能力.

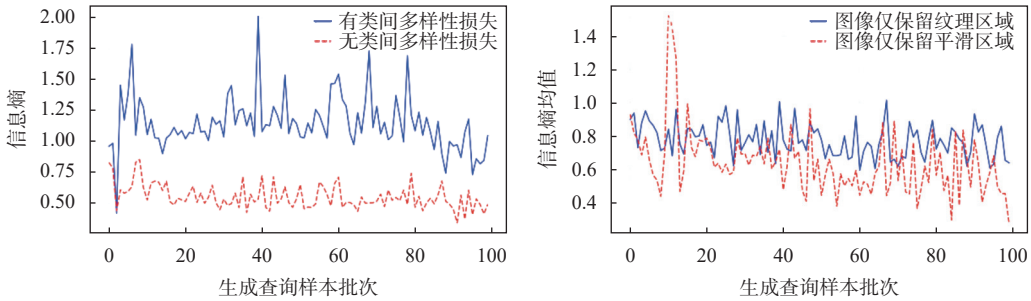


图 6　类间多样性损失和不同解码器对样本信息熵的影响

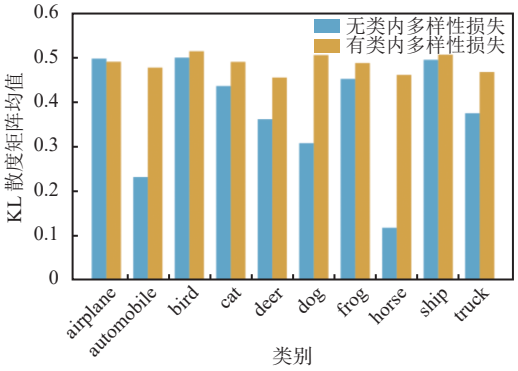


图 7　类内多样性损失对同一类别内样本 KL 散度的影响

4　总结与展望

在本文中, 我们重点讨论了无数据依赖模型窃取黑盒攻击中查询样本的特征解耦及生成. 尽管以前的方法试图以无数据的方式实现模型窃取攻击, 但他们用来训练代替模型的查询样本并不具有与真实数据相似的类间多样性和类内多样性, 导致模型窃取的有效性较低. 为了解决这些问题, 我们提出基于视觉特征解耦的无数据依赖模型窃取攻击算法 (VFDA), 利用了 3 个解码器来生成类间多样性和类内多样性的查询样本. VFDA 的 3 个解码器被用来分别生成查询样本视觉特征的纹理信息、平滑信息以及各自信息的所在区域, 完成查询样本的视觉特征解耦生成. 在得到 3 个解码器的输出后, 我们对其进行得融合到我们最终的查询样本. 此外, 我们对生成模型的训练过程中提出了基于 Logit 向量的类内多样性损失, 从而使 VFDA 方法生成的查询样本更加符合真实样本的分布. 进行了大量的实验, 结果表明, 我们所提出的 VFDA 方法相比于最先进的无数据依赖模型提取算法在有限的查询次数中达到了更加优异的攻击成功率. 此外, 通过对实验结果的进一步分析, 我们更进一步地证明了神经网络的决策输出依赖于样本的纹理信息, 而不是平滑信息. 同时, 也证明了 VFDA 方法在无数据依赖模型窃取攻击中的有效性.

受限于生成模型的生成能力, 在模型结构差异较大、查询次数较少的情况下, 无数据模型窃取攻击的成功率较低. 未来的工作可以通过引入更强大的生成模型进一步探索深度学习模型在面对模型窃取攻击时的脆弱性. 此外, 现有的模型窃取攻击方法主要集中于单一图像分类模型的研究. 然而, 随着人工智能技术的迅速发展, 多模态模型和各类大型预训练模型逐渐进入研究者的视野. 越来越多的机构和企业提供商用多模态模型接口, 而这些模型通常被视为宝贵的知识产权. 因此, 探讨更为先进的模型窃取攻击方法不仅可以推动相关技术的发展, 还能为如何有效保护模型版权提供新的思路和方法. 如何将模型窃取攻击扩展到多模态模型和预训练大模型的领域, 已成

为一个亟待解决的重要研究方向.

References:

[1] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv:1312.6199, 2014.

[2] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2015.

[3] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao/Journal of Software, 2020, 31(1): 67–81 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

[4] Ren K, Meng QR, Yan SK, Qin Z. Survey of artificial intelligence data security and privacy protection. Chinese Journal of Network and Information Security, 2021, 7(1): 1–10 (in Chinese with English abstract). [doi: 10.11959/j.issn.2096-109x.2021001]

[5] Heo B, Lee M, Yun S, Choi JY. Knowledge distillation with adversarial samples supporting decision boundary. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 3771–3778. [doi: 10.1609/aaai.v33i01.33013771]

[6] Wang L, Yoon KJ. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048–3068. [doi: 10.1109/TPAMI.2021.3055564]

[7] Haroush M, Hubara I, Hoffer E, Soudry D. The knowledge within: Methods for data-free model compression. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8491–8499. [doi: 10.1109/CVPR42600.2020.00852]

[8] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231, 2022.

[9] Wang HH, Wu XD, Huang ZY, Xing EP. High-frequency component helps explain the generalization of convolutional neural networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8681–8691. [doi: 10.1109/CVPR42600.2020.00871]

[10] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv:1607.02533, 2017.

[11] Dong YP, Pang TY, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4307–4316. [doi: 10.1109/CVPR.2019.00444]

[12] Xie CH, Zhang ZS, Zhou YY, Bai S, Wang JY, Ren Z, Yuille AL. Improving transferability of adversarial examples with input diversity. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2725–2734. [doi: 10.1109/CVPR.2019.00284]

[13] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing machine learning models via prediction APIs. In: Proc. of the 25th USENIX Security Symp. Austin: USENIX Association, 2016. 601–618.

[14] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proc. of the 2017 ACM on Asia Conf. on Computer and Communications Security. Abu Dhabi: ACM, 2017. 506–519. [doi: 10.1145/3052973.3053009]

[15] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4949–4958. [doi: 10.1109/CVPR.2019.00509]

[16] Yang PP, Wu QL, Zhang XM. Efficient model extraction by data set stealing, balancing, and filtering. IEEE Internet of Things Journal, 2023, 10(24): 22717–22725. [doi: 10.1109/JIOT.2023.3304345]

[17] He JP, Gao HC, Zhou YY. Enhancing data-free model stealing attack on robust models. In: Proc. of the 2024 Int'l Joint Conf. on Neural Networks (IJCNN). Yokohama: IEEE, 2024. 1–8. [doi: 10.1109/IJCNN60899.2024.10650742]

[18] Zhou MY, Wu J, Liu YP, Liu SC, Zhu C. DaST: Data-free substitute training for adversarial attacks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 231–240. [doi: 10.1109/CVPR42600.2020.00031]

[19] Truong JB, Maini P, Walls RJ, Papernot N. Data-free model extraction. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4769–4778. [doi: 10.1109/CVPR46437.2021.00474]

[20] Wang WX, Yin BJ, Yao TP, Zhang L, Fu YW, Ding SH, Li JL, Huang FY, Xue XY. Delving into data: Effectively substitute training for black-box attack. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4759–4768. [doi: 10.1109/CVPR46437.2021.00473]

[21] Kariyappa S, Prakash A, Qureshi MK. MAZE: Data-free model stealing attack using zeroth-order gradient estimation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13809–13818. [doi: 10.1109/CVPR46437.2021.01360]

[22] Yu MR, Sun SL. FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks. Computers & Security, 2022,

113: 102555. [doi: 10.1016/j.cose.2021.102555]
[23] Zhang J, Li B, Xu JH, Wu S, Ding SH, Zhang L, Wu C. Towards efficient data free blackbox adversarial attack. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15094–15104. [doi: 10.1109/CVPR52688.2022.01469]
[24] Dong YP, Liao FZ, Pang TY, Su H, Zhu J, Hu XL, Li JG. Boosting adversarial attacks with momentum. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9185–9193. [doi: 10.1109/CVPR.2018.00957]
[25] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083, 2019.
[26] Krizhevsky A. Learning multiple layers of features from tiny images. Department of Computer Science, University of Toronto, 2009.
[27] Stallkamp J, Schlipsing M, Salmen J, Igel C. The German traffic sign recognition benchmark: A multi-class classification competition. In: Proc. of the 2011 Int'l Joint Conf. on Neural Networks. San Jose: IEEE, 2011. 1453–1460. [doi: 10.1109/IJCNN.2011.6033395]
[28] Le Y, Yang X. Tiny imagenet visual recognition challenge. CS 231N, 2015, 7(7): 3.
[29] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

附中文参考文献:

[3] 潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67–81. http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]
[4] 任奎, 孟泉润, 闫守琨, 秦湛. 人工智能模型数据泄露的攻击与防御研究综述. 网络与信息安全学报, 2021, 7(1): 1–10. [doi: 10.11959/j.issn.2096-109x.2021001]

张锦弘(1999－), 男, 博士生, 主要研究领域为人工智能模型安全, 多媒体数据隐私保护.

董云云(1989－), 女, 讲师, CCF 专业会员, 主要研究领域为图像隐写, 数据隐私保护.

刘仁阳(1994－), 男, 博士, 主要研究领域为人工智能模型安全, 多模态大模型遗忘学习, 数据隐私保护.

周维(1974－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为网络空间安全, 分布式云计算, 智能信息处理, 交叉学科研究.

韦廷楚(1997－), 男, 博士生, 主要研究领域为深度学习, 图像隐写, 生物信息.