基于替代模型的黑盒迁移攻击方法

曾繁茂,方贤进
安徽理工大学 计算机科学与工程学院,安徽 淮南 232001

摘 要:目的 针对现有基于生成对抗网络的无数据黑盒攻击方法容易出现收敛缓慢和代价高昂的问题,提出一种新颖的黑盒迁移攻击方法。**方法** 分为两个阶段:训练数据合成与替代模型蒸馏。在训练数据合成阶段,通过优化生成器以最大化替代模型与目标模型输出的一致性,同时引入 2 种损失函数来约束生成器产生的数据分布;在替代模型蒸馏阶段,采用具有可学习参数的残差块设计替代模型,并利用生成器合成的数据来拟合目标模型的决策边界。通过交替进行这两个阶段的训练,替代模型可以更好地拟合目标模型的决策边界,进而提升攻击效果。**结果** 通过系列实验验证,针对目标模型的无目标黑盒攻击成功率可以达到 70%以上;在 CIFAR100 数据集上,该方法相较于其他黑盒攻击方法,有目标攻击成功率提高了 2%以上,且在实现相同攻击效果时,所需查询预算更低。**结论** 所提方法能够高效拟合目标模型的决策边界,具有较好的攻击效果。

关键词:对抗样本;黑盒攻击;迁移攻击;替代模型蒸馏
中图分类号:TP18;TP39.41 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0003.009

Black-box Transfer Attack Method Based on Substitute Models

ZENG Fanmao, FANG Xianjin
School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective To solve the problems of slow convergence and high cost of the existing data-free black-box attack methods based on generative adversarial networks, a novel black-box transfer attack method is proposed. **Methods** The method consists of two stages: training data synthesis and substitute model distillation. In the stage of training data synthesis, the generator is optimized to maximize the consistency between the outputs of the substitute model and the target model, and two loss functions are introduced to constrain the data distribution generated by the generator. In the stage of substitute model distillation, a substitute model is designed with residual blocks containing learnable parameters, and data synthesized by the generator is used to fit the decision boundary of the target model. By alternating between these two stages of training, the substitute model can better fit the decision boundary of the target model, thereby enhancing the attack effectiveness. **Results** Through a series of experiments, the success rate of non-targeted black-box attacks against the target model exceeded 70%. On the CIFAR100 dataset, compared with other black-box attack methods, the success rate of targeted attacks increased by more than 2%, and the required query budget was lower for achieving the same attack effect. **Conclusion** The proposed method efficiently fits the decision boundary of the target model and demonstrates good attack effectiveness.

Keywords: adversarial examples; black-box attack; transfer attack; substitute model distillation

收稿日期:2023-10-31 **修回日期:**2023-12-23 **文章编号:**1672-058X(2025)03-0070-07
基金项目:国家自然科学基金项目(52374155).
作者简介:曾繁茂(2000—),男,江西赣州人,硕士研究生,从事人工智能安全研究.
通信作者:方贤进(1970—),男,安徽舒城人,教授,博士,从事网络与信息安全研究. Email: fx12341031@163.com.
引用格式:曾繁茂,方贤进. 基于替代模型的黑盒迁移攻击方法[J]. 重庆工商大学学报(自然科学版),2025,42(3):70-76.
ZENG Fanmao, FANG Xianjin. Black-box transfer attack method based on substitute models[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(3): 70-76.

1 引　言

随着深度学习的兴起,特别是深度神经网络的成功应用,机器学习在各种任务上取得了突破性的性能。深度学习模型已被广泛应用于各种现实任务[1-3],人们也更加关注模型的安全性[4]和鲁棒性[5-6]。对抗攻击最初是由研究人员通过添加微小的、人眼难以察觉的扰动[4]来欺骗图像分类模型引起的。这种攻击展示了几乎不可察觉的扰动也足以导致模型的误分类。这一发现引发了对于模型安全性的深刻思考,各种研究成果不断涌现。对抗攻击主要可分为白盒攻击和黑盒攻击,在白盒攻击中,攻击者完全了解目标网络的结构和参数;而在黑盒攻击中,攻击者仅对目标网络进行查询,获得相应的输出。实际应用中,如模型部署在云端的场景下,黑盒攻击的威胁性更为显著。在众多黑盒攻击策略中,迁移攻击具有强大的攻击能力,其基本策略是利用与目标模型结构相似的替代模型来生成对抗样本,进而误导目标模型。

Papernot 等[7]首先采用目标模型的数据集来训练替代模型,并成功使用针对替代模型生成的对抗样本对目标模型发起攻击。但该方法明显依赖目标模型的数据集,这在实际应用中并不总是可行。为克服此限制,后续的研究逐渐开始转向使用生成对抗网络[8](Generative Adversarial Network, GAN)来获取所需数据。如 Truong 等[9]利用生成器合成数据并进一步训练替代模型,模拟目标模型的行为,虽然这在一定程度上降低了对目标模型数据集的依赖,但训练稳定性不足,模型收敛困难;Zhou 等[10]引入多分支生成器以改善样本分布,但训练效率缓慢;Wang 等[11]提出动态代理模型搜索策略,致力于更深入地学习目标模型特性,尽管如此,高效的数据合成问题依旧未得到充分解决。

上述的黑盒攻击研究,即针对那些无法获取目标模型训练数据和详细知识的情况,已取得了较大进展,但仍有优化空间。部分方法虽然能够合成多样化的数据,但资源消耗巨大,且替代模型与目标模型的拟合程度不足,从而限制了攻击效果;部分方法通过优化替代模型的结构试图提高拟合效果,但数据合成质量不佳会制约替代模型对目标模型的拟合精度;部分方法在确保数据合成质量的同时也提升了替代模型的拟合效果,但收敛较慢,易引发模型崩溃,并带来较高的查询成本。因此,为应对先前研究中出现的收敛缓慢、模型崩溃和代价高昂的问题,提出了一种新颖的黑盒迁移攻击方法。该方法包括两个关键阶段:阶段 1 的目的是在最大化替代模型与目标模型输出一致性时,确保生成器能产生多样化的数据;阶段 2 则是让替代模型利用合成数据逼近目标模型的决策边界。通过交替执行这两个阶段的训练,逐步构建出满足攻击需求的替代模型。基于该替代模型生成的对抗样本可顺利应用于目标模型。通过在多种目标网络上使用主流数据集进行黑盒攻击实验,证明了所提方法的优越性。该方法对于构建更安全的深度学习模型具有促进作用,有助于提高深度神经网络在实际应用中的安全性和可靠性。

2 相关工作

2.1 对抗攻击

现有对抗攻击方法可以分为两大类:黑盒攻击与白盒攻击。在白盒攻击中,主要有基于梯度和基于优化的攻击策略。其中,基于梯度的攻击策略更加丰富,如 FGSM[12](Fast Gradient Sign Method)攻击和 PGD[13](Projected Gradient Descent)攻击。

FGSM[12]攻击的基本思想是使用梯度上升来最大化损失函数,其生成过程如式(1)所示。式(1)中,$J(\theta,x,y)$是目标模型的损失函数,$\nabla_x J(\cdot)$是损失函数关于输入数据的梯度,ε 是控制扰动大小的超参数。虽然 FGSM 攻击的开销较小,但由于只有一次梯度更新,容易导致局部最优值。

$$x_{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta,x,y)) \tag{1}$$

PGD[13]攻击的基本思想是迭代地在输入数据上进行小步的梯度上升,其生成过程如式(2)所示。式(2)中,Π_{x+S} 表示对输入进行裁剪,确保扰动不超过 S。PGD 攻击能比 FGSM 攻击产生更优的对抗样本,但迭代过程降低了对抗扰动的迁移性。

$$x_{t+1} = \prod_{x+S}(x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta,x,y))) \tag{2}$$

尽管当前的白盒攻击方法已经取得了很好的研究成果,但其局限性在于需要详细了解目标模型的所有细节。在实际场景中,获取这些信息可能是一项困难甚至不可能的任务。然而,白盒攻击方法可以较大程度地利用本文的替代模型信息。所以本文默认使用 FGSM 攻击和 PGD 攻击,并基于替代模型生成对抗样本。

黑盒攻击方法主要分为基于查询和基于迁移的攻

击方法。基于查询的攻击依赖于查询反馈来估计梯度。尽管它因较高的攻击成功率被认为是一种有效的方法,但随着模型复杂度的提高,查询成本及梯度估计误差也相应增加。基于迁移的攻击通过训练替代模型,将黑盒场景转化为白盒问题,但其攻击效果取决于替代模型与目标模型的匹配程度。现有基于迁移的攻击研究[7,9-11]虽然取得了较好的进展,但存在合成数据质量低、替代模型拟合效果差、收敛较慢和高查询成本问题。为解决这些问题,本文在前人的基础上,设计了一个新颖的黑盒迁移攻击方法。

2.2 生成对抗网络

生成对抗网络[8](Generative Adversarial Networks, GAN)是由 Goodfellow 首次提出的一种深度学习框架,众多学者已提出了一系列的 GAN 模型。针对原始 GAN 模型生成过程过于自由的问题,Mirza 等[14]提出了 CGAN(Conditional Generative Adversarial Networks),使得图像生成能朝规定的方向进行;Radford 等[15]提出了 DCGAN(Deep Convolution Generative Adversarial Networks),首次将卷积神经网络和 GAN 结合,极大地促进了 GAN 的后续发展。鉴于 DCGAN 在图像生成任务上取得了良好的表现,本文将默认使用 DCGAN 设计生成器。

2.3 知识蒸馏

知识蒸馏[16]旨在通过教师-学生架构来对复杂的教师模型进行压缩。早期的相关研究假设有访问训练集和白盒教师模型的权限。然而,在实际应用中,训练数据往往因含有隐私信息而不被公开。针对这一挑战,Wang 等[17]在假设无法获取训练集的情况下,提出了零样本知识蒸馏,通过利用教师模型的中间层预测结果和最终预测结果来优化合成的输入样本;Chen 等[18]首次引入生成对抗网络来合成学生模型的训练样本。本文借鉴知识蒸馏思想,设计了一个替代模型,其目标是仅通过利用黑盒模型的输出来拟合该模型的决策边界。

3 黑盒迁移攻击方法设计

3.1 整体框架

在无数据黑盒攻击场景中,攻击者无权访问目标模型的原始训练数据。近期的研究逐渐转向利用 GAN 生成所需数据。虽然该策略已经引起了广泛的关注,但现有方法仍然存在不足与缺陷。例如,当依赖 GAN 生成数据时,模型可能会坍塌,导致训练过程不稳定;为了优化生成器,需要进行大量查询,这无疑增加了成本;此外,基于先验知识手动选择替代模型的结构可能会限制其对不同目标模型和任务的攻击效果。针对这些问题,通过重新思考生成器与替代模型之间的协作关系,提出一种新颖的黑盒迁移攻击方法。该方法分为两部分:训练数据合成和替代模型蒸馏,如图 1 所示。阶段 1 的目的是确保生成器在最大化替代模型与目标模型输出一致性的同时,合成尽可能多样化的数据。随后,在阶段 2 中,替代模型会尝试利用生成的数据来逼近目标模型的决策边界。通过交替进行阶段 1 和阶段 2 的训练,直到得到一个满足要求的替代模型。基于对抗样本的迁移性,针对替代模型设计的对抗样本能够迁移到目标模型。

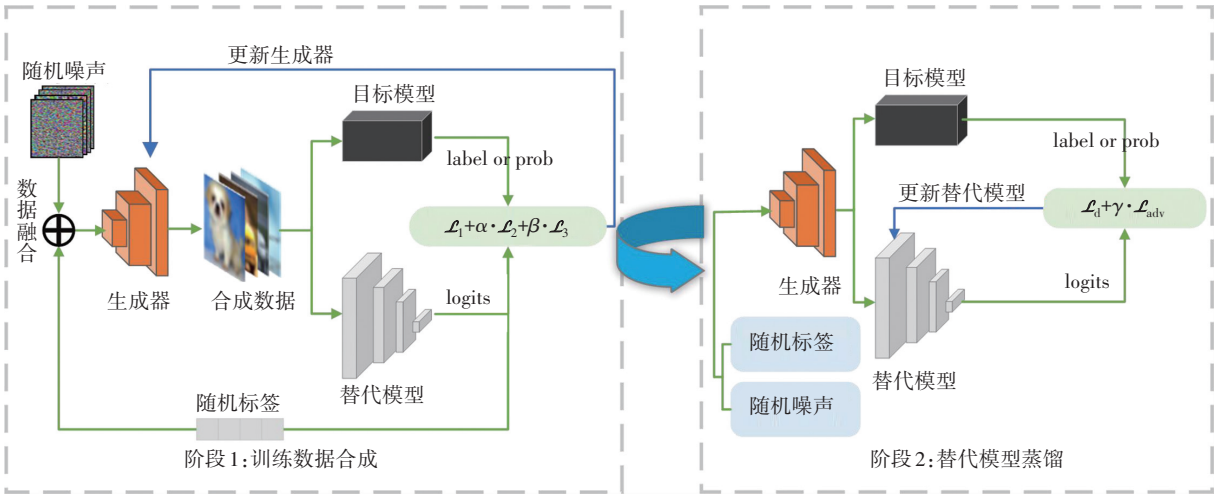

图 1 整体框架图

Fig. 1 Overall framework diagram

3.2　训练数据合成

当前,基于生成对抗网络的迁移攻击方法在训练过程中收敛缓慢,合成数据的多样性不足,这种情况会直接导致替代模型的精确度受损,进而影响到攻击的成功率。为了解决该问题,本文旨在寻找一种平衡策略,既确保生成数据的分布稳定性,又能够促进数据的多样性。

在缺乏目标模型原始训练数据的情况下,采用生成器为替代模型产生数据。与前述研究有所区别,本文不追求合成数据 X 的分布与真实数据分布完全一致。而是希望合成数据 X 的分布落在目标模型和替代模型决策边界的间隔区域,如图 2 所示。这样的数据分布将更有助于使替代模型的决策边界与目标模型的决策边界紧密对齐。

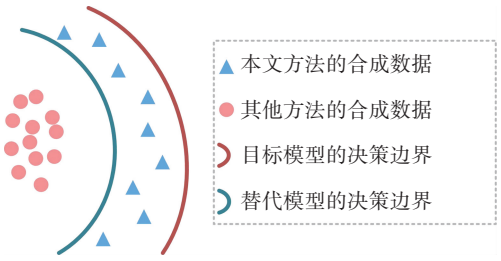

图 2　合成数据对比

Fig. 2　Comparison of synthetic data

如图 1 中的阶段 1 所示,生成器 G 从输入空间中进行随机采样,得到噪声向量 $\boldsymbol{z}$,同时接受人为定义的标签向量 $\boldsymbol{y}$。通过生成器 G,噪声向量 $\boldsymbol{z}$ 与标签向量 $\boldsymbol{y}$ 被映射为预期的合成数据 $X=G(\boldsymbol{z};\boldsymbol{y})$。进一步地,利用所生成的数据 X 探测目标模型 T 和替代模型 S,得到的输出分别为 $T(X)$ 和 $S(X)$。在此架构中,生成器 G 的目标是生成数据样本以揭示目标模型 T 和替代模型 S 之间的差异。针对生成器 G,其生成的数据 X 应能够反映出模型 T 与 S 之间的潜在差异。为实现此目标,对生成器 G 的优化策略如式(3)所示。式(3)中,$R_{\mathrm{CE}}(\cdot)$为交叉熵损失函数。$S(X)$为替代模型对于输入 X 的输出,$T(X)$表示目标模型对同一输入的输出。

$$L_1=-R_{\mathrm{CE}}(S(X),T(X)) \tag{3}$$

为确保生成数据 X 能够涵盖所有类别,引入正则化项,如式(4)所示:

$$L_2=R_{\mathrm{CE}}(S(X),\boldsymbol{y}) \tag{4}$$

为进一步增强生成器合成数据的多样性,并确保生成器能够在各个类别中合成多样化的数据,采用信息熵作为约束条件。假设存在 k 个类别,当离散随机变量 X 的概率分布为 $p(X)=\{p(X^1),p(X^2),\cdots,p(X^k)\}$时,信息熵损失 $H(X)=-1/k\sum_{i=1}^{k}p(X^i)\log p(X^i)$。据此,定义的损失函数如式(5)所示:

$$L_3=\exp(-H(\sum_{i=1}^{k}S(G(\boldsymbol{z};y_i)))) \tag{5}$$

当损失函数 L_3 最小化时,生成器 G 所生成的数据能确保均匀地分布于各个类别。这有助于使替代模型的决策边界与目标模型的决策边界更为接近。

综上所述,通过最小化式(6)的损失函数来更新生成器 G,其中,α 和 β 是超参数,本文默认都设置为 1。

$$L_{\mathrm{G}}=L_1+\alpha\cdot L_2+\beta\cdot L_3 \tag{6}$$

3.3　替代模型蒸馏

Liu 等[19]对对抗样本的迁移性进行了深入的研究。所谓对抗样本的迁移性,指的是在一个模型上产生的对抗样本在另一个模型上仍能实现欺骗的现象。为了保证对抗样本的迁移性,本文将致力于使替代模型的决策边界与目标模型的决策边界高度一致。考虑到适应多种目标模型和任务的需求,构建了一个基于 ResNet 网络架构[20]的替代模型,并融入可学习参数。具有可学习参数的残差块结构如图 3 所示,其中 μ 和 ν 是可学习参数。

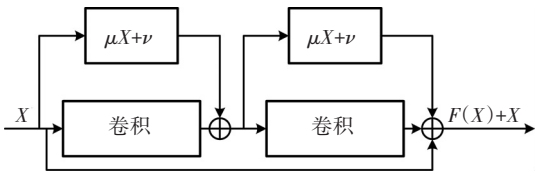

图 3　具有可学习参数的残差块

Fig. 3　The residual block with learnable parameters

基于知识蒸馏的理念,一旦生成合成数据 X,模型 S 可以模拟模型 T 的输出行为,具体如式(7)所示。式(7)中 $R_{\mathrm{MSE}}(\cdot)$ 和 $R_{\mathrm{CE}}(\cdot)$ 分别表示平均平方损失和交叉熵损失,$S(\cdot)$ 和 $T(\cdot)$ 分别表示替代模型和目标模型的输出,λ_1 和 λ_2 是超参数。对于目标模型只输出预测标签的场景,设定 $\lambda_1=0$。通过损失最小化策略,使得替代模型的表现与目标模型趋于一致。

$$L_{\mathrm{d}}=\lambda_1 R_{\mathrm{MSE}}(S(x),T(x))+\lambda_2 R_{\mathrm{CE}}(S(x),T(x)) \tag{7}$$

通过对某些原始数据施加微小扰动即可误导分类器,暗示这些原始数据位于分类器决策边界的近邻。在以往的工作中,替代模型的决策边界与目标模型的一致性往往不尽人意,导致对抗样本的迁移攻击成功率并不理想。若能针对决策边界附近的数据进行更为细致的考察,有望优化替代模型的决策边界。于是,定义的损失函数如式(8)所示。式(8)中,$R_{\mathrm{CE}}(\cdot)$为交叉

熵损失;$S(X)$和$T(X)$表示替代模型和目标模型对同一输入的输出;$I(\cdot)$是指示函数,表示当条件成立时为1,否则为0;X_{adv}是针对替代模型S使用FGSM攻击生成的对抗样本。

$$L_{adv}=R_{CE}(S(X),T(X))\times I(\text{argmax}\ S(X_{adv})\neq\text{argmax}\ S(X)) \tag{8}$$

因此,如图1中的阶段2所示,通过对损失函数最小化来更新和优化替代模型,具体损失函数如式(9)所示。其中,γ是一个超参数,本文默认设置为0.5。

$$L_S=L_d+\gamma\cdot L_{adv} \tag{9}$$

4 仿真实验与结果分析

为了评估方法的有效性,在多种目标网络上使用主流数据集进行黑盒攻击实验。值得强调的是,实验设置完全遵循黑盒攻击的场景,即目标模型仅作为一个不透明的黑盒存在,且在替代模型的训练过程中,无法获取原始的训练数据。为了深入评估方法的效果,将其与当前的主流黑盒攻击方法[10-11]进行详细对比,实验主要采用攻击成功率、查询预算等核心指标来衡量方法的性能。

4.1 实验设置

本文使用Pytorch1.10.1进行实现,网络的优化器选择Adam优化器,初始学习率设置为0.001,并从第80个epoch逐渐降至零,训练在第150个epoch停止,batch size为500,使用的显卡型号为NVIDIA GeForce RTX 3090 GPU。

为全面评估所提方法的效果,在多个目标网络上进行黑盒攻击实验。这些目标网络包括VGG16、VGG19、ResNet18及ResNet34。这些网络基于以下数据集进行训练:CIFAR10、CIFAR100以及STL10。在实验中,本文使用具有可学习参数的ResNet34网络作为替代模型,DaST方法[10]由于没有对替代模型进行改进,所以只使用普通的ResNet34网络,Dst方法[11]的替代模型则按照原文进行设计。对于生成器,本文基于DCGAN[15]的生成器进行设计,DaST[10]和Dst[11]方法的生成器则按照原文设计。对于攻击策略,本文针对替代模型采用PGD攻击策略生成可迁移的对抗样本,并将其用于目标网络的攻击。

4.2 评价指标

在无目标攻击场景中,只在被攻击模型正确分类的图像上生成对抗样本。在有目标攻击场景中,只在图像没有被分类到特定错误标签上生成对抗样本。两种场景中的实验都使用攻击成功率(Attack Success Rate, ASR)作为评价指标,设攻击成功率为f_{ASR},则攻击成功率的具体计算方法如式(10)所示,其中M和N分别是能够欺骗目标模型的对抗样本的数量和对抗样本的总数量。

$$f_{ASR}=\frac{M}{N}\times 100\% \tag{10}$$

为了进一步评估所提方法在现实任务中的性能,还在限制查询预算的情况下考察方法对目标模型的攻击成功率。

4.3 实验结果

表1展示了多种目标模型和不同数据集上的无目标攻击实验结果。实验结果表明:相较于DaST[10]和Dst[11]方法,本文在多个数据集上均获得了较高的攻击成功率。尤其在类别更为丰富的数据集如CIFAR100上,实现了明显的攻击成功率提升。这一显著提升主要得益于两点:一是合成数据落在分类边界间隔处,可以更好地帮助替代模型拟合目标模型;二是引入了带有可学习参数残差块的替代模型,为模型提供了更大的调整空间,从而能够更有效地适应并模仿不同目标模型的决策边界。需要指出的是,在小型数据集如CIFAR10上,本文在目标模型为VGG16上的攻击成功率与对比方法表现相近,原因是替代模型由于具备可学习的参数,所以具备强大的拟合能力,对较小数据集可能会产生过拟合现象。但是在现实场景中,目标模型通常基于大规模的数据集进行拟合。所以综合整体来看,本文方法仍然优于对比方法。

表1 不同黑盒攻击方法的无目标攻击成功率

Table 1 Success rates of non-target attacks using different black-box attack methods

/%

数据集	目标模型	DaST	Dst	本文方法
CIFAR10	VGG16	51.49	54.22	54.01
	ResNet18	49.45	56.43	55.95
CIFAR100	VGG19	31.16	35.21	39.47
STL10	ResNet34	54.53	66.12	70.14
	ResNet50	60.57	69.66	71.23

如表2所示,针对多种目标模型和不同数据集的有目标攻击实验表明,尽管有目标攻击本身具有一定的复杂性,但本文的攻击成功率仍优于对比方法。特别是在复杂度较高的CIFAR100数据集上,本文的攻击成功率比对比方法高出2%以上。结合表1和表2的

结果,可清晰地看出:在多数场景中,本文方法均展现出了显著的攻击成功率,验证了本文方法的优势。

表 2　不同黑盒攻击方法的有目标攻击成功率

Table 2　Success rates of targeted attacks using different black-box attack methods

/%

数据集	目标模型	DaST	Dst	本文方法
CIFAR10	VGG16	30.33	33.69	32.86
	ResNet18	40.59	43.25	44.53
CIFAR100	VGG19	12.67	17.33	19.41
STL10	ResNet34	46.35	55.91	59.85
	ResNet50	51.24	65.35	66.54

表 3 展示了不同查询预算下的攻击成功率。相较于其他方案,本文方法在相同的查询预算条件下获得了更高的攻击成功率。这主要是因为采用 GAN 结构生成数据的 Dst[10] 和 DaST[11] 方法在训练一个合格生成器时就需要大量查询,导致在受限的查询预算下其攻击成功率有所下降。而本文设计的生成器不追求完全收敛,只需生成当前替代模型和目标模型决策边界间隔处的数据,从而节省了查询预算。

表 3　不同查询预算下的攻击成功率

Table 3　Attack success rates under different query budgets

查询预算/k	DaST/%	Dst/%	本文方法/%
50	9.35	11.44	17.21
100	11.35	16.22	20.01
150	13.25	18.21	23.29
200	16.33	20.34	26.29
250	19.46	23.13	28.91
300	23.42	26.01	31.12

经过在多种目标网络上使用主流数据集进行的黑盒攻击实验验证,本文具有一定的优越性。即使在查询预算受限的条件下,本文方法仍能构建出与目标模型决策边界更为贴近的替代模型,从而为针对目标模型的攻击提供了更高的效率与成功率。

4.4　消融实验

为了深入探究本文引入的各种损失函数的影响,采用消融实验进行分析。实验的核心在于分析生成器训练过程中引入的 L_2 和 L_3 损失,以及训练替代模型时引入的 L_{adv} 损失。

在 CIFAR10 和 STL10 数据集上进行的消融实验如表 4 所示。实验结果表明:移除任何一个损失(L_2、L_3 或 L_{adv}),都会导致性能的下降,而同时移除 L_2 和 L_3 损失,则会造成更为严重的性能下降。这种现象的出现,是因为 L_2 和 L_3 损失对生成器输出数据的分布具有关键的调控作用。没有这些损失,生成器可能会产出分布极端偏斜的数据,从而导致标签不均衡,并进一步影响替代模型对目标模型的拟合效果。此外,当 L_2、L_3 和 L_{adv} 损失全部移除时,攻击成功率降至最低。这是由于缺乏 L_{adv} 损失会限制替代模型的学习能力,同时由于 L_2 和 L_3 损失的缺失,生成器产生的数据质量也受到了影响,这两个因素的共同影响显著削弱了替代模型对目标模型的拟合能力,进而导致了攻击成功率的显著下降。相反,当 L_2、L_3 和 L_{adv} 损失同时使用时,具有最高的攻击成功率,这进一步证明了这 3 种损失的兼容性和有效性。它们的结合能有效弥补单独使用时的不足,从而在本文方法中发挥了重要作用。

表 4　不同损失函数对攻击成功率的贡献

Table 4　Contribution of different loss functions to the attack success rates

/%

损失函数	CIFAR10	STL10
同时使用 L_2、L_3 和 L_{adv}	61.02	65.29
移除 L_2	55.36	58.01
移除 L_3	53.21	55.96
移除 L_{adv}	52.81	54.29
移除 L_2 和 L_3	45.87	49.68
移除 L_2 和 L_{adv}	45.04	48.49
移除 L_3 和 L_{adv}	44.28	47.97
同时移除 L_2、L_3 和 L_{adv}	40.31	43.17

5　结论与展望

通过深入分析生成器与替代模型的协同机制,本文提出了一个黑盒迁移攻击方法。该方法分为两个阶段:阶段 1 目的是确保生成器在最大化替代模型与目标模型输出一致性的同时,合成尽可能多样化的数据;阶段 2 中的替代模型利用合成数据逼近目标模型的决策边界。通过交替训练两个阶段,最终得到满足攻击要求的替代模型。基于此替代模型构建的对抗样本能迁移到目标模型。实验结果显示:本文不仅提升了攻击成功率,还有效节省了查询预算。

然而,本文仍然存在一些不足,如尽管替代模型采

用了含有可学习参数的残差块,这在一定程度上解决了手动选择替代模型结构的问题,但选择适当数量的残差块堆叠仍需依赖于特定的先验知识。在未来的工作中,将着重减少对于先验知识的依赖,同时在进一步减少查询预算的前提下,追求更高的性能。

参考文献(References):

[1] DONG Y, LIU Q, DU B, et al. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification[J]. IEEE Transactions on Image Processing, 2022, 31: 1559-1572.

[2] GALLO I, REHMAN A U, DEHKORDI R H, et al. Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images[J]. Remote Sensing, 2023, 15(2): 539-549.

[3] 夏振宇, 刘进, 亢艳芹, 等. 动态可控残差卷积神经网络的低剂量CT图像处理[J]. 重庆工商大学学报(自然科学版), 2023, 40(2): 64-72.
XIA Zhen-yu, LIU Jin, KANG Yan-qin, et al. Low dose CT image processing based on dynamic controllable residual convolution neural network[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(2): 64-72.

[4] DING Y, TAN F, GENG J, et al. Interpreting universal adversarial example attacks on image classification models[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(4): 3392-3407.

[5] MA L, LIANG L. Improving adversarial robustness of deep neural networks via adaptive margin evolution[J]. Neurocomputing, 2023, 551: 126524-126534.

[6] BIGOLIN LANFREDI R, SCHROEDER J D, TASDIZEN T. Quantifying the preferential direction of the model gradient in adversarial training with projected gradient descent[J]. Pattern Recognition, 2023, 139: 109430-109440.

[7] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. NewYork: ACM, 2017: 506-519.

[8] TREVISAIV DE SOUZA V L, MARQUES B A D, BATAGELO H C, et al. A review on generative adversarial networks for image generation[J]. Computers & Graphics, 2023, 114: 13-25.

[9] TRUONG J B, MAINI P, WALLS R J, et al. Data-free model extraction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 4771-4780.

[10] ZHOU M, WU J, LIU Y, et al. DaST: Data-free substitute training for adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 234-243.

[11] WANG W, QIAN X, FU Y, et al. DST: Dynamic substitute training for data-free black-box attack[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 14361-14370.

[12] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. IEEE, 2015.

[13] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//International Conference on Learning Representations. IEEE, 2018.

[14] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. Computer Science, 2014(10): 2672-2680.

[15] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//International Conference on Learning Representations. IEEE, 2016.

[16] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[17] WANG Z. Data-free knowledge distillation with soft targeted transfer set synthesis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(11): 10245-10253.

[18] CHEN H, WANG Y, XU C, et al. Data-free learning of student networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 3514-3522.

[19] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]//International Conference on Learning Representations. IEEE, 2016.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.

责任编辑:李翠薇