文章编号:1671-6833(2025)06-0040-09

基于基因重组知识蒸馏策略的对抗攻击方法

刘明林, 周传金, 王润泽, 王 超, 曹仰杰

(郑州大学 网络空间安全学院,河南 郑州 450002)

摘 要: 针对传统集成攻击方法存在因计算资源(包括训练数据和训练时间)需求高而在应用中受限制的问题,提出了一种基于基因重组的低计算复杂度集成攻击方法,通过生成更多样的集成模型来增强现有对抗攻击的迁移性。首先,将基因重组思想引入知识蒸馏领域,在此过程中,学生模型被视为独立个体,其参数被看作该个体的基因,每一轮的蒸馏学习视为基因的一次进化;其次,通过在进化过程中随机交换学生模型的参数,实现了人为的基因重组,从而获得更优的后代基因,通过设置不同的蒸馏温度,能够获得多个多样化的学生模型;再次,将这些多样化的学生模型与源教师模型进行集成;最后,使用集成模型生成迁移性更强的对抗样本。在 ImageNet 验证集子集上的实验结果表明:相较于其他算法,所提方法显著提高了对抗样本的迁移性。以 ResNet152 作为源模型并采用 PGD 攻击为例,所提方法在 11 种黑盒模型上的迁移攻击成功率表现最优,比基线 PGD 方法平均提高了 34.52 百分点,比 PGI 方法平均提高了 5.30 百分点,比 DGM 方法平均提高了 2.12 百分点。

关键词:集成攻击; 对抗样本; 迁移性; 基因重组; 知识蒸馏

中图分类号: TP181;TP183;TP309 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.03.007

深度卷积神经网络[1]在图像分类、目标检测、语义分割及自动驾驶等领域展现了卓越性能和巨大潜力。然而,近年来的研究表明卷积神经网络存在一定的脆弱性,即模型输出易受输入扰动的影响。通过向原始样本添加精心设计的微小扰动[2-3],能够有效诱使深度学习模型误判,这些经过设计的异常样本被称为对抗样本。

对抗样本最初起源于图像分类领域,但随着研究的不断深入,应用范畴已扩展至面部识别、视觉跟踪等计算机视觉任务。这些对抗样本对模型稳健性造成了严重影响,从而引起人们的重视[4]。

对抗样本在机器学习模型中普遍存在,常见的对抗攻击方法可分为白盒攻击和黑盒攻击[5]两大类。早期研究集中于白盒攻击[6-7],此场景下攻击者完全了解目标模型的内部结构,能够利用其弱点进行攻击。然而,在实际应用中,攻击者通常无法获取目标模型内部信息,因此,研究重点逐渐转向黑盒攻击。在黑盒攻击中,攻击者对目标模型内部结构一无所知,仅能查询输入输出信息,这使其在实际场景中的危害更为显著。基于迁移的攻击是典型的黑盒攻击方法,它是在白盒场景下攻击源模型生成对抗样本,并使用生成的对抗样本直接攻击目标黑盒模型。

早期的攻击方法如基本迭代方法(basic iterative method,BIM)[6]与投影梯度下降(projection grodient descent,PGD)方法[7]在黑盒模型攻击中表现出有限的可迁移性。为提高可迁移性,研究者利用模型的梯度信息,设计了动量积分[8]、输入多样化[9]和方差调整[10]等多种方法来提高黑盒攻击的可迁移性。

随着对抗攻击技术的发展,研究者发现集成方法在提升对抗样本迁移性方面优于传统的非集成方法。集成方法面临两大主要挑战:一是获取不同架构的模型并集成大量派生模型需要巨大的计算资源;二是可用于集成的模型多样性和数量受限。DGM(diverse gradient method)[11]是一种通过自蒸馏

收稿日期:2025-01-10;**修订日期**:2025-03-10

基金项目:国家自然科学基金资助项目(62302458);河南省自然科学基金资助项目(222300420295)

作者简介:刘明林(1991—),男,河南郑州人,郑州大学讲师,博士,主要从事图像隐写与隐写分析、数字媒体取证、AI 安全的研究,E-mail:liuminglin@zzu.edu.cn。

通信作者:曹仰杰(1976—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事计算机视觉与智能计算、人工智能、高性能计算的研究,E-mail:caoyj@zzu.edu.cn。

引用本文:刘明林,周传金,王润泽,等. 基于基因重组知识蒸馏策略的对抗攻击方法[J]. 郑州大学学报(工学版),2025,46(6):40-48. (LIU M L, ZHOU C J, WANG R Z, et al. Adversarial attack method based on genetic recombination knowledge distillation strategy[J]. Journal of Zhengzhou University (Engineering Science),2025,46(6):40-48.)

来丰富模型梯度的方法,能够在不改变模型结构的前提下,以低计算复杂度与其他方法结合,实现性能提升。它通过自蒸馏生成一系列与源模型结构相同的学生模型,使其不仅学习源模型的核心知识,还在不同程度上发展出各自的特征表现。最终,通过集成教师模型和这些学生模型进行攻击,生成迁移性更强的对抗样本。然而,单靠自蒸馏生成的学生模型多样性仍然受限。

为进一步获得更加丰富多样的学生模型群,本文提出了一种基于基因重组的自蒸馏策略,整体框架如图 1 所示。该策略在自蒸馏生成学生模型的同时对两个学生网络进行训练,将单个学生模型的若干结构分层视为控制相同性状的若干基因。在蒸馏过程中,通过不断对两个学生模型的等位基因进行重组,产生两个新基因组合的学生模型。历经多轮训练与重组迭代,得到高度多样化的学生模型群。最后,集成源模型与这些丰富多样的学生模型进行攻击,从而生成迁移性更优的对抗样本。

1　相关工作

本节概述了对抗样本[12]和对抗攻击的相关研究。首先,介绍了对抗样本的定义及其生成方法;其次,讨论了基于梯度的攻击方法,并探讨了在黑盒场景下通过对抗样本进行迁移攻击的策略;最后,介绍了几种增强对抗样本迁移性和攻击效果的算法。

1.1　对抗样本

给定一个干净样本 $\boldsymbol{x}$ 及其真实标签 $\boldsymbol{y}_{\text{true}}$、目标分类器 f,如果存在一个扰动 δ 使得分类器 f 输出分类错误,并且扰动满足 $\|\boldsymbol{x}-\boldsymbol{x}_{\text{adv}}\|_\infty < \delta$,那么这个 $\boldsymbol{x}_{\text{adv}}=\boldsymbol{x}+\delta$ 就是一个对抗样本。

1.2　对抗攻击

对抗攻击是一种通过在干净样本上添加精心设计的微小扰动,使机器学习模型产生误判的攻击方式。这类攻击利用模型对小扰动的敏感性,通过引入细微且看似无害的变化,干扰模型的正常预测和分类。

黑盒攻击可分为基于迁移的攻击和基于查询的攻击。本文主要探讨基于迁移的黑盒攻击方法,其核心在于使用替代模型代替目标黑盒模型完成攻击。首先,在替代模型上应用白盒攻击方法生成对抗样本;其次,利用这些对抗样本的良好迁移性对目标黑盒模型实施有效攻击,从而绕过访问目标模型内部信息的限制。

1.3　基于梯度的攻击

Goodfellow 等[2]提出的 FGSM(fast gradient sign method)方法通过在梯度生成方向上添加扰动,增加目标模型的损失函数,从而降低模型对真实标签的预测概率,实现对目标模型的攻击。

$$\boldsymbol{x}_{\text{adv}}=\boldsymbol{x}+\delta\cdot\text{sign}(\nabla_x J(\boldsymbol{x},\boldsymbol{y}_{\text{true}}))。\tag{1}$$

式中:$\nabla_x J$ 为损失函数相对于样本 $\boldsymbol{x}$ 的梯度。

BIM 方法是在 FGSM 的基础上引入步长的一种迭代攻击方法。PGD 是一种迭代攻击方法,通过在干净样本上多次迭代添加扰动,并在每次迭代后将生成的对抗样本投影回预设范围内,以引导目标模型产生错误分类。

$$\begin{cases}\boldsymbol{x}_{\text{adv}}^0=\boldsymbol{x};\\ \boldsymbol{x}_{\text{adv}}^{n+1}=\text{projection}_{x,\delta}\{\boldsymbol{x}_{\text{adv}}^n+\alpha\cdot\text{sign}(\nabla_x J(\boldsymbol{x}_{\text{adv}}^n,\boldsymbol{y}_{\text{true}}))\}。\end{cases}\tag{2}$$

式中:projection(·)为投影函数;n 为迭代次数。与 FGSM 相比,PGD 实现了更好的攻击效果,但在黑盒攻击中的迁移性较低。

1.4　基于迁移的攻击

在黑盒场景下,由于目标模型的内部结构和参数未知,基于梯度的攻击方法无法直接应用。Goodfellow 等[2]发现对抗样本具有迁移性,因此可

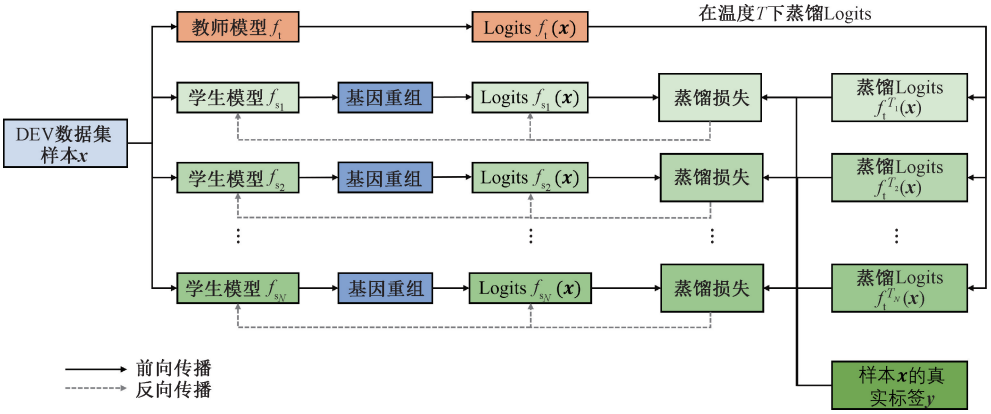

图 1　基于基因重组的自蒸馏方法框架

Figure 1　Self-distillation method framework based on genetic recombination

以利用白盒模型生成对抗样本对黑盒模型实施迁移攻击。这意味着攻击者可以通过攻击替代模型来生成对抗样本,无须直接接触目标模型,从而达到攻击效果。

为进一步提高对抗样本的迁移性,动量迭代(momentum itoration,MI)方法[9]通过在每次迭代中积累梯度方向上的动量,结合 Nestrov 动量,更准确地估计下一步的梯度方向。

$$g^{n+1}=\mu\cdot g^{n}+\frac{g_f,\boldsymbol{x}^{n}_{\mathrm{adv}}}{\|g_f,\boldsymbol{x}^{n}_{\mathrm{adv}}\|_1}。\tag{3}$$

式中:μ 为衰减因子。

NI 方法(nesterov iterative fast gradient sign method)[13]是 MI 方法的改进版。PGI 方法[14]通过在每次迭代中累积上一步在梯度方向上采样数据点的平均梯度,以稳定更新方向并避免陷入局部最大值。

在每次迭代中,DI 方法(diverse input method)[15]通过对输入图像进行随机变换,生成不同的输入模式,以增强对抗样本的可迁移性。

$$\begin{cases}\boldsymbol{x}^{n}_{\mathrm{da}}=H(\boldsymbol{x}^{n}_{\mathrm{adv}};p);\\ g^{n+1}=g_{f,x^{n}_{\mathrm{da}}}。\end{cases}\tag{4}$$

式中:$H(\boldsymbol{x}^{n}_{\mathrm{adv}};p)$ 表示在给定概率 p 下应用到 $\boldsymbol{x}^{n}_{\mathrm{adv}}$ 的数据增强方法。

TI 方法(translation-invariant method)[16]则使用预定义的卷积核 $\boldsymbol{W}$ 从一组图像中获取梯度,以生成针对防御模型的可迁移对抗样本。

$$g^{n+1}=\boldsymbol{W}\cdot\nabla_x J_f(\boldsymbol{x}^{n}_{\mathrm{adv}},\boldsymbol{y})。\tag{5}$$

VT 方法(variance tuning method)[10]利用前一迭代中的梯度方差来调整当前梯度,从而稳定更新方向并避免陷入局部最优。

$$g^{n+1}=\mu\cdot g^{n}+\frac{g_{f,x^{n}_{\mathrm{adv}}}+v^{n}}{\|g_{f,x^{n}_{\mathrm{adv}}}+v^{n}\|_1}。\tag{6}$$

式中:$v^{n}=V(\boldsymbol{x}^{n}_{\mathrm{adv}})$ 表示梯度差。

2 本文方法

本节首先介绍了遗传学中的基因重组及其对生物多样性的重要性。在此基础上,提出了一种基于基因重组的蒸馏训练策略,该策略能够生成更加多样化的模型,从而提升集成攻击的效果。

2.1 基因重组

生物多样性[17]包括遗传多样性、物种多样性和生态系统多样性,它在地球生态系统中扮演着至关重要的角色。自细菌起源以来,经过几十亿年的演化,尤其是在真核生物和有性生殖机制出现后,生物多样性显著丰富。有性生殖通过基因重组创造丰富的遗传变异,加速了生物进化,并推动了多细胞生物分化和繁衍,从而形成了多样的生物群体和复杂的生态系统。

基因重组作为生物进化的重要机制,体现在 3 个方面:第一,它通过重组不同亲本的基因产生新的遗传组合和变异,极大丰富了生物变异多样性,为进化提供了更多选择;第二,基因重组加快了生物对环境变化的适应速度,提高了生存和繁殖成功率;第三,它对多细胞生物多样性的形成至关重要,通过细胞和组织间的基因交流促进形态和功能的多样化。这些机制共同作用增加了地球上生物种类的多样性,提升了生态系统的复杂性和稳定性。

2.2 基于基因重组知识蒸馏的对抗攻击方法

Mendel[18]通过豌豆杂交实验验证并阐述了遗传学的基本理论——分离定律与自由组合定律。在第 1 次实验中,Mendel 将纯种高茎豌豆和纯种矮茎豌豆进行杂交,揭示了分离定律,即在减数分裂过程中,决定某一性状的等位基因独立分配至不同的配子并保持各自独立。随后,Mendel 以纯种黄色圆粒豌豆与纯种绿色皱粒豌豆进行杂交实验(如图 2 所示),进一步验证了自由组合定律。该定律指出,在具有两对或多对相对性状的亲本杂交时,不同染色体上的等位基因分离和组合是独立的。

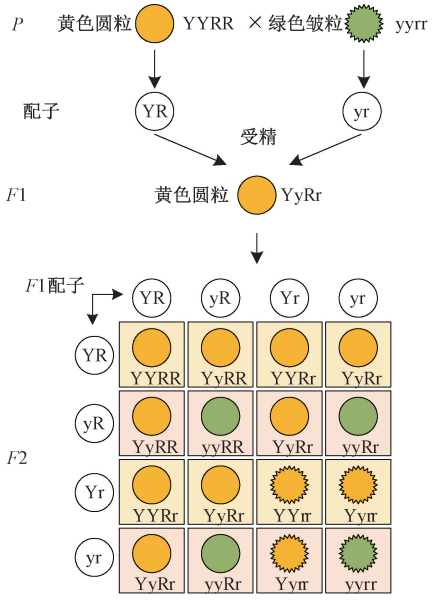

图 2 黄色圆粒豌豆与绿色皱粒豌豆杂交实验图解

Figure 2 Diagram of crossbreeding experiment between yellow round pea and green wrinkled pea

2.2.1 通过基因重组自蒸馏生成多样化学生模型

将自蒸馏过程建模为遗传进化模型,其中学生

模型被视为独立个体,其参数 θ 代表这个个体的基因。学生模型的基因初始化为 θ_0,然后通过学习教师的知识进化为第 k 代:

$$\theta_k \xleftarrow{\text{knowledge}} \theta_{k-1}, 0 < k \leqslant K。\tag{7}$$

经过 K 轮进化,学生模型不断获取教师知识,逐渐进化为最终状态。通常,一个神经网络模型由包含多层的块组成。给定一个包含 M 块的学生模型,其参数为 θ,可以将这 M 块视为拥有 M 个基因的学生个体,并将它们划分为 N 个等位基因:

$$\theta = [B^1, B^2, \cdots, B^M] = [\theta^1, \theta^2, \cdots, \theta^N], 1 < N \leqslant M。\tag{8}$$

第 i 个等位基因 θ^i 由以下基因组成:

$$\theta^i = \begin{cases} [B^{\text{Round}\frac{M(i-1)}{N}+1}, B^{\text{Round}\frac{M(i-1)}{N}+2}, \cdots, B^{\text{Round}\frac{Mi}{N}}], 0 < i < N; \\ [B^{\text{Round}\frac{M(i-1)}{N}+1}, B^{\text{Round}\frac{M(i-1)}{N}+2}, \cdots, B^{M}], i = N。 \end{cases}\tag{9}$$

式中:Round(·)表示舍入到最接近的整数。为确保最后的等位基因 θ^N 非空,进一步限制 N 的值:

$$\text{Round}\frac{M(N-1)}{N} + 1 \leqslant M \Rightarrow \text{Round}\frac{M}{N} \leqslant \frac{M-1}{N-1}。\tag{10}$$

给定两个初始化的学生模型的基因 θ_{init} 和 ϑ_{init},分别将它们的等位基因划分为 $\theta_{\text{init}} = [\theta^1_{\text{init}}, \theta^2_{\text{init}}, \cdots, \theta^N_{\text{init}}]$ 和 $\vartheta_{\text{init}} = [\vartheta^1_{\text{init}}, \vartheta^2_{\text{init}}, \cdots, \vartheta^N_{\text{init}}]$。图 3 为基因重组示例。在图 3 中,设 $N = 3$,可交换的等位基因组合为[1,2,3],每次重组后可能产生 3 种不同的情况。最终,通过 K 轮的蒸馏进化和多轮基因重组操作,可以获得更加多样化的学生模型。

2.2.2　多模型集成

每次使用基因重组策略进行自蒸馏时,会生成两个学生模型,始终将第 1 个初始化的学生模型定为 s_1,将获得 $N+1$ 个不同的模型,包括 N 个多样化的学生模型 $f_{s1}, f_{s2}, \cdots, f_{sN}$ 和一个教师模型 f_t。随后,将这 $N+1$ 个模型进行集成,以获取更为多样和鲁棒的梯度信息,从而应用基于梯度的对抗攻击方法(如 FGSM、BIM、PGD)生成对抗样本。为简化公式,教师模型 f_t 视为未经过蒸馏训练的学生模型 f_{s0},并将各种模型 $f_{s0}, f_{s1}, \cdots, f_{sN}$ 的 Logits 进行融合,使用以下的集成策略[19]来计算集成损失:

$$J_{\text{ens}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{e}_y \cdot \log\left(\sum_{k=0}^{N} \omega_k \cdot \text{Softmax}(f_{sk}(\boldsymbol{x}))\right)。\tag{11}$$

式中:$f_{sk}(\boldsymbol{x})$ 表示第 k 个模型的 Logits;$\omega_k > 0$ 为受 $\sum_{k=0}^{N} \omega_k = 1$ 约束的集合权值,本文实验设置 $\omega_k = \frac{1}{N+1}$;$\boldsymbol{e}_y$ 表示标准基向量,仅在真实标签 $\boldsymbol{y}$ 对应位置为 1,其余为 0。

此外,集成模型 $f_{\text{ens}} = \sum_{k=0}^{N} \omega_k \cdot \text{Softmax}(f_{sk}(\boldsymbol{x}))$,集成梯度 g^{n+1} 为

$$\begin{cases} \boldsymbol{x}^0_{\text{adv}} = \boldsymbol{x}; \\ g^{n+1} = g_{f_{\text{ens}}, x^n_{\text{adv}}} = \nabla_{x^n_{\text{adv}}} J_{\text{ens}}(\boldsymbol{x}^n_{\text{adv}}, \boldsymbol{y})。 \end{cases}\tag{12}$$

在获得集成梯度后,利用基于梯度的对抗攻击方法生成具有更优迁移性的对抗样本。

DGM 方法是一种通用的对抗迁移性增强策略,能够直接与基于梯度的攻击方法结合,亦可与其他迁移性增强策略相结合,以进一步提升攻击效果。由于基于基因重组的知识蒸馏策略与模型及攻击方法无关,因此所提对抗攻击方法具有广泛的适用性。

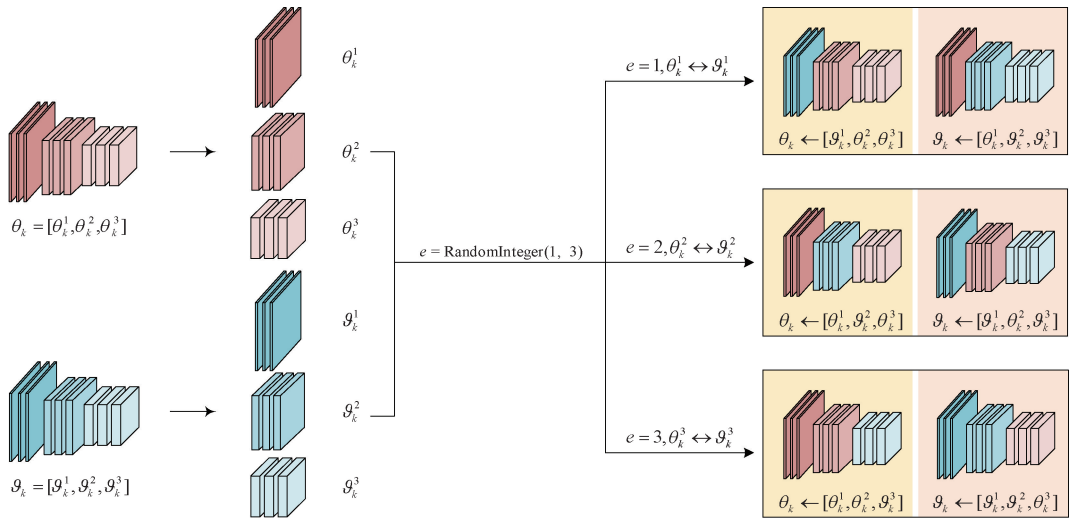

图 3　2 个学生模型进行基因重组的图解

Figure 3　Diagram of genetic recombination between two student models

3 实验结果与分析

3.1 数据集

从ImageNet验证集[20]中随机选择5 000张能够被正确分类的图像[21]进行训练。该子集涵盖了各个类别的图像,除极少数类别外,其余类别均包含5张合格图像。源模型能够对该子集的大部分图像进行正确分类。

3.2 使用模型

本文在不同的卷积神经网络模型上评估生成的对抗样本。评估对象包括9个未经防御训练的模型和4个防御模型,这些模型均为在ImageNet数据集上的预训练模型。未经防御训练的模型包括Res50、ResNet152(简写为Res152)、ResNet101(简写为Res101)[22-23]、DenseNet201(简写为DN201)[24]、SeNet154(简写为Se154)[25]、VGG19[26]、InceptionV3(简写为IncV3)[27]、Inception-V4(简写为IncV4)和inception-ResNet-V2(简写为IncRes)[28]。防御模型包括使用对抗训练的模型$IncV3_a$[7]及经过集成对抗训练的模型$IncV3_{e3}$、$IncV3_{e4}$和$IncRes_e$[29]。

3.3 对比方法

将本文方法与多种攻击方法进行对比,包括基于梯度的一步攻击方法FGSM,多步攻击方法BIM、PGD,以及基于梯度的增强方法MI、NI、PGI、DI、TI、SI等。

3.4 实现细节

本文的实验设置与DGM方法一致,采用PGD和BIM方法作为初始基线,且L_∞约束下最大约束$\varepsilon = 16/255$。在所有多步攻击方法中,迭代次数n统一设置为10,步长设置为2(在BIM中为-1)。基准方法严格遵循相关文献中的默认参数,以确保结果的可对比性和一致性。

为保证实验的准确性和可重复性,所有实验均在相同服务器环境下进行,具体配置为Intel(R) Xeon(R) CPU E5-2697 v4 @ 2.30 GHz和NVIDIA GeForce RTX 3090 GPU 24 GB。使用PyCharm作为开发平台,以Python为开发语言,在CentOS Linux 7(Core)系统上运行。代码运行的开发环境涉及多个依赖库,advertorch版本为0.2.3、pretrainedmodels版本为0.7.4、Python版本为3.8.17、torch版本为1.9.0+cu111。

在集成攻击方法中,获得不同的集成模型的成本较高。例如,获取一个初始化参数不同的预训练模型(如Res152或DN201)需在包含120万张训练图像的ImageNet数据集上训练90个轮次,耗费大量时间和计算资源。而本文方法仅需利用ImageNet数据子集(5 000张训练图像),通过知识蒸馏在24个训练轮次内即可获得一个对攻击有正向增益的模型,耗时仅需48 min,显著降低了集成攻击获得不同集成模型所需的计算资源和时间成本。

3.5 结果分析

3.5.1 一步可迁移性

对本文方法在9个无防御模型上的一步可迁移攻击性进行评估,结果如表1所示。当源模型与目标模型相同时,结果表示白盒攻击成功率。以Res152和DN201作为源模型时,本文方法在9个目标模型上均取得最高攻击成功率。以Res152作为源模型为例,所提方法在Res101、DN201和IncV4模型上的迁移攻击成功率分别为63.36%,52.72%和33.52%,均较其他方法提高了1百分点以上。实验结果表明,本文方法在DGM基础上进一步提升了FGSM对复杂白盒模型的攻击效果。此外,与迁移方法SGM相比,本文方法显著增强了FGSM在多种黑盒模型上的可迁移攻击性能。

表1 一步可迁移攻击性对比实验结果

Table 1 Comparison experiment results of one-step attack transferability

源模型	方法	攻击成功率/%								
		Res152	Res101	Res50	DN201	Se154	VGG19	IncV3	IncV4	IncRes
Res152	FGSM	72.96	47.82	47.30	37.20	26.46	41.48	27.36	21.64	20.10
	SGM	78.64	54.80	55.10	44.02	31.00	48.12	31.00	25.28	23.08
	DGM-5	82.04	62.24	61.42	51.58	36.26	52.50	37.80	32.34	29.60
	本文方法	**82.82**	**63.36**	**62.42**	**52.72**	**37.36**	**53.22**	**38.22**	**33.52**	**30.20**
DN201	FGSM	39.44	46.56	50.30	82.24	35.42	50.12	33.24	27.00	24.06
	SGM	45.74	53.32	57.22	86.36	40.38	56.32	38.02	30.24	27.44
	DGM-5	55.18	61.22	65.46	90.02	47.90	63.12	46.56	39.28	36.12
	本文方法	**56.12**	**62.36**	**66.38**	**90.68**	**48.58**	**64.32**	**47.78**	**40.54**	**36.89**

3.5.2　多步迁移性

对于多步攻击,本文在 Res152 和 DN201 源模型上,对 8 个无防御模型和 4 个防御模型进行攻击,结果如表 2 所示。

由表 2 可知,在白盒场景中,即源模型为 Res152 的第 2 列和源模型为 DN201 的第 4 列,所列方法在攻击成功率方面表现基本一致。在这 4 项实验中,本文方法与表现最佳的攻击方法相比,攻击成功率的差异最大不超过 0.04 百分点。具体而言,以 DN201 作为源模型并采用 BIM 攻击为例,本文方法的攻击成功率达到 99.96%,比 BIM 的 99.94%高 0.02 百分点,仅比最优的 SI 方法(100.00%)低 0.04 百分点。

在黑盒场景中,本文方法在大多数情况下都表现出更优越的迁移性。例如,以 Res152 作为源模型并采用 PGD 攻击为例,与原始的 PGD 攻击相比,本文方法在 11 个模型上的平均攻击成功率提高了 34.52 百分点。而在对比方法中,提升最大的是

表 2　以 Res152 和 DN201 为源模型的多步迁移性对比实验结果

Table 2　Comparison experiment results of multi-step transferability with Res152 and DN201 as the source model

源模型	方法	攻击成功率/%											
		Res152	Res101	DN201	Se154	VGG19	IncV3	IncV4	IncRes	$IncV3_a$	$IncV3_{e3}$	$IncV3_{e4}$	$IncRes_e$
Res152	PGD	99.98	83.88	52.76	29.99	46.91	23.91	21.00	19.22	12.52	10.30	9.14	5.75
	PGD+MI	99.98	91.41	75.75	54.17	66.20	50.32	43.74	42.62	32.06	28.14	25.74	19.50
	PGD+NI	**100.00**	95.34	77.91	56.26	71.74	51.98	44.86	44.54	33.32	28.54	25.48	19.00
	PGD+PGI	99.98	97.76	87.02	66.39	79.17	62.92	55.90	54.34	41.18	35.84	31.56	24.69
	PGD+DI	99.96	94.66	85.05	68.26	79.42	61.94	59.77	57.09	39.10	32.29	27.42	20.43
	PGD+TI	99.92	79.98	64.56	44.52	56.08	42.30	38.69	33.83	39.30	**37.80**	**39.08**	**33.16**
	PGD+SI	**100.00**	93.01	74.03	41.34	59.94	41.58	36.24	32.98	24.72	20.71	17.84	12.20
	PGD+DGM-5	99.96	99.74	95.24	71.68	91.40	65.06	60.64	57.96	43.56	34.52	30.40	21.54
	PGD+本文方法	99.96	**99.86**	**96.18**	**73.36**	**92.64**	**67.92**	**63.64**	**61.32**	**46.06**	36.94	33.14	23.98
	BIM	99.98	77.52	44.86	24.93	40.05	19.38	16.78	14.46	9.76	8.52	7.54	4.44
	BIM+MI	99.98	90.88	74.04	52.55	64.38	48.45	42.48	40.69	31.60	27.54	25.05	19.23
	BIM+NI	**100.00**	95.15	78.15	56.26	70.60	51.54	45.13	43.74	33.02	28.54	25.58	19.51
	BIM+PGI	99.98	97.24	85.64	65.29	76.72	60.96	54.55	52.75	40.50	**35.60**	31.31	24.84
	BIM+DI	99.92	93.17	83.87	63.80	75.97	60.15	**60.62**	**56.61**	36.85	27.92	23.80	18.11
	BIM+TI	99.90	77.16	61.45	41.51	52.25	39.06	35.52	31.24	35.97	34.31	**35.96**	**30.43**
	BIM+SI	**100.00**	89.58	66.29	33.62	52.53	33.86	29.56	26.89	19.90	16.55	14.37	9.09
	BIM+DGM-5	99.94	99.54	93.54	64.92	89.10	59.26	55.06	52.58	38.54	30.82	26.72	18.16
	BIM+本文方法	99.96	**99.68**	**94.64**	**67.56**	**90.52**	**61.34**	57.68	55.36	**41.30**	32.84	29.00	20.18
DN201	PGD	60.22	63.18	99.96	40.64	58.10	32.94	32.39	26.26	18.86	15.22	13.55	8.86
	PGD+MI	76.90	79.73	99.96	64.55	75.20	59.67	54.22	50.00	41.76	35.89	32.76	25.70
	PGD+NI	82.76	85.54	99.94	70.08	82.20	63.72	59.40	54.46	43.37	37.14	33.44	25.37
	PGD+PGI	88.43	89.80	99.94	77.39	87.01	73.08	69.28	64.31	52.65	46.22	41.14	32.49
	PGD+DI	80.07	81.68	99.89	66.92	79.98	63.82	63.30	56.25	43.28	36.80	31.95	23.74
	PGD+TI	58.18	62.63	99.88	48.00	59.28	46.36	42.60	37.10	41.92	40.34	40.84	**34.53**
	PGD+SI	76.83	78.29	**100.00**	52.76	71.12	52.11	47.46	41.60	34.80	30.68	27.48	19.41
	PGD+DGM-5	96.76	97.44	99.98	82.58	94.96	74.64	74.60	68.96	54.48	45.12	39.58	30.18
	PGD+本文方法	**97.22**	**97.90**	99.98	**84.06**	**95.88**	**76.42**	**76.18**	**70.90**	**55.40**	**46.88**	**41.32**	31.02
	BIM	52.78	55.90	99.94	33.75	50.86	27.64	26.20	21.46	15.20	12.31	10.78	7.10
	BIM+MI	75.05	78.39	99.94	63.21	73.76	57.92	53.22	48.12	40.60	35.89	32.94	25.57
	BIM+NI	82.28	84.61	99.94	68.84	80.70	62.74	58.42	53.43	43.17	37.62	33.52	25.82
	BIM+PGI	87.11	88.46	99.94	75.04	85.33	70.38	67.02	62.32	**51.19**	**45.85**	**40.76**	**32.34**
	BIM+DI	74.57	76.30	99.90	58.33	74.27	56.24	56.22	49.29	37.17	31.33	27.04	18.91
	BIM+TI	55.22	58.75	99.80	44.45	55.64	42.88	39.08	33.84	38.00	37.06	37.28	31.14
	BIM+SI	70.64	72.74	**100.00**	43.60	64.32	43.71	40.08	34.17	27.46	25.15	22.27	14.91
	BIM+DGM-5	95.36	96.04	99.96	77.34	92.94	69.20	69.20	62.68	48.60	40.68	34.84	25.44
	BIM+本文方法	**96.32**	**96.90**	99.96	**78.84**	**94.30**	**70.86**	**70.22**	**64.50**	49.42	41.38	36.32	25.70

PGI,其平均提升了 29.22 百分点。相比 PGI,本文方法平均提升了 5.30 百分点。与基础的 DGM-5 方法相比,本文方法平均提升了 2.12 百分点。

在 8 种无防御模型上,本文方法在大多数情况下显著优于其他对比方法。以 DN201 作为源模型并采用 PGD 攻击为例,与其他增强方法相比,除对源模型进行白盒攻击时的成功率低于 SI 之外,本文方法在攻击其余无防御模型时均取得了最佳效果。总体上相较于基础的 DGM 提升了 8.62 百分点。在 4 种防御模型上,本文方法取得了第 2 的优异效果。

以 Res152 为源模型,并采用 PGD 对其余 11 个模型进行迁移攻击,本文方法对这 12 个模型的平均攻击成功率达到了 66.25%,而 DGM-5 的成功率为 64.30%,效果较好的 TI 方法仅达到了 50.76%。本文方法在多个网络上的迁移攻击表现更加稳定,证明其具有较强的迁移攻击能力,这在实际应用场景尤为重要。

从实验分析可知,无论是在白盒攻击场景还是黑盒攻击场景下,本文方法相较于现有的迁移攻击和增强方法表现出显著的性能提升,并对基础的 DGM 方法进行了优化。由于本文方法与模型结构和攻击方式无关,因此可以与其他攻击方法有效结合,从而实现性能增益。结果表明,本文方法在增强攻击可迁移性方面取得了重要且有价值的进展。

4 结论

针对现有集成攻击面临的计算资源消耗大和可供集成的模型种类及数量有限等问题,本文提出了一种新颖的自蒸馏训练策略,旨在构建更为多样丰富的学生模型集合,以增强集成攻击的有效性和跨模型攻击的迁移能力。

具体而言,在源教师模型自蒸馏过程中,通过对学生模型进行基因重组,生成更加多样化的学生模型。随后,将源教师模型与生成的多样化学生模型进行集成,并利用基于梯度的攻击方法生成具有高度可迁移性的对抗样本。实验结果表明,本文方法显著增强了现有基于梯度的对抗攻击方法的可迁移性。此外,该策略作为一种通用的增强可迁移性的方法,能够与其他基于迁移的攻击方法有效结合。

尽管已取得上述进展,基于基因重组自蒸馏策略的潜能仍有待深入探索。未来的研究方向应聚焦于该策略的理解与优化,特别是基因交换的学生模型数量、模型结构、重组策略的细化以及系统性重组频率,以充分发挥该策略的潜在优势。通过对这些因素的深入探究,有望实现该策略的最大效益。

参考文献:

[1] 罗荣辉, 袁航, 钟发海, 等. 基于卷积神经网络的道路拥堵识别研究[J]. 郑州大学学报(工学版), 2019, 40(2): 18-22.
LUO R H, YUAN H, ZHONG F H, et al. Traffic jam detection based on convolutional neural network [J]. Journal of Zhengzhou University (Engineering Science), 2019, 40(2): 18-22.

[2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2015-03-20) [2024-05-10]. http://arxiv.org/abs/1412.6572.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19)[2024-05-10]. http://arxiv.org/abs/1312.6199.

[4] 赵俊杰, 王金伟. 基于 SmsGAN 的对抗样本修复[J]. 郑州大学学报(工学版), 2021, 42(1): 50-55.
ZHAO J J, WANG J W. Recovery of adversarial examples based on SmsGAN[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(1): 50-55.

[5] YUAN X Y, HE P, ZHU Q L, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[6] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2017-02-11)[2024-05-10]. https://arxiv.org/abs/1607.02533.

[7] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2019-09-04)[2024-05-10]. https://arxiv.org/abs/1706.06083.

[8] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9193.

[9] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 2725-2734.

[10] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning[C]//2021 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 1924-1933.

[11] CAO Y J, WANG H B, ZHU C X, et al. Improving the transferability of adversarial examples with diverse gradients[C]//2023 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2023: 1-9.

[12] 何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049-2070.
HE Y Z, HU X B, HE J W, et al. Privacy and security issues in machine learning systems: a survey[J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070.

[13] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[EB/OL]. (2020-02-03)[2024-05-10]. https://arxiv.org/abs/1908.06281.

[14] WANG X S, LIN J D, HU H, et al. Boosting adversarial transferability through enhanced momentum[EB/OL]. (2021-03-19)[2024-05-10]. https://arxiv.org/abs/2103.10609.

[15] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4307-4316.

[16] GAO L L, ZHANG Q L, SONG J K, et al. Patch-wise attack for fooling deep neural network[C]//European Conference on Computer Vision. Cham: Springer, 2020: 307-322.

[17] 马克平. 试论生物多样性的概念[J]. 生物多样性, 1993(1):20-22.
MA K P. Oiscussion on the concept of biodiversity[J]. Chinese Biodiversity, 1993(1):20-22.

[18] MENDEL G, LIBRARY B, PUNNETT R C. Versuche über Pflanzen-Hybriden[M]. Brünn: Im Verlage des Vereines, 1866.

[19] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. (2017-06-07)[2024-05-10]. http://arxiv.org/abs/1611.02770.

[20] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[21] WU D X, WANG Y S, XIA S T, et al. Skip connections matter: on the transferability of adversarial examples generated with ResNets[EB/OL]. (2020-02-14)[2024-05-10]. http://arxiv.org/abs/2002.05990.

[22] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[C]//European Conference on Computer Vision. Cham: Springer, 2016: 630-645.

[23] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[24] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2261-2269.

[25] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2024-05-10]. http://arxiv.org/abs/1409.1556.

[27] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.

[28] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[EB/OL]. (2016-08-23)[2024-05-10]. https://arxiv.org/abs/1602.07261.

[29] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[EB/OL]. (2020-04-26)[2024-05-10]. http://arxiv.org/abs/1705.07204.

Adversarial Attack Method Based on Genetic Recombination Knowledge Distillation Strategy

LIU Minglin, ZHOU Chuanjin, WANG Runze, WANG Chao, CAO Yangjie

(School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: To address limitations of traditional ensemble attack methods, which were constrained by high computational resource requirements, including training data and time, a low computational complexity ensemble attack method based on genetic recombination was proposed. This method aimed to enhance the transferability of existing adversarial attacks by generating a more diverse set of ensemble models. Firstly, the concept of genetic recombination was introduced into knowledge distillation. In this process, student models were treated as independent individuals, with their parameters considered as genes. Each round of distillation learning was viewed as a gene evolution. Randomly exchanging parameters among student models during the evolution process achieves artificial genetic recombination, resulting in superior offspring genes. By setting different distillation temperatures, multiple diversified student models were obtained. Next, these diverse student models were integrated with the source teacher model. Finally, the integrated model was used to generate adversarial examples with stronger transferability. Experimental results on a subset of the ImageNet validation set demonstrated that the proposed method significantly improved the transferability of adversarial samples compared to other baseline algorithms. Using ResNet152 as the source model and PGD as the attack method, the proposed method achieved the highest transfer attack success rate across 11 black-box models, outperforming the baseline PGD method by an average of 34.52 percentage point, the PGI method by an average of 5.30 percentage point, and the DGM method by an average of 2.12 percentage point.

Keywords: ensemble attacks; adversarial examples; transferability; genetic recombination; knowledge distillation

(上接第7页)

Security Performance Optimization of IRS-assisted Wireless Sensing Systems

SUN Gangcan[1,2], ZHAO Xinrui[1], HAO Wanming[2], PENG Shumin[2]

(1. Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450001, China; 2. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: To address the sensing security issue of unauthorized radar station (URS) stealing target information in multi-radar scenarios, a secure wireless sensing system model based on intelligent reflecting surface (IRS) assistance was proposed. This system deployed an IRS with sensing capabilities on the target and adopted a two-phase sensing scheme. In the first phase, the IRS sensing unit estimated the angle information of all radars. In the second phase, the IRS reflection coefficients were designed based on the estimation results to minimize the perception probability of URS. Specifically, under the constraints of ensuring the signal-to-noise ratio of the legitimate radar station (LRS) and the IRS reflection phase shift modulus, an optimization problem was formulated to minimize the maximum signal-to-noise ratio of the URS. An iterative optimization algorithm based on the Dinkelbach method and semidefinite relaxation (SDR) technique was proposed. Simulation results showed that compared to the scheme without IRS, the signal-to-noise ratio of the LRS improved by approximately 3 dB, while the signal-to-noise ratio of the URS decreased by about 12 dB, demonstrating that the proposed scheme significantly enhanced system security performance.

Keywords: intelligent reflecting surface; radar sensing; sensing security; angle-of-arrival estimation; reflection optimization