基于类别相似性驱动的动态伪目标对抗攻击方法研究

余红霞,鲁磊纪,鲍 蕾*,陈 军,张林俊
(中国人民解放军陆军兵种大学信息工程系,安徽合肥 230031)

摘 要: 本文针对深度神经网络对抗样本在黑盒攻击中迁移性受限的核心问题,提出了一种基于类别语义关联的动态伪目标对抗攻击框架. 现有方法因忽视类别间语义关联性,导致对抗扰动容易陷入模型特异性过拟合,严重制约对抗样本跨模型迁移效能. 研究表明,对抗样本在迁移过程中倾向于被误分类为语义相近的类别,而非随机类别,揭示出类别相似性是影响迁移性的关键因素. 本研究通过挖掘模型特征空间中相似类别的共享对抗子空间特性,创新性地提出了基于类别相似性驱动的动态伪目标对抗攻击方法. 首先,构建一个动态伪目标筛选策略. 在每一次扰动迭代中,根据当前模型对对抗样本的预测置信度分布,从所有非正确类别中选取预测概率最高类别作为“伪目标”. 该目标并非固定,而是随迭代过程自适应调整,确保扰动方向始终指向最具迁移潜力的语义区域. 其次,提出双梯度协同更新框架. 将原始基于真实类别的对抗损失梯度与伪目标类别的误导梯度进行线性加权融合,通过梯度场的叠加效应,扰动更新不仅能逃离源模型的决策边界,还能同时朝向多个模型共享的语义子空间推进,从而显著提升对抗样本的跨模型可迁移性. 此外,本文方法具有广泛的兼容性与可扩展性,可作为一种通用优化机制与多种主流梯度攻击策略无缝融合,在每次梯度更新计算中,通过引入动态伪目标梯度项,在不破坏原方法梯度结构的前提下,显著增强其跨模型迁移能力. 实验表明,本文方法在跨架构(Convolutional Neural Network,CNN/Transformer)、跨规模(轻量化模型)攻击场景下均展现出优越的迁移鲁棒性. 此外,该方法展现出良好的兼容性,可与多种梯度攻击策略及数据增强计算结合,在单一攻击、组合攻击与集成攻击模式下均优于现有方法. 本研究为对抗攻击提供了一种基于语义相似性的通用优化范式,为提升黑盒攻击的迁移性提供了新思路.

关键词: 对抗样本;迁移性;类别相似性;动态伪目标;双梯度协同更新
基金项目: 国家自然科学基金(No.62076252)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)08-2854-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250578

Research on Category Similarity-Driven Dynamic Fake Target Adversarial Attack Methods

YU Hong-xia, LU Lei-ji, BAO Lei*, CHEN Jun, ZHANG Lin-jun
(*Department of Information Engineering, PLA Army Services University, Hefei, Anhui* 230031, *China*)

Abstract: This paper addresses the critical limitation of adversarial example transferability in black-box attacks for deep neural networks by proposing a dynamic fake target adversarial attack framework based on categorical semantic correlations. Existing methods often overlook inter-class semantic relationships, causing adversarial perturbations to overfit to model-specific features and severely restricting the adversarial example's cross-model transferability. Studies have indicated that adversarial examples are more likely to be misclassified into semantically similar classes rather than arbitrary categories during the transfer process. This observation underscores the significance of class similarity as a pivotal factor influencing transferability. In this research, we innovatively propose a class-similarity-driven dynamic pseudo-targeted adversarial attack method by exploring the shared adversarial subspace characteristics among semantically analogous categories within the feature space. First, we establish a dynamic pseudo-target selection strategy. In each perturbation iteration, we identify the class with the highest predicted probability among all incorrect categories as the “pseudo-target”, based on the current

收稿日期:2025-07-02;录用日期:2025-08-20;责任编辑:覃怀银
*通讯作者:鲍蕾

model's confidence distribution regarding the adversarial example. This pseudo-target is not fixed, instead, it is adaptively adjusted throughout the iterative process, ensuring that the perturbation direction consistently orients toward the most transferable semantic region. Second, we introduce a dual-gradient collaborative update framework. This framework integrates the adversarial loss gradient pertaining to the true class with the misleading gradient associated with the pseudo-target class through linear weighting. Leveraging the superposition effect in the gradient field, the perturbation update not only circumvents the decision boundary of the source model but also progresses into the shared semantic subspace of multiple models, thereby significantly enhancing the cross-model transferability of adversarial examples. Furthermore, our proposed method demonstrates wide compatibility and extensibility, serving as a versatile optimization mechanism that can be seamlessly integrated with various mainstream gradient-based attack strategies. During each gradient update, the incorporation of a dynamic pseudo-target gradient term markedly amplifies cross-model transfer capability without compromising the original gradient structure of the foundational method. Experimental results illustrate that the proposed approach exhibits superior transfer robustness in cross-architecture (e.g., Convolutional Neural Networks and Transformers) and cross-scale (e.g., lightweight models) adversarial attack scenarios. Additionally, it showcases excellent compatibility, enabling effective integration with diverse gradient attack strategies and data augmentation techniques, thereby outperforming existing methodologies across single, combined, and ensemble attack settings. This study proposes a general optimization paradigm based on semantic similarity for adversarial attacks, offering novel insights to enhance the transferability of black-box attacks.

Key words: adversarial examples; transferability; category similarity; dynamic fake target; dual-gradient collaborative updating

Foundation Item(s): National Natural Science Foundation of China (No.62076252)

1 引言

近年来,深度学习领域的快速发展和显著成就,使其安全问题日益受到关注. 其中,对抗样本(adversarial examples)被公认为是深度神经网络系统最突出的安全隐患之一. 研究表明,通过添加肉眼难以察觉的细微扰动,可误导精心训练的模型产生错误预测[1-4]. 对抗样本的威胁不仅存在于单一模型,还因其跨模型迁移性而更具破坏性[5,6]. 尽管对抗样本具有跨模型迁移性,但其实际效果受限于模型间决策边界的结构性差异.

为提升黑盒攻击场景下的迁移能力,Zou等人[7]引入了Adam优化算法,根据历史梯度信息自适应调整学习率,利用tanh函数控制更新方向和幅度,保持扰动量的同时,更有效地找到误导模型的路径,提高了攻击效率和效果. Peng等人[8]则从另一个角度出发,通过整合预期数据点来稳定更新方向,旨在提高扰动生成的稳定性和攻击的连续性. Wan等人[9]引入梯度预测-修正机制,通过预测模型的梯度变化趋势自适应控制扰动更新方向,从而提高攻击的精准度和效率. Wang等人[10]提出了全局动量初始化方法,通过预收敛和全局搜索解决梯度消失问题,优化初始攻击方向的一致性,提升对抗攻击迁移性. 然而,这些方法通常聚焦于输入空间的局部扰动优化,忽略了模型间相似类别在特征空间中的决策边界重叠特性,导致生成的对抗样本难以深入共享对抗子空间. 此外,采用单一远离原始类别的梯度方向的攻击策略,容易使扰动过拟合到白盒模型的局部决策边界,缺乏动态目标引导机制,陷入模型特异性过拟合,无法有效穿透模型共享的对抗区域,黑盒迁移成功率受限.

Warde-farley等人[11]的理论研究揭示了这一问题的根源:不同模型的对抗样本空间存在共享的高维连续子空间,其决策边界在输入域中非常接近. 对抗样本倾向于在连续的大区域内集中,这与模型决策边界的平滑性和连续性有关. 不同模型的对抗子空间重叠,使得对抗样本在跨模型迁移时保持集中性,为对抗攻击提供了理论支持. 这一发现表明,尽管模型结构和参数存在差异,只要对抗扰动能使数据点越过白盒模型的决策边界足够远,这些扰动很可能对其他黑盒模型也有效. 然而,现有方法未能充分利用这一理论,导致迁移效果不佳. 进一步研究表明,对抗样本的迁移性与类别相似性密切相关. Mei等人[12]发现,模型在特征空间中对相似类别的映射具有相似性,因而对抗样本倾向于被错误分类为特定类别而非随机类别. 利用这种类别相似性可以提升对抗攻击的有效性. Ozbulak等人[13]的统计显示,71%的对抗样本在迁移攻击时会被错误分类为原始图像预测的相似类别,证实了对抗样本迁移性与模型对相似类别的映射一致性.

为利用类别相似性提升迁移性,本文提出了一种伪目标梯度协同更新策略(Fake-Targeted adversarial attack,FT). 通过实时分析模型预测置信度分布,自适应选择高误分类概率的相似伪目标类别,并采用双梯度协同更新机制——将原始分类损失梯度与伪目标对抗梯度动态融合生成对抗样本. 这种动态扰动更新机制能够突破传统方法在局部最优陷阱和模型特异性过拟合方面的局限,通过梯度场的非线性叠加效应,使扰

动向量穿透当前模型的决策边界并尽可能深地进入对抗子空间的共享区域，从而显著提升跨模型攻击效率（如图1）. 主要贡献包括：(1)提出动态伪目标驱动的攻击框架，利用相似类别间决策边界重叠特性，提升跨模型攻击效率；(2)验证了类别相似性在提升对抗样本迁移性中的关键作用，实验显示该方法在黑盒攻击中效果显著；(3)方法的通用性和兼容性，可与多种攻击方法和数据增强策略结合，增强攻击效果.

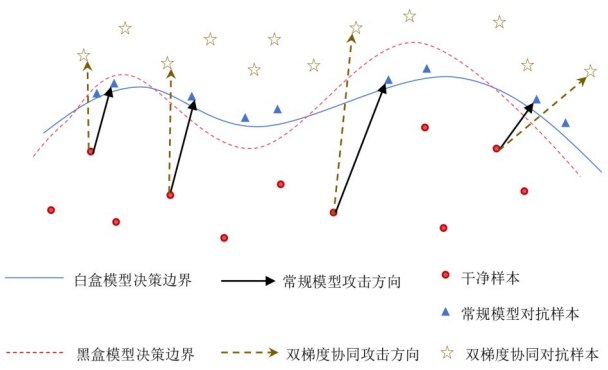


图1 伪目标攻击方法

2 本文方法

2.1 类别相似性增强机制

在对抗攻击中，单一远离白盒模型决策边界的攻击策略难以引导对抗样本越过更多模型的决策边界，Mei等人[12]的研究表明，从某一类生成的对抗样本倾向于被错误分类为特定的类，而非随机地分布于所有其他类别之中. 具体而言，若从类别i出发，通过添加精心设计的扰动来构造对抗样本，这类样本有很大概率会被错误地归类至类别j，而非其他任意类别. 这种现象的根源在于模型学到的特征表示，对于相似物理形态的类别（如“汽车”与“卡车”，“猫”与“狗”）在特征空间中的映射具有高度相似性，这些特征点靠近类间的决策边界，只需添加少量噪声即可轻松将其制造成对抗样本. 文献[14,15]都认识到特征在生成迁移性对抗样本中的关键作用，通过定位并攻击对模型决策最关键的特征，而非盲目扰动所有像素，以显著提升对抗样本在不同模型间的可迁移性. 为量化类别相似性，文献[12]采用余弦相似度作为度量方式，计算公式如下：

$$S_{ij}=\frac{\boldsymbol{w}_i\cdot\boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2\|\boldsymbol{w}_j\|_2} \tag{1}$$

其中，$\boldsymbol{w}_i$和$\boldsymbol{w}_j$分别表示类别i和类别j的权重向量，S_{ij}是它们的余弦相似度. 这里通过L2范数对其进行归一化，消除了向量长度差异的影响，仅反映向量方向的一致性，方向越接近，S_{ij}值越高，表明外观相似但类别不同的样本，在特征空间中具有相似的映射. 不同类别之间的相似度越高，导致的攻击比例就越高. 因此，对抗攻击中优先选择模型预测概率最高的错误类别作为攻击目标，可实现最小扰动下的高效攻击.

这一机制可靠的前提是相似性度量的稳定性，对此，Mei等人[12]通过控制权重向量夹角误差$|\theta|$展开敏感性分析. 实验结果表明，当$|\theta|\leqslant 0.2$时，VGG-16和ResNet-34模型中满足对称属性的类别对比例均超过95%，其参数敏感性低于5%的波动，验证了余弦相似度作为核心度量具有高度鲁棒性. 因此，选择预测概率最高的错误类别作为攻击目标，能够更有效地引导样本跨越更多模型的决策边界.

Ozbulak等人[13]的研究进一步表明，71%的对抗样本在实现模型间对抗转移时，被错误分类为原始图像预测的前5个类别之一. 这一发现揭示了对抗样本在迁移过程中倾向于被误分类为与原始类别语义相似的类别，而非随机分布的其他类别. 这与Mei等人[12]观察到的对称性现象高度一致，共同说明了类别语义相似性是驱动特定误分类模式和影响攻击迁移性的关键因素.

因此，本文在对抗攻击中利用这种类别相似性来引导扰动的生成，选择模型预测概率最高的错误类别作为攻击目标（伪目标），确保扰动方向指向与原始类别相似的方向. 在无目标攻击模式下，在跨越白盒模型的决策边界时，通过结合伪目标攻击的方式引导扰动向该相似类别方向生成，以远离决策边界，进入其与黑盒模型的共享子空间，从而提升攻击的迁移性.

2.2 双梯度协同更新方法

在无目标攻击模式下，常用攻击方法（如I-FGSM、MI-FGSM、NI-FGSM）为了跨越白盒模型的决策边界，通常采用符号化的梯度$\operatorname{sign}\left(\nabla J(\boldsymbol{x},y)\right)$更新扰动. 更新规则描述如下：

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\operatorname{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\operatorname{sign}\left(\nabla J(\boldsymbol{x}_t^{\text{adv}},y)\right)\right\} \tag{2}$$

其中，$\operatorname{sign}(\cdot)$为符号函数，$J(\boldsymbol{x},y)$为损失函数，$\nabla J(\boldsymbol{x},y)$为损失函数对输入$\boldsymbol{x}$的梯度，$y$为样本$\boldsymbol{x}$的真实标签，$\boldsymbol{x}^{\text{adv}}$为对抗样本. 如图1所示，传统基于梯度的攻击方法生成的对抗样本（蓝色三角形）通常集中在白盒模型的决策边界（实线）附近，这导致许多对抗样本未能成功越过黑盒模型的决策边界（虚线）.

为了引导对抗样本尽可能远离白盒模型的决策边界，以达到离开黑盒模型可能的决策边界，我们利用了类的相似性原理，即模型在特征空间中对相似类别的映射具有相似性，也就是对抗样本更倾向于被误分为特定的相似类别，并且该相似性原理在不同模型上具有一致性. 对此，本文通过引入类别相似性原理设定伪目标，在更新方向上融合真实标签的梯度信息和伪目标的梯度信息，远离原始类别，逼近伪目标，进而引导

生成的对抗样本越过更多模型的决策边界(如图1). 总体流程如图2所示.

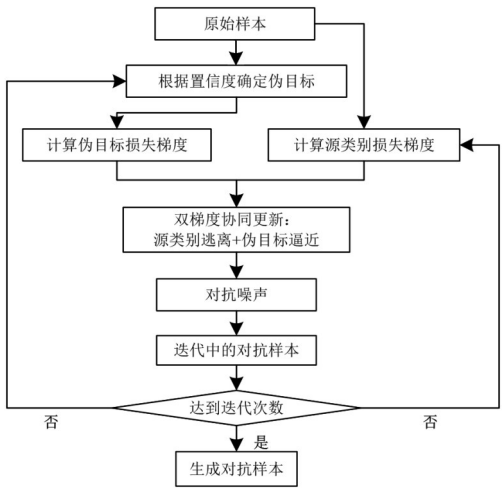


图2　对抗样本生成流程图

具体而言,给定训练好的分类器$f(\boldsymbol{x})$,分类器对输入样本$\boldsymbol{x}$的预测置信度分布记为p. 其中$p_i(\boldsymbol{x})=P(\hat{y}=i\,|\,\boldsymbol{x})$表示$\boldsymbol{x}$属于第$i$类的得分,$i=1,2,\cdots,C$,$C$为类别总数. 根据预测置信度分布情况,取置信度最高的误判类别作为伪标签y^*,满足:

$$y^*=\arg\max_{i\neq y^{\text{true}}} p_i \tag{3}$$

为了确保y^*是模型最有可能误判的类别,在迭代过程中动态选择伪标签y^*,使得其与原始类别的对抗子空间的距离始终保持最近. 通过选择最可能误分的类别,能够引导攻击方向更快地跨越白盒的决策边界;通过指导性地靠近误分类别而非仅仅远离原始类别,能够更好地远离白盒的决策边界,以达到同时跨越黑盒模型决策边界的目的.

有目标攻击模式下,以y^*为伪标签,分类器$f(\cdot)$对输入$\boldsymbol{x}$的损失$J(\boldsymbol{x},y^*)$,梯度为$\nabla J(\boldsymbol{x},y^*)$,采用符号化的梯度$\text{sign}\left(\nabla J(\boldsymbol{x},y^*)\right)$更新扰动. 更新规则描述如下:

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}-\alpha\cdot\text{sign}\left(\nabla J(\boldsymbol{x}_t^{\text{adv}},y^*)\right)\right\} \tag{4}$$

为了实现同时远离原始类别和逼近相似类别,我们对真实类别的损失梯度与伪目标对抗梯度进行了动态融合. 考虑到正确类别的得分越高,越要远离原始类别;错误类别的得分越高,越要靠近目标类别,围绕预测置信度分布情况设置梯度融合权重,得到对抗扰动更新梯度如下:

$$\boldsymbol{g}_{t+1}=\lambda_t\cdot\nabla J\left(\boldsymbol{x}_t^{\text{adv}},y\right)-\left(1-\lambda_t\right)\cdot\nabla J(\boldsymbol{x}_t^{\text{adv}},y^*) \tag{5}$$

其中,$\lambda_t=\dfrac{p_y f(\boldsymbol{x}_t^{\text{adv}})}{p_y f(\boldsymbol{x}_t^{\text{adv}})+p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})}$,$p_y f(\boldsymbol{x}_t^{\text{adv}})$为分类器$f(\cdot)$对真实标签$y$的预测置信度,$p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})$则为对伪标签$y^*$的预测置信度. 基于融合的梯度,可实现对抗样本的更新:

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}\left(\boldsymbol{g}_{t+1}\right)\right\} \tag{6}$$

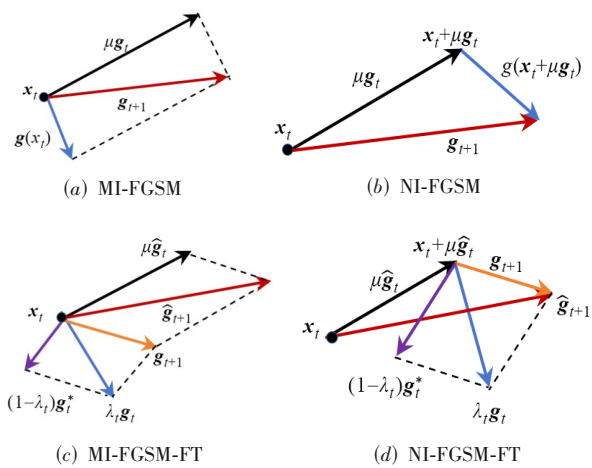
本文提出的基于伪目标梯度协同更新策略的攻击方法如算法1所示. 该方法可以与梯度攻击方法进行结合,提高其攻击成功率. 以符号化的MI-FGSM、NI-FGSM为例,其梯度更新机制如图3所示. 从图3(a)可以看出,MI-FGSM在$\boldsymbol{x}_t$点累积梯度实现对抗样本的更新;NI-FGSM使用$\boldsymbol{x}_t+\mu\boldsymbol{g}_{t-1}$点累积的梯度更新对抗样本(图3($b$));本文提出的FT策略则是在$\boldsymbol{x}_t$或$\boldsymbol{x}_t^{\text{nes}}$点先一步计算真实目标和伪目标的协同梯度,然后再采用MI-FGSM或者NI-FGSM的动量算法累积梯度,实现更新(图3(c)、图3(d)).

算法1　FGSM-FT

输入: 干净样本$\boldsymbol{x}$;标签y;损失函数J;最大迭代次数T;扰动ε

输出: 对抗样本$\boldsymbol{x}^{\text{adv}}$

1. $\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$
2. FOR t=1 TO T DO
3. 　计算置信度分布$p_i f(\boldsymbol{x}_t^{\text{adv}})=P\left(\hat{y}=i\,|\,\boldsymbol{x}_t^{\text{adv}}\right),i=1,2,\cdots,C$
4. 　确定伪标签$y^*=\arg\max_{i\neq y^{\text{true}}} p_i$
5. 　计算$\boldsymbol{x}_t^{\text{adv}}$的真实标签$y$得分$p_y f(\boldsymbol{x}_t^{\text{adv}})$
6. 　计算$\boldsymbol{x}_t^{\text{adv}}$的伪标签$y^*$得分$p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})$
7. 　计算加权系数$\lambda_t=\dfrac{p_y f(\boldsymbol{x}_t^{\text{adv}})}{p_y f(\boldsymbol{x}_t^{\text{adv}})+p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})}$
8. 　计算真实标签的损失梯度$\boldsymbol{g}_t=\nabla J(\boldsymbol{x}_t^{\text{adv}},y)$
9. 　计算伪标签的损失梯度$\boldsymbol{g}_t^*=\nabla J(\boldsymbol{x}_t^{\text{adv}},y^*)$
10. 　计算合成梯度$\boldsymbol{g}_{t+1}=\lambda_t\cdot\nabla J\left(\boldsymbol{x}_t^{\text{adv}},y\right)-\left(1-\lambda_t\right)\cdot\nabla J(\boldsymbol{x}_t^{\text{adv}},y^*)$
11. 　更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}(\boldsymbol{g}_{t+1})\right\}$
12. END FOR
13. $\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$



(a) MI-FGSM　(b) NI-FGSM　(c) MI-FGSM-FT　(d) NI-FGSM-FT

图3　梯度更新示意图

需要注意的是，当与MI-FGSM结合时，根据图3(*c*)和算法2(MI-FGSM-FT)所示，在$\boldsymbol{x}_t$点根据协同梯度更新机制计算出$\boldsymbol{g}_{t+1}$，在$\boldsymbol{g}_{t+1}$的基础上引入动量机制，更新梯度$\hat{\boldsymbol{g}}_{t+1}=\mu\boldsymbol{g}_{t+1}+\boldsymbol{g}_{t+1}/\left\|\boldsymbol{g}_{t+1}\right\|_1$，来更新对抗样本实现攻击.

算法2 MI-FGSM-FT

输入: 干净样本$\boldsymbol{x}$;标签y;损失函数J;最大迭代次数T;扰动ε

输出: 对抗样本$\boldsymbol{x}^{\text{adv}}$

$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$

FOR t=1 TO T DO

　计算置信度分布$p_i f(\boldsymbol{x}_t^{\text{adv}})=P\left(\hat{y}=i\left|\boldsymbol{x}_t^{\text{adv}}\right.\right), i=1,2,\cdots,C$

　确定伪标签$y^*=\arg\max_{i\neq y^{\text{true}}} p_i$

　计算$\boldsymbol{x}_t^{\text{adv}}$的真实标签$y$得分$p_y f(\boldsymbol{x}_t^{\text{adv}})$

　计算$\boldsymbol{x}_t^{\text{adv}}$的伪标签$y^*$得分$p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})$

　计算加权系数$\lambda_t=\dfrac{p_y f(\boldsymbol{x}_t^{\text{adv}})}{p_y f(\boldsymbol{x}_t^{\text{adv}})+p_{y^*} f(\boldsymbol{x}_t^{\text{adv}})}$

　计算真实标签的损失梯度$\boldsymbol{g}_t=\nabla J(\boldsymbol{x}_t^{\text{adv}}, y)$

　计算伪标签的损失梯度$\boldsymbol{g}_t^*=\nabla J(\boldsymbol{x}_t^{\text{adv}}, y^*)$

　计算合成梯度$\boldsymbol{g}_{t+1}=\lambda_t\cdot\nabla J\left(\boldsymbol{x}_t^{\text{adv}}, y\right)-\left(1-\lambda_t\right)\cdot\nabla J(\boldsymbol{x}_t^{\text{adv}}, y^*)$

　计算梯度$\hat{\boldsymbol{g}}_{t+1}=\mu\hat{\boldsymbol{g}}_t+\boldsymbol{g}_{t+1}/\left\|\boldsymbol{g}_{t+1}\right\|_1$

　更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}(\hat{\boldsymbol{g}}_{t+1})\right\}$

END FOR

$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

当与NI-FGSM结合时，如算法3(NI-FGSM-FT)所示，在步骤3中使用Nesterov加速，基于历史梯度计算前瞻点$\boldsymbol{x}_t^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\mu\cdot\boldsymbol{g}_t$，该点预测了沿动量方向的下一步位置，然后在Nesterov点$\boldsymbol{x}_t^{\text{nes}}$进行对抗样本梯度和伪目标梯度的协同更新(步骤4~11)，再进行后续的梯度计算和对抗样本更新.

算法3 NI-FGSM-FT

输入: 干净样本$\boldsymbol{x}$;标签y;损失函数J;最大迭代次数T;扰动ε

输出: 对抗样本$\boldsymbol{x}^{\text{adv}}$

$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$, $\hat{\boldsymbol{g}}_0=\boldsymbol{0}$

FOR t=1 TO T DO

　计算$\boldsymbol{x}_t^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\mu\cdot\hat{\boldsymbol{g}}_t$

　计算置信度分布$p_i f(\boldsymbol{x}_t^{\text{nes}})=P\left(\hat{y}=i\left|\boldsymbol{x}_t^{\text{nes}}\right.\right), i=1,2,\cdots,C$

　确定伪标签$y^*=\arg\max_{i\neq y^{\text{true}}} p_i$

　计算$\boldsymbol{x}_t^{\text{nes}}$的真实标签$y$得分$p_y f(\boldsymbol{x}_t^{\text{nes}})$

　计算$\boldsymbol{x}_t^{\text{nes}}$的伪标签$y^*$得分$p_{y^*} f(\boldsymbol{x}_t^{\text{nes}})$

　计算加权系数$\lambda_t=\dfrac{p_y f(\boldsymbol{x}_t^{\text{nes}})}{p_y f(\boldsymbol{x}_t^{\text{nes}})+p_{y^*} f(\boldsymbol{x}_t^{\text{nes}})}$

　计算真实标签的损失梯度$\boldsymbol{g}_t=\nabla J(\boldsymbol{x}_t^{\text{nes}}, y)$

　计算伪标签的损失梯度$\boldsymbol{g}_t^*=\nabla J(\boldsymbol{x}_t^{\text{nes}}, y^*)$

　计算合成梯度$\boldsymbol{g}_{t+1}=\lambda_t\cdot\nabla J\left(\boldsymbol{x}_t^{\text{nes}}, y\right)-\left(1-\lambda_t\right)\cdot\nabla J(\boldsymbol{x}_t^{\text{nes}}, y^*)$

　计算梯度$\hat{\boldsymbol{g}}_{t+1}=\mu\hat{\boldsymbol{g}}_t+\boldsymbol{g}_{t+1}/\left\|\boldsymbol{g}_{t+1}\right\|_1$

　更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}(\hat{\boldsymbol{g}}_{t+1})\right\}$

END FOR

$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

3 实验及结果分析

3.1 实验设置

数据集：根据先前的研究，实验采用ILSVRC2012验证集，从中随机抽取1 000幅图像，覆盖全部类别.

模型：实验选取7种异构模型以验证跨架构迁移性：3种CNN模型Inception-V3(IncV3)、ResNet、Vgg；2种Transformer模型Vit-b-16(Vit-b)、Swin-s；2种轻量化模型EfficientNet、ShuffleNet. 此设置覆盖了从传统CNN到前沿Transformer、从标准模型到轻量化设计的多样化架构，确保实验结论的普适性.

对比方法：在实验设计中，为全面验证动态伪目标梯度协同更新策略(FT)对对抗攻击的增强效果，本文选取了主流梯度攻击方法(I-FGSM[16]、PGD[4]、MIFGSM[17]、NIFGSM[18]、GIFGSM[10]、AIFGTM[7]、IEFGSM[8]、PCIFGSM[9])和数据增强策略(DIM[19]、TIM[20]、SIM[18]). 为清晰区分基线方法与本文改进策略，统一将结合FT的方法命名为“X-FT”，其中“X”表示基线方法(如“FGSM-FT”表示基于FGSM的FT增强方法，“DIM-FT”表示结合DIM数据增强的FT方法). 通过对比“X”与“X-FT”在黑盒攻击场景下的性能差异，定量评估FT策略对攻击迁移性的提升贡献.

为系统验证动态伪目标梯度协同更新策略(FT)对对抗攻击迁移性的增强效果，本文设计了3组对照实验：单一模型攻击、组合攻击和集成攻击，并从攻击成功率、模型架构差异、跨规模迁移性等维度对比了FT方法(X-FT)与基线方法(X)的性能差异.

3.2 单一模型攻击

表1展示了单一模型攻击场景下的黑盒成功率对比. 从中可以看出，所有攻击方法在结合FT后，其黑盒攻击成功率均有所提高，但在不同攻击方法上性能存在差异. 例如GIFGSM-FT表现最佳，在几乎所有模型上达到最高攻击成功率，尤其在Vit-b模型上生成对抗样本攻击Vgg达到86.7%(+28.2%)，攻击Swin-s达到82.5%(+20.4%)，攻击ShuffleNet达到82.2%(+17.1%). 传统方法(如I-FGSM、PGD)提升则有限. 在IncV3上生成对抗样本时，I-FGSM-FT在ResNet上仅提升6.3%，而PGD-FT在Swin-s上提升3.6%. 这可能与这些方法的梯度更新机制较为保守有关.

此外，FT策略对不同模型架构均存在适应性. 具体来说，在CNN模型上的提升幅度最大. GIFGSM-FT在IncV3上生成的对抗样本，攻击Vgg的成功率提升

15.7%,ResNet提升11.7%. 在Vit-b上的提升较明显(+8.6%),但对Swin-s的提升更高(+11.5%),表明FT对Transformer架构同样有效. 使用GIFGSM-FT算法,攻击EfficientNet时提升13.6%,攻击ShuffleNet时提升13.9%,说明FT方法对小规模模型具有强穿透力. FT策略在所有模型上平均提升约为10%~16%,尤其在跨架构(CNN→Transformer)和跨规模(标准→轻量化)场景下表现稳定,验证了其穿透多个模型决策边界的能力.

3.3 组合攻击

Rebuffi等人[21,22]指出数据增强方法可以有效提高对抗样本的迁移性,因此本文将FT策略与DIM、SIM、TIM进行组合,实验结果如表2所示. 结果表明,动态伪目标梯度协同更新策略能够显著提升组合攻击的黑盒迁移性. 例如在IncV3上生成对抗样本,DIM-FT攻击ResNet和Vgg分别提升了5.5%和4.5%,SIM-FT在Vit-b模型上提升4.3%. 数据增强类方法与FT协同效果突出,而复杂组合策略因梯度冲突提升有限(DI-TI-SI攻击Swin-s模型提升3.0%). 此外,不同模型架构的提升差异明显:FT对CNN模型优化效果更显著(DIM-FT攻击Vgg提升4.5%),而Transformer模型因全局注意力机制对梯度依赖较低,提升幅度较小(TIM-FT攻击Vit-b提升3.8%). 轻量化模型(如ShuffleNet)攻击成功率平均提升4.3%,验证了FT在跨规模攻击中的适应性.

表1　单一模型攻击场景下黑盒成功率对比　　单位:%

模型	攻击方法	IncV3	ResNet	Vgg	Vit-b	Swin-s	EfficientNet	ShuffleNet
IncV3	I-FGSM	99.2	38.4	43.2	24.8	22.0	37.9	43.6
	I-FGSM-FT	100	44.7	48.2	26.8	25.2	39.9	48.6
	PGD	99.1	36.3	40.6	23.9	21.1	34.9	42.3
	PGD-FT	99.9	42.3	46.4	26.6	24.7	39.2	48.0
	MIFGSM	99.2	46.0	50.6	28.1	27.1	43.9	50.0
	MIFGSM-FT	100	51.2	55.2	31.6	29.1	47.0	54.8
	NIFGSM	98.8	46.5	51.8	27.9	25.5	43.6	49.2
	NIFGSM-FT	99.6	51.7	56.3	31.1	28.6	47.5	54.7
	GIFGSM	99.5	57.0	58.5	33.9	30.0	51.6	57.1
	GIFGSM-FT	100	68.7	74.2	42.5	41.5	65.2	71.0
	AIFGTM	98.7	40.1	42.1	26.0	24.0	38.7	46.5
	AIFGTM-FT	99.7	44.9	50.1	28.7	26.2	41.8	49.7
	IEFGSM	99.4	49.2	53.3	30.1	26.7	47.1	52.0
	IEFGSM-FT	100	55.2	57.2	33.3	30.1	51.5	56.8
	PCIFGSM	99.1	44.2	51.5	28.9	26.3	43.2	50.2
	PCIFGSM-FT	100	50.0	52.7	31.3	29.1	46.8	54.1
Vit-b	I-FGSM	45.1	41.2	46.9	100	45.6	46.1	49.9
	I-FGSM-FT	49.0	46.8	52.3	100	50.1	48.2	53.7
	PGD	43.8	39.3	44.0	99.7	42.9	42.8	46.4
	PGD-FT	47.2	43.5	50.4	99.9	46.0	45.3	51.9
	MIFGSM	50.8	49.2	57.5	99.9	54.0	54.7	57.3
	MIFGSM-FT	55.3	54.6	62.5	100	59.5	57.3	61.8
	NIFGSM	49.9	46.9	54.9	99.3	48.0	51.9	56.7
	NIFGSM-FT	51.6	51.4	60.4	99.9	55.4	53.9	60.9
	GIFGSM	58.0	56.8	65.6	100	62.1	62.3	65.1
	GIFGSM-FT	75.8	75.0	86.7	100	82.5	81.8	82.2
	AIFGTM	50.0	48.8	53.7	99.1	55.9	54.0	55.2
	AIFGTM-FT	54.7	54.8	61.0	100	61.5	59	61.6
	IEFGSM	53.2	50.7	58.2	100	58.4	56.1	60.5
	IEFGSM-FT	56.3	56.9	64.1	100	60.9	60.3	64.1
	PCIFGSM	50.5	49.9	57.8	99.9	55.9	55.6	58.5
	PCIFGSM-FT	56.5	55.4	63.2	100	59.9	57.8	62.4

3.4 集成攻击

Liu等人[23]指出由集成模型生成的对抗样本攻击其他模型,具有更强的通用性和可迁移性. 基于这一发现,进一步验证“X-FT”方法在集成模型设置下的性能.

表2 组合攻击场景下黑盒成功率对比 单位:%

模型	攻击方法	IncV3	ResNet	Vgg	Vit-b	Swin-s	EfficientNet	ShuffleNet
IncV3	DIM	99.1	45.2	48.1	27.2	24.6	44.8	47.7
	DIM-FT	100	50.7	52.6	30.6	27.4	47.2	52.0
	TIM	98.7	33.4	38.9	24.0	20.7	32.4	41.5
	TIM-FT	99.6	38.4	44.2	27.8	22.2	37.6	46.1
	SIM	99.5	43.7	47.3	26.7	23.5	40.6	48.7
	SIM-FT	100	48.2	51.5	31.0	27.4	45.6	51.5
	DI-TI-SI	94.4	38.3	43.4	27.3	22.0	36.7	47.8
	DI-TI-SI-FT	99.4	43.2	48.5	31.2	25.0	43.0	50.6
Vit-b	DIM	58.0	54.5	59.1	99.2	59.2	62.9	60.7
	DIM-FT	61.9	61.5	63.1	100	63.1	67.5	65.3
	TIM	42.8	39.4	44.0	98.4	36.6	40.4	48.6
	TIM-FT	46.2	43.8	47.9	100	38.1	43.8	51.1
	SIM	46.5	43.1	50.0	100.0	46.5	46.7	52.8
	SIM-FT	50.7	49.0	54.4	100	50.7	50.4	55.8
	DI-TI-SI	51.7	49.6	54.0	96.5	47.0	54.6	58.2
	DI-TI-SI-FT	55.4	56.5	56.1	99.9	51.8	59.0	62.6

本文采用Dong等人[17]提出的集成策略,通过融合多个网络的逻辑回归值(logits)来构建集成攻击. 表3展示了分别集成CNN模型(IncV3、ResNet、Vgg、GoogleNet)、Transformer模型(Vit-b、Swin-b、Swin-s、Swin-t)的黑盒攻击成功率.

表3的实验结果表明,在集成模型场景下结合FT策略,对抗样本的跨模型迁移性显著增强. 例如,采用CNN模型集成时,MIFGSM-FT在ResNet和EfficientNet模型上的成功率分别达到82.3%和70.4%,较单一攻击提升31.1%和23.4%;同时,轻量化模型(如ShuffleNet)的攻击成功率平均提升约20%,验证了集成策略对跨规模模型的泛化能力. 不同攻击方法表现分化明显:MIFGSM-FT凭借动量机制在CNN模型(如Vgg)上达到88.7%的最高成功率,而传统方法(如I-FGSM-FT)因梯度干扰问题提升有限(Vit-b攻击成功率36.6%). 值得注意的是,模型架构差异显著影响攻击效果. 例如,Transformer模型(如Vit-b)因全局注意力机制对梯度敏感性较低,攻击成功率普遍低于CNN模型(最高42.2%). 此外,PCIFGSM-FT在部分模型上表现波动,且Vit-b集成攻击存在对自身模型过拟合问题,需进一步优化模型多样性.

表3 集成攻击场景下黑盒成功率对比 单位:%

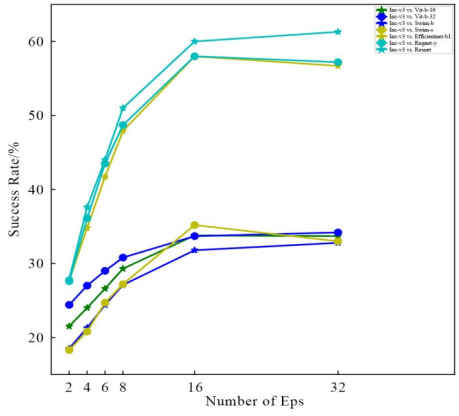
集成方法	攻击方法	IncV3	ResNet	Vgg	Vit-b	Swin-s	EfficientNet	ShuffleNet
CNN集成	I-FGSM-FT	99.5	77.9	84.9	36.6	39.9	65.1	67.5
	PGD-FT	99.1	75.5	82.8	34.5	35.8	59.6	65.6
	MIFGSM-FT	99.5	82.3	88.7	42.2	45.4	70.4	72.2
	NIFGSM-FT	99.0	79.8	87.0	41.5	43.0	68.3	69.8
	GIFGSM-FT	99.8	76.1	82.7	42.0	41.3	66.4	69.6
	AIFGTM-FT	98.9	80.0	87.3	40.4	44.2	66.8	69.2
	IEFGSM-FT	100	82.0	87.2	42.2	44.8	70.7	70.1
	PCIFGSM-FT	99.9	77.8	84.6	41.8	42.9	67.3	69.1
Transformer集成	I-FGSM-FT	47.8	51.9	59.5	97.6	99.8	54.3	57.1
	PGD-FT	46.1	49.2	58.7	97.3	99.8	51.4	54.6
	MIFGSM-FT	55.7	60.3	67.3	98.8	100	63.4	63.3
	NIFGSM-FT	53.3	57.9	65.7	96.0	99.7	59.7	61.5
	GIFGSM-FT	60.4	63.5	70.3	99.9	98.1	67.9	66.8
	AIFGTM-FT	54.9	60.5	66.8	95.4	100	63.6	62.1
	IEFGSM-FT	58.7	64.7	70.9	99.8	99.7	67.8	68.1
	PCIFGSM-FT	58.7	63.4	71.6	100	99.7	67.5	67.0

4　消融研究

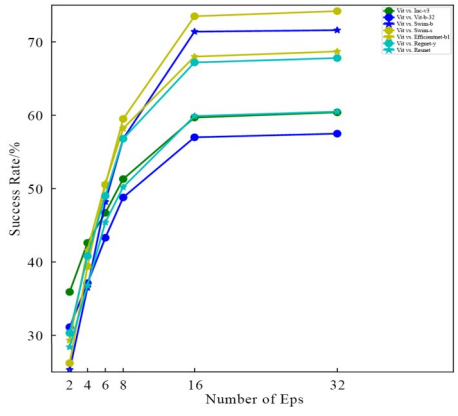
在本节中,针对超参数进行了一系列的消融实验,旨在深入理解扰动幅度ε以及预收敛迭代次数(T)对攻击性能的具体影响. 实验中,分别以IncV3和Vit-b模型为源,利用X-FT方法生成对抗样本,并将其迁移到其他目标模型上进行测试.

在探究扰动幅度ε实验中,固定迭代次数T为20,测试了ε在2、4、6、8、16、24、32等不同取值下的攻击效果. 实验结果揭示了扰动幅度ε与攻击成功率性能之间的显著关系. 无论是IncV3模型(图4(a))还是Vit-b模型(图4(b)),当ε值较小时,随着其逐渐增大,攻击成功率呈现出明显的提升趋势. 当扰动幅度ε超过16时,进一步增加ε值对提升攻击成功率的效果并不明显,甚至在某些情况下,攻击成功率不再提升. 这表明$\varepsilon=16$是该攻击方法的一个关键阈值,超过此值后增大扰动对提升攻击效果作用有限. 基于此,将$\varepsilon=16$确立为实验的基准参数值.

在确定了扰动后,聚焦于迭代次数T对X-FT方法性能的影响. 固定扰动幅度ε为16,同时在不同的迭代次数T下运行实验,以观察T值的变化如何影响攻击成功率. 如图5所示,实验结果清晰地揭示了迭代次数T与攻击成功率之间的正相关关系. 随着T值的增加,攻击成功率呈现出一致的提升趋势. 值得注意的是,当T值设定为10时,达到了最优性能. 这一结果表明,在给定的参数配置下,通过增加迭代次数,可以有效优化攻击策略,提升对抗样本的生成质量,从而提高攻击成功率.

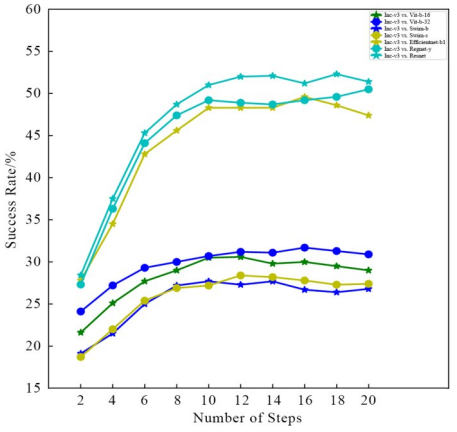


(a) IncV3模型

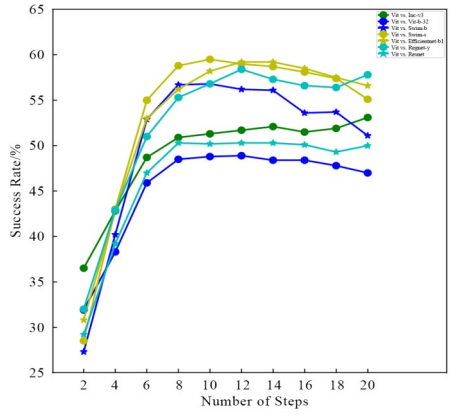


(b) Vit-b模型

图4　扰动对攻击成功率的影响



(a) IncV3模型

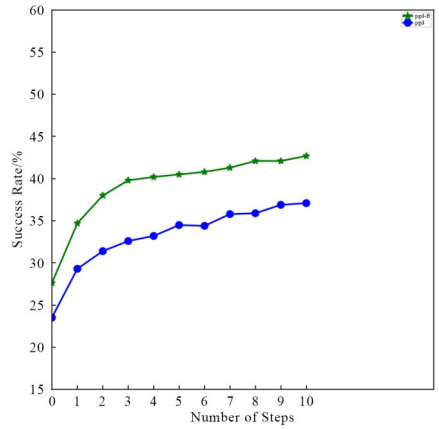


(b) Vit-b模型

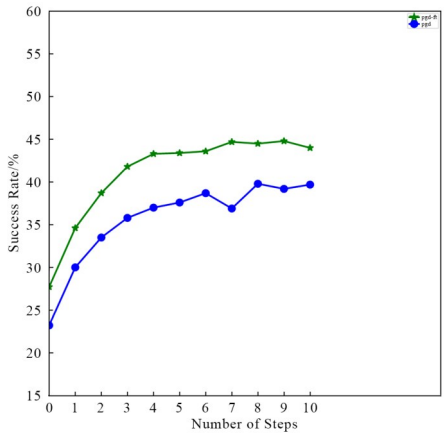
图5　迭代次数对攻击成功率的影响

5　进一步讨论

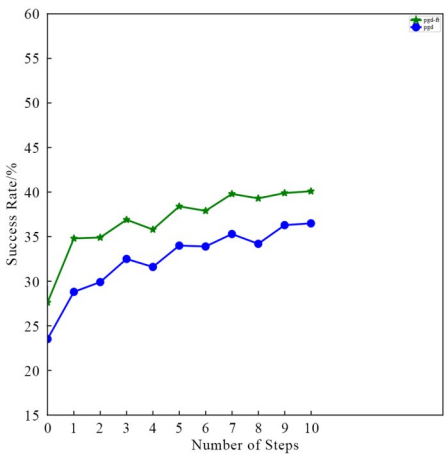
本文针对FT方法(X-FT)与传统方法(X)进行了深入分析,并对两者间的差异进行了补充性探讨. 采用PGD和PGD-FT方法,分别在IncV3、Vit-b、ShuffleNet模型上生成对抗样本,迭代次数从1到10次,然后迁移攻击ResNet模型. 如图6所示,在相同迭代次数下,PGD-FT比PGD具有更高的攻击成功率. 从另一个角度看,PGD-FT需要更少的迭代次数即可达到与PGD相同的攻击成功率. 这些结果不仅表明PGD-FT具有更好的可迁移性,同时也证明凭借动态伪目标对抗攻击,X-FT能够加速对抗样本的生成过程.



(*a*) IncV3模型



(*b*) Vit-b模型



(*c*) ShuffleNet模型

图6 攻击成功率对比

6 结论

本文提出的基于类别相似性驱动的动态伪目标对抗攻击方法(FT)通过动态选择语义相近的伪目标类别,并设计双梯度协同更新机制,不仅继承了无目标攻击的灵活性,还具备了有目标攻击的精准靶向性,在跨模型黑盒攻击中实现了突破性进展. 实验表明,FT方法在CNN、Transformer及轻量化模型上均展现出卓越的迁移鲁棒性,且与数据增强、集成策略兼容性良好. 该方法通过挖掘共享对抗子空间特性,有效穿透不同模型的决策边界,验证了类别相似性驱动策略的理论优势. 此外,FT方法具有广泛的兼容性,可无缝适配多种梯度攻击框架,为黑盒攻击提供了一种高效且通用的优化范式. 未来研究可进一步探索其在复杂防御场景中的应用潜力,推动对抗攻击与模型安全领域的深度结合.

参考文献

[1] LIU J F, LI Y S, GUO Y M, et al. Generation and countermeasures of adversarial examples on vision: A survey[J]. Artificial Intelligence Review, 2024, 57(8): 199-246.

[2] YANG B, ZHANG H W, WANG J D, et al. Adversarial example soups: Improving transferability and stealthiness for free[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1882-1894.

[3] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2015-03-20)[2025-07-01]. https://arxiv.org/pdf/1412.6572.

[4] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2017-06-19)[2025-07-01]. https://arxiv.org/abs/1706.06083?context=cs.

[5] 鲍蕾, 陶蔚, 陶卿. 结合自适应步长策略和数据增强机制提升对抗攻击迁移性[J]. 电子学报, 2024, 52(1): 157-169.
BAO L, TAO W, TAO Q. Boosting adversarial transferability through adaptive-learning-rate with data augmentation mechanism[J]. Acta Electronica Sinica, 2024, 52(1): 157-169. (in Chinese)

[6] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. (2016-02-18) [2025-07-01]. https://arxiv.org/abs/1602.02697v2.

[7] ZOU J H, DUAN Y X, LI B Y, et al. Making adversarial examples more transferable and indistinguishable[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3662-3670.

[8] PENG A J, LIN Z, ZENG H, et al. Boosting transferability of adversarial example via an enhanced Euler's method [C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

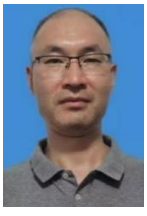
[9] WAN C, HUANG F J. Adversarial attack based on prediction-correction[EB/OL]. (2023-06-02)[2025-07-01]. https://arXiv.org/abs/2306.01809.

[10] WANG J F, CHEN Z Y, JIANG K X, et al. Boosting the transferability of adversarial attacks with global momen-

tum initialization[J]. Expert Systems with Applications, 2024, 255: 124757.
[11] WARDE-FARLEY D, GOODFELLOW I. Adversarial perturbations of deep neural networks[M]//Perturbations, Optimization, and Statistics. Cambridge: The MIT Press, 2016: 311-342.
[12] MEI S B, ZHAO C L, NI B B, et al. Towards interpreting and utilizing symmetry property in adversarial examples[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: Association for the Advancement of Artificial Intelligence, 2023: 9126-9133.
[13] OZBULAK U, PINTOR M, VAN MESSEM A, et al. Evaluating adversarial attacks on ImageNet: A reality check on misclassification classes[EB/OL]. (2021-11-22)[2025-07-01]. https://arXiv.org/abs/2111.11056.
[14] 王硕, 徐茹枝, 关志涛. 基于主特征归因的对抗样本生成方法研究[J]. 电子学报, 2023, 51(11): 3137-3145.
WANG S, XU R Z, GUAN Z T. Research on the generation of adversarial samples based on the attribution of principal features[J]. Acta Electronica Sinica, 2023, 51(11): 3137-3145. (in Chinese)
[15] 吴骥, 邵文泽, 葛琦, 等. 一种基于迭代累积梯度的多层特征重要性攻击方法[J]. 电子学报, 2024, 52(11): 3798-3808.
WU J, SHAO W Z, GE Q, et al. A multi-layer feature importance attack method based on iterative accumulated gradients[J]. Acta Electronica Sinica, 2024, 52(11): 3798-3808. (in Chinese)
[16] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[EB/OL]. (2016-11-04)[2025-07-01]. https://arxiv. org/abs/1611.01236? context=stat.ML.
[17] 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.
[18] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[EB/OL]. (2020-02-03)[2025-07-01]. https://arxiv.org/pdf/1908.06281.
[19] 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. Piscataway: IEEE, 2020: 2725-2734.
[20] 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. Piscataway: IEEE, 2020: 4307-4316.
[21] REBUFFI S A, GOWAL S, DAN C L, et al. Data augmentation can improve robustness[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 29935-29948.
[22] WANG X S, ZHANG Z L, ZHANG J P. Structure invariant transformation for better adversarial transferability[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4584-4596.
[23] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. (2016-11-08)[2025-07-01]. https://arxiv.org/abs/1611.02770.

作者简介

余红霞　女, 1984 年 10 月出生于安徽省铜陵市. 现为中国人民解放军陆军兵种大学讲师. 主要研究方向为机器学习、计算机视觉.
E-mail: genius@uestc.edu.cn



鲁磊纪　男, 1980 年 3 月出生于安徽省宣城市. 现为中国人民解放军陆军兵种大学副教授. 主要研究方向为机器学习、计算机视觉.
E-mail: 452321691@qq.com

鲍　蕾　女, 1987 年 2 月出生于安徽省芜湖市. 现为中国人民解放军陆军兵种大学讲师. 主要研究方向为机器学习、计算机视觉.
E-mail: baolei1219@sina.cn

陈　军　男, 1989 年 7 月出生于安徽省合肥市. 现为中国人民解放军陆军兵种大学讲师. 主要研究方向为机器学习、计算机视觉.
E-mail: Chenjun342423@sina.com

张林俊　男, 1987 年 2 月出生于重庆市. 现为陆军兵种大学硕士研究生. 主要研究方向为机器学习、计算机视觉.
E-mail: 493285432@qq.com