一种基于迭代累积梯度的多层特征重要性攻击方法

吴 骥[1],邵文泽[1*],葛 琦[1],孙玉宝[2]
(1. 南京邮电大学通信与信息工程学院,江苏南京 210003;2. 南京信息工程大学教育部数字取证工程研究中心,江苏南京 210044)

摘 要: 对抗样本的可迁移性对于攻击未知模型至关重要,这在实际场景中为对抗性攻击提供了可行性. 现有的迁移攻击倾向于通过不加选择地扭曲特征来降低源模型的预测精度,但是忽略了图像中目标的内在特征. 受到现有关于提取特征重要性工作的启发,本文提出一种多层累积梯度攻击方法,以破坏主导模型决策的重要目标感知特征. 具体而言,本文通过引入迭代累积梯度来获得特征重要性,这种梯度将与目标主体部分高度相关,从而帮助实现更好的迁移攻击. 进一步,本文在不同中间层进行组合攻击,最终实现了多层累积梯度攻击. 大量结果表明,相较对比实验中的最好方法,本文所提方法在正常训练模型下以更高的攻击效率取得了与之相当的攻击成功率,而在防御模型下的攻击成功率提高了2.6个百分点.

关键词: 对抗攻击;黑盒攻击;迁移性;特征重要性;迭代累积梯度
基金项目: 国家自然科学基金(No.61771250,No.61972213)
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)11-3798-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230843

A Multi-Layer Feature Importance Attack Method Based on Iterative Accumulated Gradients

WU Ji[1], SHAO Wen-ze[1*], GE Qi[1], SUN Yu-bao[2]
(1. *School of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*;
2. *Engineering Research Center for Digital Forensics Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, Jiangsu* 210044, *China*)

Abstract: The transferability of adversarial samples is crucial for attacking unknown models, providing feasibility for adversarial attacks in practical scenarios. Existing transfer attacks tend to indiscriminately distort features to degrade prediction accuracy of the source model. However, they overlook the intrinsic features of objects in the images. Inspired by existing work on feature importance extraction, this paper proposes a method termed multi-layer accumulated gradient attack, which disrupts crucial object-aware features that dominate the model decision. Specifically, this paper introduces the iterative accumulated gradients to quantify feature importance, which are highly correlated with the target object and helpful to improve transfer attacks. Furthermore, combining attacks across various intermediate layers, this paper finally achieves multi-layer accumulated gradient attack. Compared with the best performing method, experimental results demonstrate a more efficient performance of the proposed one, the attacking success rates of which are comparable as to the normally trained models while increased by 2.6 percentage points as to the defense models.

Key words: adversarial attacks; black-box attack; transferability; feature importance; iterative accumulated gradient
Foundation Item(s): National Natural Science Foundation of China (No.61771250, No.61972213)

1 引言

深度神经网络(Deep Neural Networks,DNNs)已经成为许多领域的主流工具,因此它们的脆弱性近年来引起了广泛关注. 一个明显的例子是对抗样本的存在,它们与干净样本非常相似,但能够欺骗DNNs以高置信度产生错误的预测. 一种称为迁移攻击的黑盒攻击方法可由一个已知模型生成的对抗样本成功攻击其他模型. 显然,对抗

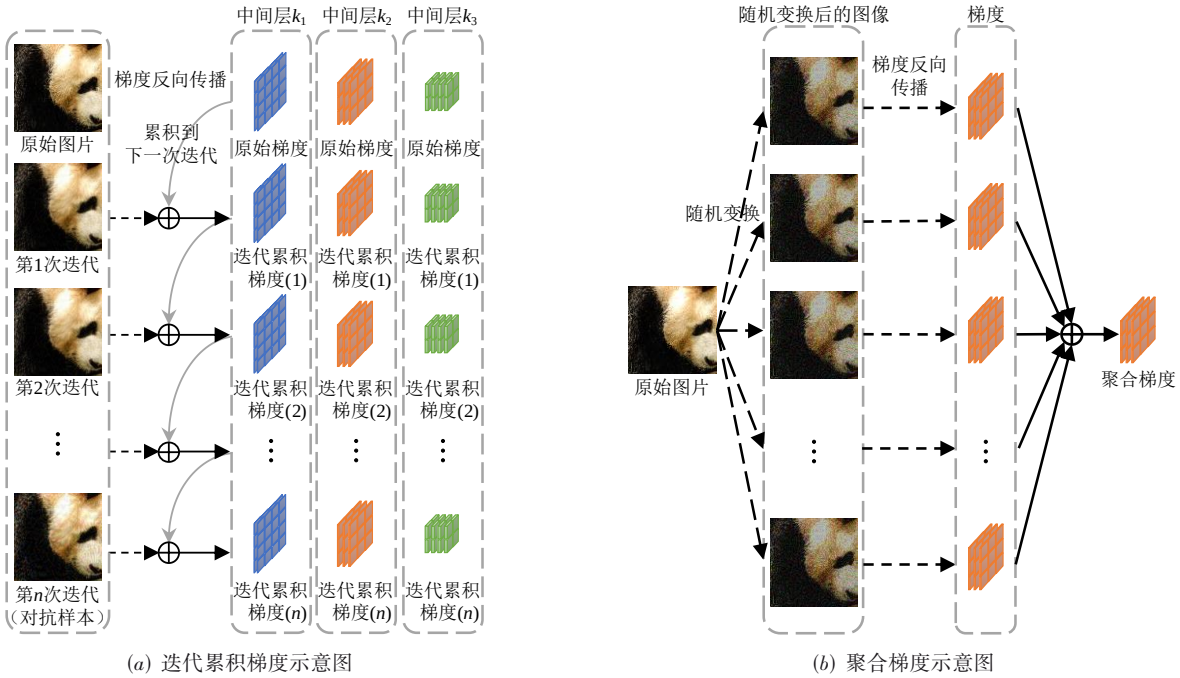
收稿日期:2023-09-07;修回日期:2024-02-03;责任编辑:梅志强
*通讯作者:邵文泽

样本的跨模型迁移性使得迁移攻击方法更加实用和灵活. 传统的攻击方法,如快速梯度符号法(Fast Gradient Sign Method,FGSM)[1]、基本迭代法(Basic Iterative Method,BIM)[2]、动量迭代法(Momentum Iterative Method,MIM)[3]等,所生成的对抗样本往往由于与源模型过拟合而表现出较弱的可迁移性. 一些后期的攻击方法[4,5]直接攻击中间层特征以增强可迁移性. 即,这些特征级攻击方法不是干扰输出层,而是最大化内部特征失真,寻求更高的攻击迁移性. 然而,中间层特征攻击往往是盲目扰动来生成对抗样本,从而容易陷入特定于模型的局部最优解.

为此,一些新近的方法转而针对图像的重要特征进行攻击,这些特征被认为是主导不同深度网络分类的共性因素,可显著提高对抗样本的迁移性. 例如,特征重要性感知攻击方法(Feature Importance-aware Attack,FIA)[6]通过对原始图像采用类似 Dropout 的随机掩码策略得到多幅内容缺失的图像,由多幅图像事先计算聚合梯度实现特征重要性引导的模型攻击. 但是,由于每次攻击需要利用多幅图像计算聚合梯度,FIA 生成对抗样本的速度受到一定影响. 与 FIA 稍有不同,注意力攻击方法(Attack On Attention,AOA)[7]直接攻击 Grad-CAM (Gradient-weighted Class Activation Mapping)[8]、SGLRP(Softmax-Gradient Layer-wise Relevance Propagation)[9]等不同解释方法得到的决策归因图,通过扰乱模型的注意力致使模型分类错误,不过决策归因图的计算会在一定程度上降低对抗攻击的效率. 与 AOA[7]类似,注意力引导转移攻击(Attention-cuided Transfer Attack,ATA)[10]同样利用源模型注意力正则化对抗样本的搜索方向,从而专注于削弱不同模型所依赖的关键特征.

本文探讨了一种新的多层特征重要性攻击方法,即多层累积梯度攻击(Multi-layer Accumulated Gradient Attack,MAGA). 简单而言,我们引入了迭代累积梯度(Iterative Accumulated Gradient,IAG)来表示特征重要性,通过破坏主导模型决策的重要目标感知特征,从而提高对抗样本的可迁移性.

如图 1 所示,不同于 FIA 基于多幅随机变换图像来计算聚合梯度以提取特征重要性的攻击策略,本文利用攻击迭代过程中所生成的对抗样本来提取特征重要性实现多层累积梯度攻击. 具体而言,本文在无目标攻击实验中发现,迭代过程生成的对抗样本会被模型判定为多种不同的类别,从这些迭代对抗样本提取到的关于中间层的梯度,会在保留一定原始目标类别信息的基础上产生关于其他类别信息的跳变. 对这些梯度信息累积和平均所得到的迭代累积梯度,其中特定于其他类别的梯度信息将会相互抑制,这些梯度信息指示的往往是目标不相关或弱相关性区域;而关于原始类别的梯度信息将得以增强,这些梯度信息指示的自然是目标对象的主体区域,或者重要目标感知特征. 因此,将迭代累积梯度作为特征重要性与中间特征图相乘构造损失函数后,即可针对性破坏主导不同神经网络模型决策的重要目标感知特征,生成更具迁移性的对抗样本. 进一步,受到 Layer-CAM[11]中可由卷积神经



(a) 迭代累积梯度示意图　　(b) 聚合梯度示意图

图 1　多层累积梯度攻击(MAGA)与特征重要性感知攻击(FIA)关于提取特征重要性的对比

网络的不同层生成可靠的类激活图的启发，不同层可提取到关于目标的不同粒度的信息，从而本文由不同深度的中间特征层入手，通过选取多个层进行组合攻击，最终得到更具迁移性的对抗样本.

本文的主要贡献概述如下：

(1)分析了现有攻击方法可迁移性相对较低的原因，通过引入迭代累积梯度，破坏主宰不同模型决策的重要目标感知特征来增强对抗样本的可迁移性.

(2)将迭代累积梯度引导的攻击方法扩展到多个特征中间层，最终得到多层累积梯度攻击方法MAGA，进一步提升了攻击效果.

(3)对不同分类模型进行的大量实验表明，与最先进的可迁移攻击方法相比，MAGA生成的对抗样本具有更好的可迁移性，并且具有更高的攻击效率.

2 无目标迁移攻击基础

对抗样本的可迁移性对于攻击未知模型(即黑盒攻击)至关重要，即使用在一个模型(源模型)上构建的对抗样本来攻击其他模型(目标模型). 具体来说，假设一个分类模型$f_\theta: x \to y$，其中x和y分别表示干净的图像和其真实标签，θ表示模型的参数. 攻击的目的是生成一个对抗样本$x^{\text{adv}}=x+\varepsilon$，是$x$被精心设计的扰动$\varepsilon$扭曲所得，以误导分类器，即$f_\theta(x^{\text{adv}})\neq y$. 通常采用$p$范数对扰动进行正则化. 因此，对抗样本的生成可以表述为如下所示的优化问题：

$$\arg\max_{x^{\text{adv}}} J\left(y, f_\theta(x^{\text{adv}})\right), \text{s.t.}\ \left|x-x^{\text{adv}}\right|_p \leqslant \varepsilon \tag{1}$$

其中，损失函数$J(\cdot,\cdot)$度量真实标签与预测标签之间的距离(即交叉熵). 上述优化问题要求显式访问f_θ的参数，而这在黑盒攻击中是不切实际的. 因此，一个可行的解决方案是在一个具有可访问参数ϕ的类似模型f_ϕ(即源模型)下进行优化：

$$\arg\max_{x^{\text{adv}'}} J\left(y, f_\phi(x^{\text{adv}'})\right), \text{s.t.}\ \left|x-x^{\text{adv}'}\right|_p \leqslant \varepsilon \tag{2}$$

从而产生可迁移的对抗样本$x^{\text{adv}'}$来攻击不同的目标模型f_θ，即使得$f_\theta(x^{\text{adv}'})\neq y$.

3 本文方法

深度模型倾向于提取任何有用的特征来最大化分类精度，即使是图像中那些本不可感知的噪声特征，这些噪声特征往往是模型特定的. 现有对抗攻击方法往往在没有意识到这些特征的情况下，通过不加选择地破坏所有感知特征来生成对抗样本，从而容易陷入特定于模型的局部最优解，影响对抗样本的可迁移性. 为此，受特征重要性[6]启发，本文进一步挖掘无目标攻击过程中的重要目标感知特征，以此寻求有助于缓解局部最优的攻击新方法. 图2概述了本文所提出的迭代累积梯度攻击方法(Iterative Accumulated Gradient Attack，IAGA)，给定输入图像，从源模型的中间层(橙色色块)提取特征图后，计算从输出反向传播到中间特征图的梯度信息，将其与之前的迭代累积梯度相加得到新的迭代累积梯度，归一化作为特征重要性；进一步对中间特征图和迭代累积梯度进行元素相乘构造对抗攻击损失函数，将优化生成的对抗样本作为输入进入下一次迭代. 简而言之，利用迭代过程不断生成的扰动噪声，引入迭代累积梯度(在3.1节中详细描述)指示特征重要性，从而能够针对性破坏主导模型决策的重要目标感知特征，以

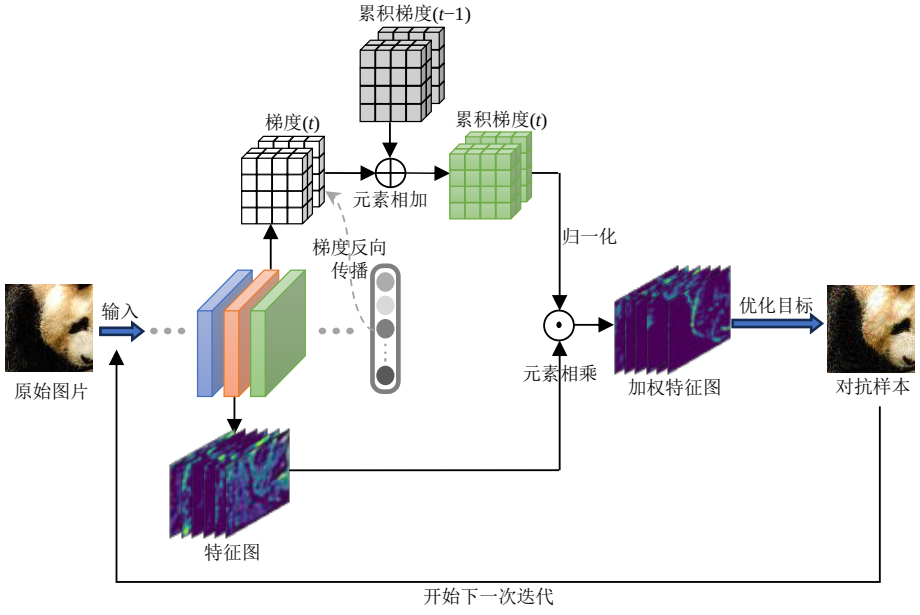


图2 迭代累积梯度攻击概述

期有效缓解局部最优. 进一步地,本文在多个中间特征层引入迭代累积梯度(在3.3节中详细描述),利用不同中间层之间信息的互补性完善了损失函数,最终提高对抗样本的迁移性.

3.1　迭代累积梯度

为了简单起见,设f表示源模型,第k层的特征图表示为$f_k(x)$. 由于特征重要性与特征对最终决策的贡献成正比,一个直观的策略是获得关于$f_k(x)$的梯度如下所述:

$$\boldsymbol{\Delta}_k^x=\frac{\partial l(x,t)}{\partial f_k(x)} \tag{3}$$

其中,$l(\cdot,\cdot)$表示对于真标签c的logit输出. 然而,原始梯度$\boldsymbol{\Delta}_k^x$将携带特定于源模型的信息,如图3所示,原始梯度图虽然在目标主体部分呈现出明显的语义信息,但在部分非目标区域也存在较强的视觉"噪声". 这种存在于不同模型、目标弱相关的"噪声",往往是深度学习模型在决策空间上的次优求解造成的.

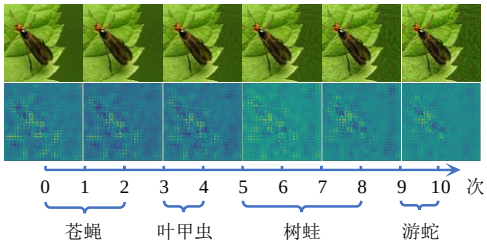


图3　不同迭代次数下对抗样本的迭代累积梯度可视化(0代表原始样本的梯度可视化)

为了抑制这些"噪声"特征,本文提出了迭代累积梯度. 对于无目标攻击而言,只需对抗样本被预测的类别与原始类别不同即可,而对抗样本最终属于哪一类不做限制. 通过实验发现,在无目标攻击的迭代过程中,不同迭代次数下的对抗样本会被源模型判定为不同的类别. 如图3所示,一个类别为苍蝇的原始图片,在经过无目标攻击的几次迭代后生成的对抗样本会被源模型识别为叶甲虫、树蛙或游蛇. 因此,从这些迭代对抗样本提取到的关于中间特征层的梯度,会在保留一定原始目标类别信息的基础上产生关于不同类别信息的跳变. 当对这些梯度信息累积和平均后,特定于不同类别的梯度信息将会相互抑制,这些梯度所指示的感知特征往往是目标不相关或弱相关的非鲁棒特征,也即那些特定于模型的"噪声"特征;而关于原始类别的梯度信息将会增强,其对应的关于目标对象的鲁棒特征也得以被指示出来.

具体来说,本文从第t次迭代后生成的图像x^t上获得关于第k层特征图$f_k(x)$的梯度$\boldsymbol{\Delta}_k$,每次迭代后都将新得到的梯度$\boldsymbol{\Delta}_k$和之前的迭代累积梯度$\boldsymbol{\Delta}_k^{t-1}$进行累积得到新的迭代累积梯度$\boldsymbol{\Delta}_k^t$,如式(4)所示:

$$\boldsymbol{\Delta}_k^t=\boldsymbol{\Delta}_k^{t-1}+\boldsymbol{\Delta}_k \tag{4}$$

那么将第t次迭代时的迭代累积梯度归一化后可以表示为

$$\bar{\boldsymbol{\Delta}}_k^t=\frac{\boldsymbol{\Delta}_k^t}{\left\|\boldsymbol{\Delta}_k^t\right\|_2} \tag{5}$$

图3所示的可视化迭代累积梯度"$\bar{\boldsymbol{\Delta}}_k^t$",随着迭代的进行,迭代累积梯度逐渐聚焦于感知目标,与原始梯度相比可以更为鲁棒地提取目标的主体信息. 由于迭代累积梯度在目标相关区域产生相对较高的强度,以迭代累积梯度作为特征重要性的对抗样本生成有望实现更好的可迁移性.

3.2　攻击算法

生成迁移性对抗样本的宗旨是抑制具有正向作用的重要特征,而增强对应于负向作用的非重要特征. 根据3.1节,基于迭代累积梯度定义的特征重要性,迭代累积梯度攻击方法(IAGA)即可通过显式地抑制重要特征来指导对抗样本x_{adv}的生成,对应的损失函数表示如下:

$$\mathcal{L}_k\left(x_{\text{adv}}\right)=\sum\left(\bar{\boldsymbol{\Delta}}_k^t\odot f_k\left(x_{\text{adv}}\right)\right) \tag{6}$$

也即,最小化损失函数式(6)将破坏主导模型决策的重要目标感知特征. 将上式代入约束优化问题,最终得到迭代累积梯度攻击(IAGA)的目标函数,表示如下:

$$\arg\min_{x_{\text{adv}}}\mathcal{L}_k\left(x_{\text{adv}}\right),\text{s.t.}\left|x-x_{\text{adv}}\right|_\infty\leqslant\varepsilon \tag{7}$$

目前文献报道了多种基于梯度的攻击算法,如BIM[2]、MIM[3]等. 考虑到引入动量的MIM算法的优越性能,本文采用MIM求解式(7),具体如算法1所示.

如图4所示,针对Vgg-16图像分类模型,算法1生成了两组具体的对抗样本示例. 图4左侧为原始图片及其热力图,右侧为对抗样本及其热力图. 不难发现,分类模型被对抗样本有效扰乱了判断的焦点,不能聚焦于目标相关的重要区域,而是错误地关注到那些目标不相关或弱相关性区域.

3.3　多层攻击

Layer-CAM可以为卷积神经网络的不同层生成可

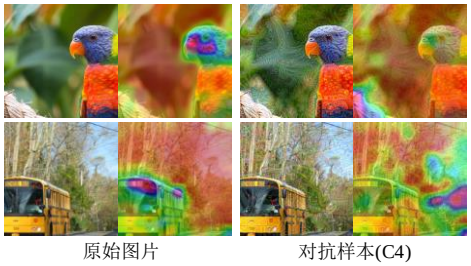


图4　原始图片和对抗样本的热力图对比(对抗样本在Vgg-16模型的Conv4层上生成)

算法1 迭代累积梯度攻击算法

输入: 原始干净图像x,真实标签c,分类模型f,中间层k,最大扰动ε,迭代次数T

输出: 对抗图像x_{adv}

初始化:$\boldsymbol{\Delta}_k^0=0, g_0=0, \mu=1, \alpha=\varepsilon/T, x_{\text{adv}}=x$

FOR t=0 TO T−1 DO

$$\boldsymbol{\Delta}_k=\frac{\partial l\left(x_{\text{adv}}^t,c\right)}{\partial f_k\left(x_{\text{adv}}^t\right)}$$

$$\boldsymbol{\Delta}_k^t=\boldsymbol{\Delta}_k^{t-1}+\boldsymbol{\Delta}_k$$

$$\overline{\boldsymbol{\Delta}}_k^t=\frac{\boldsymbol{\Delta}_k^t}{\left\|\boldsymbol{\Delta}_k^t\right\|_2}$$

构造第t次迭代优化目标:

$$\mathcal{L}_k\left(x_{\text{adv}}^t\right)=\sum\left(\overline{\boldsymbol{\Delta}}_k^t\odot f_k\left(x_{\text{adv}}^t\right)\right)$$

通过动量迭代方法更新x_{adv}^{t+1}:

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla_x\mathcal{L}_k\left(x_{\text{adv}}^t\right)}{\left|\nabla_x\mathcal{L}_k\left(x_{\text{adv}}^t\right)\right|_1}$$

$$x_{\text{adv}}^{t+1}=\text{Clip}_{x,\varepsilon}\left\{x_{\text{adv}}^t-\alpha\cdot\text{sign}\left(g_{t+1}\right)\right\}$$

END FOR

RETURN x_{adv}

靠的类激活图,不但可以从最终卷积层生成的类激活图获得粗略的空间位置,还可以从浅层生成的类激活图获得细粒度的对象细节,如图5所示,对抗样本采用单个层攻击和多个层组合攻击. C1、C2、C3、C4、C5分别表示Vgg-16模型的Conv1_2、Conv2_2、Conv3_3、Conv4_3、Conv5_3层,可以发现来自不同层的信息通常是互补的.

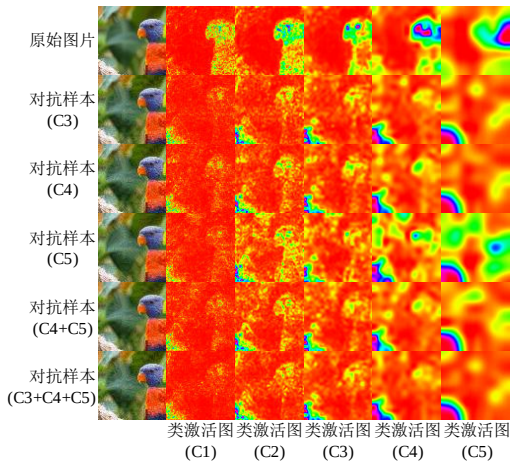


图5 原始图片和对抗样本在Vgg-16不同阶段下基于Layer- CAM从浅层到深层的类激活图展示

类似地,本文同样从不同层k_i中获得迭代累积梯度$\overline{\boldsymbol{\Delta}_{k_i}^t}$,分别计算损失函数$\mathcal{L}_{k_i}\left(x_{\text{adv}}\right)$,对于所攻击的中间层的选择,本文对每个模型的不同中间层进行了大量实验,经验性地挑选性能较好的层进行组合作为最后的攻击层,具体中间层的选择和细节将在4.2节和4.3节中介绍,最终得到多层累积梯度攻击(MAGA)的损失函数:

$$\mathcal{L}\left(x_{\text{adv}}\right)=\sum\mathcal{L}_{k_i}\left(x_{\text{adv}}\right),k_i\in K \tag{8}$$

其中,K为所选择中间层的集合. 如图5所示,相较单层攻击产生的对抗样本,多层攻击产生的对抗样本更能将模型的关注点从目标主体位置上转移,进一步散焦了热力图,从而可以获得更好的迁移性.

4 实验与分析

4.1 实验参数设置

实验采用NIPS 2017年对抗攻击竞赛的ImageNet-compatible Dataset作为数据集,其中包括1 000张图像. 将MAGA在12个最先进的分类模型上进行验证,包括7个正常训练模型和5个对抗训练模型(即防御模型). 正常训练的模型有Inception-V3(Inc-v3)[12]、Inception-V4(Inc-v4)[13]、Inception-ResNet-V2(IncRes-v2)[13]、ResNet-V1-50(Res-50)[14]、ResNet-V1-152(Res-152)[14]、VGG16(Vgg-16)[15]和VGG19(Vgg-19)[15]. 防御模型[16,17]有Adv-Inc-v3、Adv-IncRes-v2、Ens3-Inc-v3、Ens4-Inc-v3和Ens-IncRes-v2.

为了证明MAGA的有效性,本文将其与各种先进的攻击方法进行了比较,包括MIM[3]、DIM(Diverse Inputs Method)[18]、TIM(Translation Invariant Method)[19]、PIM(Patch-wise Iterative Method)[20]以及这些方法的组合版本TIDIM[19]、PIDIM[20]. 其中,DIM的变换概率为0.7,TIM的核尺寸为15,由于PIM的设置会随着目标模型和方法组合方式的不同而变化,下面将在每个相关实验中具体详细地说明其设置(即放大因子β、投射因子γ、投射核大小k_w). 此外,还将其他特征级攻击方法加入比较,如FDA(Feature Disruptive Attack)[4]、NRDM(Neural Representation Distortion Method)[5]、FIA[6]. 对于特征级攻击方法,下面选择相同层进行实验,即Inc-v3中的Mixed_5b层,Vgg-16中的Conv3_3层,Vgg-19中的Conv3_4层,Res-50中第一个block的最后一层,Res-152中第二个block的最后一层. 在所有实验中最大扰动ε=16,迭代次数T=10,步长$\alpha=\varepsilon/T$=1.6.

4.2 迁移性攻击实验

为了定量比较MAGA与其他攻击方法之间的迁移性,分别选择Vgg-16、Vgg-19、Inc-v3、Res-50、Res-152作为源模型,攻击其他正常训练的模型(表1)和防御模型(表2). 请注意,表1中没有包括TIM,因为它是为防御模型设计的.

表1 不同攻击方式对正常训练模型的攻击成功率 单位:%

源模型	攻击算法	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	Vgg-16	Vgg-19
Res-50	MIM[3]	63.7	54.3	51.6	99.8*	89.4	74.0	70.6
	DIM[18]	79.9	72.6	69.9	99.6*	94.2	86.3	85.0
	PIM[20]	67.0	53.5	49.9	**100.0***	90.0	82.0	81.2
	PIDIM[20]	82.9	77.0	77.3	99.6*	95.4	91.8	89.4
	NRDM[5]	72.8	67.3	58.6	97.0*	82.2	77.7	79.0
	FDA[4]	56.5	50.5	48.3	93.8*	70.6	70.5	67.7
	FIA[6]	86.5	81.2	77.4	99.8*	96.9	88.4	91.3
	FIA+PIDIM[6]	92.8	88.6	86.5	99.6*	97.5	96.3	95.5
	MAGA	85.0	79.2	77.8	**100.0***	96.5	90.3	89.6
	MAGA+PIDIM	**92.9**	**89.6**	**89.8**	99.7*	**98.4**	**96.5**	**95.9**
Res-152	MIM	57.0	48.2	45.8	90.7	99.8*	72.9	73.0
	DIM	80.4	72.1	72.4	95.1	99.6*	88.3	87.8
	PIM	65.7	56.5	50.8	92.2	99.8*	83.1	82.4
	PIDIM	82.0	76.7	76.7	95.9	99.7*	90.9	90.0
	NRDM	64.5	58.8	51.2	87.7	95.1*	79.4	79.1
	FDA	60.4	52.4	47.8	85.1	95.1*	74.7	75.0
	FIA	85.3	81.1	77.8	96.8	99.5*	90.2	90.2
	FIA+PIDIM	89.7	86.1	85.6	**97.7**	99.6*	**95.7**	94.5
	MAGA	82.0	76.4	75.9	96.6	**99.9***	87.7	86.4
	MAGA+PIDIM	**90.7**	**86.4**	**86.9**	**97.7**	99.8*	94.8	**94.7**
Inc-v3	MIM	**100.0***	41.4	38.9	33.1	29.6	38.5	38.2
	DIM	99.3*	64.3	59.4	40.8	36.1	47.4	46.5
	PIM	97.7*	55.8	51.4	53.2	46.4	61.6	60.5
	PIDIM	97.8*	70.4	66.5	61.7	56.0	57.6	56.1
	NRDM	98.3*	67.8	59.8	47.6	37.2	49.8	50.5
	FDA	98.8*	71.6	66.2	48.3	37.2	51.9	52.8
	FIA	98.3*	83.5	80.6	70.4	64.9	71.4	73.3
	FIA+PIDIM	98.7*	87.8	85.6	**79.6**	74.6	**82.2**	**83.5**
	MAGA	99.8*	72.3	72.2	60.6	53.6	60.1	55.2
	MAGA+PIDIM	99.3*	**88.9**	**87.3**	79.3	**75.4**	80.3	78.8
Vgg-16	MIM	80.2	81.1	74.3	88.1	83.7	99.9*	96.6
	DIM	87.0	86.9	80.8	91.6	87.5	99.6*	98.3
	PIM	83.9	81.8	75.6	90.0	85.6	99.7*	98.6
	PIDIM	89.2	88.6	84.4	93.9	90.5	99.8*	98.9
	NRDM	73.4	72.6	56.9	77.6	73.0	92.9*	90.9
	FDA	76.1	76.4	63.8	80.2	78.1	94.4*	94.9
	FIA	95.7	95.6	92.3	96.9	94.5	99.8*	99.5
	FIA+PIDIM	96.6	**97.5**	93.4	97.9	95.6	**100.0***	99.6
	MAGA	95.3	94.7	91.1	96.2	94.8	**100.0***	99.7
	MAGA+PIDIM	**96.9**	97.1	**94.3**	**98.2**	**97.2**	**100.0***	**99.9**
Vgg-19	MIM	81.3	82.1	76.4	87.0	84.1	97.0	**100.0***
	DIM	87.5	86.5	81.0	90.3	88.1	98.7	99.9*
	PIM	84.0	82.3	73.7	89.5	85.3	98.8	**100.0***
	PIDIM	90.7	89.6	85.2	93.5	91.1	99.6	99.9*
	NRDM	77.0	74.3	60.5	77.0	73.5	92.0	92.6*
	FDA	79.8	78.7	68.1	81.5	80.3	95.2	96.2*
	FIA	94.2	95.0	91.4	95.2	93.3	99.5	99.8*

续表

源模型	攻击算法	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	Vgg-16	Vgg-19
Vgg-19	FIA+PIDIM	95.7	96.1	92.5	96.4	95.6	99.7	**100.0***
	MAGA	95.2	95.2	91.8	95.9	94.8	**99.8**	99.9*
	MAGA+PIDIM	**96.8**	**96.3**	**94.3**	**97.8**	**97.0**	**99.8**	**100.0***

注:第一列为源模型,第一行列出目标模型."*"表示白盒攻击,因为目标模型是源模型.最好的结果用粗体突出显示,第二好的结果用下划线标识.

表2 不同攻击方式对防御模型的攻击成功率 单位:%

源模型	攻击算法	Adv-Inc-v3	Adv-IncRes-v2	Ens3-Inc-v3	Ens4-Inc-v3	Ens-IncRes-v2
Res-50	MIM[3]	40.5	37.0	40.3	41.4	27.0
	DIM[18]	60.6	56.3	58.7	56.5	40.4
	TIM[19]	42.5	37.1	45.0	44.7	36.8
	PIM[20]	46.3	42.7	48.8	49.6	36.8
	TIDIM[19]	54.0	48.4	54.8	57.5	45.5
	PIDIM[20]	59.7	53.3	57.2	58.5	44.0
	NRDM[5]	51.2	37.9	44.8	41.7	32.0
	FDA[4]	36.9	25.6	32.1	29.0	19.0
	FIA[6]	72.2	63.3	65.4	63.4	48.1
	FIA+PIDIM[6]	84.6	75.9	77.4	75.0	63.7
	MAGA	72.9	65.2	67.3	65.3	52.7
	MAGA+PIDIM	**85.3**	**83.0**	**82.1**	**81.6**	**73.1**
Res-152	MIM	36.8	34.6	36.1	37.3	22.0
	DIM	54.2	54.7	53.3	50.3	33.5
	TIM	41.3	37.3	43.2	47.5	34.1
	PIM	40.4	39.0	47.0	51.8	38.6
	TIDIM	52.4	48.3	57.4	60.8	46.4
	PIDIM	61.1	55.2	60.5	59.9	49.9
	NRDM	55.1	37.9	43.8	42.5	34.2
	FDA	49.0	31.5	38.5	39.2	25.4
	FIA	70.1	66.7	61.4	60.3	41.7
	FIA+PIDIM	**86.0**	82.3	80.9	80.1	69.5
	MAGA	71.9	67.5	67.4	69.1	58.4
	MAGA+PIDIM	84.9	**82.5**	**81.6**	**80.4**	**73.8**
Inc-v3	MIM	22.9	17.3	15.5	15.8	7.6
	DIM	25.7	24.2	17.8	20.7	10.0
	TIM	32.1	26.5	29.8	32.4	22.4
	PIM	34.3	30.3	33.1	38.4	25.9
	TIDIM	40.6	36.9	40.6	42.3	**30.4**
	PIDIM	36.1	28.8	25.6	28.3	15.0
	NRDM	27.0	18.8	9.3	11.5	5.5
	FDA	19.8	12.9	9.1	12.4	5.1
	FIA	54.5	54.9	**43.9**	42.0	23.5
	FIA+PIDIM	**58.6**	55.5	37.1	37.6	21.0
	MAGA	40.7	38.8	33.9	32.8	17.3
	MAGA+PIDIM	53.4	**56.1**	42.3	**43.5**	24.5
Vgg-16	MIM	64.1	60.8	64.1	64.3	44.7
	DIM	70.0	66.2	70.2	67.6	49.8
	TIM	52.6	46.2	55.2	55.4	41.7
	PIM	51.8	42.9	50.1	56.3	39.7

续表

源模型	攻击算法	Adv-Inc-v3	Adv-IncRes-v2	Ens3-Inc-v3	Ens4-Inc-v3	Ens-IncRes-v2
Vgg-16	TIDIM	59.2	47.9	59.3	60.4	47.7
	PIDIM	74.7	65.6	73.7	72.6	58.2
	NRDM	67.2	57.7	65.2	66.2	55.3
	FDA	70.8	58.5	65.1	66.7	55.8
	FIA	87.8	86.3	85.6	86.0	70.8
	FIA+PIDIM	**95.3**	90.9	92.9	91.0	86.4
	MAGA	90.6	86.7	89.3	88.9	81.9
	MAGA+PIDIM	94.9	**92.8**	**94.2**	**92.9**	**87.8**
Vgg-19	MIM	70.2	64.8	69.1	69.4	57.5
	DIM	72.0	71.5	74.9	73.7	60.7
	TIM	54.9	48.4	56.0	59.9	46.5
	PIM	62.7	55.5	59.7	70.4	53.9
	TIDIM	61.3	50.3	61.2	60.0	45.8
	PIDIM	72.0	63.5	69.2	70.9	57.0
	NRDM	66.0	57.6	64.2	65.0	53.4
	FDA	75.2	62.3	70.8	69.3	57.7
	FIA	92.0	84.9	88.4	87.2	80.5
	FIA+PIDIM	92.5	86.7	89.7	90.0	82.1
	MAGA	91.7	86.4	89.3	87.6	81.5
	MAGA+PIDIM	**94.0**	**91.0**	**93.2**	**92.7**	**86.7**

注:第一列显示源模型,第一行列出目标模型.最好的结果用粗体突出显示,第二好的结果用下划线标识.

4.2.1　对于正常训练模型的攻击实验

本文遵循文献[20]中的实验设置,PIM 中 β=10,γ=16,PIDIM 中 β=2.5,γ=2,两种方法的投射核大小 k_w=3.对于 MAGA 攻击的层选择如下:Vgg-16 中的 Conv3_3,Conv4_3,Conv5_3;Vgg-19 中的 Conv3_4,Conv4_4,Conv5_4;Inc-v3 中的 Conv2d_4a,Mixed_5b,Mixed_5c,Mixed_6a,Mixed_6b;Res-50 中 block1 和 block2 的最后一层;Res-152 中 block1 和 block2 的最后一层.

如表 1 所示,在迁移攻击中,MAGA 方法明显优于其他方法,特别是在以 Vgg-16 和 Vgg-19 为源模型时,攻击成功率始终在 90% 以上.MAGA 方法也可以很容易地应用于其他方法进一步提高迁移性,如 MAGA+PIDIM 是 MAGA 和 PIDIM 的组合(β=2.5,γ=2 和 k_w=3).可以发现往往 MAGA 与 PIDIM 结合才能得到最佳攻击效果.这是因为 PIDIM 一定程度上采用了数据增强的方式来提高对抗样本的可迁移性[21].FIA 方法通过对原始图片引入随机掩码得到多张随机变换的图片,从而计算出更为鲁棒的特征重要性来引导攻击,其本质也是通过数据增强的方式得到更好的重要性定位.而本文方法则并未采用数据增强的方式,因此与 PIDIM 方法相结合后的方法相比 FIA+PIDIM 可以得到更佳的攻击效果.

平均而言,本文的方法相较其他方法的攻击成功率提高了 9.8 个百分点,且由于每次生成对抗样本不需要对多张图片进行计算,MAGA 相比 FIA 有很大的速度优势.本文使用 NVIDIA GeForce GTX 1080Ti 显卡以 Vgg-16 为源模型进行测试时,FIA 生成 1 000 张对抗样本所需时间为 767 s,而 MAGA 用时为 366 s.

4.2.2　对于防御模型的攻击实验

由于防御模型是经过对抗性训练的,因此对对抗样本表现出很强的鲁棒性[22].在 PIM 中,本文遵循文献[20]的建议,去掉动量项,因为它可能会影响攻击防御模型的性能.PIM 及其组合的设置为 β=10,γ=16,k_w=7.对于 MAGA 攻击的层选择如下:Vgg-16 中的 Conv1_2,Conv2_2,Conv3_3,Conv4_3,Conv5_3;Vgg-19 中的 Conv1_2,Conv2_2,Conv3_4,Conv4_4,Conv5_4;Inc-V3 中的 Conv2d_4a,Mixed_5b,Mixed_5c,Mixed_6a,Mixed_6b;Res-50 中 block1 和 block2 的最后一层;Res-152 中 block1 和 block2 的最后一层.

如表 2 所示,MAGA 和相应的 MAGA+PIDIM(β=2.5,γ=2 和 k_w=7)的组合版本优于或可媲美其他传统攻击方法.平均而言,本文方法相较其他方法的攻击成功率提高了 18.7 个百分点.值得指出的是,实验结果显示,本文方法以 Inception-V3 为源模型的攻击效果相较于其他源模型有一定的性能差距,而多数对比方法在以 Inception-V3 为源模型进行攻击时同样出现了明显的性能下降,这在很大程度上印证了文献[6]中对于

模型复杂度和迁移性之间关联的结论. 也即,不太复杂的模型往往产生更具可迁移性的对抗样本(前提是模型应达到相近的分类精度),如当使用Vgg网络作为源模型时取得了较高的攻击成功率;而使用模型架构较为复杂的Inception网络时,则更难避免陷入模型的局部最优. 此外,经过对Inception-V3的攻击层选择调优后,尽管MAGA相较于FIA还有一定劣势,但组合方法MAGA+PIDIM最终与FIA+PIDIM取得了相近的攻击性能.

4.3 攻击层的选择

本文尝试对多个中间层进行组合攻击,以Vgg-16网络为例,本文首先对网络中单个层分别使用迭代累积梯度攻击方式进行攻击测试,如表3所示,C1~C5表示Vgg-16中每个阶段的最后一个卷积层.

由于模型中的浅层所提取到的细粒度特征通常是数据特定的,往往只学到了较少的有关类别的语义特征,而更深的层可能会处理这些提取到的特征以最大化模型的分类精度,使特征变得模型特定[23]. 如表3所示,中间层具有良好的类分离表示,并且它们与模型架构的相关性不高,可以在迁移攻击中取得较好的效果,而不同中间层所提取的特征信息可能存在一定的互补. 基于这样的结论,本文从中间层出发,对多个层进行组合,选择最好的组合方式作为最终的攻击层. 对于其他网络模型,本文也按照上述的方法进行了大量实验以选择最合适的攻击层的组合方式.

表3 Vgg-16模型下采用单个层攻击和多个层组合攻击的攻击成功率 单位:%

攻击层	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	Vgg-16	Vgg-19
C1	88.5	87.1	81.4	91.6	86.7	**100.0**	99.2
C2	92.4	91.7	86.5	94.7	91.2	**100.0**	99.4
C3	94.3	94.1	90.6	96.0	93.7	**100.0**	**99.8**
C4	94.0	94.5	90.0	96.0	93.2	99.9	99.7
C5	90.7	90.8	84.2	93.5	90.3	99.8	99.5
C3+C4	95.1	94.4	**91.1**	96.1	94.2	**100.0**	**99.8**
C2+C3+C4	95.1	93.7	90.4	95.9	94.0	**100.0**	**99.8**
C3+C4+C5	**95.3**	**94.7**	91.0	**96.2**	**94.8**	**100.0**	99.7
C2+C3+C4+C5	95.2	93.8	90.3	96.1	94.2	**100.0**	**99.8**
C1+C2+C3+C4+C5	94.8	93.8	90.4	96.0	94.0	**100.0**	**99.8**

注:最好的结果用粗体突出显示,第二好的结果用下划线标识.

4.4 消融实验

MAGA的关键是迭代累积梯度,它显著地提高了对抗样本的迁移性. 为了突出迭代累积梯度的贡献,下面进行消融研究去比较有迭代累积梯度和无迭代累积梯度的攻击性能,并有针对性地构造了如式(9)~(11)所示的三个损失函数,其中$\mathcal{L}_{\text{normal}}$像大多数一般方法一样在没有约束的情况下优化特征扰动,$\mathcal{L}_{\text{clean}}$使用干净的梯度$\boldsymbol{\Delta}_{\text{clean}}$,即从原始的干净图像下获得的梯度,$\mathcal{L}_{\text{IAGA}}$则为本文提出方法,使用了迭代累积梯度,本文在Vgg-16模型中的Conv3_3层使用三种目标函数分别进行了攻击实验.

表4分别显示了使用三种损失的攻击成功率. 在所有情况下,$\mathcal{L}_{\text{IAGA}}$的性能都大幅度超过其他损失函数,表明MAGA方法所提出的迭代累积梯度是有效的.

$$\mathcal{L}_{\text{normal}}=\sum f_k(x) \tag{9}$$

$$\mathcal{L}_{\text{clean}}=\sum(\boldsymbol{\Delta}_{\text{clean}}\odot f_k(x)) \tag{10}$$

$$\mathcal{L}_{\text{IAGA}}=\sum\left(\bar{\boldsymbol{\Delta}}_k^t\odot f_k\left(x_{\text{adv}}\right)\right) \tag{11}$$

表4 迭代累积梯度对攻击成功率的影响 单位:%

目标函数	Inc-v3	Inc-v4	IncRes-v2	Res-50	Res-152	Vgg-16	Vgg-19
$\mathcal{L}_{\text{normal}}$	53.5	52.7	39.0	59.9	53.4	75.5	73.9
$\mathcal{L}_{\text{clean}}$	91.8	91.1	86.6	93.2	90.6	99.9	99.0
$\mathcal{L}_{\text{IAGA}}$	**95.3**	**94.7**	**91.1**	**96.2**	**94.8**	**100.0**	**99.8**

注:$\mathcal{L}_{\text{normal}}$在没有梯度指导的情况下优化特征扰动,$\mathcal{L}_{\text{clean}}$使用原始梯度,$\mathcal{L}_{\text{IAGA}}$采用迭代累积梯度.最好的结果用粗体突出显示,第二好的结果用下划线标识.

5 结论

本文提出一种称为多层累积梯度攻击(MAGA)的迁移性黑盒攻击新方法. 该方法通过引入迭代累积梯度来指示跨分类模型的特征重要性,以此为引导有效破坏主导模型决策的重要目标感知特征. 进一步地,将上述攻击方式扩展到多层,最终实现了多层累积梯度攻击以获得更好的迁移攻击效果. 与其他文献中先进的攻击方法所进行的大量对比实验表明,MAGA生成的对抗样本普遍具有更好的迁移性,MAGA方法可作为评估各种模型鲁棒性的基准方法.

参考文献

[1] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[EB/OL]. (2015-05-20) [2023-05-01]. http://arxiv. org/abs/1412.6572v3.

[2] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Examples in the Physical World[EB/OL]. (2017-02-11) [2023-05-01]. http://arxiv.org/abs/1607.02533v4.

[3] 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.

[4] GANESHAN A, VIVEK B S, RADHAKRISHNAN V B. FDA: Feature disruptive attack[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 8068-8078.

[5] NASEER M, KHAN S H, RAHMAN S, et al. Task-generalizable adversarial attack based on perceptual metric[EB/OL]. (2019-03-26) [2023-05-01]. http://arxiv. org/abs/1811.09020v3.

[6] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 7619-7628.

[7] CHEN S Z, HE Z B, SUN C J, et al. Universal adversarial attack on attention and the resulting dataset DAmageNet[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2188-2197.

[8] IWANA B K, KUROKI R, UCHIDA S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 4176-4185.

[9] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[10] WU W B, SU Y X, CHEN X X, et al. Boosting the transferability of adversarial samples via attention[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1158-1167.

[11] JIANG P T, ZHANG C B, HOU Q B, et al. LayerCAM: Exploring hierarchical class activation maps for localization[J]. IEEE Transactions on Image Processing, 2021, 30: 5875-5888.

[12] 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.

[13] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 4278-4284.

[14] 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.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-05-01]. https://arXiv.org/abs/1409.1556v6.

[16] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[EB/OL]. (2017-02-11) [2023-05-01]. http://arxiv.org/abs/1611.01236v2.

[17] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training:Attacks and defenses[EB/OL]. (2020-04-26)[2023-05-01]. http://arxiv.org/abs/1705.0720 4v5.

[18] 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.

[19] 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.

[20] GAO L L, ZHANG Q L, SONG J K, et al. Patch-wise attack for fooling deep neural network[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 307-322.

[21] 邹军华, 段晔鑫, 任传伦, 等. 基于噪声初始化、Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法[J]. 电子学报, 2022, 50(1): 207-216.
ZOU J H, DUAN Y X, REN C L, et al. Perturbation initialization, adam-nesterov and quasi-hyperbolic momentum for adversarial Examples[J]. Acta Electronica Sinica, 2022, 50(1): 207-216. (in Chinese)

[22] 张世辉, 张晓微, 宋丹丹, 等. 基于逆扰动融合生成对抗网络的对抗样本防御方法[J]. 电子学报, 2023, 51(4): 879-884.
ZHANG S H, ZHANG X W, SONG D D, et al. Adversarial example defense method based on inverse perturbation fusing generative adversarial network[J]. Acta Electronica Sinica, 2023, 51(4): 879-884. (in Chinese)

[23] INKAWHICH N, LIANG K J, CARIN L, et al. Transferable perturbations of deep feature distributions[EB/OL]. (2020-04-27) [2023-05-01]. https://arxiv. org/abs/2004.12519.

作者简介

吴 骥 男, 1999 年 4 月出生于河北省唐山市. 现为南京邮电大学硕士研究生. 主要研究方向为神经网络可解释, 对抗攻击.
E-mail: wj233enter@163.com

邵文泽 男, 1981 年 10 月出生于江苏省连云港市. 现为南京邮电大学通信与信息工程学院教授. 主要研究方向为计算成像反问题, 神经拟态计算, 可信人工智能.
E-mail: shaowenze@njupt.edu.cn

葛 琦 女, 1984 年 1 月出生于江苏省南通市. 现为南京邮电大学通信与信息工程学院副教授. 主要研究方向为计算机视觉, 张量分解, 深度学习.
E-mail: geq@njupt.edu.cn

孙玉宝 男, 1983 年 5 月出生于江苏省连云港市. 现为南京信息工程大学计算机学院教授. 主要研究方向为深度学习, 模式识别, 高光谱遥感影像处理与分析.
E-mail: sunyb@nuist.edu.cn