

基于平移随机变换的对抗样本生成方法

李哲铭^{1,2}, 张恒巍¹, 马军强¹, 王晋东¹, 杨 博¹

(1. 中国人民解放军战略支援部队信息工程大学 密码工程学院, 郑州 450001; 2. 中国人民解放军陆军参谋部, 北京 100000)

摘 要: 基于深度神经网络的图像分类模型能够达到甚至高于人眼的识别度识别图像,但是因模型自身结构的脆弱性,导致其容易受对抗样本的攻击。现有的对抗样本生成方法具有较高的白盒攻击率,而在黑盒条件下对抗样本的攻击成功率较低。将数据增强技术引入到对抗样本生成过程中,提出基于平移随机变换的对抗样本生成方法。通过构建概率模型对原始图像进行随机平移变换,并将变换后的图像用于生成对抗样本,有效缓解对抗样本生成过程中的过拟合现象。在此基础上,采用集成模型攻击的方式生成可迁移性更强的对抗样本,从而提高黑盒攻击成功率。在 ImageNet 数据集上进行单模型和集成模型攻击的实验结果表明,该方法的黑盒攻击成功率高达 80.1%,与迭代快速梯度符号方法和动量迭代快速梯度符号方法相比,该方法的白盒攻击成功率虽然略有降低,但仍保持在 97.8% 以上。

关键词: 深度神经网络; 对抗样本; 黑盒攻击; 平移随机变换; 迁移性

开放科学(资源服务)标志码(OSID):



中文引用格式: 李哲铭, 张恒巍, 马军强, 等. 基于平移随机变换的对抗样本生成方法[J]. 计算机工程, 2022, 48(11): 152-160, 183.

英文引用格式: LI Z M, ZHANG H W, MA J Q, et al. Adversarial examples generation method based on random translation transformation[J]. Computer Engineering, 2022, 48(11): 152-160, 183.

Adversarial Examples Generation Method Based on Random Translation Transformation

LI Zheming^{1,2}, ZHANG Hengwei¹, MA Junqiang¹, WANG Jindong¹, YANG Bo¹

(1. School of Cryptographic Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; 2. Staff Department, PLA Army, Beijing 100000, China)

[Abstract] The image classification model based on Deep Neural Network (DNN) can recognize images with a recognition degree that is even higher than that of human eyes. However, it is vulnerable to attacks from adversarial examples because of the fragility of the model's structure. Existing methods for generating adversarial examples have high white-box attack rates, whereas the attack success rate of adversarial examples is low under the black-box condition. The data enhancement technique is introduced into the generation process of adversarial examples. This study proposes a method for generating adversarial examples, TT-MI-FGSM, based on random translation transformation. The random translation transformation of the original image is performed by establishing a probability model, and the transformed image is used to generate adversarial examples, which effectively alleviates over-fitting during the generation of adversarial examples. On this basis, model diversification is achieved by integrating model attacks to generate more transferability adversarial to improve the success rate of black-box attacks. The experiments of single and integrated model attacks on the ImageNet dataset show that the success rate of the black-box attack for the proposed method can be as high as 80.1%. Compared with the iterative fast gradient sign method and momentum iterative fast gradient sign method, it still exceeds 97.8%, although the success rate of the white-box attack for the proposed method is slightly reduced.

[Key words] Deep Neural Network (DNN); adversarial examples; black-box attack; random translation transformation; transferability

DOI: 10.19678/j.issn.1000-3428.0063075

基金项目: 国家重点研发计划“高安全等级移动终端关键技术”(2017YFB0801900)。

作者简介: 李哲铭(1994—),男,硕士研究生,主研方向为对抗攻击技术;张恒巍(通信作者),副教授、博士;马军强,副教授、硕士;王晋东,教授;杨 博,硕士研究生。

收稿日期: 2021-10-28 **修回日期:** 2021-12-28 **E-mail:** zhw11qd@163.com

0 概述

卷积神经网络(Convolutional Neural Network, CNN)已在图像识别和图像分类领域中得到广泛应用,并表现出良好的性能^[1-3]。但是若在原始图像中加入人类无法察觉的扰动,由此形成的对抗样本将影响CNN的性能,从而导致模型分类错误^[4]。该现象给许多实际应用系统带来了安全隐患^[5],如人脸识别系统的识别错误可能会导致准入权限管理失效^[6],自动驾驶车辆的识别错误可能会造成严重的交通事故^[7]等。因此,研究人员提出多种对抗样本攻击和防御的方法,以提高模型的鲁棒性,从而促进应用系统的安全部署和稳定使用。

现有的攻击方法主要分为白盒攻击和黑盒攻击两类。在白盒条件下,对抗样本生成方法可以针对性地生成关于某个模型的对抗样本,现有技术已经具有较高的白盒攻击成功率。但是,白盒攻击需要攻击者掌握更多的模型信息,这在现实世界中难以实现。在黑盒条件下,攻击者只需要掌握较少模型信息便可实施攻击^[8],因此,黑盒攻击具有更强的实用性。研究表明^[9],对抗样本具有良好的迁移特性,即针对某一模型生成的对抗样本对其他模型也具有一定的攻击效果。根据该性质,文献[10]通过引入

动量项、优化梯度损失函数更新方向和加快收敛速度,提高对抗样本的黑盒攻击成功率。黑盒攻击成功率低的主要原因是对于对抗样本生成过程中产生“过拟合”现象^[11],即针对已知模型生成的对抗样本仅对该模型具有较强的攻击成功率,但是攻击其他模型的成功率较低。

本文提出基于平移随机变换的对抗样本生成方法。从数据增强角度,利用平移随机变换方法优化对抗样本的生成过程。通过构建概率模型,并对原始图像进行随机变换操作,利用变换后的图像生成对抗扰动,同时逐步添加到原图像上迭代生成对抗样本,根据应用场景设计单模型攻击算法和集成模型攻击算法,提高黑盒攻击成功率。

1 相关研究

对抗样本的攻击和防御在一定意义上可以相互促进。针对深度神经网络在对抗攻击情况下表现出的脆弱性,对抗样本既可以用于攻击已训练好的模型,将攻击成功率作为一种重要指标来评价模型的鲁棒性,同时可将对抗样本作为训练数据来进一步训练模型,提高模型对对抗样本的识别能力,从而提高其对恶意攻击的防御性能。现有对抗样本生成方法和防御方法相关研究的优点和不足如表1所示。

表1 对抗攻击和对抗防御相关研究成果对比

Table 1 Comparison of research results related to adversarial attack and adversarial defense

文献	所属类别	优点	不足
文献[10]	对抗攻击	提出基于动量的迭代算法 MI-FGSM, 优化损失函数修正路径, 提高黑盒攻击成功率	由于噪声固化等原因, 使得黑盒攻击成功率仍不理想
文献[12]	对抗攻击	提出 I-FGSM 方法, 提高对抗样本的白盒攻击成功率	对抗样本的黑盒攻击成功率有所降低
文献[13]	对抗攻击	对抗攻击存在于传统机器学习中	仅适用于早期神经网络, 未涉及深度神经网络的攻击效果
文献[14]	对抗攻击	深度神经网络对对抗样本的脆弱性, 并提出 L-BFGS 方法来生成对抗样本	计算复杂度较高, 占用较多的计算资源
文献[15]	对抗攻击	提出 FGSM, 提高对抗样本的生成效率	以单步的方式优化, 白盒攻击成功率较低
文献[16]	对抗攻击	指出物理世界中对抗样本存在的可能性	对抗样本的物理攻击成功率较低
文献[17]	对抗防御	将对抗样本引入正常训练数据, 以提高训练模型的鲁棒性	对抗训练略微降低了泛化精度
文献[18]	对抗防御	提出高级表示的方法, 引导去掉扰动噪声, 从而净化对抗样本	去噪能力取决于训练集的表现力
文献[19]	对抗防御	提出集成对抗训练的方法	对单步对抗攻击有很强的鲁棒性, 但仍容易受简单的黑盒和白盒攻击
文献[20]	对抗防御	提出一种基于单步迭代的对抗防御方法	面对较强攻击时防御性较差
文献[21]	对抗防御	通过对对抗样本施加变换, 使得对抗噪声失效	依赖于梯度掩蔽技术
文献[22]	对抗防御	将其某个特征或某部分作为检测器的输入进行检测, 将检测与防御相结合	面对较强攻击时错误率较高

从表1可以看出, 现有的对抗样本攻击方法改进主要从寻找优化算法角度提高攻击性能, 如文献[12]提出迭代快速梯度符号方法(Iterative Fast Gradient Sign Method, I-FGSM), 通过将单步生成调整为多步迭代, 细化对抗样本生成过程中的损失函数更新过程。文献[10]提出向量迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method,

MI-FGSM), 通过引入动量项并稳定损失函数的更新方向, 以避免陷入局部最大值, 从而使损失函数更快达到真实最优解, 进一步提高生成对抗样本的攻击成功率。优化算法可以更高效地生成对抗样本, 但同时会加剧对抗样本与图像分类模型的过度拟合。因此, 本文从数据增强的角度提高对抗样本的泛化攻击能力, 提高黑盒攻击成功率。

2 背景知识

FGSM是典型的对抗样本生成方法。本文研究也是基于FGSM。因此,本节介绍FGSM类中的几种基本方法,定义 x 为输入的原始干净图像, y^{true} 为该图像对应的真实标签, x^{adv} 为生成的对抗样本。本文采用 θ 特征性地表示图像分类模型的结构、参数等信息。 ε 为加入对抗扰动的最大扰动值, $L(x, y^{\text{true}}; \theta)$ 为神经网络反向传播过程中的损失函数。本文采用交叉熵损失函数。

2.1 快速梯度符号方法

快速梯度符号方法^[15]是FGSM类中的初始版本。该方法中通过计算输入图像 x 的损失函数梯度,并以单步的形式沿梯度方向添加扰动,从而以最快的速度生成对抗样本。该方法具有较高的黑盒攻击成功率,由于损失函数单步变化不够精确,因此,白盒攻击成功率还有待提高,其过程如式(1)所示:

$$x^{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x L(x, y^{\text{true}}; \theta)) \quad (1)$$

2.2 迭代快速梯度符号方法

针对FGSM白盒攻击成功率低的问题,迭代快速梯度符号方法(I-FGSM)^[12]将FGSM损失函数计算时的单步计算改为多步迭代计算。该方法的实现过程如式(2)所示:

$$x_{n+1}^{\text{adv}} = \text{Clip}_x \{ x_n^{\text{adv}} + \alpha \cdot \text{sign}(\nabla_x L(x_n^{\text{adv}}, y^{\text{true}}; \theta)) \} \quad (2)$$

在初始状态时,令 $x_0^{\text{adv}} = x$ 。在之后的 T 轮迭代中,分 T 次逐步添加对抗扰动,对抗扰动的大小为 $\alpha = \varepsilon/T$,Clip函数保证生成的对抗样本不超出 x 的 ε 邻域范围。该方法使得损失函数在迭代过程中更快速且准确地达到最大值。因此,I-FGSM大幅提高白盒攻击的成功率,同时当迭代轮数较大时会出现过拟合现象,其黑盒攻击成功率相对FGSM有所下降。

2.3 动量迭代快速梯度符号方法

实验结果表明,利用I-FGSM优化损失函数,损失函数容易陷入局部最优值,这也是导致黑盒攻击成功率降低的一个原因。针对该问题,动量迭代快速梯度符号方法(MI-FGSM)^[10]将衰减因子引入到对抗样本生成过程中,使得损失函数在计算过程中保持前进的惯性,以稳定更新方向,从而突破损失函数的局部最大值,获得真实的最优解。MI-FGSM实现过程如式(3)和式(4)所示:

$$g_{n+1} = \mu \cdot g_t + \frac{\nabla_x L(x_n^{\text{adv}}, y^{\text{true}}; \theta)}{\|\nabla_x L(x_n^{\text{adv}}, y^{\text{true}}; \theta)\|_1} \quad (3)$$

$$x_{n+1}^{\text{adv}} = \text{Clip}_x \{ x_n^{\text{adv}} + \alpha \cdot \text{sign}(g_{n+1}) \} \quad (4)$$

其中: g_t 为在第 t 次迭代过程中累积的梯度值; μ 为 g_t 的衰减因子。在初始状态令 $g_t = 0, x_0^{\text{adv}} = x$ 。MI-FGSM保持了I-FGSM的白盒攻击成功率,同时提高了黑盒攻击成功率。但该方法的黑盒攻击成功率仍然不高。

3 基于平移随机变换的生成方法

FGSM类方法具有简洁易懂、生成效率高、可扩展性好等优点,但其黑盒攻击成功率较低。为此,本文利用数据增强技术,提出基于平移随机变换的对抗样本生成方法,通过将平移变换引入到对抗样本的生成过程中,并将其作为数据增强的一种典型应用,扩展了对抗样本生成过程中图像输入的多样化,有效缓解在对抗样本生成过程中存在的“过拟合”现象,从而提高攻击方法的黑盒攻击成功率。

对抗样本生成方法通常采用单模型攻击和集成模型攻击。单模型攻击是指将原始干净图像输入到单个图像分类模型中,以生成对抗样本。集成模型攻击是指将图像同时输入到多个图像分类模型中,以生成对抗样本。这两种不同的攻击方式具有不同的攻击效果。单模型攻击是针对某个单模型生成对抗样本,再攻击该模型和其他模型,攻击自身时为白盒攻击,攻击其他模型时为黑盒攻击,攻击成功率可作为对抗样本的评价指标。由于集成模型攻击采用多模型集成的方式,因此生成的对抗样本可以学习到更多的模型信息,以降低对抗样本对单模型的过拟合,从而提高对抗样本的迁移性,因此黑盒攻击成功率也更高。在实际应用中,单模型攻击只需要调用单个模型文件,占用的计算资源较少,以便对对抗样本生成方法的优劣进行比较。而集成模型攻击需要调用多个模型文件,消耗的时长和占用的计算资源较多,但是可以生成迁移性更强的对抗样本,因此,集成模型攻击常用于进一步提高对抗样本的黑盒攻击成功率。

3.1 目标优化问题

从攻击角度分析,对抗样本的目的是使图像分类模型分类出错,即在不改变人眼可辨识类别的基础上,使图像分类模型无法正常实现分类的功能。为此,本文将该问题建模为目标优化问题,如式(5)所示:

$$\arg \max_{x^{\text{adv}}} L(x^{\text{adv}}, y^{\text{true}}; \theta) \quad (5)$$

其中: y^{true} 为图像对应的真实标签; x^{adv} 为生成的对抗样本; θ 为图像分类模型信息。该优化问题的优化目标是使生成的对抗样本相对于原标签类别的损失函数达到最大,以大幅提高分类出错的概率。针对原图像标签生成的对抗样本损失函数最大,在对抗攻击的测试过程中,该图像对应其他标签类别的损失函数相对变小,分类就容易出错。

为提高攻击成功率,对抗样本需具有攻击性和隐蔽性的特点。为有效衡量对抗扰动的可感知性,本文对该优化问题的约束条件进行建模,如式(6)所示:

$$\|r\|_{\infty} = \|x^{\text{adv}} - x\|_{\infty} \leq \varepsilon \quad (6)$$

对抗样本与输入模型的原始图像的差值为添加的对抗扰动 $r = x^{\text{adv}} - x$, ε 为最大扰动值。一般用范数来度量对抗扰动的距离尺度,通过无穷范数来限定

加入的扰动大小。基于此,本文设计单模型攻击算法和集成模型攻击算法,以提升黑盒攻击强度并检验攻击效果。

3.2 单模型攻击算法

在FGSM类生成方法中,黑盒攻击成功率低的主要原因是对抗样本与图像分类模型发生了过拟合。研究发现^[10],对抗样本生成过程与神经网络模型的训练过程有相似性,对抗样本的迁移性可以与神经网络模型的泛化能力相对应。因此,本文将提高模型泛化能力的方法引入到对抗样本的生成过程中。在图像分类模型的训练过程中,数据增强常被用于提高模型的泛化能力,同样,可以将平移随机变换作为数据增强的手段,应用于对抗样本的生成过程中,以提高对抗样本的迁移性,从而有效提高对抗样本的黑盒攻击成功率。

本文将概率变换模型建模为平移随机变换过程,并设置超参数变换概率 p 和最大平移距离 D ,以精准控制生成过程,其平移随机变换函数如式(7)所示:

$$T(x_n^{\text{adv}}; p, D) = \begin{cases} T(x_n^{\text{adv}}, D), & \text{概率为 } p \\ x_n^{\text{adv}}, & \text{概率为 } 1-p \end{cases} \quad (7)$$

在平移随机变换函数 $T(\cdot)$ 中,图像将在上、下、左、右四个方向上以概率 p 进行最大平移距离为 D 像素的变换。因为此变换过程是随机的,所以能够有效提高输入图像的多样性,缓解过拟合现象。

在对单模型攻击的过程中,基于此变换函数,本文优化目标函数式(5),新的目标函数如式(8)所示:

$$\arg \max_{x^{\text{adv}}} L(T(x_n^{\text{adv}}; p, D), y^{\text{true}}; \theta) \quad (8)$$

本文提出将平移变换过程与MI-FGSM相结合的TT-MI-FGSM方法,其过程如式(9)和式(10)所示:

$$g_{n+1} = \mu \cdot g_t + \frac{\nabla_x L(T(x_n^{\text{adv}}; p, D), y^{\text{true}}; \theta)}{\|\nabla_x L(T(x_n^{\text{adv}}; p, D), y^{\text{true}}; \theta)\|_1} \quad (9)$$

$$x_{n+1}^{\text{adv}} = \text{Clip}_x \{x_n^{\text{adv}} + \alpha \cdot \text{sign}(g_{n+1})\} \quad (10)$$

基于以上研究,本文设计针对单模型攻击的算法,具体过程如算法1所示。

算法1 单个分类模型攻击算法(TT-MI-FGSM)

输入 原始输入图像 x 及其真实标签文件 y^{true} ,图像分类模型 f ,损失函数 L ,平移变换概率 p ,最大平移距离 D ,图像最大扰动值 ϵ ,最大迭代轮数 T 以及衰减因子 μ

输出 对抗样本图像 x^{adv}

1. 计算每轮添加的扰动大小: $\alpha = \epsilon/T$;
2. 初始化对抗样本生成系统: $x_0^{\text{adv}} = x, g_0 = 0$;
3. for $t = 0$ to $T-1$ do
4. 通过 $x_t^{\text{adv}} = T(x_t^{\text{adv}}; p, D)$ 求取 x_t^{adv} ;

5. 得到关于单模型的损失函数 $\nabla_x L(x_n^{\text{adv}}, y^{\text{true}}; \theta)$;

6. 通过式(4)更新 g_{t+1} ;

7. 通过式(5)更新 x_{t+1}^{adv} ;

8. 返回对抗样本 $x^{\text{adv}} = x_T^{\text{adv}}$ 。

在该算法中,平移变换概率 p 可取0~1之间的任意数值,表示平移变换概率逐渐变大。最大平移距离 D 表示在图像平移过程中,允许移动的像素最大值,例如,当图像大小为 $299 \times 299 \times 3$ (表示图像的长和宽分别为299像素,RGB三通道)时, D 的像素值变化范围为0~299。在 D 增大的过程中,对抗样本的攻击成功率会先增大后减小,这说明当平移距离较大时,会影响图像内容与图像标签的对应关系,使得式(5)中的标签值失效。因此,在实验中应选取合适的 D ,保证正常生成对抗样本。图像的最大扰动值是指允许在原图像上添加的最大扰动,为保证对抗样本在人眼视觉下不失真,通常将最大扰动值设置在40像素以下。最大迭代轮数 T 指在生成对抗样本过程中的迭代次数, T 越大,每轮添加的扰动大小,即学习率越小,但也会占用更多的计算资源。衰减因子 μ 表示历史动量的作用程度, μ 越大,表示历史动量影响越大。在模型攻击实验部分,该研究按照文献[9]中的规范将 T 设置为16, μ 设置为1,此时方便与已有方法进行对比,从而更好地体现出本文方法的成果优势。

从单个分类模型攻击算法中可以看出,在对抗样本的生成过程中,本文设置 T 轮迭代循环,在每轮开始时,都以 p 为变换概率、 D 为最大平移距离对输入图像进行平移随机变换,之后通过图像标签与基于输入图像的逻辑值计算该图像的交叉熵损失函数,并求得其梯度,由式(9)得到新一轮的累积梯度值,并根据式(10)更新对抗样本,直到满足迭代轮数要求,则完成迭代更新,输出对抗样本。当去掉算法1的第4个步骤时,该算法退化为MI-FGSM,这也体现该算法的便利性优势,当迭代轮数 T 设置为1时,该算法即可转化为TT-I-FGSM的单模型攻击算法。

3.3 集成模型攻击算法

在对抗样本防御领域中,集成对抗训练的方法可以有效提高分类模型对对抗样本的防御能力^[19]。从模型训练角度,集成模型解决了模型训练过程中的拟合问题,提高了模型的泛化能力,从而有助于提高对抗样本的防御能力。因此,在集成模型条件下的攻击算法流程如图1所示,以进一步缓解对抗样本对分类模型的“依赖”,增加模型的多样性,从而提高对抗样本的泛化能力和黑盒攻击成功率。

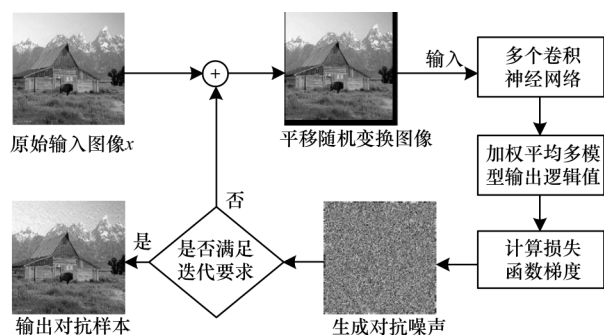


图1 集成模型攻击算法流程

Fig.1 Procedure of integrated model attack algorithm

从图1可以看出,将原始干净图像输入到对抗样本生成系统后,首先根据概率模型进行平移随机变换,变换后出现的空白区域将采用补0的方式实现边界填充,得到 $299 \times 299 \times 3$ 的变换后图像;将得到的图像输入到多个图像分类模型中,并得到各模型的输出逻辑值;根据加权平均后的逻辑值计算损失函数,并沿损失函数的梯度方向迭代更新,直到完成迭代,此时输出对抗样本图像。

在对抗样本生成过程中,本文在集成分类模型上利用TT-MI-FGSM生成对抗样本,如算法2所示。

算法2 集成分类模型攻击算法(TT-MI-FGSM)

输入 原始输入图像 x 及其真实标签文件 y^{true} , n 个图像分类模型 $f(f_1, f_2, \dots, f_n)$,各模型对应的网络逻辑值 $l(l_1, l_2, \dots, l_n)$,各模型的集成权重 $w(w_1, w_2, \dots, w_n)$,损失函数 L ,平移变换概率 p ,最大平移距离 D ,图像最大扰动值 ϵ ,最大迭代轮数 T 以及衰减因子 μ

输出 对抗样本图像 x^{adv}

1. 计算每轮添加的扰动大小: $\alpha = \epsilon/T$;
2. 初始化对抗样本生成系统: $x_0^{adv} = x, g_0 = 0$;
3. for $t = 0$ to $T-1$ do;
4. 通过 $x_t^{adv} = T(x_t^{adv}; p, D)$ 求取 x_t^{adv} ;
5. 将 x_t^{adv} 输入到不同的网络模型,并得到对应的逻辑值 $l_k(x_t^{adv}), k = 1, 2, \dots, n$;
6. 求集成逻辑值 $l(x_t^{adv}) = \sum_{k=1}^n w_k l_k(x_t^{adv})$;
7. 得到关于集成模型的损失函数 $\nabla_x L(x_n^{adv}, y^{true}; \theta)$;
8. 通过式(4)更新 g_{t+1} ;
9. 通过式(5)更新 x_{t+1}^{adv} ;
10. 返回对抗样本 $x^{adv} = x_T^{adv}$ 。

在算法2中,超参数的意义及设置的取值范围与单模型攻击算法一致,而两个算法之间的区别在于生成对抗样本的同时将图像输入到多个图像分类模型,并通过多模型输出逻辑值并权重集成的方式,实现生成对抗样本的目的。该算法可大幅提高黑盒攻击成功率,但其生成的对抗样本白盒攻击成功率略有降低。

3.4 方法的可扩展性

平移随机变换作为数据增强的一种手段,可用于改进FGSM类方法的性能,并通过与I-FGSM和MI-FGSM的结合,提高对抗样本的黑盒攻击成功率。通过调整方法中的衰减因子 μ 、最大平移距离 D 以及平移变换概率 p ,实现与现有对抗样本生成方法的转化,体现了该方法的可扩展性和便利性优势。不同对抗样本生成方法之间的转化关系如图2所示。

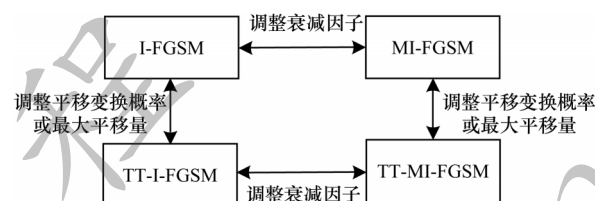


图2 不同对抗样本生成方法之间的转化关系

Fig.2 Conversion relationship among different adversarial examples generation methods

当衰减因子 μ 为0时,TT-MI-FGSM退化为TT-I-FGSM,MI-FGSM退化为I-FGSM。当 μ 不为0时,损失函数将在带有动量修正的过程中更新。当平移变换概率 p 或最大平移距离 D 为0时,TT-I-FGSM退化为I-FGSM,TT-MI-FGSM退化MI-FGSM。当平移变换的概率 p 或最大平移距离 D 不为0时,TT-MI-FGSM将先按一定概率平移变换再进行对抗样本生成。

4 实验与结果分析

本文在Intel Core i9-10900K上,利用python 3.8.5和Tensorflow 1.14.0深度学习框架对比本文提出方法TT-I-FGSM、TT-MI-FGSM与其他方法(I-FGSM^[12]和MI-FGSM^[10])的攻击成功率。目标平台的操作系统为Windows 10(专业版),内存为64 GB,主频为3.7 GHz。为提高对抗样本生成效率,实验使用NVIDIA GeForce RTX 2080Ti GPU加速完成计算过程。

4.1 实验设置

4.1.1 数据集

ImageNet数据集^[23]是由斯坦福大学李飞飞教授带领创建的计算机视觉数据集,是目前图像识别领域最大的数据库,也是评估图像分类算法性能的基准。该数据集中的每张图片都被手工标注类别,在对抗样本的生成过程中,本文利用其标签以生成对抗样本并进行攻击效果检测,并从ImageNet数据集验证集的每个类别中各随机选择图片,利用这1 000张分类正确的图片进行对抗样本的生成。

4.1.2 分类模型

在实验中共使用 7 个分类模型,其中,4 个正常训练分类模型分别是 Inception-v3^[24] (Inc-v3)、Inception-v4^[25] (Inc-v4)、Inception-ResNet-v2^[25] (IncRes-v2) 和 ResNet-v2-101^[26] (Res-101); 3 个对抗训练分类模型^[19] 分别是 ens3-adv-Inception-v3 (Inc-v3_{ens3})、ens4-adv-Inception-v3 (Inc-v3_{ens4}) 和 ens-adv-Inception-ResNet-v2 (IncRes-v2_{ens})。

4.1.3 基准方法

为更好地评估本文方法的有效性,本文选择另外两种典型的对抗样本生成方法 (I-FGSM^[13] 和 MI-FGSM^[10]) 作为对比基准,与本文提出的 TT-I-FGSM 和 TT-MI-FGSM 进行对比分析。

4.1.4 超参数设置

为了方便攻击成功率对比,本文按照动量法^[10]中

的规范设置超参数,最大扰动值 $\epsilon=16$ 像素,迭代轮数 $T=10$,步长 $\alpha=1.6$,衰减因子 $\mu=1.0$ 。本文提出的超参数:平移变换概率 $p=0.5$,最大平移距离 $D=11$ 像素。

4.2 单模型攻击实验

本文在白盒和黑盒条件下对本文方法 TT-I-FGSM 和 TT-MI-FGSM 和基准方法进行单模型攻击测试。I-FGSM、TT-I-FGSM、MI-FGSM 和 TT-MI-FGSM 方法分别在 Inc-v3、Inc-v4、IncRes-v2 和 Res-101 4 个正常训练模型上生成对抗样本,并在 7 个分类模型上 (Inc-v3、Inc-v4、IncRes-v2、Res-101、Inc-v3_{ens3}、Inc-v3_{ens4}、IncRes-v2_{ens}) 进行测试。不同对抗样本生成方法的黑盒和白盒攻击成功率如表 2 所示。表中攻击成功率为分类错误的图像数在图像总数中所占的比例,*表示白盒攻击。

表 2 不同对抗样本生成方法攻击单个模型的成功率对比

Table 2 Success rate of attack on a single model comparison among different adversarial examples generation methods %								
模型	对抗样本生成方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 _{ens3}	Inc-v3 _{ens4}	IncRes-v2 _{ens}
Inc-v3	I-FGSM	99.9*	22.7	20.3	18.2	7.1	7.6	4.2
	TT-I-FGSM	99.2*	45.8	37.8	31.3	10.1	9.2	4.9
	MI-FGSM	99.9*	48.2	47.1	39.9	15.2	14.2	7.2
	TT-MI-FGSM	99.0*	70.4	66.6	58.8	21.2	21.1	9.1
Inc-v4	I-FGSM	37.8	99.9*	26.1	21.9	8.6	8.0	5.1
	TT-I-FGSM	57.4	99.7*	42.1	35.2	11.0	9.2	6.2
	MI-FGSM	63.8	99.9*	53.7	47.7	19.7	16.8	9.4
	TT-MI-FGSM	79.0	99.4*	69.8	62.2	25.6	22.5	13.1
IncRes-v2	I-FGSM	37.3	31.9	99.6*	25.9	8.8	7.6	4.9
	TT-I-FGSM	60.7	52.8	98.8*	43.3	14.6	11.7	7.7
	MI-FGSM	68.6	61.9	99.6*	52.3	25.1	20.2	14.6
	TT-MI-FGSM	80.1	76.8	98.9*	67.8	35.9	30.1	19.9
Res-101	I-FGSM	27.8	23.4	21.4	98.2*	9.4	7.9	5.7
	TT-I-FGSM	53.7	46.6	43.9	97.9*	15.6	13.4	8.1
	MI-FGSM	52.4	48.3	45.6	98.2*	22.3	18.7	11.8
	TT-MI-FGSM	72.7	70.5	65.7	97.8*	33.6	29.5	18.2

从表 2 可以看出,在白盒条件下,本文提出的方法 TT-I-FGSM 和 TT-MI-FGSM 在各模型上攻击成功率保持在 97.8% 以上的水平。在黑盒条件下,TT-MI-FGSM 在正常训练模型上提高了攻击成功率,例如,在 Inc-v3 模型上生成的对抗样本,攻击 IncRes-v2 模型时,TT-MI-FGSM 攻击成功率比 MI-FGSM 提高 19.5 个百分点。此外,TT-MI-FGSM 在针对防御模型的黑盒攻击中也表现出较高的攻击性能,例如,在 Res-101 模型上生成的对抗样本,当攻击 Inc-v3_{ens3} 模型时,TT-MI-FGSM 算法的攻击成功率为 33.6%,比 MI-FGSM 提高 11.3 个百分点。因此,本文提出的 TT-MI-FGSM

方法可以有效提高对抗样本的迁移性。

4.3 集成模型攻击实验

集成模型通常对对抗样本具有更好的防御能力。本文通过同时攻击多个分类模型组成的集成模型来评估对抗样本生成方法的性能。本文分别利用 I-FGSM、TT-I-FGSM、MI-FGSM 和 TT-MI-FGSM 同时攻击具有相同权重的 4 个正常训练模型 (Inc-v3、Inc-v4、IncRes-v2 和 Res-101) 构成的集成模型,并通过生成的对抗样本分别在 7 个图像分类模型中测试得到攻击成功率,实验结果如表 3 所示。表中攻击成功率为分类错误的图像数在图像总数中所占的比例,*表示白盒攻击。

表3 不同对抗样本生成方法攻击集成模型的成功率对比

Table 3 Success rate of attack on intergated model comparison among different adversarial examples generation methods %

对抗样本生成方法	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3 _{ens3}	Inc-v3 _{ens4}	IncRes-v2 _{ens}	平均值
I-FGSM	99.7	96.2	91.9	86.6	18.8	15.9	9.3	59.8
TT-I-FGSM	99.7	96.2	92.7	89.4	28.7	24.5	15.6	63.8
MI-FGSM	99.8	97.7	95.1	91.2	38.4	36.0	22.4	68.7
TT-MI-FGSM	99.1	97.9	95.2	92.0	54.2	49.3	31.8	74.0

从表3可以看出,TT-MI-FGSM方法可在保持或提高白盒攻击成功率的基础上,较大幅度提高黑盒攻击的成功率。例如,在Inc-v3、Inc-v4、IncRes-v2和Res-101 4个正常训练模型的白盒攻击上,TT-I-FGSM的攻击成功率比I-FGSM提高了0.9个百分点。而在黑盒攻击中,TT-MI-FGSM比MI-FGSM的攻击成功率平均提高12.83个百分点。针对Inc-v3_{ens3}模型的攻击,TT-MI-FGSM的攻击成功率为54.2%。

不同方法生成的对抗样本图片对比如图3所示,与原图片相比,图片进行平移随机变换后生成的对抗样本,整体区域均添加了对抗扰动,在人眼视觉上未出现肉眼可见的辨识偏差。



(a)原始图片 (b)I-FGSM (c)TT-I-FGSM (d)MI-FGSM (e)TT-MI-FGSM

图3 不同方法生成的对抗样本图片对比

Fig.3 Comparison of adversarial examples images generated by different methods

4.4 超参数

本节重点分析在攻击方法中各超参数对攻击成功率的影响。本文运用TT-I-FGSM和TT-MI-FGSM攻击相同权重的4个正常训练模型(Inc-v3、Inc-v4、IncRes-v2和Res-101)组成的集成模型,通过调整超参数设置,以消融实验的方式探究超参数对实验结果的影响。超参数主要有平移变换概率 p 、最大平移距离 D 、计算梯度时的迭代轮数 T 、图像最大扰动值 ϵ 。由于动量项中的衰减因子 μ 只用于特征性地反映梯度的累积效应,因此本文不对其进行讨论。

4.4.1 平移变换概率 p 对攻击成功率的影响

其他超参数设置:最大扰动值 $\epsilon=16$ 像素,迭代轮数 $T=10$,步长 $\alpha=1.6$,衰减因子 $\mu=1.0$,最大平移距离 $D=11$ 像素,平移变换概率 p 从0以0.1的幅度增长到1。当 $p=0$ 时,表示不发生平移变换;当 $p=1$

时,表示一定发生平移变换。随平移变换概率 p 的递增,白盒攻击成功率和黑盒攻击成功率变化情况如图4所示。其中,实线表示白盒攻击成功率,虚线表示黑盒攻击成功率。

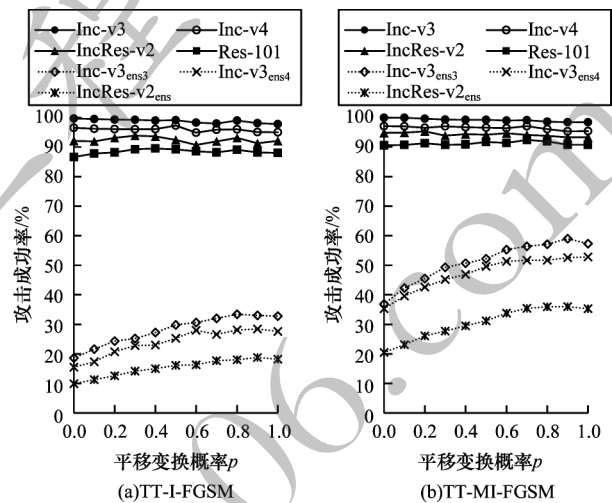


图4 平移变换概率与攻击成功率关系

Fig.4 Relationship between translation conversion probability and attack success rate

在白盒攻击中,将生成的对抗样本在4个正常分类模型上进行分类测试,TT-I-FGSM和TT-MI-FGSM白盒攻击成功率保持平稳态势,并可实现均值90%以上的攻击成功率。在黑盒攻击中,当攻击3个对抗训练模型时,随平移变换概率 p 的递增,两个方法的黑盒攻击成功率也有了较大幅度的提高。TT-I-FGSM的攻击情况如图4(a)所示,TT-MI-FGSM的攻击情况如图4(b)所示。从图4(a)和图4(b)可以看出,两折线图变化趋势大致相同,相比TT-I-FGSM,TT-MI-FGSM能够大幅提高黑盒攻击成功率,这说明带有动量的平移随机变换方法可以更好地提高对抗样本的迁移性和黑盒攻击成功率。同时,当 p 值较小时,黑盒攻击对概率 p 的变化更加敏感,即使略微增大平移变换概率,即可大幅提高攻击成功率。这种现象表明平移变换有助于黑盒攻击迁移性的提高。在对抗样本生成方法设计中,当利用对抗样本的迁移性对黑盒模型进行攻击时,可以将平移变换概率设置为最大值,即 $p=1$,具有最大的黑盒攻击成功率,而此时白盒攻击成功率仍能保持在较高水平。

4.4.2 最大平移距离 D 对攻击成功率的影响

最大平移距离 D 是指在平移随机变换过程中,图像在上、下、左、右4个方向上的最大移动距离。超参数设置:平移变换概率 $p=0.5$,最大扰动值 $\varepsilon=16$ 像素,迭代轮数 $T=10$,步长 $\alpha=1.6$,衰减因子 $\mu=1.0$ 。实验运用 TT-MI-FGSM 在 2 个最大平移距离的区间范围内进行测试,最大平移距离与攻击成功率的关系如图 5 所示。

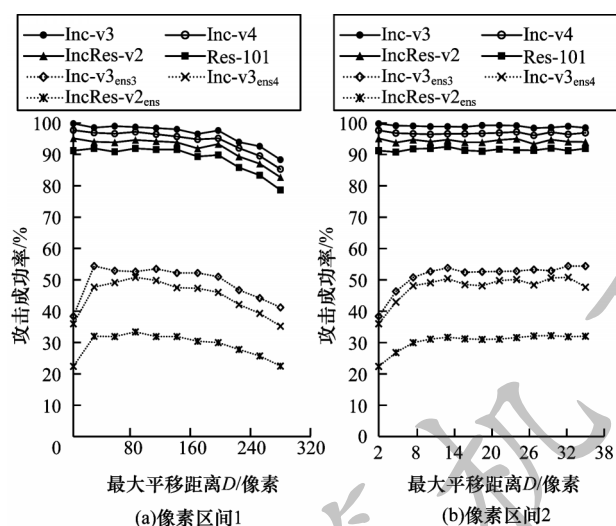


图5 最大平移距离与攻击成功率关系

Fig.5 Relationship between maximum translation distance and attack success rate

最大平移距离 D 在 0~280 像素范围内的攻击成功率变化情况为:在 0~40 像素范围内,TT-MI-FGSM 的白盒攻击成功率趋于稳定,黑盒攻击成功率快速提高,表明平移变换对提高对抗样本迁移性有较好效果,能够有效避免与生成模型的过拟合现象;在 40~200 像素的区间范围内,白盒攻击和黑盒攻击的成功率均趋于稳定;在 D 超过 200 像素以后,黑盒攻击和白盒攻击的成功率均有所下降,在此过程中,随着 D 的增大,出现了欠拟合现象,导致白盒攻击和黑盒攻击的成功率均下降。最大移动距离 D 在 0~35 像素范围内的攻击成功率变化情况如图 5(b)所示。当 $D=11$ 像素时,黑盒攻击成功率的局部最大值为 45.3%,之后便波动变化。因此,在进行对抗样本生成时,本文选用 $D=11$ 像素,此时既有效提高黑盒攻击成功率,又保证了较高的白盒攻击成功率。

4.4.3 最大扰动值 ε 对攻击成功率的影响

本文分别使用 TT-I-FGSM 和 TT-MI-FGSM 进行白盒攻击和黑盒攻击。超参数设置:迭代轮数 $T=10$,衰减因子 $\mu=1.0$,平移变换概率 $p=0.5$,图像最大扰动值从 4 像素开始,以 4 为步长增长到 32,最大扰动值与攻击成功率关系如图 6 所示。由 $\alpha=\varepsilon/T$ 可知,随着最大扰动值的增大,在迭代过程中的步长

α 也逐渐增大。从图 6 可以看出,最大扰动值 ε 在 4~16 像素范围内,随着扰动值的增大,黑盒攻击成功率得到显著提高。在黑盒攻击中,随着 ε 的增大,TT-I-FGSM 攻击成功率基本不变,而加入动量因子后 TT-MI-FGSM 的攻击成功率仍保持了上升趋势。

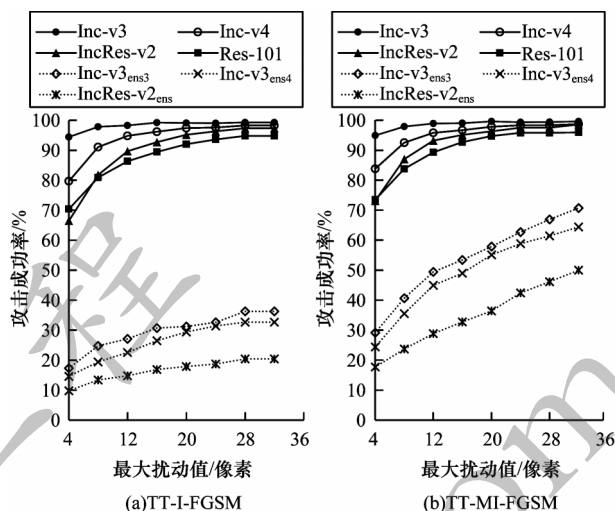


图6 最大扰动值与攻击成功率关系

Fig.6 Relationship between maximum disturbance values and attack success rate

最大扰动值 ε 的增大会带来扰动因素的增长,在生成的对抗样本中噪点也会增多,其可视化对比图如图 7 所示,当 $\varepsilon=32$ 像素时,扰动因素较大,容易被肉眼识别出来。因此,在实际对抗样本生成过程中,本文选择 $\varepsilon=16$ 像素来生成对抗样本,既具有较高的攻击成功率,又能够保证对抗样本与原图像的相似效果。

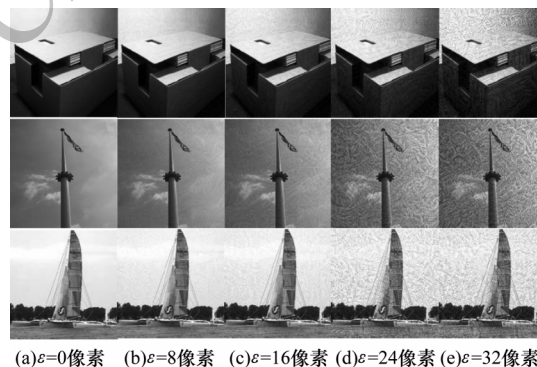


图7 最大扰动值对对抗样本图像的影响

Fig.7 Influence of maximum disturbance values on adversarial examples image

4.4.4 迭代轮数 T 对攻击成功率的影响

本文将迭代轮数从 2 开始以 4 为步长增大,最大值为 30。其他的超参数设置:最大扰动值 $\varepsilon=16$ 像素,步长 $\alpha=1.6$,衰减因子 $\mu=1.0$,平移变换概率 $p=0.5$,最大平移距离 $D=11$ 像素。迭代轮数与攻击成功率的关系如图 8 所示。从图 8(a)可以看出,随着迭代

轮数的增加,TT-I-FGSM方法的白盒攻击成功率有较大提高,黑盒攻击的成功率则略微下降,其原因随着迭代轮数增加,对抗样本与分类模型的拟合得到加强、迁移性有所下降。从图8(b)可以看出,随迭代轮数的增加,TT-MI-FGSM方法的黑盒攻击成功率也得到有效加强,说明该方法与动量法相结合具有更好的攻击效果。迭代轮数的增加将消耗更多的计算资源,因此,本文选取迭代轮数 $T=10$,既可以有效提高白盒攻击和黑盒攻击的成功率,又能够保证对抗样本的生成效率。

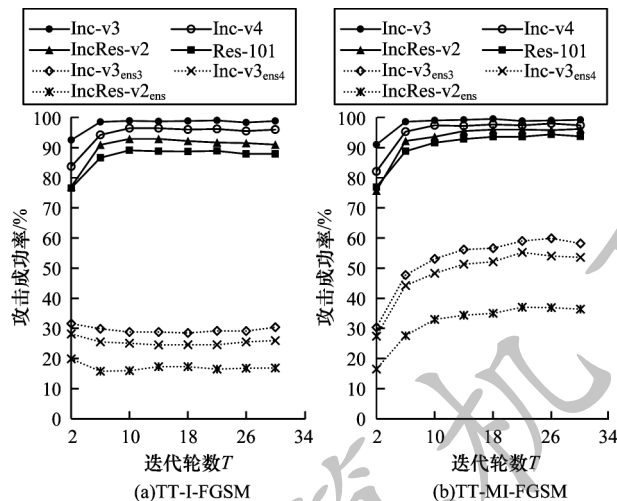


图8 迭代轮数与攻击成功率的关系

Fig.8 Relationship between iteration number and attack success rate

在以上的超参数消融实验中,本文讨论了平移变换概率 p 、最大平移距离 D 、最大迭代轮数 T 和图像最大扰动值 ϵ 的含义,并通过实验验证不同超参数对算法性能的影响。通过实验可以看出,平移变换概率 p 的增大可以提高对抗样本的攻击效果。因此,本文可使用最大平移变换概率来生成攻击性强的对抗样本。最大平移距离 D 的选取应根据原始图像的像素大小选择数值,如在 $299 \times 299 \times 3$ 尺寸的图像中,选取 $D=11$ 像素可以取得较好的攻击效果。最大扰动值可以提高攻击方法的性能,同时也会造成扰动过大,产生易被人眼识别的问题。迭代轮数的增大可以改进黑盒攻击效果,但由于会产生更多的计算开销,因此在保持一定的对抗样本生成效率时应选择适当大小的迭代轮数,通常设置在20以下。

5 结束语

本文提出一种基于平移随机变换的对抗样本生成方法TT-MI-FSGM。通过对原图像进行平移随机变换,扩展图像分类模型的输入多样性,缓解对抗样本生成过程中的过拟合现象,从而提高对抗样本的黑盒攻击成功率。此外,通过构建集成模型攻击算法,以生成迁移性更强的对抗样本。实验结果表明,与I-FGSM相比,该方法在集成模型上的黑盒攻击成功率平均提高了30.4个百分点。下一步将对物理世界中对抗样本的实

际应用进行研究,从而为神经网络模型的实际部署提供更为有效的鲁棒性测试方法。

参考文献

- [1] SADAK F, SAADAT M, HAJIYAVAND A M. Real-time deep learning-based image recognition for applications in automated positioning and injection of biological cells[J]. Computers in Biology and Medicine, 2020, 125(10): 103976.
- [2] GUO G D, ZHANG N. A survey on deep learning based face recognition[J]. Computer Vision and Image Understanding, 2019, 189: 102805.
- [3] MOPURI K R, GANESHAN A, BABU R V. Generalizable data-free objective for crafting universal adversarial perturbations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2452-2465.
- [4] 陈晓楠, 胡建敏, 张本俊, 等. 基于模型间迁移性的黑盒对抗攻击起点提升方法[J]. 计算机工程, 2021, 47(8): 162-169. CHEN X N, HU J M, ZHANG B J, et al. Black box attack adversarial starting point promotion method based on mobility between models[J]. Computer Engineering, 2021, 47(8): 162-169. (in Chinese)
- [5] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of ACM on Asia Conference on Computer and Communications Security. New York, USA: ACM Press, 2017: 506-519.
- [6] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]// Proceedings of ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2016: 1528-1540.
- [7] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning model[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 1-10.
- [8] 姜妍, 张立国. 面向深度学习模型的对抗攻击与防御方法综述[J]. 计算机工程, 2021, 47(1): 1-11. JIANG Y, ZHANG L G. Survey of adversarial attacks and defense methods for deep learning model[J]. Computer Engineering, 2021, 47(1): 1-11. (in Chinese)
- [9] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. [2021-09-20]; <https://arxiv.org/abs/1611.02770>.
- [10] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[EB/OL]. [2021-09-20]. <https://arxiv.org/pdf/1710.06081v2.pdf>.
- [11] WANG X S, HE X R, WANG J D, et al. Admix: enhancing the transferability of adversarial attacks[EB/OL]. [2021-09-20]. <http://arxiv.org/abs/2102.00436v3>.
- [12] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. [2021-09-20]. <https://arxiv.org/abs/1607.02533v4>.
- [13] BIGGIO B, CORONA I, MAIORCA D, et al. Evasion attacks against machine learning at test time[C]// Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases. New York, USA: ACM Press, 2013: 387-402.

(下转第183页)

(上接第160页)

- [14] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C]//Proceedings of International Conference on Learning Representations. Banff, Canada; [s. n.], 2014: 1-10.
- [15] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. [2021-09-20]. <https://arxiv.org/pdf/1412.6572.pdf>.
- [16] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2018: 1625-1634.
- [17] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2019: 4312-4321.
- [18] KURAKIN A, GOODFELLOW I J, SAMY B. Adversarial machine learning at scale [EB/OL]. [2021-09-20]. <https://arxiv.org/pdf/1611.01236.pdf>.
- [19] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses [EB/OL]. [2021-09-20]. <https://arxiv.org/abs/1705.07204v5>.
- [20] LIU G X, KHALIL I, KHREISHAH A. Using single-step adversarial training to defend iterative adversarial examples [C]//Proceedings of the 7th ACM Conference on Data and Application Security and Privacy. New York, USA; ACM Press, 2021: 17-27.
- [21] XIE C H, WU Y X, MAATEN L V D, et al. Feature denoising for improving adversarial robustness [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2019: 1-10.
- [22] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations [EB/OL]. [2021-09-20]. <https://arxiv.org/abs/1702.04267v1>.
- [23] 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.
- [24] SZEGEDY C, SHNATHON J, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA; IEEE Press, 2016: 2818-2826.
- [25] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA; AAAI Press, 2017: 4278-4284.
- [26] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany; Springer, 2016: 630-645.

编辑 薛晋栋