计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800087

图像对抗样本研究综述

陈梦轩[1] 张振永[1] 纪守领[2] 魏贵义[3,4] 邵 俊[1]
1 浙江工商大学计算机与信息工程学院 杭州 310018
2 浙江大学计算机科学与技术学院 杭州 310058
3 浙江工商大学信息与电子工程学院 杭州 310018
4 浙江工商大学萨塞克斯人工智能学院 杭州 310018
(s.mxuan0103@gmail.com)

摘 要 随着深度学习理论的发展,深度神经网络取得了一系列突破性进展,相继在多个领域得到了应用。其中,尤其以图像领域中的应用(如图像分类)最为普及与深入。然而,研究表明深度神经网络存在着诸多安全隐患,尤其是来自对抗样本的威胁,严重影响了图像分类的应用效果。因此,图像对抗样本的研究近年来越来越受到重视,研究者们从不同的角度对其进行了研究,相关研究成果也层出不穷,呈井喷之态。首先介绍了图像对抗样本的相关概念和术语,回顾并梳理了图像对抗样本攻击和防御方法的相关研究成果。特别是,根据攻击者的能力以及防御方法的基本思路对其进行了分类,并给出了不同类别的特点及存在的联系。接着,对图像对抗攻击在物理世界中的情况进行了简要阐述。最后,总结了图像对抗样本领域仍面临的挑战,并对未来的研究方向进行了展望。

关键词:深度学习;图像领域;对抗样本;对抗攻击;防御方法;物理世界
中图法分类号 TP391

Survey of Research Progress on Adversarial Examples in Images

CHEN Meng-xuan[1],ZHANG Zhen-yong[1],JI Shou-ling[2],WEI Gui-yi[3,4] and SHAO Jun[1]
1 School of Computer and Information Engineering,Zhejiang Gongshang University,Hangzhou 310018,China
2 College of Computer Science and Technology,Zhejiang University,Hangzhou 310058,China
3 School of Information and Electronic Engineering,Zhejiang Gongshang University,Hangzhou 310018,China
4 Sussex Artificial Intelligence Institute,Zhejiang Gongshang University,Hangzhou 310018,China

Abstract With the development of deep learning theory,deep neural network has made a series of breakthrough progress and has been widely applied in various fields. Among them,applications in the image field such as image classification are the most popular. However,research suggests that deep neural network has many security risks,especially the threat from adversarial examples,which seriously hinder the application of image classification. To address this challenge,many research efforts have recently been dedicated to adversarial examples in images,and a large number of research results have come out. This paper first introduces the relative concepts and terms of adversarial examples in images,reviews the adversarial attack methods and defense methods based on the existing research results. In particular,it classifies them according to the attacker's ability and the train of thought in defense methods. This paper also analyzes the characteristics and the connections of different categories. Secondly,it briefly describes the adversarial attacks in the physical world. In the end,it discusses the challenges of adversarial examples in images and the potential future research directions.

Keywords Deep learning,Image field,Adversarial examples,Adversarial attacks,Defense methods,Physical world

1 引言

随着深度学习理论的不断发展,特别是深度神经网络(Deep Neural Network,DNN)算法的巨大成功,人工智能在科技、产业和社会变革等方面展现了巨大潜力,受到全球的广泛关注。以深度学习为代表的人工智能技术已逐步应用于各个行业,包括图像分类(image classification)[1-2]、目标检测(object detection)[3]、语音识别(speech recognition)[4-5]、自动

到稿日期:2021-08-10 返修日期:2021-09-18
基金项目:国家重点研发计划(2019YFB1804500);国家自然科学基金(U1709217)
This work was supported by the National Key Research and Development Program of China(2019YFB1804500) and National Natural Science Foundation of China(U1709217).
通信作者:邵俊(chn.junshao@gmail.com)

驾驶[6-7]以及人脸识别[8-9]等，尤其在图像领域的应用最为深入，如自动驾驶中对交通标志图像的识别、人脸识别中对人脸图像的检测、以及场景识别中对各场景图片的分类等。

然而，随着对人工智能研究的不断深入，人们发现人工智能技术仿佛一把达摩克里斯之剑，在带来便利的同时也带来了安全隐患，如数据投毒(data poisoning)[10]、模型窃取(model theft)[11]、后门攻击(backdoor attacks)[12]、对抗样本(adversarial examples)[13]等。在这些攻击中，尤其以对抗样本攻击最为著名。对抗样本指在原有数据对象(如图像)上添加人眼无法察觉的细微扰动而产生的新数据对象。人工智能算法对这些细微扰动十分敏感，因此得出了错误结果。通过构造对抗样本对深度神经网络进行攻击的方法，一般称为对抗攻击(adversarial attack)。由于在图像领域应用广泛，图像对抗样本吸引了众多研究者的关注。

图像对抗样本的存在不仅极大地影响了图像识别分类的应用效果，而且还严重威胁到了人们的人身和财产安全。例如，在自动驾驶场景中，攻击者将路标改造成相应的对抗样本，造成自动驾驶系统对路标产生错误判断，从而导致交通事故发生。在2019－2020这两年中，图像对抗样本的相关研究成果大量涌现，甚至呈指数级增长。对这些研究成果进行梳理和总结已成为一种必然。事实上，已有多篇关于图像对抗样本的综述论文[14-19]发表在各种不同类型的期刊和会议上，但其或多或少存在以下几个问题：

(1)缺乏对最新成果的总结。已有的文献综述仅总结了2019年之前的研究结果。而如前所述，图像对抗样本研究的活跃期是在2018年以后。

(2)缺乏对成果之间联系的总结。已有的文献综述更侧重于罗列已有的研究成果，而忽视了这些成果之间存在的联系。

(3)缺乏对图像对抗样本的聚焦。很多对抗攻击综述并不专注于对图像对抗样本研究结果的总结和梳理。

针对以上问题，本文回顾并梳理了图像对抗样本攻击及防御的相关研究成果，如图1所示，主要包括2014年到2020年期间的重要研究成果。根据攻击者的能力及防御思路对其进行了分类，分析了不同类别的特点和性质，并总结了各个研究成果之间的技术逻辑关系。

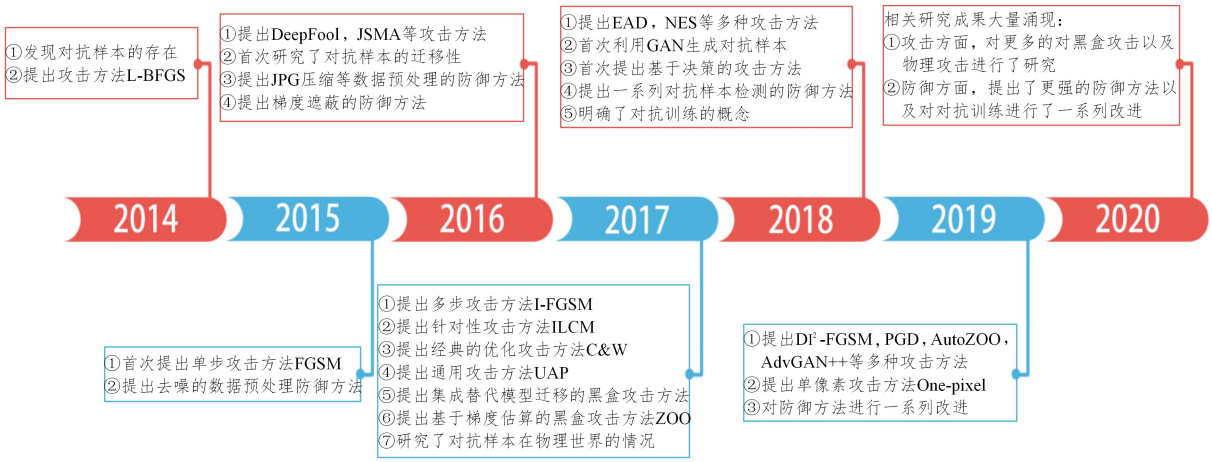

图1 对抗样本发展时间轴

Fig. 1 Timeline of adversarial examples development

本文第2节主要介绍深度学习代表算法即深度神经网络的基本概念、对抗样本的定义及相关术语；第3节分别从白盒攻击和黑盒攻击两大角度归纳对抗样本的攻击方法；第4节分类整理对抗攻击的防御方法；第5节介绍对抗样本在物理世界中的情况；最后，总结并展望对抗样本未来可能的研究方向。

2 对抗样本相关介绍

为了更好地理解对抗样本、对抗攻击以及防御方法，本节将简要介绍深度神经网络的基本概念、对抗样本的定义以及相关专业术语。

2.1 深度神经网络的基本概念

DNN起源于生物神经网络，是一个由许多名为神经元的节点按一定的层次结构连接而成的网络。图2给出了一个典型的神经网络结构，图中的每一个圆圈代表一个神经元，它是DNN中最小的单位。DNN中的操作主要为模型训练，一般包括前向传播和后向传播两个关键步骤。

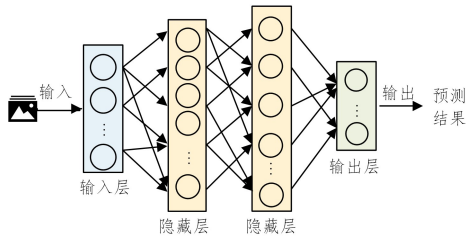

图2 深度神经网络的基本结构

Fig. 2 Basic structure of deep neural network

模型训练的前向传播过程主要是对损失函数 L 进行计算。如图3所示，某个神经元在接受多个输入 $x_1, x_2, \cdots, x_n$ 后，首先利用线性函数 $z=(w_1 * x_1 + w_2 * x_2 + \cdots + w_n * x_n) + b$ 进行加权计算，然后在激活函数 f 的作用下计算得到该神经元上的运算结果 $f(z)$。其中，$w_1, w_2, \cdots, w_n$ 代表前一个神经元到该神经元转化的权重，b 代表该神经元上的偏置，两者都是模型参数。通过各层神经元的计算，最后可得预测结果 y，该结果与真实结果 y' 之间的差异可通过损失函数 $L(y, y')$ 来衡量。

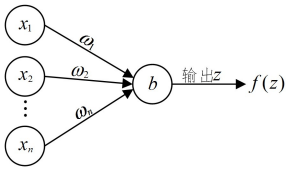

图 3 深度神经网络的前向传播过程

Fig. 3 Forward propagation of deep neural network

模型训练的反向传播过程主要是对模型参数进行优化，其最终目标是使预测结果 y 与真实结果 y' 之间的差异尽可能小，即最小化损失函数 $L(y,y')$。其优化方法通常是根据损失函数 $L(y,y')$ 的偏导(即梯度信息)对模型参数进行调整。

经过 N 次前向传播和反向传播过程，最终可得到性能理想的深度神经网络模型。

2.2 图像对抗样本的定义

图像对抗样本指在输入图像中添加精心构造且人眼难以察觉的细微干扰而形成的样本，这类样本会导致 DNN 以高置信度输出错误的分类结果。图像对抗样本生成的关键步骤在于最大化损失函数 $L(y,y')$，增大预测结果 y 与真实结果 y' 之间的差异。而这个步骤正好与 DNN 训练过程中反向传播的目标相反。如图 4 所示，一开始，DNN 能以 57.7% 的置信度将正常样本分类为大熊猫，但对正常样本添加扰动噪声生成对抗样本后，分类器却以 99.3% 的置信度将对抗样本错误分类为长臂猿。而从人眼识别来看，这两张图片中的动物都是大熊猫。在下文中，如未明确表示，对抗样本即为图像对抗样本。

图 4 对抗样本的生成[20]

Fig. 4 Generation of adversarial examples[20]

2.3 相关术语介绍

为了便于本文后续介绍对抗样本的研究成果，本节描述了对抗样本的相关术语。

(1)白盒攻击 & 黑盒攻击

根据攻击者对目标模型的了解程度，对抗样本攻击分为白盒攻击(white-box attack)和黑盒攻击(black-box attack)两类。在白盒攻击中，攻击者能够获取目标模型的所有信息，包括模型结构、参数以及训练数据集等。而在黑盒攻击中，攻击者无法获取目标模型的相关信息。由此可见，黑盒攻击的成功难度远大于白盒攻击，其攻击场景也更符合实际攻击情况。

(2)非针对性攻击 & 针对性攻击

根据攻击者对目标模型的不同攻击，对抗攻击分为非针对性攻击(non-targeted attack)和针对性攻击(targeted attack)两类。非针对性攻击指对抗样本的攻击目标仅仅导致目标模型输出错误结果；而针对性攻击要求对抗样本不仅导致目标模型输出错误结果，还要求该错误结果是某个特定类型。

(3)单步攻击 & 迭代攻击

根据对抗样本生成步骤的繁简程度，对抗攻击分为单步攻击(one-step attack)和迭代攻击(iterative attack)两类。单步攻击指攻击者只需执行一次基本操作就可获得对抗样本；而迭代攻击则需要攻击者执行多步迭代操作才能获得对抗样本。一般而言，迭代攻击产生的对抗样本比单步攻击产生的对抗样本攻击效果更好，但同时也需更多的执行时间。

(4)置信度 & 类别标签

置信度指模型将输入样本分类为某种类别的概率。一个好的对抗样本会导致目标模型以高置信度输出错误分类，而类别标签指模型的输出类别结果。

(5)对抗扰动的通用性

对抗扰动的通用性指攻击者利用该扰动既能产生针对某个目标模型的对抗样本，也能产生针对其他模型的对抗样本。显然，通用性越强，实际攻击效果就越好。

(6)对抗样本的迁移性

对抗样本的迁移性指该对抗样本不仅能使目标模型分类错误，也能使其他模型分类错误。我们可以理解为这些模型具有相似的分类边界。

(7)鲁棒性

鲁棒性(robustness)一词在各领域广泛存在，也称为健壮性。在对抗样本研究领域，鲁棒性指模型防御对抗攻击的能力。模型鲁棒性越强，防御效果越佳。

(8)模型的过拟合

模型的过拟合指模型在训练数据集上过于拟合，但在测试数据集上拟合效果较差的现象。过拟合的模型往往泛化能力较差。通常来说，造成模型过拟合的原因主要是训练数据集过少以及模型复杂度过高。

3 对抗样本的攻击方法

本节将总结和梳理对抗样本的攻击方法。虽然对抗样本生成的研究成果十分丰富，但根据攻击者对目标模型的不同了解程度，仍可将其简单地分为白盒攻击和黑盒攻击两类。

3.1 白盒攻击方法

在白盒攻击中，利用已知的目标模型信息(如模型内部结构、参数和训练数据集等)，攻击者很容易生成相应的对抗样本。根据对抗样本的生成方式，我们将白盒攻击分为以下 3 个子类：基于梯度的攻击方法、基于优化的攻击方法以及其他白盒攻击方法。

3.1.1 基于梯度的攻击方法

如上所述，生成对抗样本的关键步骤就是最大化损失函数 $L(y,y')$，因此可通过对损失函数 $L(y,y')$ 进行梯度计算，来获得对抗样本。

第一个基于梯度的攻击方法是 Goodfellow 等[20]在 ICLR 2015 会议上提出的快速梯度符号法(Fast Gradient Sign Method，FGSM)。FGSM 方法沿着梯度反方向添加扰动使损失函数 $L(y,y')$ 快速增大，最终导致模型分类错误。虽然此攻击方法可以快速生成对抗样本，但由于是单步攻击，计算所得扰动并不精准，导致攻击成功率较低。针对该问题，Kurakin 等[21]对 FGSM 方法进行了改进，并提出基础迭代

方法(Basic Iterative Method，BIM)，该方法又被称为 I-FGSM(Iterative Fast Gradient Sign Method)方法。I-FGSM 方法将 FGSM 方法中的一步扰动计算过程细分为多步，并通过裁剪操作将图像像素限制在有效区域内，从而提高了攻击成功率。然而，I-FGSM 方法生成的对抗样本容易过拟合到局部极值点，从而影响对抗样本的迁移性。对此，Dong 等[22]在 I-FGSM 方法的基础上引入动量思想，提出了 MI-FGSM(Momentum Iterative Fast Gradient Sign Method)方法，既稳定了梯度更新方向，又有效地越过了局部极值点。随后，Xie 等[23]基于图像变换手段解决了 I-FGSM 方法的过拟合问题，并将其命名为多样性攻击方法(Diverse Inputs Iterative Fast Gradient Sign Method，DI^2-FGSM)。由于 MI-FGSM 方法与 DI^2-FGSM 方法产生的对抗样本具备良好的迁移性，因此可应用于黑盒攻击中。

目前，研究者们普遍认为投影梯度下降方法(Project Gradient Descent，PGD)[24]是现阶段效果最好的基于梯度攻击的方法。PGD 方法本质上也是 I-FGSM 方法的一种改进，通过加入一层随机化处理，增加了迭代次数，极大改善了攻击效果。当然，PGD 方法还有进一步提升的空间。例如，Sriramanan 等[25]通过向损失函数引入一个松弛项，找到更合适的梯度方向，从而提高了攻击效率。

除以上 FGSM 方法及其变体的攻击方法外，研究者们也提出了其他基于梯度的攻击方法。受显著图(saliency maps)概念[26]的启发，Papernot 等[27]提出了基于雅可比矩阵的显著图攻击方法(Jacobian-based Saliency Map Attack，JSMA)。具体地，首先利用梯度信息计算出对分类结果影响最大的像素位置，然后在该像素上添加扰动，从而得到对抗样本。Cissé 等[28]针对某些情况下不能对损失函数 $L(y,y')$ 进行偏导计算的问题，通过优化手段求得近似梯度，提出了 Houdini 方法。

值得一提的是，Kurakin 等[21]在提出 I-FGSM 方法同时，通过将损失函数中的真实标签替换为目标标签形成了 ILCM(Iterative Least-likely Class Method)方法，实现了由非针对性攻击到针对性攻击的转换。这也是针对性攻击的雏形，许多基于梯度的攻击方法皆可利用这一思路实现针对性攻击。

基于梯度的攻击方法非常直观且容易理解，但其要求攻击者必须了解目标模型的梯度信息。因此，增大攻击者获取梯度信息的难度是防御这类攻击最有效的手段。

3.1.2 基于优化的攻击方法

对抗样本的生成算法本质上是寻找对抗扰动并以此产生有效对抗样本的过程。从攻击角度来看，该扰动越小越好。因此，可将对抗样本的生成算法定义为一个优化问题，并对其进行求解来实现对抗攻击。

基于优化的攻击方法的雏形最早出现在 Szegedy 等[13]在 ICLR 2014 会议上提出的 Box-constrained L-BFGS 方法中。而最经典的基于优化的攻击方法当属 Carlini 等[29]提出的 C&W 方法。该方法自定义了不同的目标函数，并通过实验数据选择出最优目标函数来实现对抗攻击。相比 Box-constrained L-BFGS 方法，C&W 方法可通过改变目标函数中的变量来增加最优解的空间大小，从而显著提高攻击成功率。

与 C&W 方法中的单个目标函数不同，Baluja 等[30]提出的 ATNs(Adversarial Transformation Networks)方法通过优化一个联合目标函数来生成对抗样本。该联合目标函数由两部分组成：一部分要求对抗样本与原图像保持相似；另一部分要求目标模型以高置信度输出错误分类。由于第一部分的存在，ATNs 方法产生的对抗样本比 C&W 方法产生的对抗样本更加自然，因此攻击效果更好。

在 AAAI 2018 会议上，Chen 等[31]将 C&W 方法扩展到 L_1 范式上，并利用弹性网(elastic net)[32]正则化优化技术解决了 L_1 范式中存在的高维特征选择问题，并将该方法命名为 EAD(Elastic-net Attacks to DNNs)方法。相比 C&W 方法，EAD 方法能找到更多有效的攻击扰动，从而具备更好的迁移性。因此，EAD 方法也常常应用于黑盒攻击中。

不同于以上基于优化的攻击方法，Su 等[33]提出了只需修改一个像素点即可攻击成功的 One-pixel 方法。其通过确定需要修改的像素位置，利用差分进化优化算法来得到对抗扰动。这种攻击方法简单高效，也无需目标模型信息，因此可应用于黑盒攻击中。

相比基于梯度的攻击方法，基于优化的攻击方法可以生成扰动更小、更精确的对抗样本，取得更高攻击成功率。

3.1.3 其他白盒攻击方法

除以上两种白盒攻击方法外，研究者们还提出了各种巧妙的白盒攻击方法。例如，基于超平面的分类思想，Moosavi 等[34]通过计算原样本与目标模型分类边界之间的最短距离，提出了 DeepFool 方法。由于直接对分类边界进行了处理，DeepFool 方法得到的对抗扰动比 FGSM 方法[20]更加精确。在 DeepFool 方法的基础上，Moosavi 等[35]提出了通用的对抗扰动方法(Universal Adversarial Perturbations，UAP)。该方法通过计算原样本与多个目标模型分类边界之间的最短距离，生成具备较强泛化能力的对抗扰动。

Laidlaw 等[36]发现，改变原始图像的某个特征功能也能生成有效的对抗样本。例如，模型会将被功能函数改变的飞机图像识别为小狗。由于此方法的本质是添加全局性扰动，相比特定位置的扰动添加方法，此方法更具不可感知性。

大部分基于优化的攻击方法只能针对某个特定的模型和图像对产生对抗样本。Sarkar 等[37]针对该问题给出了部分解决方案，其在优化方法的基础上，通过训练生成器实现可对任意图像自行生成对抗样本的功能，大大提高了对抗样本生成速度。

类似于 JSMA 方法[27]的寻找显著图思想，Phan 等[38]通过内容感知手段，并利用类激活图(Class Activation Map，CAM)[39]筛选出图像的重要特征来生成对抗样本。由于该方法只关注图像的重要特征，因此显著提高了对抗样本的生成速度，实现了低成本和高迁移的对抗攻击。

现有的对抗样本白盒攻击方法大多与图像领域中的研究成果密不可分，如图像特征提取技术的提升可以加快对抗样本的生成速度并改善其攻击效果。随着图像领域研究的不断深入，对抗样本的白盒攻击方法也会得到进一步发展。

3.2 黑盒攻击方法

由于黑盒攻击无须了解模型内部构造及相关信息，且更符合实际攻击情况，因此，黑盒攻击已逐渐成为对抗样本领域

的研究重点。当然,在黑盒攻击的实现过程中,其主要思路首先还是尽可能多地获取目标模型的信息,然后进行攻击。根据信息获取手段的不同,我们将黑盒攻击主要分为基于迁移的攻击方法、基于梯度估算的攻击方法、基于决策的攻击方法、基于 GAN 的攻击方法以及其他黑盒攻击方法 5 个子类。

3.2.1 基于迁移的攻击方法

如果通过白盒攻击获得的对抗样本具有一定迁移性,那么它很可能成为类似模型的对抗样本。而基于迁移的攻击方法正是来源于这种朴素的思想[40]。简单来说,攻击者利用白盒攻击的方法对目标模型的替代模型发起攻击,生成具有迁移性的对抗样本并成功攻击目标模型。因此,获取替代模型及提高对抗样本的迁移性成为基于迁移攻击方法的两大关键因素。

(1)替代模型的获取

查询目标模型以获取相似训练数据集,并利用该数据集训练生成替代模型是获取替代模型的主要思路[40]。因此,如何降低对目标模型的查询成本以及减轻由于训练数据集过小而带来的替代模型过拟合现象,成为获取替代模型的两个重要研究点。

Papernot 等[41]采用蓄水池算法(reservior sampling)[42]保证每个样本数据以相同概率进行扩充。实验结果表明,该方法有效地降低了对目标模型的查询成本。随后,Li 等[43]通过主动学习策略,选择信息量最大的样本进行查询,在提高模型训练质量的同时进一步降低了查询成本。

Xie 等[23]受数据增强策略的启发,通过对训练数据集进行变换处理(如裁剪、旋转等),快速有效地扩充了数据集,从而解决了替代模型的过拟合现象,提高了替代模型质量。与此方法类似,Dong 等[44]提出的转移不变攻击(translation-invariant attack)仅通过平移操作,利用原始图像及其平移变换后的集合生成了训练数据集。Wu 等[45]引入模型注意概念,将特征映射的注意力加权组合作为正则化项,进一步解决了替代模型的过拟合现象。

(2)对抗样本迁移性的提高

如果一个对抗样本能够欺骗多个模型,说明其具备较强的迁移性。Liu 等[46]提出的黑盒攻击方法以及 Dong 等[22]提出的 MI-FGSM 方法皆基于该思路。Li 等[47]发现所集成的多个替代模型之间并不需要具有较大的差异性。他们利用已有的替代模型生成多个不同的虚拟模型并进行集成,从而显著增强了对抗样本的迁移性并降低了替代模型的训练成本。Che 等[48]提出了一种新颖的替代模型集成策略(Serial-Mini-Batch-Ensemble-Attack,SMBEA)。具体来说,该方法将已有的替代模型分成不同的批次,在同一批次内通过引入 3 个集成策略来减轻对特定模型的过度拟合,提高批次内的迁移性;在批次间通过长期梯度记忆算法将先前批次所得的扰动信息保留至随后批次,提高批次间的迁移性。

通过对多个替代模型进行集成来提高迁移性的攻击方法虽然在性能上优于单个替代模型的迁移攻击方法,但此攻击方法受到替代模型数量的限制,过多的替代模型反而会降低攻击效率与成功率。此外,对抗样本在迁移过程中存在一定的损失,不可避免地降低了攻击成功率。

另外,值得注意的是,以上攻击方法或需要查询目标模型,或需要与目标模型类似的训练数据集。当攻击者对训练数据集了解不够充分时,其产生的对抗样本就不具备较好的迁移性。因此,Zhou 等[49]提出一种利用 GAN 合成样本来训练替代模型的方法,摆脱了对实际数据的需求。

3.2.2 基于梯度估算的攻击方法

模型的输出结果一般包括类别标签和置信度两种。通常来说,后者比前者更精确。由 3.2.1 节可知,在基于迁移的攻击方法中,替代模型主要依据类别标签的信息训练产生。那么能否利用置信度产生比基于迁移攻击方法更好的对抗样本?基于梯度估算的攻击方法对此给出了肯定的答案。

基于梯度估算的攻击方法主要通过查询目标模型来获取置信度,之后再进行梯度估算,最后利用估算的梯度结合白盒攻击方法生成对抗样本。与基于迁移的攻击方法相比,基于梯度估算的攻击方法不存在因迁移而造成的攻击成功率损失。然而,由于置信度的获取以及梯度的估算过程往往成本过高,因此需要对其进行优化。

Chen 等[50]提出的 ZOO(Zeroth-Order Optimization)方法是第一个基于梯度估算的攻击方法。在 ZOO 方法中,他们利用零阶优化方法提高了梯度估算的速度,但并未对置信度的获取过程进行优化。随后,Bhagoji 等[51]提出了随机特征分组(random feature grouping)和主成分分析(Principal Component Analysis,PCA)两种策略,与 ZOO 方法相比,其大幅降低了置信度的获取成本。之后,Tu 等[52]提出了一种高查询效率的黑盒攻击通用框架(Autoencoder-based Zeroth Order Optimization Method,AutoZOOM)。该框架采用了一种自适应的随机梯度估算策略和自动编码器,进一步降低了置信度获取成本,并加快了梯度估算的速度。与标准的 ZOO 攻击方法相比,AutoZOOM 方法可在更复杂的数据集上找到对抗样本。而与 ZOO 类方法不同,Ilyas 等[53]利用自然进化策略(Natural Evolutionary Strategies,NES)[54]估算梯度来降低置信度获取成本。在此基础上,Ilyas 等[55]引入了梯度先验(gradient priors)方式,加快了梯度估算速度。

3.2.3 基于决策的攻击方法

在以上两种黑盒攻击方法中,查询目标模型是必不可少的步骤。因此,当对目标模型查询受限时,以上两种攻击将无法成功。而基于决策的黑盒攻击方法通过随机游走的方式成功摆脱了对目标模型查询的依赖,更符合实际攻击情景。简单来说,攻击者首先得到扰动值较大的初始对抗样本,并以此为基础在模型决策边界(对抗性区域与非对抗性区域之间的边界)附近寻找幅度更小的扰动值来获得最终的对抗样本。

Brendel 等[56]首次强调了基于决策的攻击方法是对抗攻击的一个重要类别。之后,对于该攻击方法的研究主要集中在如何确定更小扰动值的搜索方向以及如何提高其搜索效率两个方面。Dong 等[57]通过一种简单有效的协方差矩阵自适应进化策略(Covariance Matrix Adaptation Evolution Strategy,CMA-ES)[58],对决策边界上的搜索方向进行局部几何建模,从而降低了搜索维度,提高了搜索效率。Brunner 等[59]提出一种有偏见的决策边界搜索框架,将搜索的决策边界限制在攻击成功率更高的扰动上,从而找到更好的搜索方向。Shi 等[60]通过探索初始扰动与搜索改进后扰动之间的关系,提出

了自定义搜索决策边界方法(Customized Adversarial Boundary，CAB)。实验证明，新决策攻击方法相比其他决策攻击，能够获取到更小的对抗扰动值。而 Rahmati 等[61]观察到深度神经网络的决策边界在数据样本附近通常有一个小的平均曲率，据此提出了高查询效率的决策攻击(Geometric Decision-based Attack，GeoDA)。

虽然基于决策的攻击方法更接近真实的攻击场景，但由于其攻击成功率完全取决于决策边界估计的准确性，因此该方法产生的对抗样本并不一定有效。

3.2.4 基于 GAN 的攻击方法

与上述 3 种黑盒攻击方法截然不同的是，基于 GAN 的攻击方法基本上无须了解目标模型的相关信息即可生成对抗样本。生成对抗网络(Generative Adversarial Network，GAN)[62]的主要功能是生成逼真的合成图像，因此，很自然地想到能否利用 GAN 生成攻击成功率高且视觉效果更好的对抗样本[63]。如图 5 所示，基于 GAN 的攻击方法主要包括 3 个部分：生成器、判别器和模型。其中，前两部分来自 GAN，一个用于生成对抗样本，另一个用于判断对抗样本与原始图像之间的差异L_{GAN}；而目标模型主要用于判断生成器生成的对抗样本的预测标签与真实标签之间的差异L_{Adv}。GAN 的训练目的是使L_{GAN}和L_{Adv}之间达到一个平衡。若对抗样本真实自然，足以使模型分类错误，且生成器对任意的输入图像都可稳定生成相应的对抗样本，则 GAN 训练完成，否则，生成器会根据L_{GAN}和L_{Adv}的反馈来调整相关参数进行更新。

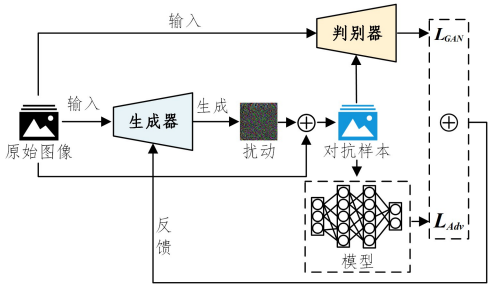

图 5 基于 GAN 的攻击方法的基本原理[63]

Fig. 5 Basic principle of adversarial attack based GAN[63]

Xiao 等[63]在 IJCAI 2018 会议上首次提出一种完整的基于 GAN 的对抗攻击方法，并将该方法命名为 AdvGAN。之后，Jandial 等[64]将生成器的输入由原始图像改为其潜在的特征向量，这一简单的改动不仅减少了 GAN 的训练时间，而且显著提高了攻击成功率。

随着 GAN 研究的不断深入，基于 GAN 的攻击方法也得到了相应的改进。例如，Zhao 等[65]和 Liu 等[66]分别基于 GAN 的改进版 WGAN 和 SN-GAN 提出了新的对抗样本，生成算法。基于 GAN 的攻击方法最大的优点在于，对抗生成网络一旦训练完毕，就可稳定生成对抗样本，实现对抗攻击，实用性更高。

3.2.5 其他黑盒攻击方法

以上 4 种黑盒攻击方法各有优缺点，那么将这四者相结合是否能获得更好的攻击效果？研究者们对此进行了回答。例如，Cheng 等[67]提出的 P-RGF(Prior-Guided Random Gradient-free)方法首先利用了基于迁移的攻击方法对目标模型进行先验查询，再进行梯度估算，实现了以较少的查询次数达到更高攻击成功率的目的。再如，Suya 等[68]提出的混合批处理攻击(hybrid batch attacks)将基于迁移的攻击方法与基于梯度估算的攻击方法结合起来，并利用种子优先级策略实现了批处理攻击，降低了查询复杂度，提高了攻击成功率。此外，文献[59，67]的研究成果表明，将基于迁移的攻击方法与基于决策的攻击方法相结合，可以达到目前为止最好的攻击效果。

还有研究者从其他角度提出了有效的黑盒攻击方法。例如，Co 等[69]提出的程序噪声(procedural noise)黑盒攻击方法，其利用程序噪声可生成自然纹理这一背景，借助贝叶斯优化(Bayesian optimization)手段[70]以较少的迭代次数生成了更真实的对抗样本。

3.3 攻击方法小结

为了便于读者理解，本文将上述对抗攻击方法在表 1 中进行了归纳整理，并在图 6 中以思维导图的模式展现了各类对抗攻击方法之间的联系及其发展情况。从图 6 中可以看出，虽然研究者们从各种角度出发研究了对抗攻击的攻击方法，但我们仍然能够依据具体的对抗样本生成方法对其进行分类，并在每个攻击类别中找到各攻击方法之间的关联性。

表 1 对抗攻击方法总结

Table 1 Summary of adversarial attack methods

攻击方式	具体分类	攻击原理	方法代表
白盒攻击	基于梯度的攻击方法	利用目标模型的梯度信息生成对抗样本	[20-25，27-28]
	基于优化的攻击方法	将对抗样本的生成算法定义为一个优化问题，对其进行优化求解	[13，29-31，33]
	其他白盒攻击方法	利用其他思想生成对抗样本，如超平面分类思想、图像特征等	[34-38]
黑盒攻击	基于迁移的攻击方法	获取单个替代模型，降低对目标模型的查询成本以及减轻替代模型过拟合现象	[23，40-41，43-45]
		对多个替代模型进行集成来提高对抗样本的迁移性	[22，46-49]
	基于梯度估算的攻击方法	查询目标模型进行梯度估算后，利用白盒攻击方法生成对抗样本	[50-53，55]
	基于决策的攻击方法	在得到扰动较大的初始对抗样本基础上，随机搜索决策边界来减小扰动	[56-57，59-61]
	基于 GAN 的攻击方法	利用生成对抗网络 GAN 生成对抗样本	[63-66]
	其他黑盒攻击方法	将以上 4 种黑盒攻击方法相结合，或从其他角度出发实现黑盒攻击	[67-69]

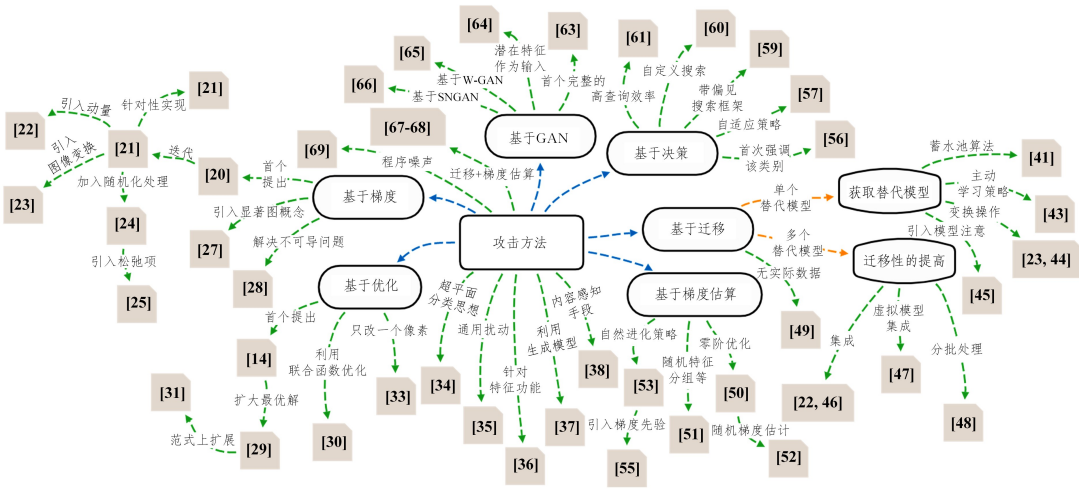

图 6 对抗攻击的联系和发展

Fig. 6 Connection and development of adversarial attacks

4 对抗样本的防御方法

对抗样本的存在不仅严重威胁了深度神经网络模型的安全，甚至使人一度怀疑深度神经网络的实用性。幸运的是，随着对抗样本研究的不断深入，许多有效的防御方法被相继提出。针对防御方法处理对抗样本的不同思路，我们将现有防御方法大致分成梯度遮蔽、数据预处理、对抗样本检测以及对抗训练 4 种类型。下面我们将根据对抗攻击防御的强度，从弱到强依次对这 4 种类型进行介绍。

4.1 梯度遮蔽方法

从第 3 节的介绍中可以发现，模型的梯度信息在对抗样本生成过程中十分重要。因此，防御对抗攻击的一个直接思路就是通过保护梯度信息来增加对抗样本生成难度，该方法被称为梯度遮蔽。

梯度遮蔽防御方法的典型代表就是 Papernot 等[71]提出的防御性蒸馏方法(defensive distillation)。该方法基于原始模型(original model)的分类结果，采用蒸馏法(distillation approach)[72]重新训练 DNN，并引入“温度”概念进行设置，得到蒸馏模型(distilled model)，从而实现了梯度的遮蔽。随后，Papernot 等[73]将不确定性引入模型训练中，对防御性蒸馏方法进行了扩充，进一步遮蔽了模型梯度信息。然而，梯度遮蔽方法并不能有效地防御无需梯度信息的对抗攻击，如 C&W 攻击[29]。

4.2 数据预处理方法

当我们无法阻止攻击者生成对抗样本时，一种直接的想法就是使对抗样本失效。由第 2 节可知，对抗样本本质上是向原始数据中添加对抗噪声生成的。因此，如果添加的对抗噪声失效，那么模型就不会对对抗样本进行错误分类，这就是数据预处理的基本思路。简单来说，数据预处理的防御方法是在图像数据输入模型前对其进行有效处理来降低对抗噪声带来的影响，如图 7 所示。

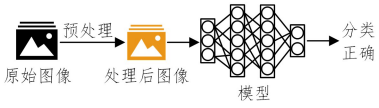

图 7 数据预处理的基本原理

Fig. 7 Principle of data pre-processing

数据预处理有两种方式，其中一种就是对输入数据进行压缩或者变换等处理来减轻对抗噪声带来的影响。Dziugaite 等[74]首次评估了 JPG 压缩算法对图像分类的影响，发现在对输入数据进行 JPG 压缩后可以逆转对抗噪声带来的威胁。而 Guo 等[75]证明了将图像输入神经网络之前应用图像变换(如位深度减少、总方差最小化等)，也可减轻对抗噪声带来的影响。

另一种数据预处理方式是在数据输入模型前或在输入过程中对其进行去噪处理。Gu 等[76]首次提出使用去噪自动编码器(Denoising Auto-encoders，DAEs)[77]去除对抗噪声来达到防御效果。而 Osadchy 等[78]通过将中值滤波器(median filter)、均值滤波器(averaging filter)以及高斯低通滤波器(gaussian lowpass filter)等组合后对输入数据进行去噪处理，在像素层面上显著消除了对抗噪声带来的影响。然而，由于 DNN 是一个高维度的网络拓扑结构，即便是非常微小的差异经过 DNN 处理后也会被放大，从而产生巨大影响。因此，去噪处理后的残余对抗噪声仍能导致模型的分类错误。针对该问题，Liao 等[79]提出高级表示指导去噪器(High Level Representation Guided Denoiser，HGD)，解决了残余噪声放大问题。此外，Prakash 等[80]通过像素偏转进行像素值的重新分配，局部破坏了对抗噪声，并基于小波变换进行去噪处理，有效地去除了对抗噪声的影响。Xie 等[81]通过在 DNN 网络模型中加入特征去噪(feature denoising)模块，对图像特征层面的数据进行去噪处理，从而更好地消除了对抗噪声。

然而，无论哪种数据预处理方法，都无法完全去除对抗噪声带来的影响。因此，对抗样本在经过预处理后，仍可能使目标模型分类错误。

4.3 对抗样本检测方法

在无法阻止对抗样本生成的情况下，检测出对抗样本的存在是另一种有效防御对抗攻击的方法。迄今为止，研究者们提出了各种对抗样本检测方法。根据检测机制的不同，可将其细分为以下两个子类：一种是差异性判断方法，即通过对抗样本与正常样本之间的差异进行不一致性判断；另一种是检测器方法，即训练额外的检测器来检测对抗样本，如图 8 所示。

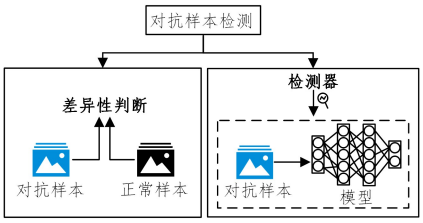

图 8 对抗样本检测的基本原理

Fig. 8 Basic principle of detecting adversarial examples

(1)差异性判断方法

Xu 等[82]提出的特征压缩手段(feature squeezing)就是一种典型的差异性判断检测方法。该方法利用不同的图像压缩算法(如颜色位深压缩、空间平滑压缩等)对输入的图像数据进行处理，通过判断压缩前后图像预测结果之间的差异来判断该输入图像是否为对抗样本。此方法基本能够防御众多典型的白盒攻击方法，如 FGSM 攻击[20]、Deepfool 攻击[34]、JSMA 攻击[27]和 C&W 攻击[29]等。Tian 等[83]根据对抗样本通常对图像变换操作更为敏感的特点，提出一种基于图像变换的对抗样本检测方法。该方法可以有效防御 C&W 攻击[29]。Pang 等[84]提出了先验训练程序和阈值测试策略相结合的方法，将对抗样本与原始数据有效区分开来。而 Yang 等[85]发现对抗样本与原始样本在特征属性上存在显著差异，引入检测框架 ML-LOO 以对特征属性进行尺度估计，并依据检测阈值来判断其是否为对抗样本。

(2)检测器方法

Zheng 等[86]通过无监督学习方法捕获 DNN 分类器中神经元之间内在的关联性，并基于对抗样本对该关联性的影响训练检测器。Ma 等[87]利用 DNN 中的来源不变量(Provenance Invariants，PI)和激活值不变量(Value Invariants，VI)构造检测器。若输入的图像数据引起 PI 或者 VI 数值的变化，则可判断该数据为对抗样本。Cintas 等[88]利用非参数扫描统计(Non-parametric Scan Statistics，NPSS)[89]来评估任何给定输入节点激活子集的异常性，通过检测该异常来判断输入的图像数据是否为对抗样本。

虽然基于对抗样本检测的防御方法能够大幅降低对抗攻击的成功率，但也存在一个致命缺陷，即当攻击者了解到目标模型的检测原理后，会相应地调整攻击策略，从而绕开检测机制，攻击成功。

4.4 对抗训练方法

数据预处理方法和对抗样本检测方法的指导思想在本质上都是一致的，即对抗样本不能作为目标模型的输入数据。但以上两种防御方法并不能百分之百成功，总存在对抗样本成功输入到目标模型的情况。因此，增强目标模型自身的鲁棒性显得尤为重要，而对抗训练(adversarial training)是目前最有效的途径[20]。如图 9 所示，对抗训练的基本思想是将对抗样本作为训练数据集的一部分重新对模型进行训练，从而获得防御能力[90]。

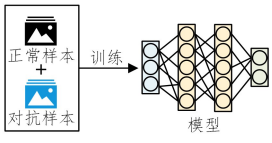

图 9 对抗训练的基本原理

Fig. 9 Basic principle of adversarial training

Goodfellow 等[20]首先通过实验发现，利用对抗样本对模型进行训练可降低模型对对抗样本的分类错误率，该发现成为了对抗训练方法的雏形。之后，Madry 等[24]在提出 PGD 攻击方法的同时，明确了对抗训练这一概念，利用 PGD 攻击方法生成的对抗样本对模型进行训练。本质上，对抗训练是一个最小最大的优化过程，可表示为：

$$\min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\max_{\|\eta\|\leqslant\varepsilon}L(\theta;x+\eta,y) \tag{1}$$

其中，θ 是神经网络模型的参数；η 是对抗扰动；L 为损失函数；(x,y)是某分布在 $\mathcal{D}$ 中的一对数据和对应的标签。对抗扰动 η 被限制在 ε 范围内($\varepsilon>0$)。从式(1)可以看出，内部 max 过程的目的是在当前模型基础上，寻找最佳的扰动 η 来生成对抗样本，使分类损失 L 最大化，从而实现对抗性。而外部 min 过程的目的是训练神经网络模型找到合适的模型参数 θ，使模型具有鲁棒性，适应这种对抗扰动，增强模型防御能力。

随着相关研究的不断深入，对抗训练防御方法的一些问题逐渐显现出来，如训练成本过高、模型准确性与鲁棒性难以权衡等。针对以上问题，研究者们提出了相应的解决方案。

(1)降低训练成本

对抗样本的生成是对抗训练的基础。然而，此过程产生的计算以及时间开销通常远大于训练模型的开销，导致该防御方法的成本难以承受。对应到最小最大的优化过程，解决该问题的关键在于如何快速求得 max 过程中的对抗扰动 η，即如何高效生成对抗样本。研究者们通过改变对抗训练过程中的计算步骤，在不影响对抗训练效果的前提下，加快了对抗训练速度，降低了成本。例如，Shafahi 等[91]提出的快速对抗训练方法 FreeAT(Free Adversarial Training)的主要思想是将旧对抗样本训练产生的梯度信息直接用于新对抗样本的产生，从而降低了生成对抗样本的计算开销。与 FreeAT 不同，Zhu 等[92]提出的 FreeLB(Free Large-Batch)方法并不通过对抗样本的训练获得新的梯度信息，而是直接在旧梯度信息上求得新梯度信息，从而减少了梯度计算步骤。而 Zhang 等[93]则通过限制前向传播和反向传播的次数来加快训练速度，该方法被称为 YOPO(You Only Propagate Once)。

(2)模型准确性和鲁棒性的权衡

研究表明，对抗训练在增强模型鲁棒性同时，会降低模型对正常图像数据的分类准确度[94]。因此需要对模型的准确性与鲁棒性进行探索，权衡两者之间的关系，即在保证 max 过程的对抗性的前提下，还需对 min 过程的鲁棒性进行优化。Zhang 等[95]提出的 TRADES 方法将鲁棒性误差分为自然误差与边界误差两部分，权衡了正常图像数据的预测准确性和模型的鲁棒性。而 Wang 等[96]的 MART(Misclassification Aware Adversarial Training)方法以及 Mao 等[97]的 TLA(Triplet Loss Adversarial)方法均采用正则化的手段对鲁棒性的误差判断进行优化来改进对抗训练。Li 等[98]通过将三元组损失(triplet loss)作为正则化项纳入对抗训练中，提出了 AT^{2}L(Adversarial Training with Triplet Loss)方法，在不牺牲模型分类准确性的情况下，显著提高了模型鲁棒性。

此外，Liu 等[99]提出的对抗训练方法以及 Wang 等[100]提出的 OAT(Once-for-all Adversarial Training)方法通过对式(1)从各方面进行优化，实现了分类准确性和模型鲁棒性之间的平衡。

研究表明，对抗训练是目前最有效的防御方法。然而，

由于这种防御方法是非自适应性的，当有全新的攻击出现时，对抗训练后的模型仍有极大的可能被攻击成功。

4.5 防御方法总结

为了方便读者更好地理解，我们将上述防御方法在表 2 中进行了归纳整理，并在图 10 中以思维导图的模式展现了各类防御方法之间的关系及其发展情况。从图 10 可以看出，防御方法层出不穷，但更多研究者将注意力集中在对抗训练防御方法的研究和改进上。

表 2 防御方法总结

Table 2 Summary of defense methods

防御方法	防御原理	代表	缺点
梯度遮蔽方法	对模型的梯度信息进行保护	[71,73]	无法防御无需梯度信息的对抗攻击，如 C&W 攻击
数据预处理方法	对输入的样本数据进行压缩、变换等处理来减弱对抗噪声的影响	[74-75]	只是在一定程度上降低了对抗噪声带来的影响，并不能 100% 消除，仍存在分类错误的可能性
	对输入的样本数据进行去噪处理	[76,78-81]	
对抗样本检测方法	差异性判断方法：根据对抗样本与正常样本之间的差异，进行不一致性判断	[82-85]	一旦攻击者了解了检测原理，就会调整攻击策略以绕开检测
	检测器方法：训练额外的检测器来检测对抗样本	[86-88]	
对抗训练方法	将对抗样本与正常样本一起对模型进行训练，提高其鲁棒性	[24,91-93,95-100]	该方法是非自适应性的，只能防御已知的攻击，对全新的攻击防御效果差甚至难以防御

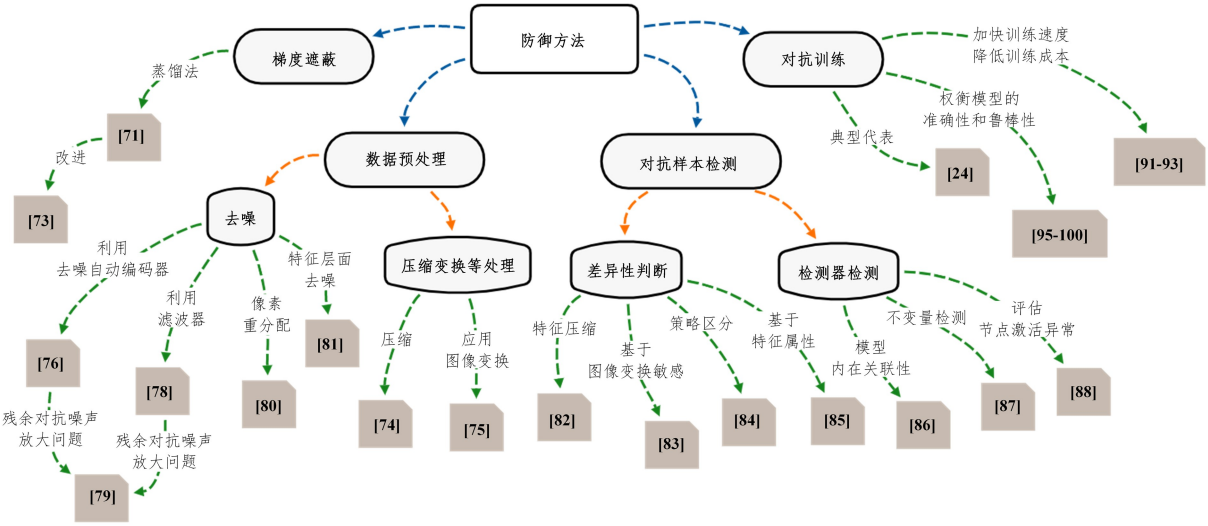

图 10 防御方法的联系与发展

Fig. 10 Connection and development of defense methods

5 物理世界中的对抗样本

由于距离、拍摄角度、光照条件以及标志遮挡等因素的影响，第 3 节中介绍的诸多对抗攻击方法在真实物理世界中攻击效果并不理想。很多研究者对此展开了研究。

Kurakin 等[21]首次讨论了物理世界(physical world)中的对抗样本，发现物理世界中存在着攻击效果明显的对抗样本。Sharif 等[101]提出一种针对面部生物特征识别系统(face biometric systems)的物理攻击方法。攻击者只需配戴添加了对抗扰动的眼镜框即可躲避人脸检测或假冒他人。Xu 等[102]也提出一种对抗人体检测器(person detectors)的物理攻击方法。攻击者只需身着带有对抗性图案的 T 恤就能成功躲避检测。此外，在物体识别分类方面，Eykholt 等[103]提出了针对停车标志(stop sign)特点的 RP2 物理攻击方法，该方法只需在标志上粘贴人为精心构造的黑白条块便可导致自动驾驶识别系统出错。而文献[104]提出的对抗补丁 AdvPatch(Adversarial Patch)可以在打印后放置到任一物理场景中。即使该补丁很小，图像分类器也会对这一物理场景照片中的物体分类错误。

然而，以上提出的对抗性物体有时过于明显，如人们很容易察觉文献[101]中颜色怪异的眼镜框。因此，物理世界中的对抗样本不仅需要提高攻击成功率，还需要具备不可感知性。Luo 等[105]考虑到人类感知系统对不同像素的敏感程度，引入了一种新的距离度量算法。该算法通过评估人眼对图像像素的敏感度来添加不同的扰动，从而产生高度不可感知的对抗样本。Liu 等[106]则利用 GAN 技术生成对抗补丁，即利用目标模型对对抗补丁的敏感度，引入一种注意力机制来预测放置补丁的关键攻击区域，从而生成更真实、更具攻击性的对抗样本。Jan 等[107]提出的方案也应用了 GAN 技术，但该方案主要研究图像从数字世界到物理世界的转变对对抗攻击的影响。而 Duan 等[108]提出的对抗伪装方法 AdvCam(Adversarial Camouflage)则注重对抗样本与原始图像的匹配度，通过添加与图像风格相似的扰动来生成不易被察觉的对抗样本。

随着各种针对物理世界的对抗攻击方法的提出，研究者们注意到了另一个问题：这些物理对抗攻击往往不易在物理世界中部署。例如，攻击者需要将对抗补丁粘贴到难以触及

的交通标志上。针对该问题，研究者们提出了多种新型物理攻击方法。Zhou 等[109]在帽子上安装特制的照射装置，将对抗扰动以红外线的方式照射在面部上，悄然绕过或改变人脸识别系统(face recognition systems)的检测。类似地，Shen 等[110]提出基于可见光的物理攻击方法(Visible Light-based Attack，VLA)，通过将对抗性可见光投射到人脸上，实现了对人脸识别系统的攻击。而 Duan 等[111]提出的攻击方法 Adv-LB(Adversarial Laser Beam)则通过操纵对抗性激光束的物理参数，将对抗性激光束照射在物体上来执行攻击。Sayles 等[112]也通过控制照射在物体上的光，使得相机拍摄的照片直接为对抗样本。

与这些基于光线控制的物理攻击方法不同，Nguyen 等[113]利用投影仪将对抗扰动投射到面部，从而达到冒充目标人物或者逃避人脸识别系统识别的目的。Lovisotto 等[114]针对自动驾驶场景下对交通标志的识别，提出了短期对抗扰动(Short-lived Adversarial Perturbations，SLAP)方法。该攻击方法同样利用了投影技术，将特制的对抗扰动投射到停车标志上实现对抗攻击。

这类利用光线或者投影的物理攻击方法具备实时可控性。根据需要打开照射或投影装置即可实现攻击，关闭后也不会留下攻击痕迹，具备良好的隐蔽性。

要生成真实物理世界中的对抗样本，所添加的扰动在视觉效果上应该是自然的，并且与图像周围环境具有强烈的感知相关性，同时还要保证其添加的位置位于图像中的感知敏感位置。此外，易于部署且具备实时可控性的对抗扰动更符合真实物理世界的复杂性。图 11 给出了物理世界中几种具有代表性的对抗样本生成方式。从上到下、从左到右的黑色虚框分别代表显眼贴纸补丁类攻击方式、不可感知类攻击方式、光线照射类攻击方式以及投影类攻击方式。

图 11 物理世界中的对抗样本

Fig. 11 Adversarial examples in physical world

目前，物理世界中的对抗样本研究受两方面的限制。一方面，现有的物理世界攻击方法大多是在极小的测试集(如仅用三四个不同的交通标志)上进行评估，这会导致攻击方法存在通用性低的问题；另一方面，物理世界中对抗样本的研究常常涉及各类硬件设备以及大规模物理图像的处理，成本较高。

6 面临的挑战及未来研究方向

对抗攻击的发展是对抗防御研究进展的基础，有效的攻击对神经网络模型鲁棒性的评估也至关重要。目前虽然在对抗样本的攻击与防御上取得了显著进展，但仍有诸多问题亟待解决，主要体现在以下 4 个方面。

(1)对抗样本出现的本质原因：对于对抗样本为何存在，许多研究者提出了假设性解释，但仁者见仁，智者见智。此问题至今还是一个开放性问题，需要进行更深入的探讨。当然，这还涉及模型的可解释性问题。研究模型的可解释性可以加深对模型内部逻辑结构的理解，有助于充分解释对抗样本存在的原因。

(2)物理世界中的对抗样本：由于物理世界的复杂性和多变性，已有物理攻击方法生成的对抗样本会受到各种环境因素的影响，因稳定性不够而导致攻击失败。因此，需要对现实生活中的对抗样本生成技术进行进一步研究，以获得既自然又稳健的对抗样本。

(3)模型鲁棒性的评估：新攻击方法的提出可以成功攻破原有防御方法，而针对新防御方法，攻击者可以找到新攻击方法。对抗样本的攻击与防御仿佛一场无止境的战争，互相博弈。因此，我们需要一种全面的模型鲁棒性评估方法来判断攻击以及防御方法的效果。现有的对模型鲁棒性的评估框架或多或少都受到一些条件的限制，缺少适用于所有攻击以及防御方法的统一评估框架。而由于每种攻击或防御方法提出时的实验条件也不尽相同，这给统一评估框架的建立带来了极大的困难。

(4)对抗样本的合理利用：对抗样本因给人工智能模型带来安全威胁而引起研究者们的重视。但事物具备两面性，已有研究工作通过将对抗样本应用在验证码中，有效防御了机器对验证码的恶意破解[115]。因此，在未来的研究工作中，如何合理利用对抗样本，并将其“变废为宝”是一个值得探讨的研究方向。

结束语 随着深度学习应用领域的不断扩大，与其安全性密切相关的对抗样本的研究工作也需要不断跟进与总结。本文针对现有研究综述存在的问题，归纳整理了图像对抗样本攻击与防御方法，对其进行了分类并给出了不同类别的特点及存在的联系。同时介绍了图像对抗攻击在物理世界中的情况，并总结分析了图像对抗样本领域仍面临的挑战及未来的研究方向。

参 考 文 献

[1] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge，MA：MIT Press，2012：1106-1114.

[2] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]// Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ：IEEE，2016：770-778.

[3] REN S Q，HE K M，GIRSHICK R B，et al. Faster r-cnn：towards real-time object detection with region proposal networks

[C]// Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2015: 91-99.

[4] MOHAMED A R, DAHL G E, HINTON G E. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(1): 14-22.

[5] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based largevocabulary speech recognition [C]// Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2016: 4945-4949.

[6] BOJARSKI M, TESTA D D, DWORAKOWSKI D, et al. End to end learning for self-driving cars [J]. arXiv: 1604.07316, 2016.

[7] TIAN Y C, PEI K X, JANA S, et al. DeepTest: Automated testing of deep-neural-network-driven autonomous cars [C]// Proceedings of the 40th IEEE International Conference on Software Engineering (ICSE). Piscataway, NJ: IEEE, 2018: 303-314.

[8] LOPES A T, AGUIAR E D, SOUZA A F D, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order [J]. Pattern Recognition, 2017, 61: 610-628.

[9] SUN Y, WANG X G, TANG X O. Deep convolutional network cascade for facial point detection [C]// Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2013: 3476-3483.

[10] MEI S K, ZHU X J. Using machine teaching to identify optimal training-set attacks on machine learners [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2015: 2871-2877.

[11] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models [C]// Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2017: 3-18.

[12] JI Y J, ZHANG X Y, WANG T. Backdoor attacks against learning systems [C]// Proceedings of the 5th IEEE Conference on Communications and Network Security (CNS). Piscataway, NJ: IEEE, 2017: 1-9.

[13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C]// Proceedings of the 2nd International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2014.

[14] AKHTAR N, MIAN A S. Threat of adversarial attacks on deep learning in computer vision: a survey [J]. IEEE Access, 2018, 6: 14410-14430.

[15] PAPERNOT N, MCDANIEL P D, SINHA A, et al. SoK: Security and privacy in machine learning [C]// Proceedings of the 3th IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2018: 399-414.

[16] 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.

[17] PAN W W, WANG X Y, SONG M L, et al. Overview of adversarial sample generation technology [J]. Journal of Software, 2020, 31(01): 67-81.

[18] WANG K D, YI P. Overview of research on model robustness in artificial intelligence confrontation environment [J]. Journal of Information Security, 2020, 5(3): 13-22.

[19] ZHANG T, YANG K W, WEI J H, et al. Survey on Detecting and Defending Adversarial Examples for Image Data [J/OL]. Journal of Computer Research and Development [2021-08-08]. http://kns.cnki.net/kcms/detail/11.1777.TP.20210607.1630.004.html.

[20] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2015.

[21] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world [J]. arXiv: 1607.02533, 2017.

[22] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum [C]// Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 9185-9193.

[23] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity [C]// Proceedings of the 32nd IEEE Conference on ComputerVision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 2730-2739.

[24] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2018.

[25] SRIRAMANAN G, ADDEPALLI S, BABURAJ A, et al. Guided adversarial attack for evaluating and enhancing adversarial defenses [C]// Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2020.

[26] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps [J]. arXiv: 1312.6034, 2014.

[27] PAPERNOT N, MCDANIEL P D, JHA S, et al. The limitations of deep learning in adversarial settings [C]// Proceedings of the 1th IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2016: 372-387.

[28] CISSÉ M, ADI Y, NEVEROVA N, et al. Houdini: fooling deep structured prediction models [J]. arXiv: 1707.05373, 2017.

[29] CARLINI N, WAGNER D A. Towards evaluating the robustness of neural networks [C]// Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2017: 39-57.

[30] BALUJA S, FISCHER I. Adversarial transformation networks: learning to generate adversarial examples [J]. arXiv: 1703.09387, 2017.

[31] CHEN P Y, SHARMA Y, ZHANG H, et al. EAD: Elastic-net

attacks to deep neural networks via adversarial examples[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park,CA:AAAI,2018:10-17.

[32] ZOU H,HASTIE T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society:Series B (Statistical Methodology),2005,67(2):301-320.

[33] SU J W,VARGAS D V,SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation,2019,23(5):828-841.

[34] MOOSAVI-DEZFOOLI S M,FAWZI A,FROSSARD P. DeepFool:A simple and accurate method to fool deep neural networks[C]// Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2016:2574-2582.

[35] MOOSAVI-DEZFOOLI S M,FAWZI A,FAWZI O,et al. Universal adversarial perturbations[C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2017:86-94.

[36] LAIDLAW C,FEIZI S. Functional adversarial attacks[C]// Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2019:10408-10418.

[37] SARKAR S,BANSAL A,MAHBUB U,et al. UPSET and ANGRI:Breaking high performance image classifiers [J]. arXiv:1707.01159,2017.

[38] PHAN H,XIE Y,LIAO S Y,et al. CAG:A real-time low-cost enhanced-robustness high-transferabilitycontent-aware adversarial attack generator[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI,2020:5412-5419.

[39] ZHOU B L,KHOSLA A,LAPEDRIZA A,et al. Learning deep features for discriminative localization[C]// Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2016:2921-2929.

[40] PAPERNOT N,MCDANIEL P D,GOODFELLOW I J,et al. Practical black-box attacks against machine learning[C]// Proceedings of the 12th ACM Asia Conference on Computer and Communications Security (AsiaCCS). New York: ACM, 2017: 506-519.

[41] PAPERNOT N,MCDANIEL P D,GOODFELLOW I J. Transferability in Machine Learning:from phenomena to black-box attacks using adversarial samples [J]. arXiv:1605.07277,2016.

[42] VITTER J S. Random sampling with a reservoir[J]. ACM Transactions on Mathematical Software (TOMS),1985,11(1): 37-57.

[43] LI P C,YI J F,ZHANG L J. Query-efficient black-box attack by active learning[C]// Proceedings of the 18th IEEE International Conference on Data Mining (ICDM). Piscataway, NJ: IEEE, 2018:1200-1205.

[44] DONG Y P,PANG T Y,SU H,et al. Evading defenses to transferable adversarial examples by translation-invariant attacks [C]// Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE,2019:4312-4321.

[45] WU W B,SU Y X,CHEN X X,et al. Boosting the transferability of adversarial samples via attention[C]// Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2020:1158-1167.

[46] LIU Y P,CHEN X Y,LIU C,et al. Delving into transferable adversarial examples and black-boxattacks[C]// Proceedings of the 5th International Conference on Learning Representations (ICLR). La Jolla,CA:LCLR,2017.

[47] LI Y W,BAI S,ZHOU Y Y,et al. Learning transferable adversarial examples via ghost networks[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). Menlo Park,CA:AAAI,2020:11458-11465.

[48] CHE Z H,BORJI A,ZHAI G T,et al. A newensemble adversarial attack powered by long-term gradient memories[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). Menlo Park,CA:AAAI,2020:3405-3413.

[49] ZHOU M Y,WU J,LIU Y P,et al. DaST:Data-free substitute training for adversarial attacks[C]// Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2020:231-240.

[50] CHEN P Y,ZHANG H,SHARMA Y,et al. ZOO:Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]// Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec). New York:ACM,2017:15-26.

[51] BHAGOJI A N,HE W,LI B,et al. Exploring the space of black-box attacks on deep neural networks[C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla,CA:LCLR,2018.

[52] TU C C,TING P,CHEN P Y,et al. AutoZOOM:Autoencoder-based zeroth order optimization method for attacking black-box neural networks[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019:742-749.

[53] ILYAS A,ENGSTROM L,ATHALYE A,et al. Black-box adversarial attacks with limited queries and information[C]// Proceedings of the 35th International Conference on Machine Learning (ICML). New York:ACM,2018:2142-2151.

[54] WIERSTRA D,SCHAUL T,GLASMACHERS T,et al. Natural evolution strategies [J]. Journal of Machine Learning Research,2014,15(1):949-980.

[55] ILYAS A,ENGSTROM L,MADRY A. Prior Convictions: Black-box adversarial attacks with bandits and priors[C]// Proceedings of the 7th International Conference on Learning Representations (ICLR). La Jolla,CA:LCLR,2019.

[56] BRENDEL W,RAUBER J,BETHGE M. Decision-based adversarial attacks:reliable attacks against black-box machine learning models[C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR,2018.

[57] DONG Y P, SU H, WU B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]// Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 7714-7722.

[58] HANSEN N, OSTERMEIER A. Completely derandomized self-adaptation in evolution strategies[J]. Evolutionary computation, 2001, 9(2): 159-195.

[59] BRUNNER T, DIEHL F, LE M T, et al. Guessing Smart: Biased sampling for efficient black-box adversarial attacks[C]// Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 4957-4965.

[60] SHI Y C, HAN Y H, TIAN Q. Polishing decision-based adversarial noise with a customized sampling[C]// Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 1027-1035.

[61] RAHMATI A, MOOSAVI-DEZFOOLI S M, FROSSARD P, et al. GeoDA: A geometric framework for black-box adversarial attacks[C]// Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 8443-8452.

[62] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] // Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2014: 2672-2680.

[63] XIAO C W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann, 2018: 3905-3911.

[64] JANDIAL S, MANGLA P, VARSHNEY S, et al. AdvGAN++: Harnessing latent layers for adversary generation[C]// Proceedings of the 17th IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 2045-2048.

[65] ZHAO Z L, DUA D, SINGH S. Generating natural adversarial examples[C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2018.

[66] LIU X Q, HSIEH C. Rob-GAN: Generator, discriminator, and adversarial attacker[C]// Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 11234-11243.

[67] CHENG S Y, DONG Y P, PANG T Y, et al. Improving black-box adversarial attacks with a transfer-based prior[C]// Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2019: 10932-10942.

[68] SUYA F, CHI J F, EVANS D, et al. Hybrid batch attacks: Finding black-box adversarial examples with limited queries[C]// Proceedings of the 29th USENIX Security Symposium (USENIX Security). Berkeley, CA: USENIX Association, 2020: 1327-1344.

[69] CO K T, MUÑOZ-GONZÁLEZ L, MAUPEOU S D, et al. Procedural noise adversarial examples for black-box attacks on deep convolutional networks[C]// Proceedings of the 26th ACM Conference on Computer and Communications Security (CCS). New York: ACM, 2019: 275-289.

[70] SNOEK J, LAROCHELLE H, ADAMS R P. Practical bayesian optimization of machine learning algorithms[C]// Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2012: 2960-2968.

[71] PAPERNOT N, MCDANIEL P D, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// Proceedings of the 37th IEEE Symposium on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2016: 582-597.

[72] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network [J]. Computer Science, 2015, 14(7): 38-39.

[73] PAPERNOT N, MCDANIEL P D. Extending defensive distillation [J]. arXiv: 1705.05264, 2017.

[74] DZIUGAITE G K, GHAHRAMANI Z, ROY D M. A study of the effect of jpg compression on adversarial images [J]. arXiv: 1608.00853, 2016.

[75] GUO C, RANA M, CISSÉ M, et al. Countering adversarial images using input transformations[C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2018.

[76] GU S X, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2015.

[77] CHEN M M, WEINBERGER K Q, SHA F, et al. Marginalized denoising auto-encoders for nonlinear representations[C]// Proceedings of the 31st International Conference on Machine Learning (ICML). New York: ACM, 2014: 1476-1484.

[78] OSADCHY M, HERNANDEZ-CASTRO J, GIBSON J S, et al. No bot expects the deepcaptcha! introducing immutable adversarial examples, with applications to captcha generation [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2640-2653.

[79] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoiser [C]// Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 1778-1787.

[80] PRAKASH A, MORAN N, GARBER S, et al. Deflecting adversarial attacks with pixel deflection[C]// Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 8571-8580.

[81] XIE C H, WU Y X, MAATEN L, et al. Feature denoising for improving adversarial robustness[C]// Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 501-509.

[82] XU W L, EVANS D, QI Y J. Feature squeezing: Detecting adversarial examples in deep neural networks[C]// Proceedings of

the 25th Network and Distributed System Security Symp (NDSS). Reston, VA: ISOC, 2018.

[83] TIAN S X, YANG G L, CAI Y. Detecting adversarial examples through image transformation[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2018: 4139-4146.

[84] PANG T Y, DU C, DONG Y P, et al. Towards robust detection of adversarial examples[C]// Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2018: 4584-4594.

[85] YANG P, CHEN J B, HSIEH C J, et al. ML-LOO: Detecting adversarial examples with feature attribution[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2020: 6639-6647.

[86] ZHENG Z H, HONG P Y. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks[C]// Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2018: 7924-7933.

[87] MA S Q, LIU Y Q, TAO G H, et al. NIC: Detecting adversarial samples with neural network invariant checking[C]// Proceedings of the 26th Network and Distributed System Security Symposium (NDSS). Reston, VA: ISOC, 2019.

[88] CINTAS C, SPEAKMAN S, AKINWANDE V, et al. Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error[C]// Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann, 2020: 876-882.

[89] MCFOWLAND E, SPEAKMAN S, NEILL D B. Fast generalized subset scan for anomalous pattern detection [J]. Journal of Machine Learning Research, 2013, 14(1): 1533-1561.

[90] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[C]// Proceedings of the 6th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2018.

[91] SHAFAHI A, NAJIBI M, GHIASI A, et al. Adversarial training for free[C]// Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2019: 3353-3364.

[92] ZHU C, CHENG Y, GAN Z, et al. FreeLB: Enhanced adversarial training for natural language understanding[C]// Proceedings of the 8th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2020.

[93] ZHANG D H, ZHANG T Y, LU Y P, et al. You only propagate once: accelerating adversarial training via maximal principle [C]// Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2019: 227-238.

[94] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[C]// Proceedings of the 7th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2019.

[95] ZHANG H Y, YU Y D, JIAO J T, et al. Theoretically principled trade-off between robustness and accuracy[C]// Proceedings of the 36th International Conference on Machine Learning (ICML). New York: ACM, 2019: 7472-7482.

[96] WANG Y S, ZOU D F, YI J F, et al. Improving adversarial robustness requires revisiting misclassified examples[C]// Proceedings of the 8th International Conference on Learning Representations (ICLR). La Jolla, CA: LCLR, 2020.

[97] MAO C Z, ZHONG Z Y, YANG J F, et al. Metric learning for adversarial robustness[C]// Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2019: 478-489.

[98] LI P C, YI J F, ZHOU B W, et al. Improving the robustness of deep neural networks via adversarial training with triplet loss [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann, 2019: 2909-2915.

[99] LIU C H, JÁJÁ J. Feature prioritization and regularization improve standard accuracy and adversarial robustness[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann, 2019: 2994-3000.

[100] WANG H T, CHEN T L, GUI S P, et al. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free[C]// Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT Press, 2020.

[101] 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 the 23rd ACM Conference on Computer and Communications Security (CCS). New York: ACM, 2016: 1528-1540.

[102] XU K, ZHANG G, LIU S, et al. Adversarial t-shirt! evading person detectors in a physical world[C]// European Conference on Computer Vision (ECCV). Springer, Cham, 2020: 665-681.

[103] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification [C]// Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 1625-1634.

[104] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch [J]. arXiv: 1712. 09665, 2017.

[105] LUO B, LIU Y N, WEI L X, et al. Towards imperceptible and robust adversarial example attacks against neural networks [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2018: 1652-1659.

[106] LIU A S, LIU X L, FAN J X, et al. Perceptual-sensitive gan for generating adversarial patches[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 1028-1035.

[107] JAN S T K, MESSOU J, LIN Y C, et al. Connecting the digital and physical world: Improving the robustness of adversarial at-

tacks[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2019: 962-969.

[108]DUAN R J, MA X J, WANG Y S, et al. Adversarial camouflage: Hiding physical-world attacks with natural styles[C]// Proceedings of the 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 997-1005.

[109]ZHOU Z, TANG D, WANG X, et al. Invisible mask: Practical attacks on face recognition with infrared[J]. arXiv: 1803.04683, 2018.

[110]SHEN M, LIAO Z, ZHU L, et al. Vla: A practical visible light-based attack on face recognition systems in physical world[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019, 3(3): 1-19.

[111]DUAN R, MAO X, QIN A K, et al. Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink[C]// Proceedings of the 34st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 16062-16071.

[112]SAYLES A, HOODA A, GUPTA M, et al. Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Shutter Effect[C]// Proceedings of the 34st IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 14666-14675.

[113]NGUYEN D L, ARORA S S, WU Y, et al. Adversarial light projection attacks on face recognition systems: A feasibility study[C]// Proceedings of the 33st IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 814-815.

[114]LOVISOTTO G, TURNER H, SLUGANOVIC I, et al. SLAP: Improving physical adversarial examples with short-lived adversarial perturbations[C]// Proceedings of the 30th USENIX Security Symposium (USENIX Security). Berkeley, CA: USENIX Association, 2021.

[115]SHI C H, JI S L, LIU Q J, et al. Text captcha is dead? A large scale deployment and empirical study[C]// Proceedings of the 27th ACM Conference on Computer and Communications Security (CCS). New York: ACM, 2020: 1391-1406.

CHEN Meng-xuan, born in 1996, postgraduate, is a member of China Computer Federation. Her main research interests include AI security and adversarial examples.

SHAO Jun, born in 1981, Ph.D, professor, is a member of China Computer Federation. His main research interests include applied cryptography, blockchain and AI security.

(责任编辑:喻藜)