计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：图像分类模型的对抗样本攻防研究综述
作者：闫嘉乐，徐洋，张思聪，李克资
网络首发日期：2022-08-31
引用格式：闫嘉乐，徐洋，张思聪，李克资．图像分类模型的对抗样本攻防研究综述[J/OL]．计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.TP.20220830.1507.002.html

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

网络首发时间：2022-08-31 09:33:02
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20220830.1507.002.html

图像分类模型的对抗样本攻防研究综述

闫嘉乐，徐洋，张思聪，李克资

贵州师范大学 贵州省信息与计算科学重点实验室，贵阳 550001

摘　要：深度学习模型在图像分类领域的能力已经超越了人类，但不幸的是，研究发现深度学习模型在对抗样本面前非常的脆弱，这给它在安全敏感的系统中的应用带来了巨大挑战。图像分类领域对抗样本的研究工作被梳理和总结，以期为进一步的研究该领域建立基本的知识体系，首先介绍了对抗样本的形式化定义和相关术语，然后介绍了对抗样本的攻击和防御方法，特别是新兴的可验证鲁棒性的防御，并且讨论了对抗样本存在可能的原因。为了强调在现实世界中对抗攻击的可能性，回顾了相关的工作。最后，在梳理和总结文献的基础上，分析了对抗样本的总体发展趋势和存在的挑战以及未来的研究展望。

关键词：图像分类；对抗样本；深度学习；对抗攻击；对抗防御
文献标志码: A　**中图分类号**:TP183；TN192　doi：10.3778/j.issn.1002-8331.2205-0520

Survey of research on adversarial examples attack and defense in image classification model

YAN Jiale, XU Yang, ZHANG Sicong, LI Kezi

Key Laboratory of Information and Computing Science of Guizhou Province, Guizhou Normal University, Guiyang 550001, China

Abstract：Deep learning models have surpassed human capabilities in the field of image classification, but unfortunately, research has found that deep learning models are very vulnerable to adversarial examples attacks, which poses a great challenge for its application in security-sensitive systems. The research work on adversarial examples in the field of image classification is sorted out and summarized in order to establish a basic knowledge system to further study the field, firstly, the formal definition of adversarial examples and related terms are introduced. Then, the methods of adversarial examples attack and defense are introduced, especially the emerging defense of certified robustness, and the possible reasons for the existence of adversarial examples are discussed. To highlight the possibility of adversarial attacks in the real world, related work is reviewed. Finally, based on summarizing and combing the literature, the general trends and challenges of adversarial examples and future research outlook are analyzed.

Key words：image classification; adversarial examples; deep learning; adversarial attack; adversarial defense

近年来，深度学习作为人工智能的核心技术被广泛使用在大量的场景和应用中，它在许多任务中取得先进的性能和快速的发展，例如计算机视觉[1]，自然语言处理[2]，语音辨识[3]，自动驾驶[4]，医疗诊断[5]等

基金项目：国家自然科学基金项目（U1831131）；中央引导地方科技发展专项资金（黔科中引地〔2018〕4008）；贵州省科技计划项目（黔科合支撑[2020]2Y013 号）；贵州省研究生科研基金项目（黔教合 YJSKYJJ〔2021〕102）。

作者简介：闫嘉乐(1996-),CCF 学生会员,男,硕士研究生,研究方向为深度学习、人工智能安全,E-mail: jialeyan@gznu.edu.cn;徐洋(1983-),CCF 高级会员,男,通信作者,博士,教授,研究方向为网络空间安全、机器学习,E-mail: xy@gznu.edu.cn;张思聪(1989-),CCF 普通会员,男,博士,讲师,研究方向为网络空间安全、机器学习;李克资(1997-),CCF 学生会员,男,硕士研究生,研究方向为深度学习、语音识别、人工智能安全。

任务，有些领域甚至超过人类的处理能力。在计算机视觉领域，自从 Krizhevsky 等人[6]利用 AlexNet 网络在图像分类任务上取得了划时代的突破以来，卷积神经网络成为该领域最先进的模型，虽然深度学习模型性能很优越，但不幸的是，Szegedy 等人[7]研究发现深度神经网络模型在对抗样本面前非常的脆弱，此后的研究工作发现深度学习模型在语音识别[8]、文本分类[9]、恶意软件检测[10]等不同的任务中也存在对抗样本现象。由于图像分类模型的对抗攻击在文献中是最常见的，因此这是本文梳理的重点。

卷积神经网络模型和其他的深度学习模型在面对对抗攻击时的脆弱性，促使机器学习社区重新审视与模型构建相关的所有过程，试图找到模型缺乏鲁棒性可能的原因。对抗攻击与防御之间的军备竞赛最终形成了一个对抗机器学习的最新研究领域，该领域致力于构建更可信、更鲁棒的深度学习模型。

图像分类的对抗机器学习目前是一个非常活跃的研究领域，它占据了该领域的大部分研究工作，几乎每天都有新的论文发表，但迄今为止还没有一个有效的解决方案来确保深度学习模型的安全性。本文梳理和总结了图像分类领域的研究工作，整理了该领域的核心分类体系，以便为读者更好的探索该领域打下坚实的基础。

虽然对抗样本领域已经存在多篇在图像领域的综述论文[11-13]，但本文相比这些综述文章对该领域的介绍更加全面，可以让读者对图像分类领域的全貌有一个清晰的了解。本文相比 2020 年文献[11]梳理总结了对抗样本防御相关的工作，相比 2021 年文献[12]梳理总结了解释对抗样本存在性的相关工作，以期为读者抛砖引玉，相比 2021 年文献[13]梳理总结了物理世界的对抗样本相关的工作。本文梳理和回顾了图像分类领域对抗样本攻击和防御的发展历程，如图 1 所示。

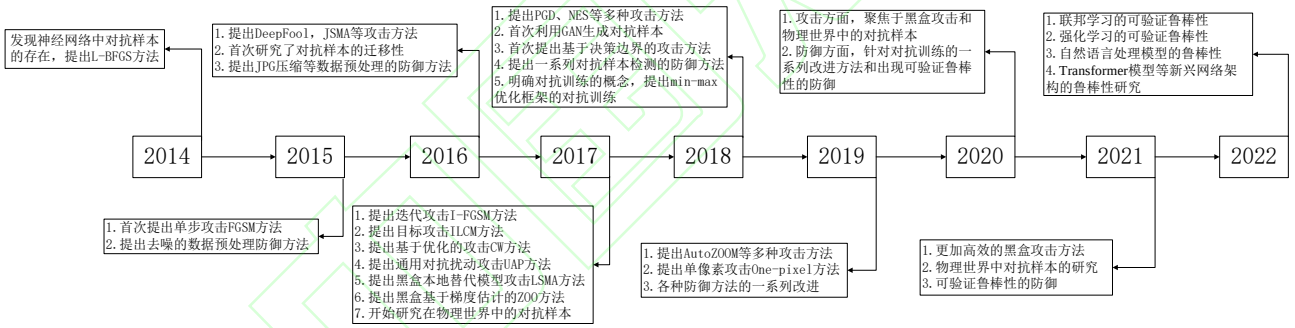

图 1　对抗样本的发展历程

Fig.1　The development process of AE

本文的其余部分的结构如下：第 1 节介绍了对抗样本的定义和理解本文所梳理工作所需要的重要概念和相关术语。第 2 节按照白盒攻击、黑盒攻击和物理世界中的攻击进行分类，梳理了对抗样本的攻击方法。第 3 节梳理了对抗样本存在性相关的解释。第 4 节按照增强模型鲁棒性的防御、输入预处理的防御、对抗样本的检测防御和可验证鲁棒性的防御的分类法回顾了文献中的对抗防御方法。第 5 节在梳理和总结文献的基础上，讨论分析了对抗样本的总体发展趋势和未来研究展望。第 6 节对全文进行了总结。

1　对抗样本相关概念的介绍

1.1　对抗样本的定义

本小节以图像分类模型为例，介绍对抗样本的定义。一个图像对抗样本可以形式化定义如下：f 表示由正常图像训练得到的分类模型，$\boldsymbol{x}$ 表示正常的输入图像，敌手(Adversary)寻找一个对抗扰动 $\boldsymbol{\delta}$ (Adversarial Perturbation, AP)，使得 $\boldsymbol{x}'=\boldsymbol{x}+\boldsymbol{\delta}$，$\boldsymbol{x}'$ 即为对抗样本(Adversarial Example, AE)，其中对抗扰动 $\boldsymbol{\delta}$ 使得 $\boldsymbol{x}$ 跨越了分类模型 f 的决策边界导致 $f(\boldsymbol{x})\neq f(\boldsymbol{x}')$，如图 2 所示。综上，对抗样本的简单形式化定义如下公式(1)。

$$\text{Find AP } \boldsymbol{\delta},\quad \text{s.t. } f(\boldsymbol{x}')\neq f(\boldsymbol{x}), \boldsymbol{x}'=\boldsymbol{x}+\boldsymbol{\delta}, \|\boldsymbol{\delta}\|_p<\varepsilon \tag{1}$$

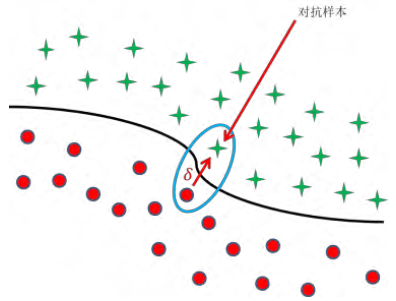

图 2　对抗样本在 2 维决策空间下示例

Fig.2　Example of AE in 2D decision space

图 3 展示了一个图像对抗样本生成的示例，敌手在分类模型的测试阶段使用FGSM攻击[14]方法精心制作一个对抗扰动，添加在正常样本上生成对抗样本，输入给分类模型，使得本来以 57%置信度分类为大熊猫的图像被以 99%的置信度误分类为长臂猿。

x　　δ　　x'

正常样本　　对抗扰动　　对抗样本

图 3　图像对抗样本的生成[14]

Fig.3　Genneration of image AE[14]

1.2　相关术语介绍

(1) 白盒攻击和黑盒攻击

根据敌手对目标模型的先验知识掌握情况，攻击可以被分类为白盒攻击(White-box Attack)和黑盒攻击(Black-box Attack)。在白盒攻击中，敌手可以完全访问目标模型的训练数据、参数和结构，甚至是防御的参数和结构。在黑盒攻击中，敌手并不知道目标模型和训练模型的参数，以及防御方法的有关信息。

(2) 目标攻击和无目标攻击

根据对抗攻击是否设置目标结果，攻击被分为目标攻击(Targeted Attack)和无目标攻击(Non-targeted Attack)。目标攻击是敌手旨在诱导目标模型将输入样本分类为特定目标结果。无目标攻击是敌手旨在诱导目标模型将输入样本分类为非正常样本的真实结果的其它任何结果。

(3) 单步攻击和迭代攻击

根据生成对抗样本的计算复杂度，攻击被分为单步攻击(One-step Attack)和多步攻击(Iterative Attack)。单步攻击使用梯度一步计算得到对抗扰动，迭代攻击利用更多的迭代步骤来制作和微调对抗扰动。

(4) 置信度和类别标签

置信度(Confidence)是输入样本经过模型分类为某种类别的概率。类别标签(Class Label)指图片分类模型得到的类别标签结果。

(5) 目标模型和替代模型

目标模型(Target Model)是被敌手攻击的模型。替代模型(Substitute Model)是一个由敌手设法训练的模型，用来重现目标模型的预测行为。

(6) 对抗样本的迁移性

对抗样本的迁移性(Transferability)是指对抗样本在原始被计算的模型之外的模型进行泛化的能力。

(7) 数字攻击和物理世界攻击

数字攻击(Digital Attack)是可以完全访问模型的实际数字输入的攻击。物理世界攻击(Physical-wrold Attack)是攻击真实世界的系统。

(8) 扰动约束度量

由于对于人类视觉系统来说，很难定义一个不易察觉的标准，L_p 范数经常被用来控制添加到图像中扰动的大小。L_1 和 L_2 范数分别表示正常图像和对抗图像在输入空间中的曼哈顿距离和欧几里得距离，L_0 范数表示对抗图像在正常图像中修改的像素数量，L_∞ 范数衡量的是正常图像和对抗图像之间对应位置上所有像素的最大差异。

2　对抗样本的攻击方法

本节主要梳理和总结了深度学习图像分类领域对抗样本的相关攻击方法，总体上按白盒和黑盒两大类攻击进行介绍，最后为了强调物理世界中对抗样本的严峻性，回顾了相关研究工作。图 4 根据敌手对模型的不同访问程度描述了攻击的大致分类。在本节中，白盒（基于梯度的）攻击和黑盒（基于迁移的、基于置信度分数的和基于决策边界的攻击）会被详细的介绍。表 1 和表 2 中对白盒攻击进行了总结和分析。

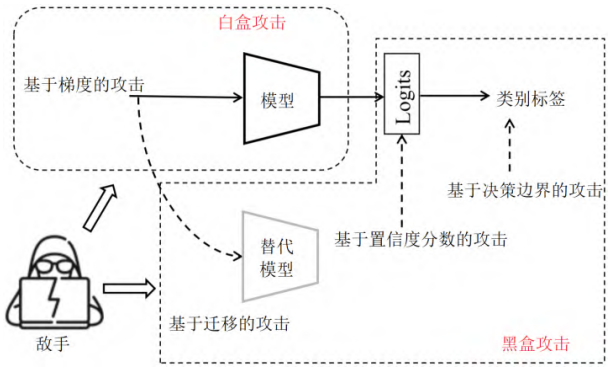

图 4 敌手对模型不同访问程度的攻击类别分类

Fig.4 Classification of attack categories of adversaries with different access degrees to the model

2.1 白盒攻击

本小节根据对抗样本的制作机理，将白盒攻击总结为 3 个方向进行介绍分别是：基于梯度的攻击、基于优化的攻击和其它的白盒攻击方法。

2.1.1 基于梯度的攻击

基于梯度的攻击算法主要利用目标模型关于给定输入的梯度信息来寻找一个使模型损失值更大的对抗扰动，从而使加入该对抗扰动的正常图像导致模型误分类，这种攻击方式在文献中使用的最多。由于基于梯度的攻击通常需要获取目标模型内部结构的信息，因此绝大多数基于梯度的攻击都是白盒攻击。

Goodfellow 等人[14]开创性的提出了基于梯度的单步攻击方法：快速梯度符号方法(Fast Gradient Sign Method, FGSM)。FGSM 方法在给定上限范数约束 ε 的一次迭代中，沿着正常样本梯度的反方向添加扰动来最大化目标模型的训练损失误差，降低分类的置信度，增加类间混淆的可能性，使得模型分类错误。给定一个输入图像 $\boldsymbol{x}$，FGSM 根据公式(2)生成一个对抗样本 $\boldsymbol{x}'$。FGSM 算法简单有效，在图像攻击领域发挥着重要的作用，很多后续的研究都是基于该算法进行的。

$$\boldsymbol{x}' = \boldsymbol{x} + \varepsilon \cdot sign(\nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x};\boldsymbol{\theta}), y)) \tag{2}$$

鉴于单步攻击的 FGSM 的扰动较大，成功率较低。Kurakin 等人[15]在 FGSM 方法的原理上进行优化，提出了基本迭代攻击方法(Basic Iterative Method, BIM)，有些文献中也称为 I-FGSM 方法(Iterative Fast Gradient Sign Method, I-FGSM)，该攻击在本质上是迭代的 FGSM 算法，将 FGSM 的单次计算对抗扰动转换为迭代小步计算对抗扰动，BIM 攻击通过迭代公式(3)来生成对抗图像，该攻击方法是引入物理世界攻击的一个有影响力的贡献。

$$\boldsymbol{x}'_{i+1} = clip_{\varepsilon}(\boldsymbol{x}'_i + \alpha \cdot sign(\nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}'_i;\boldsymbol{\theta}), y))) \tag{3}$$

虽然 I-FGSM 方法提高了对抗样本攻击的成功率，但是该方法生成的对抗图像容易陷入优化的局部极值点，且易过拟合到攻击模型上，因此会减弱生成的对抗样本的迁移性。针对该问题，Dong 等人[16]在 I-FGSM 的基础上添加一个动量项，从而加速了收敛以及避免落入优化的局部极小值，形成了 MI-FGSM 方法(Momentum Iterative Fast Gradient Sign Method)，该方法添加动量的巧妙思路解决了以往迭代攻击的缺点：随着迭代次数的增加，黑盒攻击的可迁移性减弱，该思路不仅增强了对白盒模型的攻击能力，而且提高了对于黑盒模型的攻击成功率。MI-FGSM 攻击方法的非定向攻击可以被归纳为公式(4)所示，其中 $\boldsymbol{g}_i$ 的初始值为 0，且 $\boldsymbol{g}_i$ 使用衰减因子 μ 累积前 i 次迭代的梯度，从而稳定了梯度的更新。

$$\begin{aligned} \boldsymbol{x}'_0 &= \boldsymbol{x} \\ \boldsymbol{g}_{i+1} &= \mu \cdot \boldsymbol{g}_i + \frac{\nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}'_i;\boldsymbol{\theta}), y)}{\left\| \nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}'_i;\boldsymbol{\theta}), y) \right\|_1} \\ \boldsymbol{x}'_{i+1} &= \boldsymbol{x}'_i + \alpha \cdot sign(\boldsymbol{g}_{i+1}) \end{aligned} \tag{4}$$

I-FGSM 由于对有效对抗扰动的多次搜索，因此被认为是强大的攻击之一，但它计算代价高昂。后来，Madry 等人[17]提出了投影梯度下降(Projected Gradient Descent, PGD)方法，本质是 I-FGSM 的一种变体，与 I-FGSM 相比，PGD 使用均匀的随机噪声初始化，增加攻击的迭代轮数，并且提出在 I-FGSM 中对梯度进行投影，而不是对梯度进行裁剪操作。经过大量实验

验证，PGD 攻击被对抗机器学习领域顶级学术会议的学者们广泛认为是最强大的一阶攻击。PGD 攻击的非定向攻击如下公式(5)所示。

$$x'_{i+1} = proj_{x,\epsilon}(x'_i + \alpha \cdot sign(\nabla_x \ell(f(x'_i;\theta), y))) \quad (5)$$

表 1　对抗样本的攻击方法总结

Table 1　Summarization of adversarial attack methods

方法名称	攻击原理	威胁模型		攻击类型		场景		扰动限制	攻击强度	领域
		白盒	黑盒	目标	无目标	数字	物理			
FGSM[14]	基于梯度	✓	×	✓	✓	✓	×	L_∞	弱	图像分类
I-FGSM[15]	基于梯度	✓	×	✓	✓	✓	×	L_∞	中	图像分类
MI-FGSM[16]	基于梯度	✓	×	✓	✓	✓	×	L_∞	中	图像分类
PGD[17]	基于梯度	✓	×	✓	✓	✓	×	L_2, L_∞	强	图像分类
VMI-FGSM[18]	基于梯度	✓	×	✓	✓	✓	×	L_2, L_∞	强	图像分类
JSMA[19]	基于梯度	✓	×	✓	✓	✓	×	L_0	弱	图像分类
L-BFGS[7]	基于优化	✓	×	✓	✓	✓	×	L_2	弱	图像分类
CW[20]	基于优化	✓	×	✓	✓	✓	×	L_0, L_2, L_∞	强	图像分类
Homotopy-Attack[21]	基于优化	✓	×	✓	✓	✓	×	L_0, L_1, L_2, L_∞	强	图像分类
DeepFool[22]	基于超平面	✓	×	×	✓	✓	×	L_2, L_∞	中	图像分类
UAP[23]	基于超平面	✓	×	×	✓	✓	×	L_2, L_∞	强	图像分类
LSMA[24]	基于迁移	×	✓	✓	✓	✓	×	L_∞	弱	图像分类
Ensemble-LSMA[25]	基于迁移	×	✓	✓	✓	✓	×	L_∞	中	图像分类
Curls&Whey[26]	基于迁移	×	✓	×	✓	✓	×	L_2	中	图像分类
TREMBA[27]	基于迁移	×	✓	✓	✓	✓	×	L_∞	强	图像分类
FIA[28]	基于迁移	×	✓	✓	✓	✓	×	L_∞	强	图像分类
ZOO[29]	基于查询	×	✓	✓	✓	✓	×	L_2	弱	图像分类
NES[30]	基于查询	×	✓	✓	✓	✓	×	L_∞	弱	图像分类
AutoZOOM[31]	基于查询	×	✓	✓	✓	✓	×	L_2	中	图像分类
One-pixel[32]	基于查询	×	✓	✓	✓	✓	×	L_0	弱	图像分类
Meta Attack[33]	基于查询	×	✓	✓	✓	✓	×	L_2	强	图像分类
Simulator Attack[34]	基于查询	×	✓	✓	✓	✓	×	L_2, L_∞	强	图像分类
Boundary Attack[35]	基于决策	×	✓	✓	✓	✓	×	L_2	强	图像分类
Opt-Attack[36]	基于决策	×	✓	✓	✓	✓	×	L_2	强	图像分类
HSJA[37]	基于决策	×	✓	✓	✓	✓	×	L_2, L_∞	强	图像分类
QEBA[38]	基于决策	×	✓	✓	✓	✓	×	L_2	强	图像分类
NonLinear-BA[39]	基于决策	×	✓	✓	✓	✓	×	L_2	强	图像分类
RP_2[40]	基于优化	✓	×	✓	✓	×	✓	L_2, L_∞	弱	无人驾驶
PhysGAN[41]	基于 GAN	✓	×	×	✓	×	✓	L_2	中	无人驾驶
Fake Point Spoofing[42]	基于设备	✓	×	×	✓	×	✓	不适用	弱	激光雷达
LiDAR-Adv[43]	基于梯度	✓	×	×	✓	×	✓	不适用	中	激光雷达
Pedestrian Detection Attack[44]	基于梯度	✓	×	×	✓	×	✓	图像域	强	行人检测
AdvPatch[45]	基于梯度	✓	×	✓	×	×	✓	图像域	强	图像分类
AdvCam[46]	基于优化	✓	×	✓	✓	✓	✓	L_2, L_∞	强	图像分类

虽然基于梯度的攻击方法在白盒环境中取得了令人难以置信的成功率，但大多数现有的基于梯度的攻击方法在黑盒环境中往往表现出较弱的可迁移性，特别是在攻击具有防御机制的模型的情况下。针对该问题，Wang 等人[18]提出了 VMI-FGSM(Variance tuning MI-FGSM)方法，以增强基于梯度的迭代攻击方法类的可迁移性。具体来说，在梯度计算的每次迭代中，不直接使用当前的梯度进行动量积累，而是进一步考虑

之前迭代的梯度方差来调整当前梯度，以稳定更新方　　向摆脱糟糕的局部最优值。

表 2　对抗攻击方法分析

Table 2　Analysisof adversarial attack methods

威胁模型	攻击方法分类	攻击机制	优势	局限性	适用场景
白盒	基于梯度	主要利用目标模型的损失函数对于输入图像的梯度信息来生成对抗扰动	简单直接的强大方法，依靠模型的梯度进行攻击	掩蔽模型的梯度或采用不可微分技术便可使其失效	1.白盒攻击 2.黑盒攻击（基于迁移）
	基于优化	主要将生成对抗样本的过程形式化为一个优化问题，通过求解优化问题来求得对抗扰动	1.生成的对抗扰动相比于基于梯度的要小很多 2.攻击成功率高	计算量过大导致对抗样本生成速度慢	1.白盒攻击 2.攻击基于梯度掩蔽的防御方法
	其它的白盒攻击	主要利用深度学习的超平面分类思想以及图像的特征等思路来生成最小化的对抗扰动	基于超平面思路简单直接，可以找到较小的对抗扰动，相对基于优化的方法，迭代次数更少，时间效率更高	需要重复的计算样本点到分类超平面的距离，对于深度学习的高维数据计算量很大	1.白盒攻击 2.寻找较小扰动的高质量对抗样本场景
黑盒	基于迁移	主要基于少量训练数据训练替代模型，基于替代模型利用白盒攻击生成对抗图像迁移到目标模型上进行攻击	攻击仅需要给定目标模型的少量训练数据和查询	在现实世界中，可以通过简单的限制对目标模型的查询数量来规避攻击。	1.黑盒攻击 2.攻击基于梯度掩蔽的防御方法
	基于置信度分数查询	主要利用查询目标分类器输出的置信度分数信息来近似梯度信息，从而使用估计的梯度生成对抗扰动	1.攻击仅需要给定目标模型输出的置信度分数， 2.适合攻击包含不可微分的模型或包含随机策略的防御	1.在现实世界中，可以通过简单的隐藏目标模型的置信度分数(概率)来规避攻击 2.攻击需要较多查询，收敛时间较长	1.黑盒攻击 2.攻击基于梯度掩蔽的防御方法 3.攻击包含随机化策略的防御
	基于决策边界	初始基于较大扰动的对抗图像，通过查询目标模型的预测标签，沿着模型的决策边界进行随机游走来逐步减小对抗扰动	1.攻击仅需要模型的预测类别 2.可以与其它方法一起使用，如基于梯度的攻击	1.通常需要较多的迭代次数才能收敛 2.通常需要更多对目标模型的查询来优化扰动	1.黑盒攻击 2.攻击基于梯度掩蔽的防御方法

前边介绍的基于梯度的攻击都集中在从整体上扰动图像，并且限制扰动的 L_2 或者 L_∞ 范数，Papernot 等人[19]提出 JSMA 方法(Jacobian-based Saliency Map Attack)则将扰动限制在图像的一个较小的区域。JSMA 攻击引入显著性映射来评估每个输入特征对模型类预测的影响，利用该信息来筛选在改变模型预测时最有影响力的像素，通过扰动一些显著特征来引起模型的错误分类，该攻击方法倾向于找到稀疏的对抗扰动，生成的对抗样本的视觉质量很高。

综上所述，目前基于梯度的攻击中，主要有基于 FGSM 攻击进行发展和改进的路线(I-FGSM、MI-FGSM、PGD、VMI-FGSM)，以及基于稀疏性扰动的发展路线(JSMA 等)。FGSM 计算成本低，但是生成的对抗扰动通常比迭代攻击（例如 I-FGSM、MI-FGSM、PGD、VMI-FGSM 等）生成的对抗扰动更大，生成的对抗样本的视觉质量较差，并且对模型的欺骗效果更差。MI-FGSM 和 PGD 都改进了 I-FGSM 方法，MI-FGSM 在优化过程中使用动量，增强了对抗样本的可迁移性，PGD 嘈杂的初始点和投影梯度产生了更强的攻击，VMI-FGSM 改进 MI-FGSM，使用梯度的方差调整更新，进一步增强了对抗样本的可迁移性。

2.1.2　基于优化的攻击

生成对抗样本的核心问题是如何找到有效的对抗扰动，寻找对抗扰动可以被形式化为一个优化问题，因此可以通过对优化问题的求解来实现攻击。通常来说，基于优化的攻击相比于基于梯度的攻击生成的对抗扰动添加在正常图像上视觉效果要好，并且生成的扰动范数更小，但相比基于梯度的方法更耗时。

2014 年 Szegedy 等人[7]首次发现了深度神经网络模型在对抗扰动下的脆弱性，首次引入对抗样本的概念，其工作是对抗样本领域的开山之作，文中提出了 L-BFGS 攻击算法，通过寻找导致神经网络误分类的最小损失函数加性扰动项，将问题转化为凸优化问题，形式化为公式(6)所示的优化问题来寻找对抗扰动。L-BFGS 攻击寻找人类感知最小的对抗扰动，因此生成对抗扰动的计算开销很大，速度很慢，并且攻击的成功率不高。文中也证明了对抗样本在不同的深度神经网络分类模型之间具有很好的迁移性。

$$\min\|\boldsymbol{\delta}\|_2,\ \text{s.t. } f(\boldsymbol{x}+\boldsymbol{\delta})=\boldsymbol{y}';\boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^m \tag{6}$$

Carlini 和 Wagner 提出了基于优化的相对较强的一阶攻击 CW 攻击[20]，本质是基于 L-BFGS 攻击的改进，具体的 CW 攻击相对于 L-BFGS 攻击有以下三个改进：1)使用模型中实际输出的梯度，而不是经过 softmax 操作后的梯度。2)应用不同的扰动约束度量：L_0，L_2，L_∞ 范数。3)应用不同的目标函数 $\ell(\cdot)$，通过实验分析选择出了最优的目标函数来生成对抗样本。相比 L-BFGS 攻击，CW 方法可以改变目标函数中的超参数来扩大最优解的搜索空间，进而显著提高对抗攻击的成功率。由于 CW 攻击需要对算法的一些参数进行优化，速度极慢，并且不具有黑盒可迁移性，但是一种非常强的白盒攻击方法，可以攻破防御性蒸馏防御方法。

稀疏的对抗攻击只扰动几个像素来欺骗深度神经网络。与逐像素的整体扰动相比，高度稀疏的对抗攻击更危险，因为不易被人眼所察觉。Zhu 等人[21]通过提出一种同伦(Homotopy)优化算法在一个统一的框架内同时解决了对抗扰动的稀疏性和扰动约束问题。同伦攻击(Homotopy-Attack)方法利用不同区域的特性，施加不同程度的 L_∞ 范数扰动上界，其中该上界的计算依赖于不同坐标轴的像素饱和度水平，以最小化的对抗样本与正常样本之间的 L_0 距离。实验表明该方法与最先进的方法相比，可以产生非常稀疏的对抗扰动，同时保持相对较低的扰动强度。

2.1.3 其他白盒攻击方法

除了基于梯度和基于优化的攻击方法外，活跃的研究者们还想出了 DeepFool 和 UAP 这种基于超平面思想的攻击方法，以下将简要介绍。

为了解决 FGSM 攻击中扰动大小 ε 不确定的问题，Moosavi-Dezfoolli 等人[22]提出 DeepFool 方法，该方法思路是计算正常样本和目标模型分类边界之间最小距离来生成对抗扰动，该方法是一种基于 L_2 范数的非目标攻击方法，该方法巧妙利用以直代曲、化繁就简、迭代解决的思路，将正常图像周围的类边界线性化，形成一个凸多面体，然后向最优方向更新一小步，将正常图像推向最近的分类超平面，直到其跨过分类超平面改变类标签，具体如图 5 所示。由于 DeepFool 攻击产生近似最小扰动，因此生成的扰动相比于基于梯度和基于优化的攻击方法都要小，相比于基于优化的攻击方法速度更快。

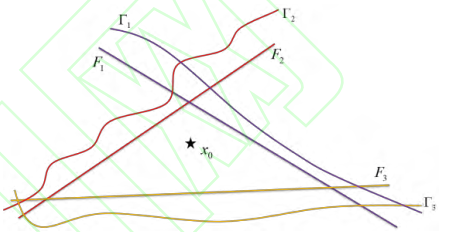

图 5 DeepFool 分类超平面示例

Fig.5 DeepFool classification hyperplane example

上边介绍的对抗攻击生成的对抗扰动仅可以在特定的图像上欺骗目标模型，是否存在图像上的通用对抗扰动呢？Moosavi-Dezfoolli 等人经过研究提出了 UAP[23](Universal Adversarial Perturbations)，攻击者只需在相同分布的所有样本中添加 UAP 算法生成的扰动即可生成对抗样本。UAP 方法利用分类超平面思想依次迭代推导出每个样本的扰动向量 $\Delta\boldsymbol{v}_i$，将 $\Delta\boldsymbol{v}_i$ 进行聚合，最终产生一个扰动 $\boldsymbol{v}$ 让所有的样本 $\boldsymbol{x_i}$ 跳出分类决策边界 $\mathcal{R}_i$ 之外生成对抗样本，如图 6 所示。

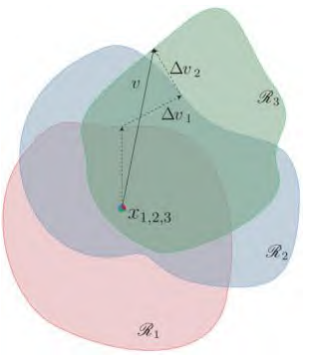

图 6 通用对抗扰动示意图[23]

Fig.6 Schematic diagram of UAP[23]

2.2 黑盒攻击

从现实的对抗的角度来看，黑盒攻击是最实用的一类，因为黑盒攻击假设不了解目标模型相关的信息，它的实用性使得在对抗机器学习社区中非常受欢迎。本节总结和梳理形成了 3 个方向的黑盒攻击方法，主要是：基于迁移的攻击、基于置信度分数查询的攻击

和基于决策边界的攻击。表 1 和表 2 中对黑盒攻击进行了总结和分析。

2.2.1 基于迁移的攻击

基于迁移的攻击允许攻击者进行目标模型的查询和访问目标模型的一部分训练数据集，然后攻击者使用这些信息构建一个合成模型，攻击者在合成模型上使用白盒攻击生成对抗样本，最后将该对抗样本迁移到目标模型上进行攻击。基于迁移的攻击是介于黑盒攻击和白盒攻击之间的一种攻击。这种攻击的条件假设较强，因此不贴合真实场景，对抗样本更好的可迁移性是基于迁移的攻击研究的一个重要目标。

Papernot 等人提出了最早的黑盒攻击被称为本地替代模型攻击(Local Substitute Model Attack, LSMA)[24]，在该攻击中，敌手被允许访问用于训练分类模型的部分原始训练数据以及对分类模型的查询访问。LSMA 攻击通过生成替代模型(Substitute Model)来模拟被攻击模型的近似决策边界，并基于当前的替代模型生成对抗样本，这些对抗样本最终被用于攻击原始目标模型。在训练过程中，雅可比矩阵(Jacobian Matrix)被用来有效利用，以减少目标模型的查询次数。LSMA 方法使梯度掩蔽防御策略无效，因为它不需要梯度信息。

后来，Liu 等人[25]在 LSMA 方法中引入了集成(ensemble)的思想，即同时选择多个模型并结合其损失值来生成相应的对抗样本。该方法考虑了不同模型之间决策边界的相似性，从而首次实现了在不同模型之间大范围迁移对抗性样本的目标。

Shi 等人[26]为了增强黑盒攻击场景下对抗样本的多样性和可迁移性，受 MI-FGSM 攻击方法的启发，提出了 Curls&Whey 黑盒攻击方法。Curls&Whey 攻击方法通过在替代模型上生成对抗样本，然后运用在黑盒模型中，主要包含两个步骤：1)利用卷曲迭代法(Curls Iteration)沿梯度上升方向或下降方向添加对抗扰动到原始正常图像，优化迭代轨迹多样性和适应性。2)Whey 优化主要用来在对抗样本中去除过多的冗余对抗扰动，提升了对抗图像的视觉质量。

此外，Huang 和 Zhang 结合了基于迁移和基于置信度分数的攻击思想提出了 TREMBA[27]。该方法首先通过替代模型在白盒攻击中生成一个初始的对抗样本，然后以这个初始的对抗样本为搜索起点，继续使用基于置信度分数的攻击方法进行查询，最后迭代出迁移效果较好的对抗样本。这种方法有效减少了查询次数，同时提高了黑盒攻击的成功率。

在过去提出的攻击中，分类模型对图片中的像素点一视同仁，没有区别对待，学到了很多缺乏迁移性的噪声特征，这很容易导致局部最优。Wang 等人[28]提出了特征重要性感知攻击(Feature Importance-aware Attack, FIA)，用梯度来表示特征的重要性，通过抑制重要特征和促进琐碎特征来优化加权特征映射，使模型决策错误，从而获得更强的可迁移性对抗样本。

2.2.2 基于置信度分数查询的攻击

基于置信度分数查询的攻击相对于基于迁移的攻击拥有更强的假设，不需要任何关于数据集的知识，它会反复查询看不见的分类器，得到分类器输出的置信度向量，以尝试生成合适的对抗扰动来完成攻击。基于置信度分数查询的攻击相对于基于迁移的攻击更加符合现实场景。

Chen 等人[29]开创性的发展了基于梯度估计的黑盒攻击方法，即零阶优化(Zeroth Order Optimization, ZOO)来估计目标模型的梯度，以此来生成对抗图像。ZOO 方法受 CW 方法的启发，其优化的方案一致，由于黑盒攻击方法，不能获得模型梯度，因此使用对称差商的零阶优化方法来估计梯度和 Hessian 矩阵，如下公式(7)所示。因为深度学习中输入样本的维度较高，因此 ZOO 方法的近似计算开销较大，需要较多的模型查询次数，后续的研究工作也都进一步朝着降低计算开销的方向改进。

$$\hat{\boldsymbol{g}} := \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \approx \frac{f\left(\boldsymbol{x}+h\boldsymbol{e}_i\right)-f\left(\boldsymbol{x}-h\boldsymbol{e}_i\right)}{2h} \tag{7}$$

为了解决 ZOO 攻击方法的估算梯度开销较大的问题，Ilyas 等人[30]巧妙利用投影梯度下降(PGD)和自然进化策略(Natural Evolution Strategies, NES)，估算梯度来降低置信度获取成本，在黑盒场景下构造对抗样本，成功的攻击了当时谷歌的云视觉 API。

为了进一步的降低置信度分数的获取成本以及加快梯度的估算速度，Tu 等人[31]提出了基于自动编码器的零阶优化方法(Autoencoder-based Zeroth Order Optimization Method, AutoZOOM)，该方法是一个通用的查询效率高的黑盒攻击框架，它可以在黑盒场景下有效地产生对抗性样本。AutoZOOM 利用自适应随机梯度估计策略来降低查询的次数和减小扰动的失真度，同时，使用未标记的数据离线训练自动编码器，从而加快了对抗性样本的生成速度。AutoZOOM 方法与标准的 ZOO 攻击方法相比，可以大大减少模型的查询次数，同时保持攻击的有效性以及对抗性样本的视觉质量较高。

Su 等人提出的 One-pixel 攻击[32]和前边介绍的 JSMA 攻击一样，将对抗扰动限制在图像较小的区域内，只需要扰动几个或单个像素点便可以获得较好的攻击效果。为了提高攻击像素点的查找效率，引入了差分进化(Differential Evolution)的查找策略，使得攻击简单高效。

Du 等人[33]提出了 Meta Attack，该攻击使用基于自动编码器结构的元学习来近似梯度，并使用爬虫元学习(Reptile Meta-learning)训练方法进行训练，通过训练元攻击者并将其纳入优化过程，该方法可以在不降低攻击成功率和失真度的情况下大幅减少所需的查询次数。

最近，Ma 等人[34]提出了查询更加高效的 Simulator Attack，其主要做法是训练一个模拟器(Simulator)，其中基于知识蒸馏(Knowledge Distillation)的均方误差损失函数被应用于元学习中的内部和外部更新，以学习许多不同网络模型的输出，从而可以模拟任何未知模型的输出。一旦训练完成，模拟器只需要少量的查询数据进行微调，就可以准确地模拟未知网络的输出，这使攻击需要的大量查询转移到模拟器上，有效地降低了攻击中目标模型的查询次数，使得攻击更加符合现实场景。

2.2.3 基于决策边界的攻击

基于决策边界的攻击既不依赖于替代模型，也不需要置信度分数向量。相比于基于置信度分数查询的攻击，基于决策的攻击代表了一个更受限制的对抗场景，即只需要来自黑盒分类器输出的类别标签便可以成功攻击。这种攻击更加符合真实世界的场景，因此更具研究价值，但攻击难度更大，通常需要较多的查询次数。

为了更加符合现实世界中的黑盒场景限制，Brendel 等人[35]提出了基于决策边界的攻击的开山之作 BoundaryAttack，该攻击只依赖于分类模型输出的类别标签，无需梯度或者置信度分数等信息。Boundary Attack 生成对抗样本的具体的示例如图 7 所示，Boundary Attack 的思路是寻找与原始图像 $\boldsymbol{x}$ 相似的对抗图像 $\boldsymbol{x}'$，主要的做法是反复扰动一个初始对抗图像 $\boldsymbol{x}_0'$，$\boldsymbol{x}_0'$ 和 $\boldsymbol{x}$ 属于不同的类别，然后沿着 $\boldsymbol{x}$ 和 $\boldsymbol{x}_0'$ 所属类别之间的决策边界进行随机游走，使用拒绝采样进行优化，仅需要对对抗图像 $\boldsymbol{x}_i'$ 查询模型输出的类别标签，直到最小化原始图像 $\boldsymbol{x}$ 和对抗图像 $\boldsymbol{x}'$ 之间的距离度量 $d(\boldsymbol{x},\boldsymbol{x}')$ 即可生成对抗图像。由于拒绝采样优化方式的蛮力性质，因此 Boundary Attack 需要较多的迭代搜索次数（例如数十万次迭代）才能找到高质量的对抗图像，因此后来对于 Boundary Attack 的研究主要集中在如何找到更小扰动值的搜索方向和如何加快其搜索速度两个方面的工作。

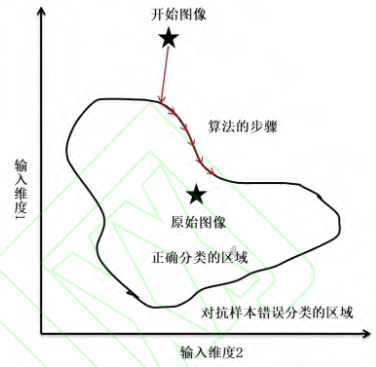

图 7 边界攻击的示意图[35]

Fig.7 Schematic diagram of boundary attack[35]

为了提高攻击的查询效率，Cheng 等人[36]提出了基于决策边界的 Opt-Attack，由于只能获得目标模型输出的类别标签，因此攻击的目标函数不是连续的，故难以进行优化，作者等人重新将问题形式化为实数值优化问题，使得目标函数变得连续，因此可以使用任何的零阶优化算法求解，解决了 Boundary Attack 需要超多次的模型查询以及无法保证收敛性的问题，使得攻击具有更高的查询效率。

Chen 等人[37]为了解决边界攻击的查询次数较多的问题，提出了一种基于决策边界的攻击 HSJA(Hop Skip Jump Attack)，HSJA 攻击在 Boundary Attack 的基础上进行改进，由于在模型决策边界的边缘实现了梯度估计技术，解决 Boundary Attack 的查询次数较多的问题，因此可以更有效的生成对抗样本，具有较高的成功率和较低的查询次数。

减少基于决策边界的攻击所需的查询次数的挑战是如果不进行多次查询，就很难探索高维数据的决策边界。Li 等人提出 QEBA[38](Query-Efficient Boundary-Based Blackbox Attack)试图通过向图像添加扰动来生成查询，从而减少所需的查询次数。因此，探测决策边界被简化对每个生成的查询搜索一个更小的、有代表性的子空间，基于子空间的梯度估计与原始空间的估计相比是最佳的。QEBA 大大减少了模型所需的查询次数，针对离线模型和现实世界中的在线 API 均能产生高质量的对抗样本。

Li 等人创新性地克服了黑盒攻击的梯度不可获得，提出了一种查询高效的 NonLinear-BA[39](Nonlinear Black-box Attack)方法，该方法是一种基于非线性梯度

投影的边界黑盒攻击，通过利用矢量投影进行梯度估计。高维梯度估计的计算成本很高，需要大量的查询，NLBA 将梯度投射到低维的支持物上，极大地提高了估计效率，可以高效的生成对抗样本。

2.3 物理世界中的对抗样本

本文第 2.1 和 2.2 小节详细介绍了在实验室环境下数字世界的白盒和黑盒的关于图像分类场景下的对抗攻击方法，人工智能研究社区很多乐观的研究者认为对抗样本仅能存在于数字世界中，物理世界中受光照条件、距离、拍摄角度、曝光程度、设备差异以及标志遮挡等诸多因素都会导致对抗样本的失效，确实第 2 节中的很多对抗攻击方法直接用在物理世界中的攻击效果大多都不理想，但不幸的是，人工智能安全研究者经过研究发现对抗样本在物理世界中也能成功地攻击深度学习模型。因此，本小节从自动驾驶、通用目标检测以及图像分类等方面梳理了物理世界中对抗样本的工作，以期为后续的研究者提供一些研究灵感。表 1 对物理世界中的对抗样本攻击进行了总结。

Etimov 等人[40]在无人驾驶系统中提出了一种在物理世界中生成对抗扰动的白盒攻击方法，被称为鲁棒物理扰动(Robust Physical Perturbations, RP_2)，该方法在一系列的动态物理环境中产生鲁棒的对抗扰动，其中动态物理环境包括距离、角度和分辨率等物理条件的变化，该方法在无人驾驶系统的道路标识识别系统中实现了很高的欺骗率。文中采用两种方法攻击路标分类模型，分别是海报(Poster)攻击和贴纸(Sticker)攻击，如图 8 所示从左图到右图依次为两种类型的攻击方法，深度学习模型的分类模型会将“停车”路标识别为限速“60km/h”的路标，这些黑白贴纸模仿了生活中常见的涂鸦，不容易引起人们的注意，具有很强的隐蔽性，这给使用深度学习模型的无人驾驶系统带来了巨大挑战，可能会导致严重的交通事故。

图 8 无人驾驶系统的两种路牌识别攻击[40]

Fig.8 Two types of road sign recognition attacks for autonomous driving systems[40]

Kong 等人[41]提出一种基于 GAN[47]的 PhysGAN 方法来生成范数约束的对抗图像，打印后的图像显示出对物理世界条件的鲁棒性，例如光照条件、视角等变化。PhysGAN 方法专门被设计用来欺骗无人驾驶车的转向模型，该模型是一个基于回归公式的角度预测问题。PhysGAN 计算驾驶视频的视觉特征流的扰动，但忽略场景的背景，这样的策略允许对动态场景条件进行有效的扰动，从而避免了文献[40]中静态场景假设的需要。

保证自动驾驶安全的重要感知系统主要由摄像机和激光雷达等组成，以往的对抗攻击的研究大多集中在基于摄像头的感知上，基于激光雷达(LiDAR)感知的探索很少。Cao 等人首次展示了激光雷达在白盒场景下实现的假点欺骗(Fake Point Spoofing)攻击[42]，该攻击在模拟场景中取得了很好的攻击效果，然而在真实的道路场景中，实现该攻击需要将攻击装置动态地对准受害者汽车上的激光雷达，因此对激光发射装置的精度要求非常高。同年内，Cao 等人使用基于梯度的优化方法，即 LiDAR-Adv 攻击方法[43]生成一个 3D 可打印的物理对抗物体，可以导致激光雷达无法检测到打印出的 3D 对抗物体，因此给激光雷达系统检测障碍物带来了新的挑战。

Thys 等人[44]成功的将对抗样本攻击应用到目标检测模型中，提出了一种行人检测攻击(Pedestrian Detection Attack)，在白盒环境下，作者使用对抗补丁(Adversarial Patch)攻击部署的 YOLOv2[48]目标检测器，具体做法是使用 40×40 大小的对抗补丁贴在人的身上，就可以成功避开目标检测器的检测，具体的示例如图 9 所示，图像左边没有对抗补丁的人被目标检测器成功检测出来，图像右边拿着对抗补丁的人成功的攻击了目标检测器，使其被目标检测器忽略掉。

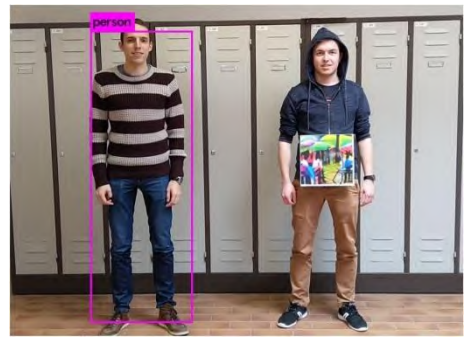

图 9 利用对抗补丁的行人检测攻击[44]

Fig.9 Pedestrian detection attacks using adversarial patches[44]

Ho 等人[49]实验研究表明，在对物理世界的物体进行成像时，随着相机的抖动和姿势的变化，可以获得轻易骗过深度学习模型的图像，该研究中扰动的不可

感知性通过语义的形式呈现，即上下文的不可感知性图像中的抖动和姿势看起来很自然，不易被人察觉。

对抗补丁攻击[45]也是一种发动物理世界攻击的有效方法，一个对抗补丁通常是一个清晰可见的图案，它可以放置在被攻击对象旁边从而导致模型输出错误的结果，具体效果如图 10(c)所示，在桌子上香蕉的旁边放置对抗补丁，图像分类模型以很高的置信度将香蕉误分类为烤面包机。近来，Duan 等人[46]提出了一种基于神经风格迁移技术[50]的 AdvCam 方法来计算不受约束的扰动，这种扰动以伪装目标对象的形式来进行物理世界的攻击，该方法能够生成比先前工作[40]如图 10(a)和[45]如图 10(c)更隐秘的扰动，因为生成的扰动对人眼来说更自然，如图 10(b)所示，AdvCam 方法以伪装自然污渍的形式对停车标志添加对抗扰动来攻击深度学习分类模型，如图 10(d)所示，AdvCam 方法在香蕉旁边以伪装产品标签的形式来达到攻击的目的。AdvCam 方法生成的物理世界的对抗扰动不仅具有很强的隐蔽性，同时对物理世界的各种条件具有很强的鲁棒性和适应性。

(a) RP_2 (b) AdvCam (c) AdvPatch (d) *AdvCam*

图 10 三种物理世界对抗攻击的效果(a)[40](b)[46](c)[45](d)[46]

Fig.10 The effect imageof the three physical worlds adversarial attack(a)[40](b)[46](c)[45](d)[46]

3 对抗样本存在性相关的解释

对抗样本自发现以来，就受到研究人员的广泛关注，文献中有大量的假设来解释深度神经网络的对抗脆弱性，但是很多的解释都不能很好的泛化，并且很多解释之间互相冲突，到目前为止，对于对抗样本存在的原因还没有达成共识。研究人员普遍认为对抗样本现象仍未被充分的理解，关于其成因方面的工作仍然具有吸引力，本小节回顾和梳理了该方向的贡献和主要的假设。

(1) 高维非线性假设

Szegedy 等人[7]认为对抗样本是数据流形(Data Manifold)上形成的低概率的盲区(Pockets)，这些盲区通常很难通过简单的随机抽样被找到，他们认为难以采样的盲区正是深度神经网络的高度非线性导致的，因此模型的泛化能力较差，如图 11 所示，样本空间中的类别○和+被分类模型很好地分开，但是每个类别的每个元素周围都密布着另外一个类别的元素，因为低概率的对抗性盲区密集地分布在图像空间中。Gu 等人[51]和 Song 等人[52]认为这种盲区的出现主要是由于目标函数、训练过程以及训练样本的多样性和数据集的规模受限等的一些缺陷导致的，进而导致神经网络模型的泛化性较差。

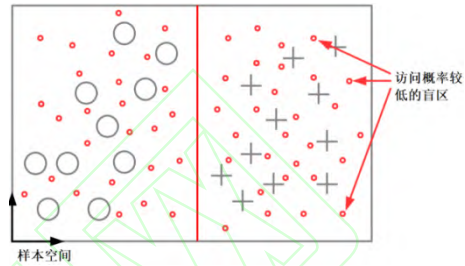

图 11 对抗样本存在的非线性解释[56]

Fig.11 Non-linear explanation for the existence of AE[56]

(2) 线性假设

Goodfellow 等人[14]的假设与 Szegedy 等人的非线性假设相反，他们认为深度神经网络中对抗样本的存在恰恰是线性原因导致的，由于深度神经网络学习到是一个高维特征空间，因此在输入上的微小扰动经过高维空间的变换后会导致最终的输出结果大相径庭。Fawzi 等人[53]表明神经网络模型的对抗鲁棒性是独立于训练过程的，深度神经网络表示的高阶模型中类别之间的距离比线性分类器更大，他们认为更深层次的模型中更难找到对抗样本，这种解释也与 Szegedy 等人的非线性假设相违背。除此之外也有研究结果与线性假设相反，Tabacof 等人[54]发现对抗样本现象可能是一个更复杂的问题，实验结果表明浅层的模型比深层模型更容易受到对抗样本的影响。虽然有些研究对线性假设提出了批评，但一些相关的攻击(例如 FGSM[14]和 DeepFool[22])和防御(例如 Thermometer Encoding[55])都是建立在线性假设的基础上的。

(3) 边界倾斜假设

Tanay 等人[56]否定了 Goodfellow 等人提出的线性假设，认为其不充分且没有说服力，他们提出了一个边界倾斜假设来解释对抗样本现象，具体的假设是深度神经网络虽然学习能力很强，但是通常学到的训练数据的类边界与训练数据的数据流形并非完全重合，而是存在一个倾斜的角度，因此在正常样本上添加的微小扰动容易导致对抗样本的产生。随着倾斜度的降低，所需的扰动量也更小，生成的对抗样本也具有更高的置信度和误导率，作者认为这种效果可能是模型过拟合的结果，如图 12 展示了边界倾斜假设的示意图，

对抗样本存在于倾斜的边界之间，即训练数据学到的类边界和训练数据的数据流形之间。

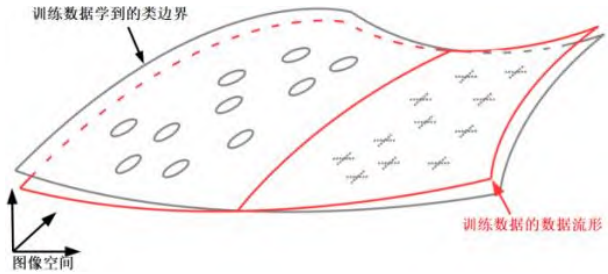

图 12 对抗样本存在的边界倾斜解释[56]

Fig.12 Boundary tilting explanation for the existence of AE[56]

(4) 高维流形假设

Gilmer 等人[57]、Mahloujifar 等人[58]、Shafahi 等人[59]和 Fawzi 等人[60]一致认为对抗样本现象是数据流形高维性导致的结果，为了提供证据，Gilmer 等人创建了一个合成数据集来训练神经网络模型，以便更好地控制实验，模型被训练好以后，作者观察到被模型正确分类的输入接近于附近被错误分类的对抗性输入，基于该实验结果 Gilmer 等人否认对抗样本和正常样本的数据分布不同的假设[52,61]。

(5) 缺乏足够的训练数据

Schmidt 等人[62]认为经过训练学习到的模型必须具有很强的泛化性，需要借助鲁棒优化实现对抗鲁棒性。作者观察到对抗样本的存在并非神经网络分类模型的缺陷，而是统计学习的场景下无法避免的结果，迄今为止仍然没有可行的策略来实现模型的对抗鲁棒性。作者通过实验认为现存的数据集规模太小，不足以支撑训练鲁棒的神经网络模型。

(6) 非鲁棒特征的假设

Ilyas 等人[63]提出了一种不同的假设解释，他们认为对抗样本是神经网络的基本数据特征，而不是没有根据的错误。他们证明了对抗样本的存在可以归因于非鲁棒的特征，与标准训练框架无关，这与研究人员普遍认为的结论相反。非鲁棒的特征在数据集中普遍存在，这些特征有可能是深度学习模型实现更高准确性的有效来源。作者研究还证明了解耦鲁棒和非鲁棒特征的可能性，并且表明鲁棒的特征比非鲁棒的特征更加符合人类的感知。

4 对抗攻击的防御方法

对抗样本的存在严重威胁了深度学习模型在现实场景中的大量应用，甚至打击了研究者对于深度学习前景的信心。幸运的是，自对抗样本发现以来有许多研究工作来针对对抗攻击进行防御，总结起来可以大致分为 4 个大的方向：增强模型鲁棒性的防御、输入预处理的防御、对抗样本的检测防御和可验证鲁棒性的防御。表 3 对本小节防御方法进行了分类总结。

4.1 增强模型鲁棒性的防御

4.1.1 对抗训练

对抗训练框架被学术界普遍认为是对抗攻击的最强有力的防御策略，该策略的主要的做法是让模型在训练过程中暴露在对抗样本中，以获得对对抗样本的免疫力。对抗训练最初在文献[7]和[14]中使用，Madry 等人[17]首次从理论上研究并且通过鲁棒优化的视角来表述它，随后很多研究工作专注于对抗训练。

Ding 等人[64]指出对抗训练对训练数据分布很敏感。Song 等人[65]强调对抗训练有很差的泛化性。对抗训练虽然有不足之处，但仍然被许多研究者关注。在近来的几年里，改进对抗训练的多种变体出现。Wang 等人[66]提出了 MART(MisclassificationAware adveRsarial Training)方法，将正常样本的错误分类结果在训练过程中的显著影响考虑在内，进一步提升了对抗训练模型的鲁棒性。Vivek 等人[67]提出了一种 Dropout 调度方法，用单步方法提高对抗训练的有效性。Song 等人[68]提出了对抗训练的鲁棒局部特征(Robust Local Features for Adversarial Training, RLFAT)，在训练过程中使用输入的随机块洗牌，提升了对抗训练模型的泛化能力。Zheng 等人[69]提出在训练过程的多个迭代中使用相同的对抗扰动，在减少整个训练过程计算量的同时取得可接受的性能，使得对抗训练更加有效。

考虑对抗训练的进一步变体，Dong 等人[70]提出了对抗分布训练(Adversarial Distributional Training)，该方法也将对抗训练形式化为极大极小化问题，但不同的是内部最大化的目的是在熵正则化下学习对抗分布，外部的最小化问题使最坏情况下对抗分布的损失最小化。Jia 等人[71]提出了可学习攻击策略的对抗训练 LAS-AT 方法，通过学习自动生成攻击策略，在不同的训练阶段采用不同的攻击策略来提升对抗训练方法的鲁棒性。

文献中关于对抗训练还有很多侧重于分析对抗训练，而不是设计其变体。Xie 等人[72]报告了对抗训练的一些有趣的特性，其中最引人注意的是对正常图像和对抗图像进行单独的批量归一化(Batch Normalization)导致对抗鲁棒性的改善，以及不受视觉模型中网络深度的限制，更深模型的对抗鲁棒性会持续地改善。

Wong 等人[73]研究表明，FGSM 攻击结合随机初始化的对抗训练相比于 PGD 攻击的对抗训练同样有效。但 Andriushchenko 等人[74]对 FGSM 攻击结合随机初始化的对抗训练方面的改进表示了否定。

文献中还包含对抗训练被使用来满足特定任务的需要。Wu 等人[75]提出了一种对抗训练方法，该方法中的对抗样本是专门针对物理世界的攻击生成的。作者指出，通常用于数字世界攻击的对抗训练和随机平滑对于物理世界的攻击效果不佳。

表 3　图像分类领域对抗样本的防御方法总结

Table 3　Summary of adversarial defense methods in image classification field

防御方法分类	具体防御原理	代表性工作	优势	局限性	适用场景
增强模型鲁棒性的防御	对抗训练使用对抗样本和正常样本一起训练来提升模型本身的鲁棒性	[7]、[14]、[17]	改善模型的决策边界，防御白盒迭代攻击非常有前景的防御	1.较大的计算开销 2.防御策略是非自适应性的，仅对已知的攻击防御效果好，对全新的攻击和自适应攻击的防御效果要差	在白盒迭代攻击场景下，增强模型鲁棒性
	其它的增强鲁棒性的防御，通过正常的训练数据改变模型的相关结构，例如使用新型损失函数替换经典的交叉熵损失函数	[76]、[77]	即插即用，可以和对抗训练结合使用来进一步增强模型的对抗鲁棒性	设计思路较为复杂，模型的训练更加复杂，且很多策略会进一步引入计算开销	与对抗训练结合使用进一步增强模型鲁棒性
输入预处理的防御	对输入模型的数据样本进行压缩、变换等操作来缓解对抗扰动的影响	[78-84]	1.该防御往往计算开销较小 2.可以在一定程度上消除非最优的冗余对抗扰动 3.可以和其他防御结合使用	该防御不能百分之百消除对抗扰动的影响，未消除的对抗扰动仍然会导致模型错误分类，在白盒或者更强的攻击下没有足够的对抗鲁棒性	1.非最优的对抗图像 2.黑盒攻击 3.可以集成设计在更复杂的防御中
	对输入模型的数据样本进行去噪处理来消除对抗扰动	[85-87]			
对抗样本的检测防御	为预先训练的模型提供相应的机制或者模块，来检测对抗样本，以保护模型免受对抗攻击	[88-91]	1.计算开销小 2.防御策略模型无关 3.可以和其他防御结合使用	敌手一旦预测到防御策略，调整相应的攻击策略很容易绕过此类防御	1.远离模型决策边界的对抗图像检测 2.可以集成设计在更复杂的防御中
可验证鲁棒性的防御	可验证的鲁棒性防御试图保证在正常图像的L_p范数球内目标模型不存在对抗攻击可以使模型出错	[92-107]	此防御可以在一定程度解决无休止的攻防循环，为模型提供预测准确率的下界，在一定条件下可以防御任何攻击	可验证鲁棒性的防御其评估通常缺乏通用性，且目前无法扩展到像 ImageNet 的大规模数据集上	需要理论保证的可验证防御要求下

4.1.2　其它增强鲁棒性的防御

除了对抗训练通过专注于对抗样本来修改模型的权重外，还有许多方法通过正常的训练数据改变模型的相关结构，从而增强了模型的对抗鲁棒性。Pang 等人[76]建议使用最大化马氏中心损失(MaxMahalanobis Center Loss)替换 softmax 的交叉熵损失，以此来增强模型的对抗鲁棒性。Xiao 等人[77]提出使用 k-Winner-Takes-All (k-WTA)的不连续激活函数替换 ReLU 激活函数,从而保护模型不受基于梯度的攻击。

4.2　输入预处理的防御

基于输入预处理的防御旨在通过输入的变换来清除或者减轻对抗扰动对输入模型的影响。文献[78]-[80]研究了基于 JPEG 的输入压缩来消除图像中的对抗扰动，经过压缩处理的对抗图像显著失去了他们的模型欺骗能力。通常来说，输入预处理的防御的优点在于它可以很容易与其他防御机制结合使用，例如与对抗训练模型结合使用。Raff 等人[81]提出将多个输入变换随机组合，从而确保他们对自适应攻击的防御，但作者也发现更多的输入变换会导致模型在正常图像上的

性能显著下降。

输入数据的随机化变换有助于提升对抗鲁棒性。Xie 等人[82]研究表明随机调整对抗样本的大小会降低攻击能力，作者还发现在对抗样本中添加随机填充会降低攻击性能。Wang 等人[83]使用单独的数据变换模块对模型的输入数据进行变换来消除图像中可能存在的对抗扰动。在文献[84]中发现神经网络模型的训练过程中的高斯数据增强也有助于提升对抗鲁棒性，虽然效果很微小。

Samangouei 等人[85]首次使用 GAN 进行输入的变换，他们的方法 Defense-GAN 学习正常图像的分布，在推理阶段，它计算一个接近输入图像的输出，来消除潜在的对抗扰动。Gupta 等人[86]提出了一种基于去噪的防御，它有选择地去噪图像的显著影响区域，确保模型正常的输出结果。Akhtar 等人[87]提出了一种针对通用扰动[23]生成的对抗图像的防御框架，该框架在目标网络添加了相关的预输入层，这些预输入层被训练来修正经过扰动的图像，从而使得分类器获得正确的预测。

4.3 对抗样本的检测防御

检测防御技术主要是为预先训练的模型提供相应的机制或者模块，来检测对抗样本，以保护模型免受对抗攻击。在大多数情况下，这些方法仅限于在模型的推理阶段检测输入的对抗样本的存在。

Xu 等人[88]认为输入特征空间大的过于冗余，因此提出特征压缩(Feature Squeezing)机制，通过剔除不必要的输入特征来减少敌手可用的冗余特征自由度。特征压缩机制是一种简单的对抗样本检测框架，它将模型对原始正常图像的预测与压缩后的模型预测进行比较，如果前后两次的预测差异高于指定的阈值，则判定该图像是对抗图像，因此将被丢弃。作者的工作提出了两种技术，分别是压缩颜色位(Squeezing Color Bits)和空间平滑(Spatial Smoothing)，前者可以在不损失太多信息的情况下显著降低比特深度，后者是一种降低图像噪声的处理技术。

Meng 等人[89]提出了神经网络检测对抗样本的框架 MagNet，该框架由两个部分组成，分别是检测器(Detector)和重整器(Reformer)，前者主要用于拒绝远离不同类别决策边界的样本，后者主要用于给定一个输入样本，寻找输入样本的近似值，该近似值靠近正常样本的决策边界，使用上述的构件来进行对抗样本的检测。

Liang 等人[90]将图像的对抗扰动视为噪声，采用标量量化和空间平滑滤波检测对抗扰动。Feinman 等人[91]提出利用不确定性估计和在神经网络的特征空间中执行密度估计来检测对抗扰动。

4.4 可验证鲁棒性的防御

虽然文献中有很多防御方法，但随着后来的研究表明存在更强的攻击可以击败现有的防御方法[108]。尽管对抗训练被研究者广泛认可是一种相对出色的防御策略，但其也存在缺陷，例如文献[109]研究发现在 L_∞ 范数约束的扰动下对抗训练的模型对于 L_p 范数约束的扰动下的攻击仍然很脆弱，其中的 $p \neq \infty$。可验证的鲁棒性防御试图保证在正常图像的 L_p 范数球内目标模型不存在对抗攻击可以使模型出错。这个可验证的鲁棒性保证要么是提供打破给定防御的最小的 L_p 范数扰动[92-93]，要么是提供范数的下界[94-96]。还有一些工作，旨在提高网络的鲁棒性，并推动产生更适合鲁棒性验证技术的模型[97-98]。目前大多数的可验证防御仅能证明针对一种范数约束扰动的鲁棒性，却很难同时证明针对多种范数约束扰动的鲁棒性，但也存在一些工作可以同时证明针对多种范数约束扰动的鲁棒性[99-100]。

Croce 等人[99]对使用 ReLU 激活函数的神经网络提出了一个正则化方法，以此来增强模型对于 L_1 和 L_∞ 攻击的鲁棒性，并且表明了它的结果对于任何的 L_p 范数($p \geq 1$)都是可证明的，都能保证构建可证明的鲁棒模型。相较于为模型的 top-1 预测提供可验证的鲁棒性，Jia 等人[101]使用高斯随机平滑(Gaussian Randomized Smoothing)方法为模型的 top-k 预测推导出 L_2 范数约束扰动下的严格鲁棒性，该方法建立在文献[102]介绍的随机平滑概念上。Zhai 等人[103]也基于随机平滑的想法提出了一种模型的 MACER (MAximizing the CErtified Radius)方法，该方法可以扩展到大型模型上。Fischer 等人[104]扩展了随机平滑的概念，将平移、旋转等参数化转换纳入其中，并且验证了模型在参数空间的鲁棒性。Zhang 等人[105]将随机分类器中的高斯平滑噪声扩展至非高斯噪声，他们设计了一个非高斯的平滑分布族，该工作对 L_1、L_2 和 L_∞ 攻击的防御更加有效。文献[106-107]中研究了更多的针对对抗补丁攻击[45]的可验证防御。

综上所述，近来可验证鲁棒性防御的研究方向在对抗机器学习领域逐渐变得热门起来，吸引了机器学习社区的众多研究人员的关注，该领域未来是极有前景的研究方向，必将促进可信机器学习领域的进一步的发展。

5 对抗样本的总体发展趋势和未来研究展望

5.1 对抗攻击

最近的对抗攻击方法通常旨在进一步减小对抗扰动的范数约束大小以及增强黑盒攻击中对抗样本的可迁移性，使攻击对现实场景中更具威胁性。近来的黑盒攻击研究比较活跃，根据基于迁移的黑盒攻击的相关文献的报告，它在具有相似网络架构的模型之间可以更好迁移。基于决策边界的攻击相比于基于置信度分数查询的攻击更受欢迎。通常基于置信度分数查询的攻击优化两个相互矛盾的目标：(a)通过使用更少的查询次数获得更高的错误率；(b)限制对抗扰动的范数约束尽可能的小，以确保不可察觉性。通常基于决策边界的攻击最常用的策略是先查询黑盒模型得到较大的扰动，然后在保持错误预测的同时，通过微调来减小扰动的范数。

5.2 对抗防御

尽管针对攻击的防御层出不穷，但对抗机器学习社区公认的相对有前途的防御仍然是对抗训练。比较有趣的是，在对抗样本的开山之作 Szegedy 等人[7]的工作中，对抗样本的概念和对抗训练的概念同时被提出，后来的关于防御方面的大多数文献都偏离了最初的增强模型自身鲁棒性的想法，大多防御策略依赖于特定的规则和启发式的方法，因此在更强的攻击或者不同的攻击条件下，它们中的许多会被攻破。事实上，Tramer 等人[110]表明顶级学术会议上的十三种不同的防御措施可以被适应性攻击(Adaptive Attack)所攻破。对抗机器学习社区从防御角度来看更关注于对抗训练和可验证鲁棒性的防御上，因为这两个方向相对来说最有发展前景，但是也存在经过训练的鲁棒模型在正常图像上的精度下降的挑战，在文献中很容易观察到一般经受住较强攻击的方法在正常图像上的精度会降低。

5.3 未来展望

自对抗样本发现以来，其论文就如雨后春笋般大量涌现，世界各国都推出相应的可信人工智能计划，为构建可信人工智能做好了顶层设计，为人工智能在安全至关重要场景中的大量应用铺平了道路。由于该领域仍有很多悬而未决的重大挑战，因此可以预见该方向仍然会是非常活跃的研究领域，以下将探讨一些未来有前景的研究方向。

深度神经网络在安全关键领域的大量应用使得模型需要更好的鲁棒性，大多数的对抗防御策略仅限于经验评估，并不声称对未知攻击具有鲁棒性，可验证鲁棒性的防御虽然提供了鲁棒性的下限，确保被评估的防御系统的性能不会低于下限，但可验证鲁棒性防御的评估通常缺乏通用性和可扩展性[111]，因此研究其通用性和可扩展性是有前景的方向。

在对抗机器学习领域，即使在小的数据集上，生成高质量对抗样本的计算代价依然很高，因此对于响应时间要求很高的应用系统来说，攻击算法的效率至关重要[112]。对于防御而言，一个优秀的防御评估需要测试大量的攻击，因为计算效率的问题，这些攻击在给定数据集(例如 ImageNet 数据集)上的计算是不可行的，因此设计强大且高效的攻击算法势在必行。

目前存在大量针对经典的卷积神经网络架构攻击和防御的研究，但鲜有针对新颖网络架构的研究，例如二值化神经网络模型(Binarized Neural Networks, BNNs)、神经常微分方程模型(Neural Ordinary Differential Equations, Neural ODEs)以及在计算机视觉领域越来越受欢迎 Transformer 模型及其变体。近来的研究指出，BNN 模型[113]、Neural ODEs 模型[114]以及 Transformer 模型[115-116]都可以在保证正常样本上精度的前提下，在对抗鲁棒性上超越传统的卷积神经网络模型。虽然有一些积极的结果，但是并没有被大家接受的答案来解释这些模型鲁棒性优越的根本原因，该方向的深入研究有利于设计更高效更鲁棒的模型，因此是一个充满希望的研究领域。

基于前文的文献梳理和总结可以发现对抗样本研究中还存在一些具有挑战性的研究问题，例如关于对抗样本存在性和可迁移性公认的可证明解释[117]，寻找具有内在鲁棒性的新型网络架构模型[118]，模型对抗鲁棒性和准确性的平衡[119]，模型对抗鲁棒性和公平性的平衡[120]，有前景的对抗训练相关问题的深入研究[121]以及对抗样本和可解释性相结合方向的研究[122]，希望未来有工作可以解决这些有趣且重要的开放问题。

事物都有两面性，对抗样本样本也不例外。近期的许多研究表明对抗样本也可以有积极的作用。例如对抗样本被用来进行个人隐私保护[123]，被用来生成数据或模型水印[124]和指纹[125]保护深度学习模型免受知识产权侵害。对抗样本也可以被很好地利用来提升不平衡学习的性能[126]，适当设计的对抗样本也是有效的数据增强工具，可以同时提高模型的泛化性和对抗鲁棒性[127]。此外，值得注意的是对抗样本也可以被用来设计更加鲁棒的文本验证码[128]。综上，在未来的研究中，如何合理的利用对抗样本来达到“对抗向善”的效果也是一个非常有前景的研究方向。

基础模型采用与任务无关的大规模数据预训练来进行表征学习，然后根据特定下游任务进行微调适应，因为基础模型的缺陷会被下游的所有适应性模型继承，因此如何将对抗鲁棒性纳入基础模型的预训练，以及如何从预训练到微调最大化迁移对抗鲁棒性是至关重要的。Fan 等人[129]和 Wang 等人[130]在元学习和对比学习中对抗鲁棒性保存和迁移方面展示了有前景的结果。基础模型的快速增长和日益强烈的需求创造了一个独特的机会，其对抗性鲁棒性被倡导作为下一代可信人工智能原生属性，因此基础模型对抗鲁棒性的研究是一个有前景的研究问题。

对抗样本的出现影响了深度神经网络在安全至关重要领域的部署，导致一些研究者甚至出现了对深度学习发展持悲观的态度，从目前来看，对抗样本的痛点似乎要长期留在深度学习研究中，解决对抗样本问题仍然任重而道远。

6 结束语

本文旨在总结和梳理对抗样本在图像分类领域的攻击和防御方法，以期促进对抗机器学习领域构建更加可信、更加鲁棒的深度学习模型。为了增强研究人员对现实世界中的对抗样本的重视，简单梳理了物理世界中的对抗样本。对抗样本为什么会存在这个问题，迄今为止没有达成一个有理论依据的统一共识，这仍然是一个有趣且具有挑战性的问题，因此梳理和总结了解释对抗样本存在性相关的工作，以期为后续的研究者解决该问题提供一些概况和灵感。最后，基于梳理的大量文献，思考和分析了对抗样本的总体发展趋势和面临的挑战以及未来研究展望。

参考文献：

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 770-778.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems,2017, 30.

[3] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of the International Conference on Machine Learning. PMLR,2014: 1764-1772.

[4]CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision,2015: 2722-2730.

[5] LIAO Y, VAKANSKI A, XIAN M. A deep learning framework for assessing physical rehabilitation exercises[J].IEEE Transactions on Neural Systems and Rehabilitation Engineering,2020, 28(2): 468-477.

[6] KRIZHEVSKY A,SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J].Advances in Neural Information Processing Systems,2012, 25.

[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the International Conference on Learning Representations, 2014.

[8] CARLINI N, WAGNER D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//2018 IEEE Security and Privacy Workshops (SPW). IEEE,2018: 1-7.

[9] EBRAHIMI J, RAO A, LOWD D, et al. Hotflip: White-box adversarial examples for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),2018.

[10] GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial examples for malware detection[C]//European Symposium on Research in Computer Security. Springer, Cham,2017: 62-79.

[11] 潘文雯,王新宇,宋明黎,陈纯.对抗样本生成技术综述[J].软件学报,2020,31(1):67-81.
PAN WW, WANG XY, SONG ML,et al. Survey on generating adversarialexamples[J].JournalofSoftware,2020,31(1):67–81.

[12] 陈梦轩,张振永,纪守领,魏贵义,邵俊.图像对抗样本研究综述[J].计算机科学,2022,49(2):92-106.
CHEN MX, ZHANG ZY, JI SL, et al.Survey of research progress on adversarial examples in images[J].Computer Science,2022,49(2):92-106.

[13] 白祉旭,王衡军,郭可翔.基于深度神经网络的对抗样本技术综述[J].计算机工程与应用,2021,57(23):61-70.
BAI ZX, WANG HJ, GUO KX. Summary of Adversarial Examples Techniques Based on Deep Neural Networks[J]. Computer Engineering and Applications, 2021, 57(23): 61-70.

[14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations,2015.

[15] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of the International Conference on Learning Representations Workshop track,2017.

[16] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018: 9185-9193.

[17] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// Proceedings of the International Conference on Learning Representations,2018.

[18] WANG X, HE K. Enhancing the transferability of adversarial attacks through variance tuning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021: 1924-1933.

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

[20] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). IEEE,2017: 39-57.

[21] ZHU M, CHEN T, WANG Z.Sparse and imperceptible adversarial attack via a homotopy algorithm[C]// Proceedings of the International Conference on Machine Learning. PMLR,2021: 12868-12877.

[22] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 2574-2582.

[23] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 1765-1773.

[24] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017: 506-519.

[25]LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]// Proceedings of the International Conference on Learning Representations,2017.

[26] SHI Y, WANG S, HAN Y. Curls & whey: Boosting black-box adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019: 6519-6527.

[27] HUANG Z, ZHANG T. Black-box adversarial attack with transferable model-based embedding[C]//Proceedings of the International Conference on Learning Representations,2020.

[28] WANG Z, GUO H, ZHANG Z, et al.Feature importance-aware transferable adversarial attacks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2021: 7639-7648.

[29] 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,2017: 15-26.

[30] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//Proceedings of the International Conference on Machine Learning, PMLR, 2018: 2137-2146.

[31] 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 AAAI Conference on Artificial Intelligence. 2019, 33(1): 742-749.

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

[33] DU J, ZHANG H, ZHOU J T, et al. Query-efficient meta attack to deep neural networks[C]//Proceedings of the International Conference on Learning Representations,2020.

[34]MA C, CHEN L, YONG J H. Simulating unknown target models for query-efficient black-box attacks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021: 11835-11844.

[35] BRENDEL W, RAUBER J, BETHGE M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models[C]//Proceedings of the International Conference on Learning Representations,2018.

[36] CHENG M, LE T, CHEN P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[C]//Proceedings of the International Conference on Learning Representations,2019.

[37] CHEN J, JORDAN M I, WAINWRIGHT M J. Hopskipjumpattack: A query-efficient decision-based attack[C]// Proceedings of the IEEE Symposium on Security and Privacy (SP),IEEE,2020:1277-1294.

[38] LI H, XU X, ZHANG X, et al. Qeba: Query-efficient boundary-based blackbox attack[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2020:1221-1230.

[39] LI H, LI L, XU X, et al. Nonlinear projection based gradient estimation for query efficient blackbox attacks[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, 2021: 3142-3150.

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

[41] KONG Z, GUO J, LI A, et al. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020: 14254-14263.

[42]CAO Y, XIAO C, CYR B, et al. Adversarial sensor attack on lidar-based perception in autonomous driving[C]//

Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security,2019: 2267-2281.

[43]CAO Y, XIAO C, YANG D, et al. Adversarial objects against lidar-based autonomous driving systems[J]. arXiv:1907.05418,2019.

[44]THYS S, VAN RANST W, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,2019: 0-0.

[45]BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J].Advances in Neural Information Processing Systems,2017,30.

[46]DUAN R, MA X, WANG Y, et al. Adversarial camouflage: Hiding physical-world attacks with natural styles[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020: 1000-1008.

[47] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J].Advances in Neural Information Processing Systems,2014,27.

[48] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[49] HO C H, LEUNG B, SANDSTROM E, et al. Catastrophic child's play: Easy to perform, hard to defend adversarial attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9229-9237.

[50] JING Y, YANG Y, FENG Z, et al. Neural style transfer: A review[J]. IEEE Transactions on Visualization and Computer Graphics,2019,26(11):3365-3385.

[51] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[C]//Proceedings of the International Conference on Learning Representations Workshop,2015.

[52] SONG Y, KIM T, NOWOZIN S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[C]//Proceedings of the International Conference on Learning Representations, 2018.

[53] FAWZI A, FAWZI O, FROSSARD P. Fundamental limits on adversarial robustness[C]//Proceedings of the International Conference on Machine LearningWorkshop on Deep Learning,PMLR,2015.

[54] TABACOF P, VALLE E. Exploring the space of adversarial images[C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2016: 426-433.

[55] BUCKMAN J, ROY A, RAFFEL C, et al. Thermometer encoding: One hot way to resist adversarial examples[C]// Proceedings of the International Conference on Learning Representations,2018.

[56] TANAY T, GRIFFIN L. A boundary tilting persepective on the phenomenon of adversarial examples[J]. arXiv:1608.07690,2016.

[57] GILMER J, METZ L, FAGHRI F, et al. Adversarial spheres[C]//Proceedings of the International Conference on Learning Representations,2018.

[58] MAHLOUJIFAR S, DIOCHNOS D I, MAHMOODY M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2019,33(1):4536-4543.

[59]SHAFAHI A, HUANG W R, STUDER C, et al. Are adversarial examples inevitable?[C]//Proceedings of the International Conference on Learning Representations, 2019.

[60] FAWZI A, FAWZI H, FAWZI O. Adversarial vulnerability for any classifier[J].Advances in neural information processing systems,2018,31.

[61] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]//Proceedings of the International Conference on Learning Representations, 2018.

[62] SCHMIDT L, SANTURKAR S, TSIPRAS D, et al. Adversarially robust generalization requires more data[J].Advances in Neural Information Processing Systems, 2018, 31.

[63] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[J].Advances in Neural Information Processing Systems,2019,32.

[64] DING G W, LUI K Y C, JIN X, et al. On the Sensitivity of Adversarial Robustness to Input Data Distributions[C]// Proceedings of the International Conference on Learning Representations,2019.

[65] SONG C, HE K, WANG L, et al. Improving the generalization of adversarial training with domain adaptation[C]// Proceedings of the International Conference on Learning Representations,2019.

[66] WANG Y, ZOU D, YI J, et al. Improving adversarial robustness requires revisiting misclassified examples[C]// Proceedings of the International Conference on Learning Representations,2020.

[67] VIVEK B S, BABU R V. Single-step adversarial training with dropout scheduling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,IEEE,2020:947-956.

[68] SONG C, HE K, LIN J, et al. Robust local features for improving the generalization of adversarial training[C]// Proceedings of the International Conference on Learning Representations,2020.

[69] ZHENG H, ZHANG Z, GU J, et al. Efficient adversarial training with transferable adversarial examples[C]// Proceedings of the IEEE/CVF Conference on Computer Vi-

sion and Pattern Recognition,2020:1181-1190.
[70] DONG Y, DENG Z, PANG T, et al. Adversarial distributional training for robust deep learning[J].Advances in Neural Information Processing Systems,2020,33: 8270-8283.
[71]JIA X, ZHANG Y, WU B, et al. LAS-AT: Adversarial training with learnable attack strategy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022:13398-13408.
[72] XIE C, YUILLE A. Intriguing properties of adversarial training at scale[C]//Proceedings of the International Conference on Learning Representations,2020.
[73] WONG E, RICE L, KOLTER J Z. Fast is better than free: Revisiting adversarial training[C]//Proceedings of the International Conference on Learning Representations, 2020.
[74] ANDRIUSHCHENKO M, FLAMMARION N. Understanding and improving fast adversarial training[J]. Advances in Neural Information Processing Systems,2020, 33:16048-16059.
[75] WU T, TONG L, VOROBEYCHIK Y. Defending against physically realizable attacks on image classification[C]// Proceedings of the International Conference on Learning Representations,2020.
[76] PANG T, XU K, DONG Y, et al. Rethinking softmax cross-entropy loss for adversarial robustness[C]// Proceedings of the International Conference on Learning Representations,2020.
[77] XIAO C, ZHONG P, ZHENG C. Enhancing adversarial defense by k-winners-take-all[C]//Proceedings of the International Conference on Learning Representations, 2020.
[78] LIU Z, LIU Q, LIU T, et al. Feature distillation: Dnn-oriented jpeg compression against adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019: 860-868.
[79] DAS N, SHANBHOGUE M, CHEN S T, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression[J].arXiv:1705.02900,2017.
[80] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[C]//Proceedings of the International Conference on Learning Representations, 2018.
[81] RAFF E, SYLVESTER J, FORSYTH S, et al. Barrage of random transforms for adversarially robust defense[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019: 6528-6537.
[82] XIE C, WANG J, ZHANG Z, et al. Adversarial examples for semantic segmentation and object detection[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:1369-1378.
[83] WANG Q, GUO W, ZHANG K, et al. Learning adversary-resistant deep neural networks[J].arXiv:1612.01401, 2016.
[84] ZANTEDESCHI V, NICOLAE M I, RAWAT A. Efficient defenses against adversarial attacks[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 39-49.
[85] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]//Proceedings of the International Conference on Learning Representations, 2018.
[86] GUPTA P, RAHTU E. Ciidefence: Defeating adversarial attacks by fusing class-specific image inpainting and image denoising[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:6708-6717.
[87] AKHTAR N, LIU J, MIAN A. Defense against universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3389-3398.
[88] XU W, EVANS D, QI Y. Feature squeezing: Detecting ad-versarial examples in deep neural networks[C]// Proceedings of Network and Distributed System Security Symposium,2018.
[89] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security,2017:135-147.
[90] LIANG B, LI H, SU M, et al. Detecting adversarial examples in deep networks with adaptive noise reduction[J]. arXiv:1705.08378,2017
[91] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[J].arXiv:1703.00410, 2017.
[92] KATZ G, BARRETT C, DILL D L, et al. Reluplex: An efficient SMT solver for verifying deep neural networks[C]//Proceedings of the International Conference on Computer Aided Verification,Springer,Cham,2017: 97-117.
[93] TJENG V, XIAO K, TEDRAKE R. Evaluating robustness of neural networks with mixed integer programming[C]//Proceedings of the International Conference on Learning Representations,2019.
[94] HEIN M, ANDRIUSHCHENKO M. Formal guarantees on the robustness of a classifier against adversarial manipulation[J].Advances in Neural Information Processing Systems,2017,30.
[95] RAGHUNATHAN A, STEINHARDT J, LIANG P. Certified defenses against adversarial examples[C]// Proceedings of the International Conference on Learning Representations,2018.
[96] WONG E, KOLTER Z. Provable defenses against adversarial examples via the convex outer adversarial polytope[C]//Proceedings of the International Conference on

Machine Learning,PMLR,2018:5286-5295.
[97] MIRMAN M, GEHR T, VECHEV M. Differentiable abstract interpretation for provably robust neural networks[C]//Proceedings of the International Conference on Machine Learning,PMLR,2018:3578-3586.
[98] XIAO K Y, TJENG V, SHAFIULLAH N M, et al. Training for faster adversarial robustness verification via inducing relu stability[C]//Proceedings of the International Conference on Learning Representations,2019.
[99] CROCE F, HEIN M. Provable robustness against all adversarial lp-perturbations for p≥1[C]//Proceedings of the International Conference on Learning Representations,2020.
[100] TRAMER F, BONEH D. Adversarial training and robustness for multiple perturbations[J].Advances in Neural Information Processing Systems,2019,32.
[101] JIA J, CAO X, WANG B, et al. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing[C]//Proceedings of the International Conference on Learning Representations,2020.
[102] CAO X, GONG N Z. Mitigating evasion attacks to deep neural networks via region-based classification[C]// Proceedings of the 33rd Annual Computer Security Applications Conference,2017:278-287.
[103] ZHAI R, DAN C, HE D, et al. Macer: Attack-free and scalable robust training via maximizing certified radius[C]//Proceedings of the International Conference on Learning Representations,2020.
[104] FISCHER M, BAADER M, VECHEV M. Certified defense to image transformations via randomized smoothing[J].Advances in Neural Information Processing Systems,2020,33:8404-8417.
[105] ZHANG D, YE M, GONG C, et al. Black-box certification with randomized smoothing: A functional optimization based framework[J].Advances in Neural Information Processing Systems,2020,33:2316-2326.
[106] CHIANG P Y, NI R, ABDELKADER A, et al. Certified defenses for adversarial patches[C]//Proceedings of the International Conference on Learning Representations, 2020.
[107] AWASTHI P, JAIN H, RAWAT A S, et al. Adversarial robustness via robust low rank representations[J]. Advances in Neural Information Processing Systems, 2020, 33:11391-11403.
[108] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[C]//Proceedings of the International Conference on Machine Learning,PMLR,2018:274-283.
[109] SCHOTT L, RAUBER J, BETHGE M, et al. Towards the first adversarially robust neural network model on MNIST[C]//Proceedings of the International Conference on Learning Representations,2019.
[110] TRAMER F, CARLINI N, BRENDEL W, et al. On adaptive attacks to adversarial example defenses[J].Advances in Neural Information Processing Systems,2020,33: 1633-1645.
[111] PAUTOV M, TURSYNBEK N, MUNKHOEVA M, et al. CC-Cert: A probabilistic approach to certify general robustness of neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2022,36(7): 7975-7983.
[112] GUESMI A, KHASAWNEH K N, ABU-GHAZALEH N, et al. ROOM: Adversarial machine learning attacks under real-time constraints[J].arXiv:2201.01621,2022.
[113] DUNCAN K, KOMENDANTSKAYA E, STEWART R, et al. Relative robustness of quantized neural networks against adversarial attacks[C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN). IEEE,2020:1-8.
[114] HUANG Y, YU Y, ZHANG H, et al. Adversarial robustness of stabilized neural ode might be from obfuscated gradients[C]//Proceedings of the Mathematical and Scientific Machine Learning,PMLR,2022:497-515.
[115] MAHMOOD K, MAHMOOD R, VAN DIJK M. On the robustness of vision transformers to adversarial examples[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:7838-7847.
[116] BHOJANAPALLI S, CHAKRABARTI A, GLASNER D, et al. Understanding robustness of transformers for image classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021: 10231-10241.
[117] SHAMIR A, MELAMED O, BENSHMUEL O. The dimpled manifold model of adversarial examples in machine learning[J].arXiv:2106.10151,2021.
[118] SUN J, JIANG T, LI C, et al. Searching for robust neural architectures via comprehensive and reliable evaluation[J].arXiv:2203.03128,2022.
[119] CAO G, WANG Z, DONG X, et al. Vanilla feature distillation for improving the accuracy-robustness trade-off in adversarial training[J].arXiv:2206.02158,2022.
[120] SUN H, WU K, WANG T, et al. Towards fair and robust classification[C]//Proceedings of the IEEE 7th European Symposium on Security and Privacy (EuroS&P),IEEE, 2022:356-376.
[121] WANG H, ZHANG A, ZHENG S, et al. Removing batch normalization boosts adversarial training[C]// Proceedings of the International Conference on Machine Learning,PMLR,2022:23433-23445.
[122] 董胤蓬,苏航,朱军.面向对抗样本的深度神经网络可解释性分析[J/OL].自动化学报:1-14[2022-01-18].
Dong Y P, SU H, ZHU J. Towards interpretable deep neural networks by leveraging adversarial examples[J/OL].Acta

Automatica Sinica : 1-14 [2022-01-18].

[123]HUANG H, MA X, ERFANI S M, et al. Unlearnable examples: Making personal data unexploitable[C]// Proceedings of the International Conference on Learning Representations,2021.

[124]ARAMOON O, CHEN P Y, QU G. Don't Forget to Sign the Gradients![C]//Proceedings of Machine Learning and Systems,2021,3:194-207.

[125]WANG S, WANG X, CHEN P Y, et al. Characteristic examples: high-robustness, low-transferability fingerprinting of neural networks[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence,2021:575-582.

[126]ZHANG J, ZHANG L, LI G, et al. Adversarial examples for good: adversarial examples guided imbalanced learning[J].arXiv:2201.12356,2022.

[127]HSU C Y, CHEN P Y, LU S, et al. Adversarial examples can be effective data augmentation for unsupervised machine learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2022.

[128]SHAO R, SHI Z, YI J, et al. Robust text captchas using adversarial examples[J].arXiv:2101.02483,2021.

[129]FAN L, LIU S, CHEN P Y, et al. When does contrastive learning preserve adversarial robustness from pretraining to finetuning?[J].Advances in Neural Information Processing Systems,2021,34:21480-21492.

[130]WANG R, XU K, LIU S, et al. On fast adversarial robustness adaptation in model-agnostic meta-learning[C]//Proceedings of the International Conference on Learning Representations,2021.