DOI:10.7544/issn1000-1239.20220507
59(10):2164-2189,2022

面向图像分类的对抗鲁棒性评估综述

李自拓　孙建彬　杨克巍　熊德辉
(国防科技大学系统工程学院　长沙　410073)
(lizituo0926@163.com)

A Review of Adversarial Robustness Evaluation for Image Classification

Li Zituo, Sun Jianbin, Yang Kewei, and Xiong Dehui
(*College of Systems Engineering, National University of Defense Technology, Changsha* 410073)

Abstract　In recent years, artificial intelligence algorithms represented by deep learning have been successfully used in the fields such as financial security, automatic driving, medical diagnosis. However, the emergence of adversarial attacks has brought huge security risks to the application of image classification, which is a basic visual task in the above fields. Improving the ability of deep learning model to resist adversarial attacks (i.e., the adversarial robustness) has become a feasible technique to effectively alleviate this problem. In order to evaluate the adversarial robustness of deep learning model scientifically and comprehensively, many scholars have carried out in-depth research on adversarial robustness evaluation from the perspectives of benchmark evaluation and index evaluation. This paper reviews the adversarial robustness mainly from the perspective of index evaluation. Firstly, we introduce the concepts related to adversarial examples and the reasons for their existence, and summarize the evaluation criteria that should be followed in the evaluation of adversarial robustness. Secondly, we focus on sorting out existing adversarial robustness evaluation indicators from two aspects of attacked model and test data. Then, the mainstream image classification datasets and the adversarial attack-defense integration tools are analyzed and summarized to lay a foundation for the follow-up relative research. Finally, the advantages and disadvantages of the current research and the potential future research direction are discussed. This paper aims to provide practitioners or learners in related fields with a comprehensive, systematic and objective overview of adversarial robustness evaluation index for image categorization.

Key words　adversarial robustness; evaluation indicators; adversarial attack and defense; image classification; deep learning

摘　要　近年来,以深度学习为代表的人工智能技术在金融安防、自动驾驶、医疗诊断等领域取得了较为成功的应用.然而,图像分类作为上述应用中的一项基础视觉任务,正遭受着对抗攻击等技术手段带来的巨大安全隐患.提高深度学习模型抵御对抗攻击的能力(即对抗鲁棒性)成为有效缓解该问题的可行技术途径.为了科学、全面地提升深度学习模型的对抗鲁棒性,众多学者从基准评估和指标评估2个角度围绕对抗鲁棒性评估开展了大量研究.该研究着重对上述指标评估相关研究进行综述:首先,介绍对

收稿日期:2022-06-11;**修回日期**:2022-08-11
基金项目:国家自然科学基金项目(72071206,71901212);湖南省科技创新项目(2020RC4046)
This work was supported by the National Natural Science Foundation of China (72071206, 71901212) and the Science and Technology Innovation Program of Hunan Province (2020RC4046).
通信作者:孙建彬(sunjianbin@nudt.edu.cn)

抗样本相关概念以及存在的原因,总结提出进行对抗鲁棒性评估时需要遵循的评估准则;其次,从被攻击模型和测试数据2个维度,重点梳理和对比分析现有的主要对抗鲁棒性评估指标;而后,分析总结现阶段主流的图像分类数据集和对抗攻防集成工具,为后续开展对抗鲁棒性评估奠定基础;最后,探讨当前研究的优势和不足,以及未来潜在的研究方向.旨在为相关领域从业人员或学习者提供一个较为全面的、系统的和客观的面向图像分类的对抗鲁棒性评估指标综述.

关键词 对抗鲁棒性;评估指标;对抗攻防;图像分类;深度学习

中图法分类号 TP391

2019年瑞莱智慧RealAI团队对人脸照片进行算法处理,将照片打印并粘贴到镜框上,通过佩戴眼镜成功攻破19款商用手机的人脸解锁[1];2020年美国东北大学团队[2]设计了一款印有特殊图案的T恤,可使穿戴者躲避智能摄像头的监测;2021年腾讯科恩实验室[3]通过在路面部署干扰信息,导致特斯拉Model S车辆经过时对车道线做出错判,致使车辆驶入反向车道……

由此可见,尽管深度学习在执行各种复杂任务时取得了出乎意料的优异表现,但在安全应用领域仍有很大的局限性.Szegedy等人[4]发现,深度学习对于精心设计的输入样本是很脆弱的.这些样本可以轻易用人类察觉不到的微小扰动,欺骗一个训练好的深度学习模型,使模型做出错误的决策.现在,深度学习中的对抗攻击技术受到了大量关注,以面向图像分类为主的对抗攻击算法[5-10]不断涌现.

在此背景下,越来越多的研究者开始关注如何提升模型抵御对抗攻击的能力,即增强模型的对抗鲁棒性,并探索出了一系列的对抗防御手段,如梯度遮蔽[11-12]、对抗训练[6,13]、数据处理[14-15]和特征压缩[16]等.尽管这些方法对于改善模型的对抗鲁棒性是有效的,但是目前针对模型对抗鲁棒性的评估框架尚未完善,主要是通过不断改进攻防算法,反复进行对抗,定性给出模型鲁棒性好坏的基准,或者使用分类准确率等指标单一地衡量模型的对抗鲁棒性.此外,许多攻击算法或多或少会受到实验条件的限制,难以适用于所有的深度学习模型,这些问题为模型的对抗鲁棒性评估(adversarial robustness evaluation)带来了挑战.

目前,面向图像分类的对抗鲁棒性评估领域还有很大的发展空间,如何正确、科学、定量且全面地评估模型的对抗鲁棒性,正在吸引业界和学术界的关注.为了更好地探究对抗鲁棒性评估问题,本文系统梳理并分析总结了面向图像分类的对抗鲁棒性评估方法,以促进该领域的研究.

1 对抗样本相关介绍

生成对抗样本是开展对抗鲁棒性评估工作的基础.为了更好地理解对抗鲁棒性评估,本节首先简要介绍对抗样本的概念和相关专业术语,并探讨对抗样本存在的原因.为行文方便,对本文使用的符号进行说明,如表1所示:

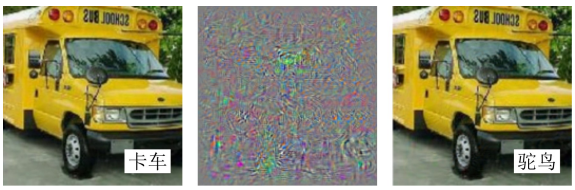
Table 1 Notations Used in this Paper

表1 本文符号说明

符号	描述
$F(\cdot)$	用于训练、测试与评估的模型
X	原始样本集
x_i	第i个原始样本
y_i	第i个原始样本的标签
ε	向原始样本添加的扰动
T	测试集
x_i^{adv}	第i个原始样本对应的对抗样本
y_i^{adv}	第i个对抗样本的标签
N	测试集中样本的数量
$\|\cdot\|$	样本间的距离度量
T'	用于评估模型测试充分性的测试集
x	用于评估模型测试充分性的测试样本
$N_{总}$	评估模型测试充分性时神经元总数
n	对抗样本未能误导模型分类错误的数量
m	对抗样本成功误导模型分类错误的数量

1.1 对抗样本及相关术语

概念1. 对抗样本.最早提出这一概念的是Szegedy等人[4],他们在原始样本上添加肉眼难以察觉的微小扰动,愚弄了当时最先进的深度神经网络(deep neural networks, DNNs),诱导模型分类错误.如图1所示,通过在原始样本上添加图中的扰动,就能让模型将卡车错误地识别成鸵鸟.



原始样本 扰动 对抗样本

Fig. 1 Illustration for the generation of adversarial example[4]

图 1 对抗样本生成示意图[4]

这些经过精心设计的样本被称为对抗样本(adversarial example),构造对抗样本的过程被称为对抗攻击.数学语言描述为:给任意一种模型 $F(\cdot)$ 以及原始样本集 X 中标签为 y_i 的原始样本 x_i 添加微小扰动 ε 生成相应的对抗样本 x_i^{adv}.此时,对抗样本 x_i^{adv} 应满足:

$$\begin{aligned} &F(x_i^{adv})\neq y_i,\\ &\text{s.t. } \|x_i-x_i^{adv}\|\leqslant\varepsilon, \end{aligned} \tag{1}$$

其中,$\|\cdot\|$ 指的是原始样本与对抗样本之间的距离度量,最常用的是 l_p 范数[9,17-18].

概念 2. 对抗攻击目标.Biggio 等人[19]指出对抗攻击的目标是根据需求实现损失函数最小化或最大化.从实际攻击效果来看,也就是通过添加精心设计的微小扰动实现模型的错误分类.根据不同的攻击目的,可以将对抗攻击目标划分为非目标攻击和目标攻击.非目标攻击指的是对抗样本诱导模型分类错误,但不指定错分为哪一种类别,而目标攻击限定了模型将标签为 i 的样本错分成第 j 类,数学语言描述分别为 $F(x_i^{adv})\neq y_i$ 和 $F(x_i^{adv})=y_j$.

概念 3. 对抗攻击知识.它指的是攻击者所掌握的相关信息,包括训练样本、模型结构和模型输出等.针对攻击者对智能系统了解情况的多少,可以将攻击划分为白盒攻击、灰盒攻击和黑盒攻击,攻击难度依次增大.由于灰盒攻击的边界难以界定,目前研究大多以白盒攻击和黑盒攻击为主,本文不对灰盒攻击进行相关介绍.

概念 4. 对抗攻击能力[20-21].指攻击者修改训练数据或测试数据的能力.在针对图像分类任务开展对抗攻击时,攻击者的能力往往仅限于对测试集数据进行修改,不考虑通过数据投毒等手段,影响模型的训练过程,这种攻击被称为探索性攻击.与之对应的诱导性攻击,指的是通过修改训练集,破坏原有训练数据的概率分布,使模型无法达到理想的分类效果.由此可见,诱导性攻击从根本上实现了对模型的攻击,比探索性攻击的攻击性更强.

通过分析图像分类全过程各环节[22]的特点,从上述提到的攻击目标、知识以及能力 3 个维度对对抗攻击方法进行分类,形成如图 2 所示的对抗攻击分类框架.诱导性攻击主要对原始数据输入以及数据处理阶段进行攻击,探索性攻击是在模型训练完成后,针对分类阶段进行攻击;倘若攻击者无法获取模型训练及训练前各阶段的信息,则开展的攻击为黑盒攻击,否则为白盒攻击;在最终的分类阶段,针对攻击者能否精确控制分类器对测试样本的分类结果,可以将对抗攻击划分为目标攻击和非目标攻击 2 类.

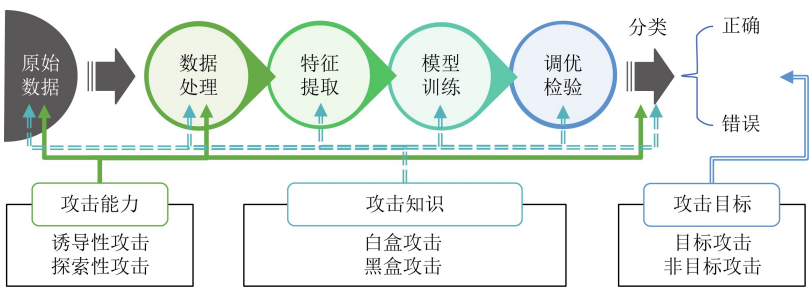


Fig. 2 Classification framework of adversarial attack

图 2 对抗攻击分类框架

1.2 对抗样本存在的解释

对抗鲁棒性评估与对抗样本密切相关,深入理解对抗样本产生的机理有助于从根本上提出科学的评估方法与指标.然而关于对抗样本存在的解释仍有许多争议,目前还没有得出一个准确统一的结论.

Szegedy 等人[4]认为网络模型的非线性特性是导致对抗样本存在的原因.具体而言,他们从正负实数分类问题中发现,由于无理数的个数要远多于有理数,训练集中无理数和有理数的比例难免失去平衡,基于此数据集训练的模型可能无法对有理数进行正确的分类.但不可否认的是,有理数是的确存在且稠密的.对抗样本好比有理数,模型的非线性特性

使得模型训练时对高维特征的提取不充分,仅仅学习到局部子空间的特征,可能导致一直存在但被发现的概率很低的对抗样本难以被观察到,进而影响了模型的决策.遗憾的是,文献[4]并没有给出相关的数学说明.

相反,Goodfellow 等人[23]恰恰认为对抗样本存在的原因与高维空间的线性模型相关,并以此提出快速梯度符号(fast gradient sign method, FGSM)攻击算法.假设模型的权重向量为 $\boldsymbol{w}$,则有

$$\boldsymbol{w}^{\mathrm{T}}\cdot x_i^{\mathrm{adv}}=\boldsymbol{w}^{\mathrm{T}}\cdot x_i+\boldsymbol{w}^{\mathrm{T}}\cdot\varepsilon. \tag{2}$$

当 $\boldsymbol{w}$ 的维数 n 非常大时,即使扰动 ε 很小,$\boldsymbol{w}^{\mathrm{T}}\cdot\varepsilon$ 可能会是一个比较大的数值.因此,得出结论:在高维空间的线性模型中,微小扰动也能使输出产生较大的变化,以致模型进行错误分类.不过,该观点也存在一些局限性,张思思等人[24]指出线性不是生成对抗样本的充分条件.还有研究表明[25]高维空间的线性模型并没有更容易发现对抗样本,不是所有的线性模型都会存在对抗样本.

此外,Moosavi-Dezfooli 等人[8]证明了全局通用的对抗扰动的存在,并提出对抗样本的存在是由于分类器决策边界的几何相关性.之后,Moosavi-Dezfooli 等人[26]又进一步证明了存在共享子空间,决策边界沿该空间正弯曲,添加扰动的样本可以越过边界,成功愚弄模型.Tanay 等人[25]也从模型决策边界角度提出了分类决策边界超出样本的子流形导致产生对抗样本的观点.

2 对抗鲁棒性评估

科学、有效地评估模型的对抗鲁棒性对于构建对抗鲁棒模型、提高智能系统安全性具有重要意义.然而,至今尚未形成一个公正、统一的对抗鲁棒性评估指标或方法.现阶段面向图像分类的对抗鲁棒性评估主要分为基准评估和指标评估 2 类.前者通过提出并改进各种攻防算法[27-31],反复进行对抗,以排名基准[32]的形式反映对抗鲁棒性的强弱;后者从对抗样本的角度出发提出一系列评估指标,旨在通过全面、合理的指标对模型的对抗鲁棒性进行评估.相比前者,后者的优势在于能够以客观可量化的方式衡量模型的对抗鲁棒性,为增强模型的对抗鲁棒性提供可解释的科学依据.

2.1 基本概念

在深度学习领域,鲁棒性(robustness)指的是智能系统在受到内外环境中多种不确定因素干扰时,依旧可以保持功能稳定的能力.而对抗鲁棒性(adversarial robustness)[12,33]专指对抗环境下模型抵御对抗攻击的能力,即模型能否对添加微小扰动的对抗样本做出正确分类的能力.以任意攻击方法在原始样本上添加扰动,模型正确识别该样本的概率越高,说明模型的对抗鲁棒性越强.从数据空间的角度来看,添加的扰动可以被描述为对抗扰动距离[7](即原始样本和对抗样本之间的距离),距离范围内的样本都能够被正确分类.因此也可以说,最小对抗扰动距离(minimal adversarial perturbation)越大,则允许添加的扰动范围越大,模型的对抗鲁棒性越强.

可以看出,对抗鲁棒性评估的关键是计算最小对抗扰动距离.如果可以计算出最小对抗扰动距离的精确值,那么最小对抗扰动距离的值将可以作为模型对抗鲁棒性评估的指标.然而,由于神经网络模型是大型、非线性且非凸的,对抗鲁棒性等模型属性的验证问题已被证明是一个 NP 完全(non-deterministic polynomial-complete, NP-C)问题[33-35].作为与对抗鲁棒性相关的指标,最小对抗扰动距离难以被精确求解.因此,许多研究转向使用最小对抗扰动的上界或下界去近似精确值[36].当扰动距离大于上边界距离时,说明至少有 1 个添加了该扰动的样本被模型误分类;当扰动距离小于下边界距离时,则任意添加了该扰动的样本都能被模型正确分类,如图 3 所示.通过最大下边界距离或最小上边界距离逼近最小对抗扰动距离,从而实现对模型对抗鲁棒性的评估.

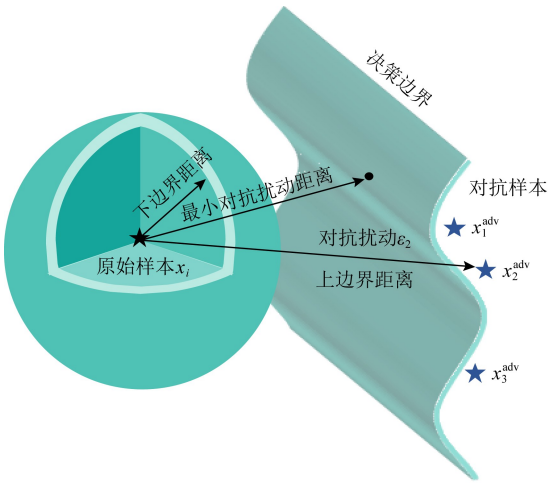


Fig. 3 Schematic diagram of upper and lower boundary of adversarial disturbance

图 3 对抗扰动上、下边界示意图

2.2 评估准则

对抗鲁棒性评估是一个比较困难的问题,执行不合理的实验会导致评估无效.比如向原始样本添加的扰动过大,人眼即可判别生成样本与原始样本,在此基础上得到的评估结果是无意义的[37].诸如此类的错误常常被研究人员忽略.本文梳理了进行对抗鲁棒性评估时需要遵守的 3 个评估准则[37],以便指导研究人员开展正确的评估.

1) 合理使用攻击算法进行评估.部分攻击算法是针对某种特定模型而设计的,若将其应用于其他模型,很难体现出模型是否具有抵御这种攻击的能力,也无法说明模型具有对抗鲁棒性.另外,在进行对抗鲁棒性评估时,需要保证评估的普适性,这就意味着不能仅仅使用带有默认超参数的对抗攻击进行评估,应该排列组合所有参数,达到不同程度的攻击效果,从而评估模型的对抗鲁棒性.

2) 保证模型在正常环境下的良好性能.实验表明,对抗训练会导致智能系统中神经网络模型的准确率下降[38].显然,以牺牲模型对原始样本的准确率来增强模型对抗鲁棒性的做法是不可取的.因此,开展对抗鲁棒性评估,应说明模型对原始样本的分类准确率.被攻击模型保持正常环境下的分类准确率是正确评估对抗鲁棒性的前提.在满足该前提的条件下,如果被攻击模型能够正确识别对抗样本,才能说明模型具有较好的对抗鲁棒性.

3) 结合实际需求选择评估结果分析方法.理论上,评价智能系统的对抗鲁棒性应分析模型受到攻击威胁时的最坏结果.在现实情况中,往往从统计学角度以平均效果衡量鲁棒性的好坏.针对不同需求,应适当调整评估思维,给出符合实际的评估结果.进行统计学分析时,还需对分类准确率等结果进行方差计算,避免出现较高的偏差,使评估结果失去稳定性.

2.3 评估指标体系

当前大多数研究通过分类准确率、攻击次数或扰动强度这 3 种指标进行对抗鲁棒性评估.这些指标能够直观上反映模型对抗鲁棒性的强弱,但未能全面地考虑到影响对抗鲁棒性的因素.针对该问题,本节梳理了现有研究中所涉及的 30 余种对抗鲁棒性评估指标,并从被攻击模型和测试数据 2 个角度对指标进行分类.

2.3.1 面向模型的指标

在图像分类全过程中,被攻击模型的训练与决策是影响最终鲁棒性评估效果的关键环节.进行对抗鲁棒性评估时,可以从训练、决策阶段关注模型的结构、行为特征,通过挖掘模型结构相关指标,深入理解模型在对抗环境下的反应,总结与模型决策相关的评价指标,帮助研究人员实现模型对抗鲁棒性的评估.因此,本节将面向模型的指标进一步划分为基于模型行为的指标和基于模型结构的指标.

1) 基于模型行为的指标

模型行为可以理解为模型对测试样本做出的反应.基于模型行为的指标主要是依据对抗环境下模型的输出结果进行对抗鲁棒性度量.

① 原始样本分类准确率(OA)

尽管我们的目的是解决对抗环境下的模型鲁棒性评估问题,但是为了保证模型的基本性能,也需要对正常环境下模型的行为表现进行测试,避免出现模型的对抗鲁棒性表现优异而在原始样本上的分类效果很差的情况,以致无法应用于现实场景中.

原始样本分类准确率(original accuracy, OA)经常被用来衡量正常环境下模型的分类性能,其含义是被分类正确的原始样本数量占总体样本数量的百分比.

$$OA(F,T)=\frac{1}{N}\sum_{i=1}^{N}count(F(x_i)=y_i), \quad (3)$$

其中,$count(\cdot)$表示括号内等式为真则为 1,否则为 0.OA 值越大,说明模型在正常环境下的分类性能越好.

② 对抗样本分类准确率(ACA)

除描述正常环境下模型分类性能的指标外,还有一些公认的、被广泛接受和使用的、用以描述对抗环境下模型分类性能的指标,如对抗样本分类准确率(adversarial classification accuracy, ACA).它指的是被正确分类的对抗样本的数量占总体样本数量的百分比,数学表达式为

$$ACA(F,T)=\frac{1}{N}\sum_{i=1}^{N}count(F(x_i^{adv})=y_i). \quad (4)$$

由式(4)可见,ACA 值越大,说明模型在对抗环境下的分类性能越好,在一定程度上可以说明模型的对抗鲁棒性越好.

③ 对抗样本攻击准确率(AA)

与对抗样本分类准确率相对应的是对抗样本攻击准确率(adversarial accuracy, AA).它指的是被错误分类的对抗样本的数量占总体样本数量的百分比,数学表达式为

$$AA(F,T)=\frac{1}{N}\sum_{i=1}^{N}count(F(x_i^{adv})\neq y_i). \quad (5)$$

由式(5)可见,AA 值越大,说明模型在对抗环境下的分类性能越差,在一定程度上可以说明模型的对抗鲁棒性越差.

针对白盒攻击与黑盒攻击 2 种不同的攻击,还可以将攻击准确率进一步划分为白盒攻击准确率(adversarial accuracy on white-box attacks, AAW)和黑盒攻击准确率(adversarial accuracy on black-box attacks, AAB)[39].

④ 白盒攻击准确率(AAW)

在非目标攻击场景中,白盒攻击准确率指的是通过白盒攻击技术生成的对抗样本 $x_i^{\text{adv}-W}$ 经过模型分类,被错分成其他类别的数量占总体样本数量的百分比,而在目标攻击场景中,则指的是被错分成第 j 类对抗样本数占总体数量的百分比.

$$AAW(F,T)=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}count(F(x_i^{\text{adv}-W})\neq y_i),\\ \quad\text{非目标攻击},\\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}count(F(x_i^{\text{adv}-W})=y_j),\\ \quad\text{目标攻击}.\end{cases}\tag{6}$$

由式(6)可见,AAW 值越大,说明模型越容易将由白盒攻击方法生成的对抗样本错分成其他类别,在一定程度上可以说明模型的对抗鲁棒性越差.

⑤ 黑盒攻击准确率(AAB)

黑盒攻击准确率(AAB)[39]与白盒攻击准确率(AAW)类似,不同之处在于对抗样本是由黑盒攻击算法生成.

$$AAB(F,T)=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}count(F(x_i^{\text{adv}-B})\neq y_i),\\ \quad\text{非目标攻击},\\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}count(F(x_i^{\text{adv}-B})=y_j),\\ \quad\text{目标攻击}.\end{cases}\tag{7}$$

由式(7)可见,AAB 值越大,说明模型越容易将由黑盒攻击方法生成的对抗样本错分成其他类别,在一定程度上可以说明模型的对抗鲁棒性越差.

⑥ 正确类别平均置信度($ACTC$)

Ling 等人[40]还提出使用正确类别平均置信度(average confidence of true class, $ACTC$)衡量模型的对抗鲁棒性.该指标具体含义为对抗样本 x_i^{adv} 被分类成正确类别时模型置信度的平均值.

$$ACTC(F,T)=\frac{1}{n}\sum_{i=1}^{n}P(F(x_i^{\text{adv}})=y_i),\tag{8}$$

其中,n 表示对抗样本攻击失败的数量,$P(\cdot)$表示模型分类的置信度.$ACTC$ 值越高,说明模型正确识别对抗样本类别的能力越强,对抗鲁棒性越强.

⑦ 对抗类别平均置信度($ACAC$)

相比于原始样本分类准确率和对抗样本攻击准确率,对抗类别平均置信度(average confidence of adversarial class, $ACAC$)[40]能够更加详细地描述模型分类器识别对抗样本的能力.对于 N 个对抗样本 $x_i^{\text{adv}}(i=1,2,\cdots,N)$,对抗样本分类置信度可以定义为模型分类器将对抗样本进行错误分类的置信度的平均值.

$$ACAC(F,T)=\begin{cases}\dfrac{1}{m}\displaystyle\sum_{i=1}^{m}P(F(x_i^{\text{adv}})\neq y_i),\\ \quad\text{非目标攻击},\\ \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}P(F(x_i^{\text{adv}})=y_j),\\ \quad\text{目标攻击}.\end{cases}\tag{9}$$

其中,m 表示对抗样本攻击成功的数量,$P(\cdot)$表示模型分类的置信度.$ACAC$ 值越高,说明模型识别对抗样本类别的能力越差,对抗鲁棒性越差.

⑧ 噪声容忍估计(NTE)

主流攻击算法以最大化模型错分为除正确类别外的某一类别的概率为目标,很少关注除正确类别和错分类别外的其他类别的情况.降低模型将样本分类成其他类别的概率是提高对抗样本稳健性的有效手段.Luo 等人[41]致力于最大化目标类的概率与所有其他类的最大概率之间的差距,提出噪声容忍估计(noise tolerance estimation, NTE)衡量对抗攻击的鲁棒性:

$$NTE(F,T)=\frac{1}{m}\sum_{i=1}^{m}\left[P_{y_j}(F(x_i))-\max(P_{y_{k\neq j}}(F(x_i)))\right],\tag{10}$$

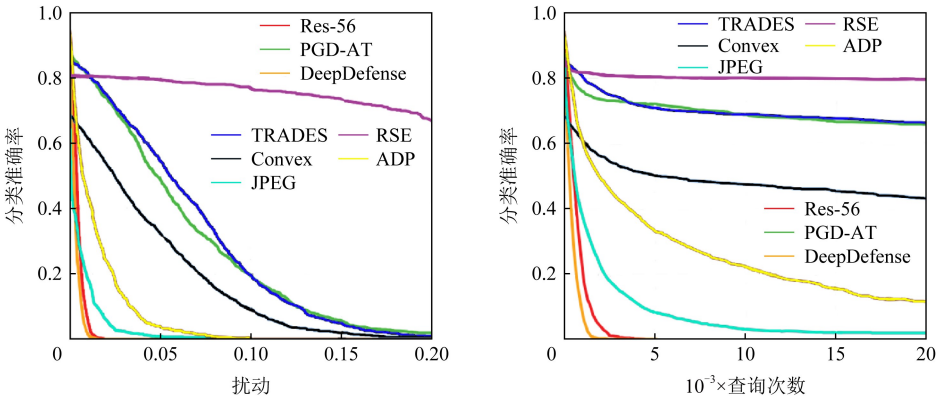
其中,$P_{y_j}(\cdot)$表示模型将样本错分成类别 j 的置信度,$P_{y_{k\neq j}}(\cdot)$表示模型将样本错分成除类别 j 外的其他类别的置信度.NTE 值越大,说明对抗样本的鲁棒性越强.将 NTE 值取平均,平均 NTE 值越小,模型容易混淆除正确类别外的多种类别,在一定程度上可以说明模型的对抗鲁棒性越差.

⑨ 互补鲁棒性曲线(CRC)

除以上指标外,许多研究通过绘制曲线图,表达了多种指标在一定范围内变化时模型分类器的不同响应,更加直观地展示了对抗鲁棒性评估的结果.如

Dong 等人[42]采用 2 条互补鲁棒性曲线(complementary robustness curves, *CRC*),展示了受到对抗攻击的深度学习模型的鲁棒性.互补鲁棒性曲线如图 4 所示.

图 4(a)是"扰动-准确率"曲线,Dong 等人[42]通过计算所有扰动下的分类准确率绘制而成,横坐标是扰动大小,纵坐标是分类准确率.他们选取基于自然进化策略(natural evolution strategy, NES)的梯度估计算法作为攻击算法,Res-56(Resnet-56)模型等作为被攻击的模型.通过图 4(a),我们可从全局视角对比 8 种模型抵御 NES 攻击的鲁棒性.如使用 NES 算法攻击 TRADES 和 Convex 模型,相同扰动下 TRADES 模型的准确率始终比 Convex 模型高,在一定程度上说明 TRADES 模型抵御 NES 攻击的鲁棒性更好.图 4(b)是"查询次数-准确率"曲线,横坐标是模型查询次数,纵坐标是分类准确率.其中,模型查询次数亦可替换成不同攻击方法的迭代次数.该曲线能够显示攻击的效率,例如尽管查询次数为 20 000 时,Res-56 与 DeepDefense 模型的分类准确率都下降到 0,但是 Res-56 模型受攻击后分类准确率下降得更慢、曲线更靠右,说明 Res-56 模型抵御 NES 攻击的能力更强.



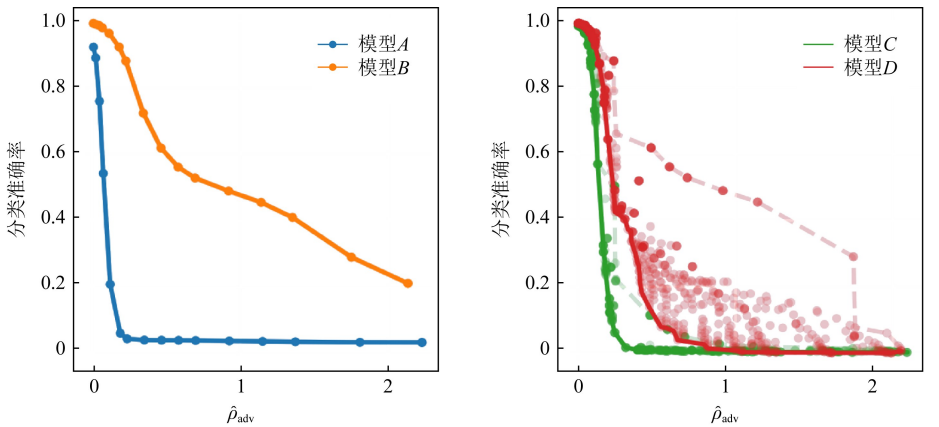
(a) 分类准确率随扰动的变化 (b) 分类准确率随查询次数的变化

Fig. 4 Complementary robustness curves[42]

图 4 互补鲁棒性曲线[42]

⑩ 精确扰动曲线(*APC*)

相比于互补鲁棒性曲线(*CRC*),Šircelj 等人[43]提出的精确扰动曲线(accuracy-perturbation curves, *APC*)优势在于展示了被攻击模型受不同扰动影响的最差性能.首先组合各种超参数,生成添加了不同大小扰动的对抗样本,并计算不同模型对它们的分类精度;其次根据式(11)计算这些对抗样本的子集(它们的原始图像已被模型正确分类)添加的扰动大小;最后将每组扰动幅度 $\hat{\rho}_{\text{adv}}$ 和分类精度绘制为"扰动-准确率"曲线,如图 5 所示:



(a) 模型A, B的分类准确率随扰动幅度的变化 (b) 模型C, D的分类准确率随扰动幅度的变化

Fig. 5 Accuracy-perturbation curves[43]

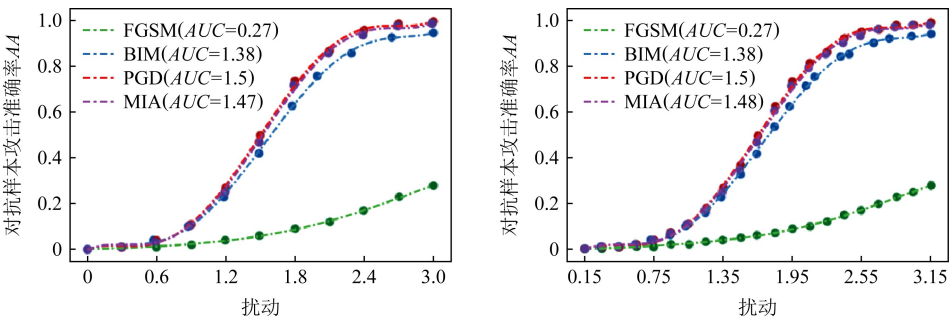
图 5 精确扰动曲线[43]

$$\hat{\rho}_{\mathrm{adv}}(F)=\frac{1}{N}\sum_{i=1}^{N}\frac{\|\varepsilon\|_2}{\|x_i\|_2}. \tag{11}$$

图 5(a)横坐标为扰动大小,纵坐标为分类准确率.显然,曲线 B 对应模型的对抗鲁棒性更强.然而,在使用多参数组合的方法时通常得到分散的"扰动-准确率"点,这为评估对抗鲁棒性的强弱带来了困难.为此,Šircelj 等人[43]进一步设计了一种最小包络法(minimum wrap),将每个扰动对应的最小准确率的点连接成线,如图 5(b)中的曲线 C 和曲线 D,这 2 条曲线体现了对抗样本在 2 种模型上分类的最坏情况估计.通过对比发现,相比于模型 C,模型 D 的对抗鲁棒性更强.

⑪ 分段采样曲线(PSC)

上述 CRC 和 APC 这 2 种评估指标通过绘制"扰动-准确率"曲线,定性给出模型的评估结果.Wu 等人[44]提出了一种分段采样曲线(piece-wise sampling curves, PSC)框架,如图 6 所示.该框架通过曲线下面积(area under curve, AUC)[45]实现了测量结果的可量化,避免了人为比较带来的主观偏差,同时尽可能多地适用于不同攻击方法.



(a) 针对MNIST的非目标攻击(r=10, d=5) (b) 针对MNIST的非目标攻击(r=20, d=5)

Fig. 6 Piece-wise sampling curves[44]

图 6 分段采样曲线[44]

PSC 框架包括 3 个步骤,分别是选择扰动范数 l_p、划分扰动范围与绘制分段采样曲线.以图 6 为例,Wu 等人首先确定扰动范围 $\varepsilon=[0,3]$,将分辨率 r 分别设为 10 和 20(将扰动范围划分成 r 段);而后对范围内 r 个分段采样扰动点上的对抗样本进行测试,获得该扰动下的攻击准确率,通过调整参数多次实验,获得 r 个扰动点上的最佳攻击准确率;最后根据选择的拟合函数阶数 $d=5$,将采样点进行拟合,得到最终的 5 阶分段采样曲线.图 6(a)使用 4 种非目标攻击算法攻击一个 4 层卷积神经网络在 MNIST 数据集上训练的模型,FGSM 算法对应的 AUC 最小,说明随着扰动的增大,FGSM 攻击准确率提升的速率最慢,模型抵御 FGSM 攻击的能力最好.图 6(b)对扰动范围进行了更细致的划分,更加准确地描述了模型在不同攻击下的性能,但同时也会花费更多的计算资源.

此外,Wu 等人[44]构建了 PSC 工具包,用户只需在特定设置下上传添加不同扰动时模型的分类准确率结果,就能生成对应的分段采样曲线.与使用单个或多个扰动点进行评估的方式相比,该方法可以量化对抗鲁棒性的强弱.

在上述①~⑪种指标中,①~⑧种指标从不同的角度描述了受到各种攻击时模型对原始样本和对抗样本的响应.⑨~⑪种指标在①~⑧种指标的基础上,通过绘制多指标曲线,站在全局视角定性或定量地评估模型的对抗鲁棒性.计算原始样本分类准确率、对抗样本攻击准确率、绘制分段采样曲线等指标,能够快速给出最直观的评估结果,获取模型抵御攻击的显性知识,是一种简单、有效的对抗鲁棒性评估方法.

2) 基于模型结构的指标

与基于模型行为的指标不同的是,这一类指标关注模型内部神经元、损失函数等与模型结构相关的信息,观察模型内部对于对抗样本的反应,进而衡量模型的对抗鲁棒性.

① 神经元敏感度(NS)

对抗鲁棒性意味着模型对对抗样本保持一定的钝感,尽管受到扰动的影响,原始样本和对抗样本在模型的隐藏层中仍具有相似的表示,从而输出正确的分类结果.因此,一些研究人员尝试通过计算原始样本与对抗样本在隐藏层中特征表示的偏差,衡量模型的对抗鲁棒性.

Zhang 等人[46]率先提出神经元敏感度(neuron sensitivity, NS),从内部神经元的角度解释模型的

敏感性.具体而言,给定一个原始样本 x_i 和对应的对抗样本 x_i^{adv},可以得到对偶对集合 $\overline{D}=\{(x_i,x_i^{\text{adv}})\}$,计算神经元敏感度($NS$):

$$NS(\boldsymbol{F}_l^m,\overline{D})=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{dim(\boldsymbol{F}_l^m(x_i))}\left\|\boldsymbol{F}_l^m(x_i)-\boldsymbol{F}_l^m(x_i^{\text{adv}})\right\|_1, \quad (12)$$

其中,$\boldsymbol{F}_l^m(x_i)$,$\boldsymbol{F}_l^m(x_i^{\text{adv}})$分别表示神经元 $\boldsymbol{F}_l^m$ 在正向过程中对原始样本 x_i 和对抗样本 x_i^{adv} 的输出.$dim(\cdot)$表示向量的维度.NS 值越小,模型的对抗鲁棒性越强.

② 神经元不确定性(NU)

在关键安全应用领域,模型预测的不确定性被广泛研究.陈思宏等人[47]研究发现预测不确定性越大,模型对抗鲁棒性越强,通过增加模型预测的不确定性,以达到提高对抗鲁棒性的目的.预测结果的方差是衡量不确定性最直观的方法.

受此启发,Liu 等人[39]使用神经元 $\boldsymbol{F}_l^m$ 的方差来计算神经元不确定性(neuron uncertainty, NU):

$$NU(\boldsymbol{F}_l^m)=\frac{1}{N}\sum_{i=1}^{N}variance(\boldsymbol{F}_l^m(x_i)), \quad (13)$$

其中,$variance(\cdot)$是方差计算函数.NU 值越大,模型的对抗鲁棒性越强.

③ 冲突上下限(ΔK)

Dempster-Shafer 证据理论常被用来处理不确定信息,辅助完成智能决策任务.苏记柱[48]对 CNN 模型进行重构,结合证据推理方法,分析信息冲突值随扰动强度的变化关系,提出了 2 个对抗鲁棒性评估指标:冲突上限 $\Delta k_{\max}$ 和冲突下限 $\Delta k_{\min}$,计算公式分别为

$$\Delta k_{\max}=\frac{1}{N}\sum_{i=1}^{N}(k_{Ni}-k_{0i}), \quad (14)$$

$$\Delta k_{\min}=\frac{1}{N}\sum_{j=1}^{N}(k_{\min j}-k_{0j}), \quad (15)$$

其中,$k_0=\{k_{01},k_{02},\cdots,k_{0N}\}$为每个样本特征之间的冲突值,$k_{\max}=\{k_{N1},k_{N2},\cdots,k_{NN}\}$表示在 FGSM 攻击下,当样本逐渐添加扰动至被模型误分类时,每个样本特征之间的冲突值,$\Delta k_{\max}$为模型可接受的样本特征之间冲突的最大变化范围.$\Delta k_{\max}$越大,表示模型的鲁棒性越强,反之,模型的鲁棒性越差.$k_{\min}=\{k_{\min 1},k_{\min 2},\cdots,k_{\min N}\}$为每个样本所产生的对抗样本中最小的冲突值.$\Delta k_{\min}$越大表明模型越鲁棒.

不难发现,冲突上限是冲突下限的特殊情况.冲突上限限制了攻击方法和扰动强度大小,冲突下限适应的场景更加宽泛.

④ 经验边界距离(EBD)

Liu 等人[49]以模型的决策边界为切入点,对输入样本周围邻域进行全面分析,提出经验边界距离(empirical boundary distance, EBD)指标.该指标为所有样本决策边界距离的平均值,计算公式如下:

$$EBD=\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{\boldsymbol{v}_{ik}}, \quad (16)$$

$$\varepsilon_{\boldsymbol{v}_{ik}}=\min RMS(\boldsymbol{v}_{ik}), \quad (17)$$

其中,$\boldsymbol{v}_{ik}$表示点 x_i 在 k 方向上添加的扰动向量,各个方向的向量之间彼此正交.$RMS(\cdot)$为均方根运算.$\varepsilon_{\boldsymbol{v}_{ik}}$表示模型对添加扰动向量后的点 x_i' 分类发生变化时向量长度的临界值,即点 x_i 在 k 方向上距离决策边界的最短距离.EBD 值越大,对抗扰动上边界距离越大,最小对抗扰动距离越大,模型的对抗鲁棒性越强.

该指标计算了模型对不同方向输入扰动的样本的分类置信度,并以此作为估计样本到决策边界距离的标准,从模型结构的角度挖掘了模型决策区域的信息,为鲁棒性评估提供了一种新的思路.

⑤ 经验边界距离(EBD-2)

在 EBD 的基础上,Liu 等人[39]又进一步提出了 EBD-2 评估指标,该指标的优势在于可以计算所有类别的决策边界的最小距离:

$$\varepsilon'_{\boldsymbol{v}_{ik}}=\min RMS(\boldsymbol{v}_{ik}). \quad (18)$$

与经验边界距离 EBD 不同的是,这里的 $\varepsilon'_{\boldsymbol{v}_{ik}}$ 是模型对添加扰动向量后的点分类成 j 类时向量长度的临界值.

⑥ CLEVER score

Weng 等人[50]发现,使用最小对抗扰动距离的上界或下界去估计模型的对抗鲁棒性时,距离下界与局部梯度的最大范数相关,故将鲁棒性评估问题转化为局部 Lipschitz 常数估计问题.

为了有效、可靠地估计局部 Lipschitz 常数,Weng 等人提出了一种称为 CLEVER 的鲁棒性度量指标,利用极值理论[51]去估计最小对抗扰动距离下边界,并给出了相应的数学证明.该指标的优势在于独立于攻击进行鲁棒性度量,适用于任何模型.CLEVER 数值越大,模型的对抗鲁棒性越强.

⑦ 二阶 CLEVER score

在 CLEVER 的基础上,Weng 等人[52]进一步提出二阶 CLEVER,将 CLEVER 扩展为能够评估具有不可微分输入转换的网络的鲁棒性,使其可评估部署了多种基于梯度掩蔽的模型.

⑧ CROWN 框架

通过寻找最小对抗扰动距离下界去估计模型的对抗鲁棒性,存在计算耗时、估计误差大等问题.为此,Weng 等人[53]利用 ReLU 网络进一步提出 Fast-Lin 和 Fast-Lip 算法,计算出较高精度的距离下界,用以评估模型的对抗鲁棒性.同时,相比于基于线性规划的方法和 Reluplex[33]方法,Fast-Lin 和 Fast-Lip 计算速度要快 2～4 个数量级.尽管 Fast-Lin 和 Fast-Lip 能够以高精度估计对抗边界,但是仍有一些模型需要 ReLU 以外的函数激活,例如 RNN 和 LSTM 等.

为解决激活函数受限的问题,Zhang 等人[54]提出了适用于具有一般激活函数的模型的鲁棒性评估框架 CROWN,该框架允许灵活地选择激活函数的上界或下界,从而减小评估对抗鲁棒性时难以避免的近似误差.实验表明,与文献[53]相比,CROWN 的认证下界提高了 26%,并且可以将 CROWN 扩展到具有各种激活函数的大型神经网络中.此外,对于具有超过 10 000 个神经元的模型,可以在大约 1 min 内利用 1 个 CPU 内核求出鲁棒性下界.

⑨ 经验噪声不敏感度(ENI)

Xu 等人[55]率先提出算法鲁棒性的概念,这一概念的提出受益于 2 个相似样本的测试误差往往非常接近的思想.受此启发,Liu 等人[49]从 Lipschitz 常数的角度,提出了 ε 经验噪声不敏感度(ε-empirical noise insensitivity, ENI).具体而言,将原始样本和对抗样本输入模型,计算模型损失函数之间的差异,以损失值大小来衡量模型对约束 ε 的广义噪声的不敏感性和稳定性:

$$ENI_f(\varepsilon)=\frac{1}{N\times M'}\sum_{i=1}^{N}\sum_{j=1}^{M'}\frac{|l_f(x_i\mid y_i)-l_f(\mu_{ij}\mid y_j)|}{\|x_i-x_i^{\mathrm{adv}}\|_\infty},\quad \text{s.t. } \|x_i-x_i^{\mathrm{adv}}\|_\infty\leqslant\varepsilon, \tag{19}$$

其中,M'是使用各种攻击算法依次生成对抗样本的数量,$l_f(\cdot\mid\cdot)$表示模型 f 的损失函数.ENI 能够衡量模型对包括对抗扰动在内的广义噪声的鲁棒性.ENI 值越小,说明模型的对抗鲁棒性越强.

与基于模型行为的指标相比,以上 9 种指标的优势在于能够刻画图像分类过程中模型的内部特性,挖掘更多关于模型结构的隐性知识,从模型本身而非直接的模型结果的角度评估对抗鲁棒性.

2.3.2 面向数据的指标

在进行对抗鲁棒性评估时,应从图像分类全过程出发,分析各阶段影响评估结果的因素,进而实现全面、科学且准确的评估.然而,由于习惯上过度关注于智能系统的高精度性能结果,研究者常常忽略了测试过程是否充分以及测试数据质量的好坏带来的影响.因此,在进行对抗鲁棒性评估时,同样需要关注模型测试的充分性以及测试数据的质量.考虑到对抗样本的特点,即测试数据的质量与图像的不可感知性相关,本文将面向数据的指标进一步划分为基于测试充分性的指标和基于视觉不可感知性的指标.

1) 基于测试充分性的指标

测试充分性[56]这一概念最早是在软件工程领域提出的,它是指使用一组有代表性的数据对软件进行全面测试的程度.Pei 等人[57]将其类比到模型测试中,提出了一种神经元覆盖率(neuron coverage, $NCov$)指标来量化神经元被激活的比例.

$$NCov(T',x)=\frac{|\{n_e\mid\forall x\in T',out(n_e,x)>t\}|}{|N_{总}|}, \tag{20}$$

其中,T'为测试集,x 为测试样本,$N_{总}$ 为神经元总数,$out(n_e,x)$指的是给定一测试样本返回模型中神经元 n_e 的输出值,t 为设定的阈值,当输出值大于阈值时神经元被激活.然而,Ma 等人[58]经实验发现,$NCov$ 无法体现测试集中原始样本和对抗样本之间的差异.为此,他们提出了多层次多粒度覆盖的一系列测试标准 DeepGauge,使用 k 节神经元覆盖率(k-multisection neuron coverage, $kMNCov$)、神经元边界覆盖率(neuron boundary coverage, $NBCov$)、强神经元激活覆盖率(strong neuron activation coverage, $SNACov$)这 3 种指标来衡量对抗环境下模型测试的充分程度.

① k 节神经元覆盖率($kMNCov$)

为体现原始样本和对抗样本的差异,Ma 等人将进行测试时神经元输出范围划分为主要功能区和极端案例区.当测试样本与原始样本分布相近时,神经元输出值落入主要功能区$[low_n,high_n]$,否则落入极端案例区$(-\infty,low_n)\cup(high_n,-\infty)$,也就是对应对抗样本所在的区域.将区域$[low_n,high_n]$划分成 k 个长度相等的节段,$kMNCov$ 为测试集 T 覆盖的节数与总节数之比,用以衡量覆盖主要功能区程度,数学公式描述为

$$kMNCov(T,k)=\frac{1}{k\times|N_k|}\sum_{n\in N}|\{S_i^n\mid \exists x\in T:\varnothing(x,n_e)\in S_i^n\}|, \tag{21}$$

其中,T 为测试集,x 为测试样本,N_k 为各节神经元数,S_i^n 表示第 i 部分的覆盖率.$kMNCov$ 值越大,说明主要功能区被覆盖的部分越多,原始样本在模型中的测试效果越充分.

② 神经元边界覆盖率($NBCov$)

神经元边界覆盖率 $NBCov$ 是测量给定测试集 T' 覆盖极端案例区面积大小的指标,定义为极端案例区被覆盖的神经元数量与神经元总数的比率:

$$NBCov=\frac{|UpperCornerNeuron|+|LowerCornerNeuron|}{2\times|N'|},\tag{22}$$

其中,$UpperCornerNeuron$ 是极端案例区上边界神经元被覆盖的数量,$LowerCornerNeuron$ 是极端案例区下边界神经元被覆盖的数量,N' 表示上、下边界神经元数.$NBCov$ 值越大,说明极端案例区被覆盖的部分越多,对抗样本在模型中的测试效果越充分.

③ 强神经元激活覆盖率($SNACov$)

在研究 DNN 可解释性问题时,过度活跃的神经元可能有助于 DNN 模型的学习.因此,Ma 等人[58]又提出了侧重于活跃神经元的测试指标,即描述极端案例区上边界情况的指标 $SNACov$.它被用来测量测试集 T' 在极端案例区中上边界被覆盖的情况,描述为覆盖的极端案例上边界神经元数与总神经元数之比:

$$SNACov(T')=\frac{|UpperCornerNeuron|}{|N_{总}|}.\tag{23}$$

值得注意的是,提高测试充分性并不能有效改善模型的对抗鲁棒性[59-60],但是测试越充分说明评估结果的可信度越高.因此,在进行对抗鲁棒性评估时,可以考虑通过①~③指标来检验结果的可信度或准确性.

2) 基于视觉不可感知性的指标

相比于原始样本,对抗样本最鲜明的特点是添加了肉眼难以感知的扰动.如果添加的扰动太大导致样本过于模糊不清,肉眼即可判别出样本为对抗样本,便失去了评估对抗鲁棒性的意义.本节总结了多种指标来刻画扰动的不可感知性.

① 平均 l_p 失真度(ALD_p)

目前大多数研究采用 l_p 范数($p=0,1,\cdots,\infty$)来衡量原始样本与对抗样本之间的扰动距离,Liu 等人[39]提出将 ALD_p 作为扰动不可感知性的度量指标.从数学含义上讲,该指标是 l_p 失真度的平均归一化:

$$ALD_p=\frac{1}{N}\sum_{i=1}^{N}\frac{\|x_i^{adv}-x_i\|_p}{\|x_i\|_p},\tag{24}$$

其中,N 表示样本的数量.ALD_p 越小,失真越小,对抗扰动越小,肉眼区分原始样本和对抗样本的难度越大.

② 平均结构相似度(ASS)

在实践中,通常需要对整个图像进行单一的整体质量测量,Wang 等人[61]提出的结构相似度(structural similarity, $SSIM$)就是评估图像整体质量的指标.针对对抗环境下的评估问题,平均结构相似度(average structural similarity, ASS)[39]是在 $SSIM$ 的基础上进行改进,被用来描述原始样本与对抗样本之间的结构相似度.

$$ASS(x_i,x_i^{adv})=\frac{1}{M}\sum_{j=1}^{M}SSIM(x_{ij},x_{ij}^{adv}),\tag{25}$$

其中,x_{ij} 和 x_{ij}^{adv} 是第 i 个样本第 j 个局部窗口的图像内容,M 是图像中局部窗口的数量.ASS 越高,原始样本和对抗样本越相似.

③ 扰动敏感度(PSD)

基于对比掩蔽理论[62-64],Liu 等人[39]提出了扰动敏感度(perturbation sensitivity distance, PSD)指标来评估人类对扰动的感知.因此,PSD 定义为

$$PSD=\sum_{i=1}^{N}\sum_{j=1}^{M}\delta_{ij}*Sen(x_{ij}),\tag{26}$$

$$Sen(x_{ij})=1/SD(x_{ij}),\tag{27}$$

$$SD(x_{ij})=\sqrt{\frac{1}{n^2}\sum_{x_{ij}\in S_i}(x_{ij}-\mu)^2},\tag{28}$$

其中,M 是像素总数,δ_{ij} 是第 i 个样本的第 j 个像素,x_{ij} 表示周围正方形区域,S_i 是由 $n\times n$ 区域中的像素组成的集合,μ 是区域内像素的平均值.当像素具有低方差时,扰动灵敏度高.PSD 越大,添加于对抗样本中的扰动越容易被察觉.

还有学者从特征数据的角度挖掘样本特征差异与鲁棒性的相关性.董一帆[65]认为模型对原始样本和对抗样本进行特征提取后,两者的特征子集差距越小,说明模型的对抗鲁棒性越强.因此,引入特征子集一致性[66]指标来衡量模型的对抗鲁棒性,计算公式为

$$C(A,B)=\frac{rd-k^2}{k(w-k)},\tag{29}$$

其中,A 和 B 分别为 2 个子集,X 为特征总空间,w 为特征总数,令 $|A|=|B|=k$,$0<k<|X|=w$,$r=|A\cap B|$.当 $C>0$ 时,表示 2 个子集是相似的;当 $C=0$ 时,表示 2 个子集是不相关的;当 $C<0$ 时,表示 2 个子集是不相似的.

除以上指标外,还有许多与常规图像处理相关的评估指标,如高斯模糊鲁棒性(robustness to Gaussian blur, RGB)[40]、图像压缩鲁棒性(robustness

to image compression, RIC)[40],以及与腐蚀扰动相关的评估指标,如自然噪声平均差值(mCE)、自然噪声相对差值($RmCE$)和连续噪声分类差别(mFR)[49,67]等.从某种意义上来说,以上这些指标的确可以衡量模型的对抗鲁棒性.例如,Corrupt 算法[39]中添加的扰动是自然噪声,可以根据($RmCE$)的大小来衡量对抗扰动的距离.但是对于大多数对抗攻击算法,这些指标更多地是衡量其他特定环境下的鲁棒性,并非严格的对抗鲁棒性.因此,在进行指标评估时,应针对特定的需求选取合适的指标.

本文按照指标类别、攻击方式和扰动类别等维度,对以上 30 余种评估指标进行详细的描述,如表 2 所示.考虑到实际应用过程中往往仅选取部分指标进行评估,根据指标的含义,给出了每一种指标

Table 2 Evaluation Indexes of Adversarial Robustness

表 2 对抗鲁棒性评估指标

大类	子类	序号	评估指标	白盒攻击	黑盒攻击	对抗扰动	其他扰动	相关度
基本指标		1	扰动距离[9]	√	√	√	√	★★★★
		2	攻击强度[9]	√	√	√		★★★
		3	查询次数[9]		√	√		★★★★
模型	行为	4	原始样本准确率(OA)	√	√			★
		5	对抗样本分类准确率(ACA)	√	√	√		★★★
		6	对抗样本攻击准确率(AA)	√	√	√		★★★
		7	白盒攻击准确率(AAW)[39]	√		√		★★★
		8	黑盒攻击准确率(AAB)[39]	√	√	√		★★★
		9	正确类别平均置信度($ACTC$)[40]	√	√	√		★★★★
		10	对抗类别平均置信度($ACAC$)[40]	√		√		★★★★
		11	噪声容忍估计(NTE)[41]	√		√		★★
		12	分段采样曲线(PSC)[44]	√	√	√		★★★★
		13	互补鲁棒性曲线(CRC)[42]	√	√	√		★★★★
		14	精确扰动曲线(APC)[43]	√	√	√		★★★★
	结构	15	神经元敏感度(NS)[46]	√		√		★★★★
		16	神经元不确定性(NU)[39]	√		√	√	★★★★
		17	冲突上下限(ΔK)[48]	√		√		★★★★
		18	经验边界距离(EBD)[49]	√		√	√	★★★
		19	经验边界距离(EBD-2)[39]	√		√		★★★★
		20	CLEVER score[50]	√		√		★★★
		21	二阶-CLEVER score[50]	√		√		★★★★
		22	CROWN 框架[54]	√		√	√	★★★★
		23	经验噪声不敏感(ENI)[49]	√	√	√	√	★★★★
数据	测试充分性	24	k 节神经元覆盖($kMNCov$)[58]	√		√	√	★
		25	神经元边界覆盖率($NBCov$)[58]	√		√	√	★
		26	强神经元激活覆盖率($SNACov$)[58]	√		√	√	★
	视觉不可感知性	27	平均 l_p 失真度(ALD_p)[39]	√	√	√		★
		28	平均结构相似度(ASS)[61]	√	√	√		★
		29	扰动敏感度(PSD)[39]	√	√	√		★
	其他	30	特征子集一致性[65]	√	√	√		★★★
		31	高斯模糊鲁棒(RGB)[40]	√	√		√	★★
		32	图像压缩鲁棒性(RIC)[40]	√	√		√	★★
		33	自然噪声平均差值(mCE)[49,67]	√	√		√	★★
		34	自然噪声相对差值($RmCE$)[49,67]	√	√		√	★★
		35	连续噪声分类差别(mFR)[49,67]	√	√		√	★★

注:“√”表示被评估模型所遭受攻击的攻击方式与扰动类别.其他扰动指的是除对抗扰动以外的干扰.“★”的数量表示指标与对抗鲁棒性的相关程度.“★”表示指标与对抗鲁棒性不相关;“★★”表示指标所评估的并非严格意义上的对抗鲁棒性;“★★★”表示指标与对抗特性较为相关;“★★★★”表示指标能够有效反映对抗鲁棒性的强弱.

和对抗特性的相关度,并将相关度划分为 4 个等级,使用不同数量的“★”进行表示.“★”的数量越多,说明指标与对抗特性越相关.其中,“★”表示虽然指标无法评估对抗鲁棒性,但在进行评估时需要通过这些指标进行测试,以保证评估结果的有效性;“★★”表示指标和对抗样本相关,常被用于评估对抗样本或衡量对抗扰动以外的其他扰动(自然噪声、堕化扰动等),能够反映模型的鲁棒性,而并非严格意义上的对抗鲁棒性;“★★★”表示指标与对抗特性较为相关,但由于评估视角单一、片面,往往需要结合多种指标综合给出评估结果;“★★★★”表示能够从全局角度或基于模型内部机理评估模型的对抗鲁棒性,指标的大小在一定程度上反映了对抗鲁棒性的强弱.

显然,3 种基本指标以及面向模型的指标与对抗鲁棒性评估的相关程度较高,而面向数据的指标的相关程度普遍较低,这与对抗鲁棒性评估的本质有关.深度学习模型的黑盒特性与不可解释性[68]为对抗鲁棒性的评估带来了困难,挖掘更多模型内在的信息有助于了解对抗环境下模型面对不确定干扰因素的反应.因此,进行对抗鲁棒性评估可以采用 OA,$ACTC$,PSC,EBD 等面向模型的指标,同时也要结合实际需求,考虑是否将面向数据的指标纳入评估范围内.

3 对抗攻防工具与数据集

近年来,对抗攻防研究发展迅速,相应算法层出不穷.自 2016 年始,许多研究单位推出了集成众多主流算法的对抗攻防工具,以提高研究者与开发人员的测评效率,助力推动智能系统安全领域的发展.此外,在使用对抗攻防工具进行实验的过程中,各学者还应用了多种不同的数据集.本节将介绍主流的数据集与对抗攻防集成平台,方便后续开展对抗鲁棒性评估研究.

3.1 常用数据集

目前,针对不同领域、不同应用场景的图像数据集层出不穷,本文选取图像分类领域较为经典的、被广泛使用的 6 种数据集进行介绍,具体信息如表 3 所示:

Table 3 Common Image Classification Data Sets

表 3 常用的图像分类数据集

数据集	图像数量	类别数	各类图像数量	图像尺寸	评估指标
MNIST[69]	7 000	10	7 000	28×28	9-14,17,18,20-22,24-26,31,32
ImageNet[70]	14 197 122	21 841	1 000	约 500×400	12,15,18,23-26,33-35
Caltech101[71]	9 145	101	>40	约 300×200	
Caltech256[72]	30 607	256	>80	约 300×200	
CIFAR10[73]	60 000	10	6 000	32×32	4,7-22,27-29,31,32
CIFAR100[73]	60 000	100	600	32×32	

注:评估指标一列中的数字代表使用该数据集进行对抗鲁棒性评估的指标序号(见附录表 A1).

1) MNIST

MNIST 数据集[69]是图像分类领域常用的数据集之一.该数据集由美国国家标准与技术研究所(National Institute of Standards and Technology)组织整理,收集了来自 250 个人的 0~9 手写数字图片.该数据集总计 70 000 组图片数据,具体包括训练集图片 60 000 张及对应标签 60 000 个、测试集图片 10 000 张及对应标签 10 000 个,每张图片像素大小为 28×28.共有 16 种指标在该数据集上开展对抗鲁棒性评估.

2) ImageNet

ImageNet 数据集[70]于 2007 年开始收集,直到 2009 年以论文形式发布,后续被 Kaggle 公司继续维护,是世界上图像分类、识别、定位领域最大的数据库.截至目前,ImageNet 数据集总共有 14 197 122 张图像,涵盖 21 841 个类别.使用率最高的子集是 ImageNet 大规模视觉识别挑战赛(ILSVRC)2012—2017 图像分类和定位数据集.该数据集跨越 1 000 个对象类,包含 1 281 167 个训练图像、50 000 个验证图像和 100 000 个测试图像,每张图片像素大小约为 500×400.共有 10 种指标在该数据集上开展对抗鲁棒性评估.

3) Caltech101/256

Caltech101 数据集[71]是加利福尼亚理工学院收集整理的数据集,由 9 146 张图像组成.每张图像都标有 1 个对象,包含 101 个类别对象以及 1 个

额外的背景杂波类别.每个类约40～800张图像,大多数类别有大约50张图像,每张图片像素大小约为300×200.Caltech256数据集[72]在2006年被发布,包括256类目标图像和1类背景图像,共257类.与Caltech101相比主要变化表现在:图像总数达到30 608张,且每类最少含有80幅图像,最多含有827幅图像,目前暂无相关研究基于此数据集开展对抗鲁棒性评估.

4) CIFAR10/100

CIFAR10数据集[73]是由Hinton的学生Alex Krizhevsky和Ilya Sutskever整理的一个用于识别普适物体的小型数据集.该数据集由10个类别的60 000个像素为32×32彩色图像组成,每个类别包含6 000个图像,其中有50 000个训练样本和10 000个测试样本.CIFAR100数据集与CIFAR10数据集类似,不同的是CIFAR100数据集具有100个类别,每类包含600张图像,其中训练样本500个,测试样本100个.CIFAR100数据集[73]中的100个类分为20个超类,每张图像都带有一个"精细"标签(它所属的类)和一个"粗略"标签(它所属的超类),共有22种指标在CIFAR10上开展评估.

综上所述,在面向图像分类的对抗鲁棒性评估研究中,使用最多的数据集有CIFAR10,MNIST和ImageNet.为便于进行评估结果的对比,研究人员可使用以上3种数据集开展对抗鲁棒性评估实验.

3.2 对抗攻防集成工具

对抗攻防集成工具实现了将攻防算法模块化,为攻防对抗实验提供了可组合、可操作、可更新的框架.为方便研究者快速了解各种工具并使用,现总结国内外主流的对抗攻防工具内嵌的攻防算法以及适用框架等信息,如表4所示.考虑到篇幅长度,我们将每一种工具的具体细节整理到附录表A2～A10.

Table 4 Mainstream Adversarial Attack and Defense Integration Tools

表4 主流对抗攻防集成工具

工具	推行年份	开发单位	支持框架													数据集					攻击算法	防御算法	评估指标	提供模型	应用领域
			PyTorch	TensorFlow1	TensorFlow2	JAX	Theano	Lasagne	Keras	MXNet	Scikit-learn	XGBoost	LightGBM	CatBoost	GPy	PaddlePaddle	Caffe2	MNIST	CIFAR10	CIFAR100	ImageNet	其他					
CleverHans[74-75]	2016	宾夕法尼亚州立大学	√		√	√												√	√				16	1			图像分类
Foolbox[76]	2017	图宾根大学	√		√	√																	48		(2)(3)(27)		图像分类
ART[77]	2018	IBM爱尔兰研究院	√	√	√				√	√	√	√	√	√	√			√				√	38	>30	(31)～(36)		图像分类 目标检测 目标跟踪 语音识别
DEEPSEC[40]	2019	浙江大学	√															√	√				16	13	(2)～(15)	(4)(10)～(12)	图像分类
AdverTorch[78]	2019	Borealis AI	√																				21	7			图像分类
Ares[42]	2020	清华大学	√																√		√		19	10	(1)(3)	(4)(13)～(27)	图像分类
AdvBox[79]	2020	百度	√	√	√				√	√						√	√	√			√		10	6		(4)(7)～(9)	图像分类 目标检测
DeepRobust[80]	2020	密歇根州立大学	√															√	√			√	10 9	8 6		(1)～(5)	图像领域 图领域
AISafety[39]	2021	北京航空航天大学	√																√		√		20	5	(1)(3)～(30)	(4)～(6)	图像分类
RobustBench[81]	2021	图宾根大学																	√	√	√					(120+)	图像领域 图领域

注:"√"代表每种工具所支持的框架或数据集;没有括号的数字代表每种工具内嵌攻防算法的数量;括号里面的数字代表使用该工具进行对抗鲁棒性评估的指标或模型的序号(见附录表A2～A10,这些表于2022年3月统计,大部分工具仍在不断补充新的攻防算法,实际支持框架、内嵌算法数量等以更新情况为准).

1) CleverHans①

CleverHans[74-75]是最早被提出的机器学习模型攻防库,它集成了16种攻击算法和1种防御算法.基于该库,研究人员可以快速研发更强的对抗攻击与防御算法,实现模型的对抗训练以及鲁棒性基准测试.在最初版本v0.1中CleverHans仅支持Tensor-Flow1[82],部分模型可支持Keras[83],从v4.0.0开始,CleverHans支持的框架分别有JAX,TensorFlow2和PyTorch[84].

2) Foolbox②

Foolbox[76]是一种Python工具箱,用于生成对抗扰动并量化和比较机器学习模型的鲁棒性.Foolbox v0.8.0提供了48种对抗攻击算法,所有攻击算法都能够通过调整内部超参数生成最小的对抗扰动.此外,针对大多数攻击算法适用于特定的深度学习框架,Foolbox v0.8.0允许在诸多机器学习框架上运行,如PyTorch,Keras,TensorFlow,Theano[85],Lasagne和MXNet[86],引入原始样本误分类准确率、正确类别分类置信度、对抗类别分类置信度等对抗鲁棒性评估指标,同时支持自定义评估指标.目前,Foolbox更新到3.0版本,它现在建立在EagerPy之上,并对PyTorch,TensorFlow,JAX框架提供支持.

3) ART③

对抗性鲁棒性工具集(adversarial robustness toolbox, ART)[77]是用于机器学习安全性的Python库.ART提供的工具使开发人员和研究人员能够防御和评估机器学习模型和应用程序,以抵御逃避、中毒、提取和推理等对抗性威胁.ART支持许多流行的机器学习框架(TensorFlow, Keras, PyTorch, Scikit-learn[87], MXNet, XGBoost[88], LightGBM[89], CatBoost[90],GPy等9种框架)、多种数据类型(图像、表格、音频、视频等)和机器学习任务(分类物体检测、语音识别、生成模型、认证等).截止目前,ART集成了近40种攻击算法、30多种防御算法以及包括CLEVER score在内的6种对抗鲁棒性评估指标.

4) DEEPSEC④

DEEPSEC[40]是同时具有对攻击和防御算法进行比较研究、验证各种攻击和防御有效性以及评估机器学习模型鲁棒性等功能的平台.它全面、系统地集成了各种对抗攻击、防御算法和相关评估指标,其中攻击算法有16种,防御算法有13种以及评估指标有15种.用户可在平台上使用MNIST和CIFAR10数据集训练DenseNet[91],AlexNet[86],ResNet56[92]等模型,但是该平台仅支持PyTorch.

5) AdverTorch⑤

AdverTorch[78]是用于对抗性鲁棒性研究的Python工具箱,包含21种攻击算法和7种防御算法模块,以及用于对抗性训练的脚本.AdverTorch构建在PyTorch上,可利用动态计算图的优势实现简洁有效的攻防测试.

6) Ares⑥

Ares[42]是一个专注于对对抗鲁棒性进行基准测试的Python库.基于Ares算法平台,RealAI等后继推出基于深度学习模型的对抗攻防基准平台adversarial robustness benchmark,此基准平台可以更加公平、全面地衡量不同攻防算法的效果,提供简单高效的鲁棒性测试工具.同样,该平台对主流的攻防算法实现了模块化的设计,支持数十种主流攻防算法的实现.

7) AdvBox⑦

AdvBox[79]是由百度开源的一系列AI模型安全工具集,可以在PaddlePaddle[93],PyTorch,Caffe2[94],MXNet,Keras,TensorFlow中生成欺骗神经网络的对抗样本,也可以对机器学习模型的鲁棒性进行基准测试.与之前的工作相比,该平台不仅支持对抗样本的生成、检测和保护,还适用于许多攻击场景,例如人脸识别攻击、真假人脸检测和"隐形T恤".

8) DeepRobust⑧

DeepRobust[80]与以上平台工具不同,它是适用于图像领域和图领域的对抗性学习库.目前DeepRobust包含图像领域的10种攻击算法和8种防御算法以及图域的9种攻击算法和6种防御算法,具体实验在Pytorch上实现.

9) AISafety⑨

AISafety[39]是一个用于对抗攻击全流程评测算法学习研究的Python库,其主要研究内容为集成对抗攻击和噪声攻击相关的攻击算法、评测算法

① https://github.com/cleverhans-lab/cleverhans

② https://github.com/bethgelab/foolbox

③ https://github.com/Trusted-AI/adversarial-robustness-toolbox

④ https://github.com/ryderling/DEEPSEC

⑤ https://github.com/BorealisAI/advertorch

⑥ https://github.com/thu-ml/ares

⑦ https://github.com/advboxes/AdvBox

⑧ https://github.com/DSE-MSU/DeepRobust

⑨ https://git.openi.org.cn/OpenI/AISafety

和加固防御算法.该平台可灵活测试数据集质量、算法训练、评估和部署等算法各项指标.目前 AISafety 已集成的数据集有 CIFAR10,ImageNet 数据集,针对 CIFAR10 数据集集成了 ResNet20[95],FP_ResNet[96],VGG16[97] 测试模型,针对 ImageNet 数据集有 VGG19 模型.此外,用户可按照模型扩展要求,实现并上传自定义模型.

10) RobustBench①

RobustBench[81] 是一个图像分类领域评估攻防算法鲁棒性的基准平台,由图宾根大学的团队发布.除了标准化的测试基准之外,RobustBench 还提供了最庞大模型的存储库,其中包含 120 多个模型.

通过统计谷歌学术中以上 10 种工具所对应文献的引用次数及占比发现②,目前最常用的 3 种对抗攻防集成工具是 Foolbox,CleverHans 和 ART,如图 7 所示.研究人员可参考上述结论,结合实际需求,选择合适的对抗鲁棒性评估工具.

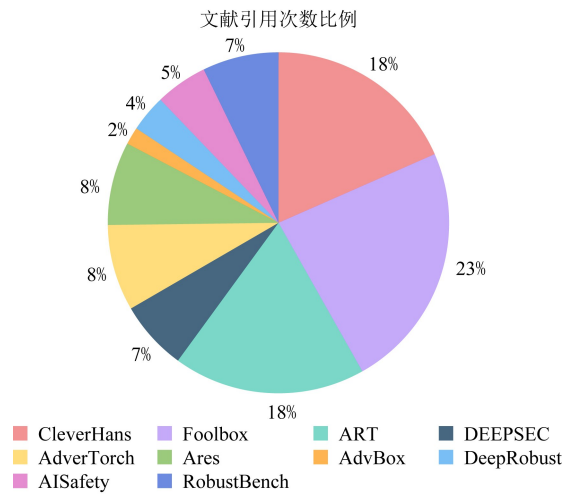


Fig. 7 Proportion of citation times of 10 tools

图 7 10 种工具对应文献的引用次数比例

4 未来研究方向

对抗攻击技术近年来获得了业界内的广泛关注,取得了许多突破性进展,但关于对抗鲁棒性评估的研究仍处于起步阶段,依然面临许多的挑战.基于本文对对抗鲁棒性评估研究现状的深入分析,未来该领域的研究需要重点关注 4 个方向:

1) 深入开展对抗样本存在机理、深度学习模型脆弱性和可解释性等理论研究.目前国内外学术界关于对抗样本的存在原因尚未达成共识,缺乏完备的数学理论支撑,对于如何正确解释深度学习模型的内在逻辑与扰动下的决策行为尚在探索当中.这些难题与对抗鲁棒性的评估有着紧密关联,开展相关理论研究有助于理解对抗鲁棒性评估问题的本质,把握影响模型对抗鲁棒性的关键因素,能从根本上解决对抗环境下模型鲁棒性的评估问题,是未来对抗鲁棒性评估问题在理论层面上值得研究的方向之一.

2) 提出一种或一组通用的、可量化的、综合的对抗鲁棒性评估指标.无论是面向数据的指标还是面向模型的指标,分析视角单一,给出的评估结果很难被直接采纳.此外,影响模型对抗鲁棒性的因素众多,采用一种或几种指标难以准确、完整地评估对抗鲁棒性的强弱.参考其他研究领域指标评估的方法,梳理影响对抗鲁棒性的全部因素,抓住关键要素,提出一种或一组通用的、可量化的、综合的指标,全面评估模型的对抗鲁棒性,是未来对抗鲁棒性评估问题在方法层面上值得研究的方向之一.

3) 构建科学、统一、规范、完备的对抗鲁棒性评估框架.面向图像分类的数据集种类繁多,攻击方法不断被创新,评估指标与评估方法不尽相同,尽管对抗攻防集成工具涵盖多种攻防算法,但也无法保证进行对抗鲁棒性评估的实验条件和度量标准是一致的,这为模型与模型之间、模型防御前后对抗鲁棒性的比较带来了困难.搭建对抗鲁棒性评估框架,全面综合各种攻防算法、数据集与评估指标,在标准对抗环境下从多层次、细粒度分析图像分类全过程模型抵御对抗攻击的能力,是未来对抗鲁棒性评估问题在流程层面上值得研究的方向之一.

4) 重点研究黑盒、非目标的融合攻击环境下的对抗鲁棒性评估方法.物理场景中难以获取模型的全部信息,针对白盒、目标攻击的评估方法难以应用于实际智能系统模型的对抗鲁棒性评估任务,且由于目前黑盒、非目标攻击的性能远低于人们的预期,无法保证使用该攻击进行评估的效果.更重要的是,现实环境中攻击者可能融合对抗扰动、自然噪声等多种类型干扰或多种攻击方法开展对抗攻击,亦或利用智能系统在动态环境下依据时间、空间等信息进行决策的漏洞,设计融合多元信息干扰的对抗攻

① https://github.com/RobustBench/robustbench

② 本文将对抗攻防工具所对应文献的引用次数作为该领域研究人员或学习者使用这 10 种对抗攻防工具开展评估的次数,进而分析出最受业界欢迎与认可的 3 种对抗攻防工具.

击方法,这给对抗鲁棒性评估带来了新的契机与挑战.如何评估模型在黑盒、非目标的融合攻击环境下的对抗鲁棒性,是未来对抗鲁棒性评估问题在实际应用层面上值得研究的方向之一.

5 总 结

面对对抗攻击等各种威胁,增强模型的对抗鲁棒性是保障智能系统安全的重要方式和手段.评估对抗鲁棒性是指导提升模型对抗鲁棒性的基础.然而,关于对抗鲁棒性评估的研究还停留在初级阶段,仅仅依靠排名基准或简单指标无法准确衡量模型抵御对抗攻击的能力.因此,本文在调研和分析国内外对抗鲁棒性评估研究的基础上,针对图像分类这一基础视觉任务,从对抗样本存在原因、对抗鲁棒性评估准则、对抗鲁棒性评估指标等方面对现有研究成果进行了归类、总结和分析.同时,梳理了现阶段主流的图像分类数据集和对抗攻防集成工具.最后,指出了对抗鲁棒性评估未来可能的研究方向,旨在为该领域研究的进一步发展和应用提供一定借鉴与帮助.

作者贡献声明:李自拓负责文献调研、内容设计、论文撰写和最后版本修订;孙建彬负责提出指导意见、框架设计和全文修订;杨克巍负责论文审核与修订;熊德辉负责提出指导意见以及论文修订.

参 考 文 献

[1] RealAI. AI face changing, stealth attack, face unlocking... How terrible is AI security? [EB/OL]. (2019-07-26) [2022-05-24]. https://mp.weixin.qq.com/s/oNmPilXb7EaVmz-3aNcU7Q (in Chinese)
(RealAI. AI换脸、隐身攻击、破解人脸解锁…AI安全到底有多可怕?[EB/OL]. (2019-07-26) [2022-05-24]. https://mp.weixin.qq.com/s/oNmPilXb7EaVmz-3aNcU7Q)

[2] Xu Kaidi, Zhang Gaoyuan, Liu Sijia, et al. Adversarial T-shirt! evading person detectors in a physical world [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2020: 665-681

[3] Jing Pengfei, Tang Qiyi, Du Yuefeng, et al. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations [C] //Proc of the 29th USENIX Security Symp. Berkeley, CA: USENIX Association, 2021: 3237-3254

[4] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint, arXiv: 1312.6199, 2013

[5] Deb D, Zhang Jianbang, Jain A K. Advfaces: Adversarial face synthesis [C] //Proc of the 2020 IEEE Int Joint Conf on Biometrics (IJCB). Piscataway, NJ: IEEE, 2020: 1-10

[6] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv: 14126572, 2014

[7] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2574-2582

[8] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1765-1773

[9] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 2017 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 39-57

[10] Su Jiawei, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841

[11] Papernot N, McDaniel P. Extending defensive distillation [J]. arXiv preprint, arXiv: 1705.05264, 2017

[12] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] //Proc of the 2016 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2016: 582-597

[13] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [J]. arXiv preprint, arXiv: 170507204, 2017

[14] Gu Shixiang, Rigazio L. Towards deep neural network architectures robust to adversarial examples [J]. arXiv preprint, arXiv: 14125068, 2014

[15] Das N, Shanbhogue M, Chen Shang-Tse, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression [J]. arXiv preprint, arXiv: 170502900, 2017

[16] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [J]. arXiv preprint, arXiv: 170401155, 2017

[17] Chen Pinyu, Sharma Y, Zhang Huan, et al. Ead: Elastic-Net attacks to deep neural networks via adversarial examples [J]. arXiv preprint, arXiv: 170904114, 2017

[18] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [J]. arXiv preprint, arXiv: 170606083, 2017

[19] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [C] //Proc of Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2013: 387-402

[20] Yuan Xiaoyong, He Pan, Zhu Qile, et al. Adversarial examples: Attacks and defenses for deep learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824

[21] Qayyum A, Usama M, Qadir J, et al. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward [J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 998-1026

[22] Li Minghui, Jiang Peipei, Wang Qian, et al. Adversarial attacks and defenses for deep learning models [J]. Journal of Computer Research and Development, 2021, 58(5): 909-926 (in Chinese)
(李明慧, 江沛佩, 王骞, 等. 针对深度学习模型的对抗性攻击与防御[J].计算机研究与发展, 2021, 58(5): 909-926)

[23] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv: 1412.6572, 2014

[24] Zhang Sisi, Zuo Xin, Liu Jianwei. The problem of the adversarial examples in deep learning [J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904 (in Chinese)
(张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904)

[25] Tanay T, Griffin L. A boundary tilting persepective on the phenomenon of adversarial examples [J]. arXiv preprint, arXiv: 160807690, 2016

[26] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Analysis of universal adversarial perturbations [J]. arXiv preprint, arXiv: 170509554, 2019

[27] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world [J]. arXiv preprint, arXiv: 160702533, 2016

[28] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9185-9193

[29] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the 2016 IEEE European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2016: 372-387

[30] Sarkar S, Bansal A, Mahbub U, et al. UPSET and ANGRI: Breaking high performance image classifiers [J]. arXiv preprint, arXiv: 170701159, 2017

[31] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling deep structured prediction models [J]. arXiv preprint, arXiv: 170705373, 2017

[32] Maho T, Bonnet B, Furony T, et al. RoBIC: A benchmark suite for assessing classifiers robustness [C] //Proc of the 2021 IEEE Int Conf on Image Processing (ICIP). Piscataway, NJ: IEEE, 2021: 3612-3616

[33] Katz G, Barrett C, Dill D L, et al. Reluplex: An efficient SMT solver for verifying deep neural networks [C] //Proc of Int Conf on Computer Aided Verification. Berlin: Springer, 2017: 97-117

[34] Sinha A, Namkoong H, Duchi J. Certifiable distributional robustness with principled adversarial training [J]. arXiv preprint, arXiv: 171010571, 2017

[35] Kantchelian A, Tygar J D, Joseph A. Evasion and hardening of tree ensemble classifiers [C] //Proc of the Int Conf on Machine Learning. Long Beach, CA: PMLR, 2016: 2387-2396

[36] Ji Shouling, Du Tianyu, Deng Shuiguang, et al. Robustness certification research on deep learning models: A survey [J]. Chinese Journal of Computers, 2022, 45(1): 190-206 (in Chinese)
(纪守领, 杜天宇, 邓水光, 等. 深度学习模型鲁棒性研究综述[J]. 计算机学报, 2022, 45(1): 190-206)

[37] Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness [J]. arXiv preprint, arXiv: 190206705, 2019

[38] Lechner M, Hasani R, Grosu R, et al. Adversarial training is not ready for robot learning [C] //Proc of the 2021 IEEE Int Conf on Robotics and Automation (ICRA). Piscataway, NJ: IEEE, 2021: 4140-4147

[39] Liu Aishan, Liu Xianglong, Guo Jun, et al. A comprehensive evaluation framework for deep model robustness [J]. arXiv preprint, arXiv: 210109617, 2021

[40] Ling Xiang, Ji Shouling, Zou Jiaxu, et al. DEEPSEC: A uniform platform for security analysis of deep learning model [C] //Proc of the 2019 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 673-690

[41] Luo Bo, Liu Yannan, Wei Lingxiao, et al. Towards imperceptible and robust adversarial example attacks against neural networks [C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1652-1659

[42] Dong Yinpeng, Fu Qian, Yang Xiao, et al. Benchmarking adversarial robustness on image classification [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 321-331

[43] Šircelj J, Skočaj D. Accuracy-perturbation curves for evaluation of adversarial attack and defence methods [C] //Proc of the 25th Int Conf on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2021: 6290-6297

[44] Wu Jing, Zhou Mingyi, Zhu Ce, et al. Performance evaluation of adversarial attacks: Discrepancies and solutions [J]. arXiv preprint, arXiv: 210411103, 2021

[45] Kunhardt O, Deza A, Poggio T. The effects of image distribution and task on adversarial robustness [J]. arXiv preprint, arXiv: 210210534, 2021

[46] Zhang Chongzhi, Liu Aishan, Liu Xianglong, et al. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity [J]. IEEE Transactions on Image Processing, 2020, 30: 1291-1304

[47] Chen Sihong, Shen Haojing, Wang Ran, et al. Relationship between prediction uncertainty and adversarial robustness [J]. Journal of Software, 2022, 33(2): 524-538 (in Chinese)
(陈思宏, 沈浩靖, 王冉, 等. 预测不确定性与对抗鲁棒性的关系研究[J]. 软件学报, 2022, 33(2): 524-538)

[48] Su Jizhu. Robustness evaluation of convolutional neural network models based on prediction uncertainty [D]. Beijing: Beijing Jiaotong University, 2021 (in Chinese)

(苏记柱. 基于预测不确定性的卷积神经网络模型鲁棒性评估[D]. 北京: 北京交通大学, 2021)

[49] Liu Aishan, Liu Xianglong, Yu Hang, et al. Training robust deep neural networks via adversarial noise propagation [J]. IEEE Transactions on Image Processing, 2021, 30: 5769-5781

[50] Weng Tsui-wei, Zhang Huan, Chen Pinyu, et al. Evaluating the robustness of neural networks: An extreme value theory approach [J]. arXiv preprint, arXiv: 180110578, 2018

[51] De Haan L, Ferreira A. Extreme Value Theory: An Introduction [M]. Berlin: Springer, 2006

[52] Weng Tsui-wei, Zhang Huan, Chen Pinyu, et al. On extensions of clever: A neural network robustness evaluation algorithm [C] //Proc of the 2018 IEEE Global Conf on Signal and Information Processing (GlobalSIP). Piscataway, NJ: IEEE, 2018: 1159-1163

[53] Weng Tsui-wei, Zhang Huan, Chen Hongge, et al. Towards fast computation of certified robustness for ReLU networks [C] //Proc of Int Conf on Machine Learning. Sweden: PMLR, 2018: 5276-5285

[54] Zhang Huan, Weng Tsui-wei, Chen Pinyu, et al. Efficient neural network robustness certification with general activation functions [J]. Advances in Neural Information Processing Systems, 2018, 31: 4939-4948

[55] Xu Huan, Mannor S. Robustness and generalization [J]. Machine Learning, 2012, 86(3): 391-423

[56] Myers G J, Sandler C, Badgett T. The Art of Software Testing [M]. New York: John Wiley & Sons, 1979

[57] Pei Kexin, Cao Yinzhi, Yang Junfeng, et al. DeepXplore: Automated whitebox testing of deep learning systems [C] // Proc of the 26th Symp on Operating Systems Principles. New York: ACM, 2017: 1-18

[58] Ma Lei, Xu Juefei, Zhang Fuyuan, et al. DeepGauge: Multi-granularity testing criteria for deep learning systems [C] // Proc of the 33rd ACM/IEEE Int Conf on Automated Software Engineering. Piscataway, NJ: IEEE, 2018: 120-131

[59] Dong Yizhen, Zhang Peixin, Wang Jingyi, et al. An empirical study on correlation between coverage and robustness for deep neural networks [C] //Proc of the 25th Int Conf on Engineering of Complex Computer Systems (ICECCS). Piscataway, NJ: IEEE, 2020: 73-82

[60] Yan Shenao, Tao Guanhong, Liu Xuwei, et al. Correlations between deep neural network model coverage criteria and model quality [C] //Proc of the 28th ACM Joint Meeting on European Software Engineering Conf and Symp on the Foundations of Software Engineering. New York: ACM, 2020: 775-787

[61] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[62] Legge G E, Foley J M. Contrast masking in human vision [J]. JOSA, 1980, 70(12): 1458-71

[63] Liu Anmin, Lin Weisi, Paul M, et al. Just noticeable difference for images with decomposition model for separating edge and textured regions [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(11): 1648-1652

[64] Lin Weisi, Li Dong, Ping Xue. Visual distortion gauge based on discrimination of noticeable contrast changes [J]. IEEE Transactions On Circuits and Systems for Video Technology, 2005, 15(7): 900-909

[65] Dong Yifan. Robustness of intrusion detection methods in a dversarial environment [D]. Changsha: National University of Defense Technology, 2018 (in Chinese)
(董一帆. 对抗环境下入侵检测方法的鲁棒性研究[D]. 长沙: 国防科技大学, 2018)

[66] Kuncheva L I. A stability index for feature selection [C] // Proc of the 25th Int Multi-Conf on Artificial Intelligence and Applications. New York: ACM, 2007: 390-395

[67] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations [J]. arXiv preprint, arXiv: 190312261, 2019

[68] Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning Interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096)

[69] Li Deng. The MNIST database of handwritten digit images for machine learning research [J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142

[70] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database [C] //Proc of the 2009 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248-255

[71] Li Feifei, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories [C] //Proc of the 2004 Conf on Computer Vision and Pattern Recognition Workshop. Piscataway, NJ: IEEE, 2004: 178-178

[72] Griffin G, Holub A, Perona P. Caltech-256 object category dataset [EB/OL]. [2022-03-15]. http: //authors. library. caltech.edu/7694

[73] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images [D]. Toronto, Canada: University of Toronto, 2009

[74] Papernot N, Faghri F, Carlini N, et al. Technical report on the CleverHans v2. 1.0 adversarial examples library [J]. arXiv preprint, arXiv: 161000768, 2016

[75] Papernot N, Goodfellow I, Sheatsley R, et al. CleverHans v2.0.0: An adversarial machine learning library [J]. arXiv preprint, arXiv: 161000768, 2016

[76] Rauber J, Brendel W, Bethge M. Foolbox: A Python toolbox to benchmark the robustness of machine learning models [J]. arXiv preprint, arXiv: 170704131, 2017

[77] Nicolae M I, Sinn M, Minh T N, et al. Adversarial robustness toolbox v0.2.2 [J]. arXiv preprint, arXiv: 180701069, 2018

[78] Ding Guangwei, Wang Luyu, Jin Xiaomeng. AdverTorch v0.1: An adversarial robustness toolbox based on Pytorch [J]. arXiv preprint, arXiv: 190207623, 2019

[79] Goodman D, Hao Xin, Wang Yang, et al. AdvBox: A toolbox to generate adversarial examples that fool neural networks [J]. arXiv preprint, arXiv: 200105574, 2020

[80] Li Yaxin, Jin Wei, Xu Han, et al. DeepRobust: A Pytorch library for adversarial attacks and defenses [J]. arXiv preprint, arXiv: 200506149, 2020

[81] Croce F, Andriushchenko M, Sehwag V, et al. RobustBench: A standardized adversarial robustness benchmark [J]. arXiv preprint, arXiv: 201009670, 2020

[82] Abadi M, Barham P, Chen Jianmin, et al. TensorFlow: A system for large-scale machine learning [C] //Proc of the 12th USENIX Symp on Operating Systems Design and Implementation (OSDI 16). Berkeley, CA: USENIX Association, 2016: 265-283

[83] Ketkar N. Introduction to Keras [M] //Deep Learning with Python. Berlin: Springer, 2017: 97-111

[84] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library [J]. Advances in Neural Information Processing Systems, 2019, 32: 8026-8037

[85] Al-Rfou R, Alain G, Almahairi A, et al. Theano: A Python framework for fast computation of mathematical expressions [J]. arXiv preprint, arXiv: 1605.02688, 2016

[86] Chen Tianqi, Li Mu, Li Yutian, et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems [J]. arXiv preprint, arXiv: 151201274, 2015

[87] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python [J]. The Journal of Machine Learning Research, 2011, 12: 2825-2830

[88] Chen Tianqi, Guestrin C. Xgboost: A scalable tree boosting system [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794

[89] Ke Guolin, Meng Qi, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree [J]. Advances in Neural Information Processing Systems, 2017, 30: 3149-3157

[90] Dorogush A V, Ershov V, Gulin A. CatBoost: Gradient boosting with categorical features support [J]. arXiv preprint, arXiv: 181011363, 2018

[91] Iandola F, Moskewicz M, Karayev S, et al. DenseNet: Implementing efficient convnet descriptor pyramids [J]. arXiv preprint, arXiv: 14041869, 2014

[92] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[93] Ma Yanjun, Yu Dianhai, Wu Tian, et al. PaddlePaddle: An open-source deep learning platform from industrial practice [J]. Frontiers of Data and Domputing, 2019, 1(1): 105-115

[94] Markham A, Jia Yangqing. Caffe2: Portable high-performance deep learning framework from Facebook [EB/OL]. (2017-04-19) [2022-05-24]. http://dublincore.org/documents/2003/08/26/usageguide

[95] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Identity mappings in deep residual networks [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 630-645

[96] Grüning P, Martinetz T, Barth E. Feature products yield efficient networks [J]. arXiv preprint, arXiv: 200807930, 2020

[97] Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition [C] //Proc of the 8th Annual Computing and Communication Workshop and Conf (CCWC). Piscataway, NJ: IEEE, 2018: 169-175

Li Zituo, born in 1999. Master candidate. His main research interests include adversarial robustness evaluation, test and evaluation, and deep learning.

李自拓,1999 年生.硕士研究生.主要研究方向为对抗鲁棒性评估、试验鉴定和深度学习.

Sun Jianbin, born in 1989. PhD, associate professor, master supervisor. His main research interests include system test and evaluation, decision and analysis under uncertainty.

孙建彬,1989 年生.博士,副教授,硕士生导师.主要研究方向为系统试验与评价、不确定性决策分析.

Yang Kewei, born in 1977. PhD, professor, PhD supervisor. His main research interests include test and evaluation, defense acquisition and system of systems requirement modeling.

杨克巍,1977 年生.博士,教授,博士生导师.主要研究方向为试验鉴定、国防采办与体系需求建模.

Xiong Dehui, born in 1990. PhD, lecturer. His main research interests include deep learning and big data analysis.

熊德辉,1990 年生.博士,讲师.主要研究方向为深度学习与大数据分析.

附录 A

Table A1 Corresponding Number of Model and Evaluation Index

表 A1 模型及评估指标对应编号

编号	模型	评估指标
1	CNN	*CA*
2	GCN	*MR*
3	DenseNet	*ACAC*
4	ResNet	*ACTC*
5	VGG	ALD_p
6	FP_ResNet	*ASS*
7	YOLOv3	*PSD*
8	SSD300	*NTE*
9	Retina-Resnet50	*RGB*
10	AlexNet	*RIC*
11	DenseNet	*CC*
12	original DL model	*CAV*
13	PGD-AT	*CRR*/*CSR*
14	DeepDefense	*CCV*
15	TRADES	*COS*
16	Convex	*MCE*
17	JPEG	*RMCE*
18	RSE	*MFP*
19	ADP	*MT*5*D*
20	Inc-v3	*SNS*
21	Ens-AT	*ENI*
22	ALP	*BD*
23	FD	*BD*2
24	JPEG	*CAR*
25	Bit-Red	*NC*
26	R&P	*KMNC*
27	RandMix	*NBC*
28		*SNAC*
29		*TKNC*
30		*TKNP*
31		*Clique Method Robustness Verification*
32		*Loss Sensitivity*
33		*Empirical Robustness*
34		*CLEVER*
35		*Wasserstein Distance*
36		*Pointwise Differential Training Privacy*
37		*Criterion*

注:表 4 中 RobustBench 所含模型的数量过多,此表不一一列举.

Table A2 Integration Algorithm Details of CleverHans v4.0.0

表 A2 CleverHans v4.0.0 集成算法明细

攻击算法	防御算法
Carlini & Wagner-L2 Attack(CW)	Resampling
Fast Gradient Sign Method(FGSM)	
Hop Skip Jump Attack	
Projected Gradient Descent(PGD)	
Spatial Transformation Method	
Sparse L1 Descent Attack	
Virtual Adversarial Method	
Fast Feature Adversaries	
Saliency Map Method	
Elastic Net Method	
DeepFool	
LBFGS	
Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation(SPSA)	
Basic Iterative Method(BIM)	
Madry Et Al Attack	
Momentum Iterative Method(MIM)	

Table A3 Integration Algorithm Details of Foolbox v3.0.0

表 A3 Foolbox v3.0.0 集成算法明细

攻击算法	防御算法
L2 Contrast Reduction Attack	
Virtual Adversarial Attack	
DDN Attack	
Projected Gradient Descent Attack(L2/Linf)	
Basic Iterative Attack(L2/Linf)	
Fast Gradient Attack(L2/Linf)	
L2 Additive Gaussian Noise Attack	
L2 Additive Uniform Noise Attack	
L2 Clipping Aware Additive Gaussian Noise Attack	
L2 Clipping Aware Additive Uniform Noise Attack	
Linf Additive Uniform Noise Attack	
L2 Repeated Additive Gaussian Noise Attack	
L2 Repeated Additive Uniform Noise Attack	
L2 Clipping Aware Repeated Additive Gaussian Noise Attack	
L2 Clipping Aware Repeated Additive Uniform Noise Attack	
Linf Repeated Additive Uniform Noise Attack	
Inversion Attack	
Binary Search Contrast Reduction Attack	
Linear Search Contrast Reduction Attack	
Hop Skip Jump Attack	
L2 Carlini Wagner Attack	
Newton Fool Attack	
EAD Attack	
Gaussian Blur Attack	
DeepFool Attack(L2/Linf)	

续表 A3

攻击算法	防御算法
Salt and Pepper Noise Attack	
Linear Search Blended Uniform Noise Attack	
Binarization Refinement Attack	
Dataset Attack	
Boundary Attack	
Brendel Bethge Attack(L0/L1/L2)	
Linfinity Brendel Bethge Attack	
FMN Attack(L0/L1/L2/Linf)	
Pointwise Attack	
FGM	
FGSM	
PGD(L2/Linf)	

Table A4 Integration Algorithm Details of ART

表 A4 ART 集成算法明细

攻击算法	防御算法
Adversarial Patch	InverseGAN
Auto Attack	DefenseGAN
Auto Projected Gradient Descent (Auto-PGD)	Video Compression
Boundary Attack/Decision-Based Attack	Resampling
Brendel and Bethge Attack	Thermometer Encoding
Carlini and Wagner Attack(L0/L2/Linf/ASR)	MP3 Compression
Decision Tree Attack	Total Variance Minimization
DeepFool	PixelDefend
DPatch	Gaussian Data Augmentation
Robust DPatch	Feature Squeezing
Elastic Net Attack	Spatial Smoothing
Fast Gradient Method (FGM)	JPEG Compression
Feature Adversaries	Label Smoothing
Frame Saliency Attack	Virtual Adversarial Training
Geometric Decision Based Attack	Reverse Sigmoid
High Confidence Low Uncertainty Attack	Random Noise
Hop Skip Jump Attack	Class Labels
Imperceptible ASR Attack	High Confidence
Basic Iterative Method (BIM)	Rounding
Projected Gradient Descent (PGD)	General Adversarial Training
Laser Attack	Madry's Protocol
LowProFool	Fast is Better Than Free
NewtonFool	Defensive Distillation
Malware Gradient Descent	Neural Cleanse
Over The Air Flickering Attack	Basic Detector Based on Inputs
Pixel Attack	Detector Trained on the Activations of a Specific Layer
Threshold Attack	Detector Based on Fast Generalized Subset Scan
Jacobian Saliency Map Attack (JSMA)	Detection Based on Activations Analysis
Shadow Attack	Detection Based on Data Provenance

续表 A4

攻击算法	防御算法
Shape Shifter Attack	Detection Based on Spectral Signatures
Simple Black-box Adversarial Attack	
Spatial Transformations Attack	
Square Attack	
Targeted Universal Perturbation Attack	
Universal Perturbation Attack	
Virtual Adversarial Method	
Wasserstein Attack	
Zeroth-Order Optimization (ZOO) Attack	

Table A5 Integration Algorithm Details of DEEPSEC

表 A5 DEEPSEC 集成算法明细

攻击算法	防御算法
Fast Gradient Sign Method (FGSM)	Naive Adversarial Training(NAT)
Random Perturbation with FGSM (R+FGSM)	Ensemble Adversarial Training(EAT)
Basic Iterative Method (BIM)	PGD-based Adversarial Training(PAT)
Projected LoGradient Descent Attack (PGD)	Defensive Distillation(DD)
Un-targeted Momentum Iterative FGSM (U-MI-FGSM)	Input Gradient Regularization(IGR)
DeepFool (DF)	Ensemble Input Transformation(EIT)
Universal Adversarial Perturbation Attack (UAP)	Random Transformations Based Defense(RT)
OptMargin (OM)	Pixel Defense(PD)
Least Likely Class Attack (LLC)	Thermometer Encoding Defense(TE)
Random Perturbation with LLC (R+LLC)	Region-based Classification(RC)
Iterative LLC Attack (ILLC)	Local Intrinsic Dimensionality Based Detector(LID)
Targeted Momentum Iterative FGSM (T-MI-FGSM)	Feature Squeezing Detector(FS)
Box-constrained L-BFGS Attack (BLB)	MagNet Detector(MagNet)
Jacobian-based Saliency Map Attack (JSMA)	
Carlini and Wagner's Attack (CW)	
Elastic-Net Attacks to DNNs (EAD)	

Table A6 Integration Algorithm Details of AdverTorch

表 A6 AdverTorch 集成算法明细

攻击算法	防御算法
Gradient Attack	Conv Smoothing 2D
Gradient Sign Attack	Average Smoothing 2D
Fast Feature Attack	Gaussian Smoothing 2D
L2 Basic Iterative Attack	Median Smoothing 2D
Linf Basic Iterative Attack	JPEG Filter
PGD Attack	Bit Squeezing
Linf PGD Attack	Binary Filter
L2 PGD Attack	
L1 PGD Attack	
Sparse L1 Descent Attack	
Momentum Iterative Attack	
Linf Momentum Iterative Attack	

续表 A6

攻击算法	防御算法
L2 Momentum Iterative Attack	
Carlini Wagner L2 Attack	
Elastic Net L1 Attack	
DDN L2 Attack	
LBFGS Attack	
Single Pixel Attack	
Local Search Attack	
Spatial Transform Attack	
Jacobian Saliency Map Attack	

Table A7 Integration Algorithm Details of Ares

表 A7 Ares 集成算法明细

攻击算法	防御算法
FGSM	RST
BIM	TRADES
PGD	FS-AT
CW	Pre-Training
DeepFool	AT-HE
MIM	Robust Overfitting
DIM	FastAT
TIM	AWP
SI-NI-FGSM	HYDRA
VIM	Label Smoothing
SGM	
CDA	
AutoAttack	
Boundary	
SPSA	
Evolutionary	
NES	
Nattack	
TTA	

Table A8 Integration Algorithm Details of AdvBox

表 A8 AdvBox 集成算法明细

攻击算法	防御算法
L-BFGS	Feature Squeezing
FGSM	Spatial Smoothing
BIM	Label Smoothing
ILCM	Gaussian Augmentation
MI-FGSM	Adversarial Training
JSMA	Thermometer Encoding
DeepFool	
CW	
Single Pixel Attack	
Local Search Attack	

Table A9 Integration Algorithm Details of DeepRobust

表 A9 DeepRobust 集成算法明细

攻击算法	防御算法
NATTACK	FGSM Training
C&W	Fast(an Improved Version of FGSM Training)
DeepFool	PGD Training
FGSM	YOPO(an Improved Version of PGD Training)
LBFGS	TRADES
Onepixel	Thermometer Encoding
PGD	LID-based Adversarial Classifier
BPDA	Base Defense
Universal	AdvTraining(Graph)
YOPOPGD	GCN Module(Graph)
FGA(Graph)	GCN_preprocess Module(Graph)
RL-S2V(Graph)	PGD Module(Graph)
Nettack(Graph)	Prognn Module(Graph)
IG-Attack(Graph)	R_GCN Module(Graph)
RND(Graph)	
Metattack(Graph)	
PGD(Graph)	
Min-Max(Graph)	
DICE(Graph)	

Table A10 Integration Algorithm Details of AISafety

表 A10 AISafety 集成算法明细

攻击算法	防御算法
Basic Iterative Method(BIM)	Ensemble Adversarial Training(EAT)
Boundary Attack(BA)	New Adversarial Training (NAT)
Box-constrained L-BFGS Attack(BLB)	Original Adversarial Training (OAT)
Carlini & Wagner-L2 Attack(CW)	PGD Adversarial Training(PAT)
Corrupt	RAND
Deepfool	
Elastic-Net Attacks to DNNs(EAD)	
Fast Gradient Sign Method(FGSM)	
Iterative Least Likely Class Attack(ILLC)	
Jacobian-based Saliency Map Attack(JSM)	
Least-Likely-Class Iterative Methods(LLC)	
Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation(SPSA)	
Nature Evolutionary Strategies(NES)	
OPTMARGIN Attack(OM)	
Projected Gradient Descent(PGD)	
RAND-FGSM (R-FGSM)	
Random Least Likely Class Attack(RLLC)	
Universal Adversarial Perturbation Attack(UAP)	
Utargeted Momentum Iterative Fast Gradient Sign Method(UMI-FGSM)	
Zeroth Order Optimization Based Black-box(ZOO)	