计算机科学
COMPUTER SCIENCE

计算机视觉领域对抗样本检测综述

张鑫, 张晗, 牛曼宇, 姬莉霞

引用本文

张鑫, 张晗, 牛曼宇, 姬莉霞. 计算机视觉领域对抗样本检测综述[J]. 计算机科学, 2025, 52(1): 345-361.

ZHANG Xin, ZHANG Han, NIU Manyu, JI Lixia. Adversarial Sample Detection in Computer Vision:A Survey [J]. Computer Science, 2025, 52(1): 345-361.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300080

计算机视觉领域对抗样本检测综述

张 鑫[1] 张 晗[1,2] 牛曼宇[1] 姬莉霞[1,3]
1 郑州大学网络空间安全学院 郑州 450001
2 智能警务四川省重点实验室 四川 泸州 646000
3 四川大学计算机学院 成都 610065
(geekxin@gs.zzu.edu.cn)

摘 要 随着数据量的增加和硬件性能的提升,深度学习在计算机视觉领域取得了显著进展。然而,深度学习模型容易受到对抗样本的攻击,导致输出发生显著变化。对抗样本检测作为一种有效的防御手段,可以在不改变模型结构的前提下防止对抗样本对深度学习模型造成影响。首先,对近年来的对抗样本检测研究工作进行了整理,分析了对抗样本检测与训练数据的关系,根据检测方法所使用特征进行分类,系统全面地介绍了计算机视觉领域的对抗样本检测方法;然后,对一些结合跨领域技术的检测方法进行了详细介绍,统计了训练和评估检测方法的实验配置;最后,汇总了一些有望应用于对抗样本检测的技术,并对未来的研究挑战进行展望。

关键词:深度学习;对抗样本攻击;对抗样本检测;人工智能安全;图像分类
中图分类号 TP391

Adversarial Sample Detection in Computer Vision:A Survey

ZHANG Xin[1],ZHANG Han[1,2],NIU Manyu[1] and JI Lixia[1,3]
1 School of Cyber Science and Engineering,Zhengzhou University,Zhengzhou 450001,China
2 Intelligent Policing Key Laboratory of Sichuan Province,Luzhou,Sichuan 646000,China
3 School of Computer Science,Sichuan University,Chengdu 610065,China

Abstract With the increase in data volume and improvement in hardware performance,deep learning(DL) has made significant progress in the field of computer vision. However,deep learning models are vulnerable to adversarial samples,causing significant changes in the output. As an effective defense method,adversarial sample detection can prevent adversarial samples from affecting the deep learning model without changing the model structure. This paper organizes the research work on adversarial example detection in recent years,analyzes the relationship between adversarial example detection and training data,classifies them according to the characteristics used in the detection method,and systematically and comprehensively introduces adversarial sample detection methods in the field of computer vision. Then,some detection methods that combine cross-domain technologies are introduced in detail,and the experimental configurations for training and evaluating detection methods are statistically analyzed. Finally,some technologies that are expected to be applied to adversarial sample detection are summarized,and future research challenges and development directions are prospected.

Keywords Deep learning,Adversarial sample attacks,Adversarial sample detection,AI security,Image classification

1 引言

得益于对网络结构以及训练策略的改进,深度学习(Deep Learning,DL)模型的精度不断提高,甚至表现出了超越人类的工作能力。各种深度学习模型如 AlexNet[1],VGG16[2],ResNet[3]和 Faster R-CNN[4]等不断涌现,它们在大规模视觉识别任务上取得了优异的性能。然而,早期的视觉模型在设计时未考虑对抗样本对模型的威胁,很容易受到对抗样本攻击。对抗样本由 Szegedy 等[5]于 2013 年首次提出,他们发现在输入数据中添加微小的、人眼几乎无法察觉的扰动,可以导致深度神经网络的预测结果发生显著变化。自此,各种对抗样本攻击方法开始出现。Goodfellow 等[6]在 2014 年提出了

到稿日期:2024-03-12 返修日期:2024-08-10
基金项目:国家自然科学基金青年科学基金(62302458);河南省重大科技专项(231100210200);河南省高等学校重点科研项目(24A520047);智能警务四川省重点实验室 2024 年度开放课题项目(ZNJW2024KFQN005)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(62302458),Major Science and Technology Project of Henan Province(231100210200),Key Scientific Research Project of Universities in Henan Province(24A520047) and 2024 Open Subject Project of Sichuan Provincial Key Laboratory of Intelligent Policing(ZNJW2024KFQN005).
通信作者:姬莉霞(jilixia@zzu.edu.cn)

快速梯度符号方法(Fast Gradient Sign Method,FGSM),该方法启发了后续许多具有代表性的对抗样本生成算法,包括基于梯度的攻击[7]、基于优化的攻击[8]、基于元学习的攻击等[9]。这些对抗性攻击方法已经不只局限于图像分类任务[5-6],还扩展到了其他领域,如目标检测[10-11]、语音识别[12-13],甚至物理世界中的场景[14-15]。

为应对对抗样本对深度学习模型构成的威胁,研究人员开始探索各种检测和防御对抗样本的方法。最初的防御方法主要集中在对模型进行训练的过程中增加对抗样本,这种方法被称为对抗训练[6],但实验表明对抗训练会影响模型精度。随着研究的深入,人们开始探索更复杂的防御策略。一种常见的防御策略是使用模型的集成来提高鲁棒性[16],通过结合多个模型的预测来提高对抗攻击的鲁棒性。然而,这种方法也有其局限性,因为它需要大量的计算资源,会导致模型的复杂性增加。

为了应对上述挑战,越来越多的研究者开始关注对抗样本检测技术,试图在对抗样本影响模型之前预测并拒绝对抗样本。例如,基于统计分布的检测方法[17]通过比较样本间的统计特征来检测对抗样本;基于特征变换的方法[18]通过分析模型预测的不一致性来发现对抗样本。随着研究的深入,一些研究者开始研究使用频率分析来检测对抗攻击的方法[19-20]。

图 1 给出了通过 Web of science[1)] 检索"对抗样本检测"与"计算机视觉"相关文献的结果(截至 2023 年 10 月)。

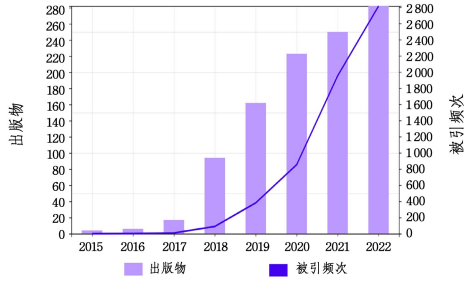


图 1 对抗样本检测领域出版物数量的变化趋势

Fig. 1 Trend in the number of publications in the field of adversarial sample detection

可以看出,近年来对抗样本检测领域的研究呈现出快速增长的趋势,但是,相关的总结性研究仍相对较少,使得初学者难以系统理解这个领域。目前已有学者撰写了有关对抗样本检测方法的综述[21-23]。文献[21]主要介绍了对抗样本防御技术,其中涉及的对抗样本检测方法较少且分类不全面。文献[22]虽然回顾了各种对抗样本检测方法并设计了实验进行对比,但是其对方法的介绍较为简略。文献[23]侧重方法的介绍,但其分类结构复杂,并且涉及的文献相对较早,它们都采用了相对复杂的分类结构。与之不同,本文对分类结构进行了合理的调整。例如,将文献[22]中基于去噪和特征压缩的无监督检测整合为基于特征变换的检测,因为其原理上都是通过特征变换后模型预测的不一致来进行检测。类似地,文献[23]中基于对抗训练和神经网络特性的有监督检测方法均使用图像的原始特征或神经网络中间特征来训练分类器进行检测,可被归为一类。

1) https://webofscience.clarivate.cn

与之前的相关综述相比,本文的主要贡献如下:

1)分类清晰,文献完备。为避免分类结构的复杂性,本文总结了之前工作[22-23]中一些技术原理相似的检测方法,并对分类原理进行了系统介绍。此外,之前工作中介绍的最新方法截至 2021 年,本文在此基础上新增了近两年最新的对抗样本检测研究,并总结和分析了它们的优势和局限性。

2)整合跨领域技术,提供新思路。本文梳理了最新的研究成果,介绍了这些新颖的检测方法如何与其他领域的先进技术相结合,包括与生成模型、语义分割等技术结合的检测方法。同时,还总结了有望提高对抗样本检测性能的技术,分析了它们在对抗样本检测中的可行性,为未来研究提供了新的方向和思路。

3)深度剖析检测方法和样本之间的联系,并以此为依据,将本文所介绍方法划分为无监督和有监督两类。进一步根据方法所依赖的特征进行细分,从模型(包括激活通道和中间层特征等模型内部特征)、数据(包括像素值分布和颜色分布等数据统计特征)及其他角度(邻域比较和特征变换)等方面进行分析,剖析检测方法的运作原理,突显不同方法之间的原理差异。

鉴于当前对抗样本的研究主要针对图像数据,本文的讨论也集中于图像分类领域。第 2 章介绍相关定义和背景,回顾对抗样本的定义和生成方法;第 3 章和第 4 章按照优化过的分类介绍对抗样本检测方法;第 5 章详细阐述最新的研究成果,重点介绍它们如何与跨领域先进技术相结合;第 6 章介绍检测方法的评估,包括相关数据集、评价指标和实验设置;第 7 章探讨对抗样本检测领域中的一些前沿,即尚未在相关文献中广泛讨论的技术或原理;最后总结未来的研究挑战。本文具体组织结构如图 2 所示。

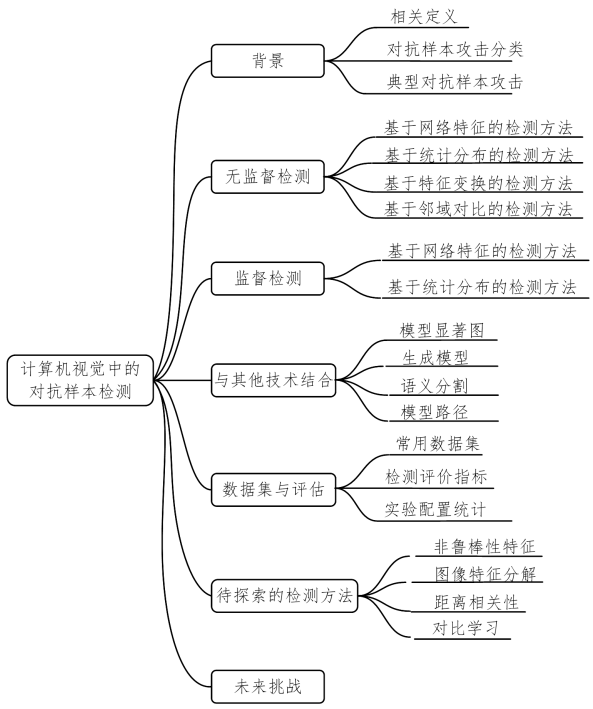


图 2 文章组织结构

Fig. 2 Organizational structure of this paper

2　背景

本章首先介绍了计算机视觉领域中对抗样本的基本概念以及各种分类,然后列举了一些典型的对抗样本攻击算法。

2.1　相关定义

表 1 列出了对抗样本所涉及的相关定义。

表 1　相关定义描述

Table 1　Related definition descriptions

定义	描述
对抗攻击	生成对抗样本的算法
对抗性扰动	添加到正常图像中的噪音,能使其成为对抗样本
目标模型	对抗攻击所针对的模型
对抗样本	对正常样本的修改版本,添加了对抗性扰动
欺骗率	攻击算法产生的对抗样本中,成功欺骗目标模型的样本占比
迁移性	在一个模型上生成的对抗样本欺骗其他模型的能力
对抗训练	将对抗样本加入模型训练集,以提高模型的鲁棒性
对抗样本防御	一系列方法或策略,旨在增强模型鲁棒性,减轻或阻止对抗样本对模型的负面影响
检测器	一种算法或者模型,仅用于检测图像是否是对抗样本

2.2　对抗样本攻击方法

深度神经网络(Deep-Learning Neural Network,DNN)是一类用于解决模式识别和机器学习问题的人工神经网络模型。然而,深度神经网络模型对于对抗样本(Adversarial Examples)表现出脆弱性。在给定模型 f 和输入样本(x,y)的情况下,存在对抗样本 x',满足 $\|x'-x\|<\varepsilon$,并且可使预测函数 $f(x)\neq f(x')$,其中 ε 是最大允许扰动的范围。

对抗样本攻击的目标是通过在原始输入 x 中添加极小的扰动,来使得模型产生错误的预测结果。这些对抗性扰动靠人眼难以察觉,但对深度学习模型却能产生显著影响,导致模型输出错误的预测结果。对抗性扰动的生成方式主要有 4 种:基于梯度,基于优化,基于边界,基于生成模型。

基于梯度的攻击以 FGSM[6] 为基础,其基本攻击原理如图 3 所示,通过模型反向传播得到的梯度计算对抗性扰动,并将其添加至图像,使得深度学习模型分类错误。PGD[24] 对单次攻击的步长(最大允许扰动)进行限制,提出了迭代的攻击方式,通过"少量多次"地向图像中添加对抗性扰动,可以在相同大小的扰动下提高其攻击性。

基于优化的对抗性扰动生成方式将攻击过程视为一个逐步优化的过程,不仅要确保扰动足够小以至于无法被人眼察觉,而且要确保能够使模型进行错误分类。优化目标函数通常如式(1)所示:

$$\begin{aligned}&\text{Minimize}\ \|\delta\|_2^2\\&\text{s.t.}\ \ f(x+\delta)=y^{\text{tar}},x+\delta\in[0,1]^R\end{aligned}\tag{1}$$

其中,$f,(\cdot)$是深度神经网络的预测函数,x 是干净图片,δ 是添加的全局扰动,y^{tar}是目标标签。

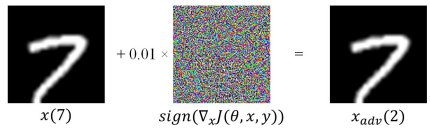


图 3　对抗样本攻击的作用原理

Fig. 3　Principle of adversarial sample attack

基于边界的攻击采用了高维超平面分类的思想,即通过迭代地调整样本,使其逐渐接近模型的决策边界并最终越过边界,从而被误分类。例如,DeepFool[25] 将分类边界和样本之间的距离定义为改变样本分类标签的最小扰动;在二分类问题中,它计算的对抗扰动相当于样本到分类边界的距离。DeepFool 在每次迭代中调整扰动,将原始样本逼近决策边界,直到成功跨越边界。在相同攻击成功率下,与 FGSM 相比,DeepFool 所需的扰动量更小。

生成对抗网络(Generative Adversarial Networks, GAN)可以用来生成对抗样本。例如,Xiao 等[26] 提出了一种基于 GAN 的攻击框架,称为 AdvGAN。该框架包含生成器 G、鉴别器 D 和目标模型 f。生成器 G 的输入是原始样本 x,输出是对抗扰动 $G(x)$。将对抗样本 $x+G(x)$ 输入到鉴别器 D 和目标模型 f 中。作者在 GAN 损失的基础上添加了对抗损失,鉴别器 D 采用 GAN 损失使对抗样本更接近原始样本,在目标模型 f 中采用对抗损失使其预测结果偏离正确标签。

这些方法可以针对不同的攻击目标、攻击场景和模型特性来生成具有攻击性的样本。

此外,对抗样本还可以从多个方面进行分类和研究。

1)基于攻击者知识的分类。根据攻击者对目标模型的了解程度,可以将对抗样本攻击分为白盒攻击、灰盒攻击和黑盒攻击。白盒攻击指攻击者完全了解目标模型的结构和参数,能够直接访问模型的内部信息;灰盒攻击指攻击者对目标模型有一定的了解,但无法获得完整的模型信息;黑盒攻击指攻击者对目标模型一无所知,只能通过模型的输入和输出进行攻击。图 4 展示了白盒攻击和黑盒攻击的基本过程,白盒攻击通过反向传播得到梯度并计算扰动,而黑盒攻击需要查询受害模型并标注数据,使用标注数据训练替代模型,再对替代模型进行白盒攻击得到对抗样本。

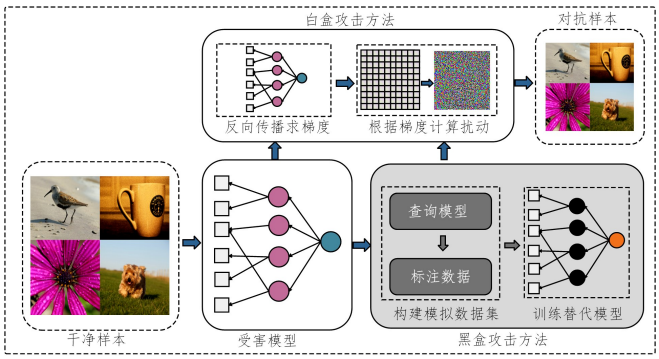


图 4　对抗样本攻击过程

Fig. 4　Adversarial sample attack process

2)基于扰动范围的分类。根据对原始输入图像进行扰动的范围和强度,可以将对抗样本攻击分为全局扰动攻击和局部扰动攻击。全局扰动攻击指对整个输入图像进行扰动,导致模型产生错误的预测结果;局部扰动攻击指只对图像的特定区域进行扰动,以误导模型对该区域的识别。

3)基于攻击目标的分类。根据攻击者希望模型产生的错误结果类型,可以将对抗样本攻击分为定向攻击和非定向攻击。定向误导攻击旨在使模型将输入图像误认为特定的目标类别;非定向攻击旨在使模型将输入图像分类为任意错误的类别。

4)基于攻击频次的分类。根据攻击者对模型的持续攻击方式,可以将对抗样本攻击分为单次攻击和迭代攻击。单次攻击指攻击者只进行一次扰动操作以生成对抗样本;而迭代攻击指攻击者通过多次迭代操作来逐渐优化对抗样本的效果。表 2 列出了一些著名攻击方法。

表 2 著名对抗样本攻击方法

Table 2 Famous adversarial sample attack methods

对抗攻击	模型知识	扰动范围	攻击目标	攻击频次
FGSM[6]	白盒	全局	非定向	单次
C&W 攻击[8]	白盒	全局	定向	迭代
PGD[24]	白盒	全局	非定向	迭代
DeepFool[25]	白盒	全局	非定向	迭代
AdvGAN[26]	白盒	部分	定向	迭代
JSMA[27]	白盒	部分	定向	迭代
One-Pixel[28]	黑盒	部分	定向	迭代
ZOO[29]	黑盒	部分	定向	迭代
GAP++[30]	白盒	全局	定向	单次

2.3 对抗样本检测方法与样本之间的关系

本节将介绍对抗样本检测方法分类的基准。首先,通过分析检测方法的数据需求,将本文所介绍的方法划分为无监督和有监督两大类。然后,从多个角度深入分析这些方法所依赖的特征,剖析检测方法的运作原理,并进行详细分类。

2.3.1 对抗样本与正常样本的特征差异

对抗样本检测通常被视为二分类问题,需要将输入样本分为两类:正常样本和添加恶意扰动的对抗样本。通过人眼观察,对抗样本与正常样本非常相似,但它们之间存在微小的特征差异,这些差异特征是检测对抗样本的关键。但由于这些差异特征微小且难以被准确提取,因此需要通过其他特征来间接反映它们的存在,以区分对抗样本与正常样本。本文从不同角度来考虑这些特征。

从模型的角度,对抗样本可能会导致网络中的中间层特征发生显著变化,因此可以考虑使用模型激活通道、中间层特征等作为指标。这些特征可以捕获对抗样本与正常样本之间的差异。

从数据的角度,使用一系列数据统计特征,如 softmax 分布、主成分分析(Principal Components Analysis,PCA)分量等。这些统计特征可以通过多种不同的统计量来体现对抗样本与正常样本之间的不同,因此,可以分析像素值的分布、颜色分布或纹理特征的统计数据,以检测对抗样本的存在。

还有一些研究从其他角度进行分析。例如,有研究[31]表明对抗样本在子空间中的近邻样本与正常样本有显著区别,这在近邻分布上体现了对抗样本与正常样本的差异。此外,一些研究者发现,一些图像增强操作,如去噪和特征压缩等处理,会使对抗样本的预测结果发生明显变化[32-33]。

从这些角度得到的特征可以通过不同的方法进行比较,根据是否需要对抗样本作为训练数据来训练检测模型,可将检测方法分为两种:无监督检测方法和监督检测方法。

2.3.2 无监督检测方法

设计过程中无需对抗样本作为训练数据的检测方法简称无监督检测方法,其核心思想是利用对抗样本与正常样本之间微小而难以察觉的特征差异来进行阈值分类。这些方法根据原理的不同,分为以下 4 类。

1)基于网络特征的检测:这类方法集中关注深度学习模型的特征表示。它们使用模型的中间层特征,如卷积层的输出,来捕获对抗样本与正常样本之间的差异。由于对抗样本可能导致模型中间特征发生显著变化,因此通过提取这些特征并设置适当的阈值,可以筛选出对抗样本。

2)基于统计分布的检测:这类方法通过比较对抗样本与正常样本之间的统计分布来检测对抗样本。它们使用各种统计量,如像素值分布、颜色分布或纹理特征的统计数据,来捕获对抗样本的特殊模式。这些统计特征通常在对抗样本中表现出与正常样本不同的特征。

3)基于邻域对比的检测:这类方法利用对抗样本子空间近邻数据与正常样本的差异进行检测。对抗样本局部近邻的类别与对抗样本有明显不同,因此可以通过样本与附近样本之间的差距来检测对抗样本。

4)基于特征变换的检测:这类方法通过对样本进行变换,根据模型预测的不一致性来检测对抗样本。对抗样本对于去噪、特征压缩等操作通常表现不鲁棒,基于这一性质,可以检测出对抗样本。

2.3.3 监督检测方法

监督检测方法需要对抗样本作为训练数据训来练检测模型。相比无监督检测方法,监督检测方法在差异特征的基础上通过深度学习模型提取更深层次的差异,通常对于特定类型的对抗样本能够实现更高的检测精度,特别是对于那些具有高度迁移性的黑盒攻击。因此,监督检测方法通常使用特定类型的对抗样本进行训练,从而有针对性地提高模型的安全性。监督检测方法根据所使用特征的不同,通常分为两大类,即基于网络特征的检测和基于统计分布的检测。

1)基于网络特征的检测。从模型角度出发,关注正常样本和对抗样本在神经网络中特征或行为的差异性,基本原理是使用网络内部的特征如梯度和神经元激活状态等特征训练有监督模型。

2)基于统计分布的检测。从数据的角度出发,将输入数据的统计量,如局部固有维数(Local Intrinsic Dimensionality,LID),作为输入数据训练分类器,使用分类器对样本进行分类。

2.3.4 两类方法的差异

需要强调的是,尽管监督方法所使用的特征在某种程度上与无监督检测方法相似,但它们的检测原理存在根本性的不同。无监督检测方法通过特征提取和设置适当的阈值来筛选对抗样本,而无需训练分类器;监督检测方法通常需要生成

对抗样本以构建训练数据，然后使用这些数据或对应的特征来训练二分类器或其他类型的分类器，这些分类器可以区分对抗样本和正常样本。

本文通过区分监督和无监督检测方法，以及根据它们所使用的特征，将对抗样本检测方法细分。这种分类有助于理解不同方法之间的原理差异，以及它们在提高模型的安全性方面的特点。

3　无监督检测

无监督检测方法仅使用正常样本来设计检测器，由于不需要先验知识来识别对抗样本，因此在面对新的、未知的对抗攻击时更有优势。在无监督检测方法中，根据检测原理将其分为 4 个子类：基于网络特征的检测方法，基于统计分布的检测方法，基于特征对齐的检测方法和基于特征变换的检测方法。

本章将详细介绍这 4 种无监督检测方法子类及原理，并分析其优势和限制，以更好地理解它们在对抗样本检测中的作用。

3.1　基于网络特征的检测方法

对抗样本通过微小但有针对性的扰动对原始数据进行处理，这会导致神经网络内部和输出的特征发生变化，例如神经元的激活状态、特征映射。这些特征的变化提供了一种识别对抗样本的方式，即提取神经网络中的这些特征差异，并设定阈值来检测对抗样本。其检测原理如图 5 所示，步骤如下：

1）将待测样本输入至深度学习模型；

2）获得待测样本的中间特征、激活状态或者 Softmax 等网络特征；

3）计算待测样本和正常样本的网络特征差异，如果大于设定阈值，则将待测样本判定为对抗样本。

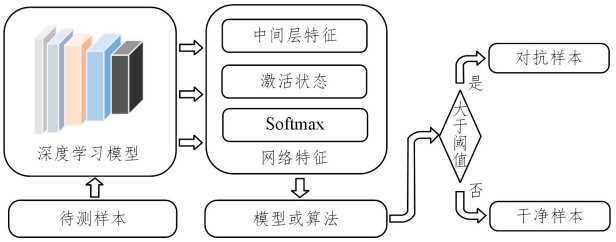


图 5　基于网络特征的检测方法的原理流程

Fig. 5　Principle flow of detection method based on network features

3.1.1　基于网络不变量

Ma 等[34]通过分析各种攻击下 DNN 模型的内部结构，发现攻击会使模型的内部激活值产生显著变化。这些变化主要表现在两个方面：起源通道和激活值分布通道。起源通道指那些对最终分类结果产生重大影响的神经元构成的通道；而激活值分布通道则描述了每一层神经元中会对最终结果造成影响的分布情况。基于这些发现，他们提出了一种检测对抗样本的方法，该方法侧重于分析起源通道和激活值分布通道中的不变量。他们引入了起源不变量（Provenance Invariance，PI）和激活值不变量（Value Distribution Invariance，VI），通过监测这些不变量是否发生变化来检测模型是否受到对抗样本的攻击。

Ma 等针对多种攻击方法进行了实验，其中对于通过 ResNet50 生成的 ImageNet 对抗样本平均检测成功率在 95% 以上；此外，该方法对于可绕过检测的攻击方法也有出色的效果，因为该检测方法基于整个模型的特征，绕过该检测方法需要对样本添加极大的扰动，这会使产生的对抗样本失去隐蔽性。这种方法的高度可解释性和精确性，使其在对抗样本检测领域具有独特价值。通过深入分析模型内部的 PI 和 VI，可以更清晰地了解对抗样本的影响机制。然而，这种方法并非适用于所有类型的神经网络模型，特别是对于那些具有特殊结构或注意力机制的模型，如 SE（Squeeze-and-Excitation）模块[35]，提取这些不变量会变得复杂且困难。此外，该方法的复杂性随着模型规模的增大而增加，因此在处理大型模型时可能效率较低，需要权衡检测的准确性和计算成本。

3.1.2　基于多层中间特征

Aldahdooh 等[36]提出了基于多层中间特征的对抗样本检测技术（Selective and Feature based Adversarial Detection，SFAD），其结合了最近的不确定性方法 SelectiveNet[37]并集成了 3 个检测模块。首先是选择性检测模块，它基于阈值检测，使用 SelectiveNet 通过不确定性拒绝异常输入。然后是置信度检测模块，其同样是基于阈值的检测，利用 SFAD 分类器的 Softmax 概率进行评估，检测对抗样本。SFAD 分类器通过分析最后 n 层的特征，通过自动编码、上下采样和添加噪声块来提取输入数据的鲁棒特征。最后一个模块是集成预测，其通过蒸馏的方式将多个模型的输出结合起来，以识别对抗样本。

SFAD 结合了不确定性方法 SelectiveNet，并集成了多个检测模块，包括选择性检测、置信度检测和集成预测，每个模块都有其独特的功能，可以相互补充和强化。这种多模块设计可以提高对抗样本检测的准确性。然而，模块之间的依赖关系会导致后期模块的推理速度相对较慢，因为后续模块需要等待前期模块的输出结果，这种依赖关系会导致更长的推理时间。

3.2　基于统计分布的检测方法

正常样本通常具有与对抗样本不同的分布特性。基于这一观察，可以采用多种方法来分析和利用正常输入数据的不同统计特性，从而构建检测器。其基本原理如图 6 所示，步骤如下：

1）对正常样本进行分布建模，得到正常样本分布；

2）计算待测样本与正常样本分布的分布距离，如果大于阈值，则将待测样本判定为对抗样本。

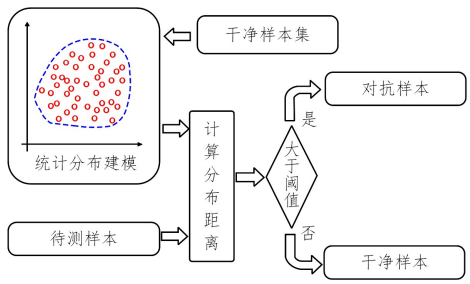


图 6　基于统计分布的检测方法的原理流程

Fig. 6　Principle flow of detection method based on statistical distribution

3.2.1 基于 Softmax 分布

Softmax 函数用于将原始分数(Logits)转化为概率分布。对于一个具有 N 个类别的分类问题,Softmax 函数表示如下:

$$P(i)=\frac{e^{z_i}}{\sum_{j=1}^{N}e^{z_i}} \tag{2}$$

其中,$P(i)$表示类别 i 的概率,z_i 表示原始分数。分母部分是对所有类别的原始分数进行指数化和求和,以确保所有概率之和等于 1,这使得 Softmax 函数的输出可以被解释为每个类别的概率。

Hendrycks 等[38]发现正常样本与对抗样本的 Softmax 分布是不同的。正常样本的 Softmax 向量通常拥有更大的最大概率,而对抗样本各类别的概率分布更均匀,这是因为模型在正确分类时更加自信,更有把握将概率分配给正确的类别,因此他们测量均匀分布与 Softmax 分布之间的 Kullback-Leibler 散度[39],然后利用该散度进行阈值检测。

尽管 Hendrycks 等的方法提供了一种独特的思路,通过比较待测样本 Softmax 分布与正常分布来检测对抗样本,计算 Kullback-Leibler 散度通常相对较快,可以在较短的时间内进行对抗样本检测,但其实际的有效性和适用性受到了多种因素的制约,包括对抗样本生成的扰动大小、攻击类型以及模型的置信度。此外,有研究人员[40]指出,这种方法更适用于如 JSMA[27]等对抗样本置信度较低的攻击算法,对于那些能够生成高置信度对抗样本的攻击算法,这种方法效果并不明显。

3.2.2 基于 PCA 特征分布

PCA 是一种常见的数据分析方法,用于降低数据的维度并找到数据中的主要特征。Liu 等[16]观察到,对抗样本后期 PCA 分量的方差明显大于正常输入。因此他们提出了一种检测器,该检测器计算样本后期 PCA 分量的方差并与预定阈值进行比较,如果方差大于阈值,则输入数据可能是对抗样本。这种方法简单有效,不需要复杂的训练或模型更新,因此易于实现和部署,并且不依赖对抗样本,实验表明其可以有效检测到 FGSM[6]等一些常见对抗攻击。

基于 PCA 的对抗样本检测方法简单有效,但需要注意阈值选择。这种方法对多种对抗攻击类型有效,但并非适用于所有情况。随着对抗攻击的不断演化,该方法需要不断改进或结合其他检测技术以应对新的威胁。

3.2.3 基于高斯混合模型

高斯混合模型(Gaussian Mixture Model, GMM)是一种统计模型,可以用来描述复杂数据的总体分布,特别是那些包含多个不同模式或群集的数据分布。Zheng 等[41]观察到,当深度神经网络分类器错误地将一个特定的类标签分配给对抗样本时,这些对抗样本的内部隐藏状态与同一类的正常样本明显不同。为了利用这一现象,他们提出了 I-defender,用于对抗样本检测。

I-defender 的核心思想是使用高斯混合模型来近似每个类别的内在隐态分布(Intrinsic Hidden State Distribution, IHSD),然后针对每个样本的预测标签,选择相应的检测器和阈值(正常样本对应的内在隐态分布)。通过比较样本的 IHSD 与阈值,可以对样本进行分类:如果 IHSD 小于阈值,则将待测样本判定为对抗样本;如果 IHSD 大于阈值,则将待测样本判定为正常样本。这种方法利用了深度神经网络内部隐藏状态的特性,可以有效提取对抗样本与正常样本的区别,提高检测性能。

I-defender 的优势在于它是一种无监督的检测方法,具有较强的可移植性,可以轻松应用于不同领域,而且可以与各种现有的防御策略相结合。然而,该方法需要为每个类别单独确定阈值,这个过程比较耗时,难以找到最佳的阈值,因此需要进行大量参数调整工作。

3.3 基于邻域对比的检测方法

基于邻域对比的对抗样本检测方法是一种利用邻域关系来识别对抗样本的技术。该方法首先寻找与待测样本在嵌入特征空间中接近的邻居,然后通过邻居的属性和特征判别对抗样本。与基于统计分布的检测方法相比,基于邻域对比的检测方法提供了一种更加局部化的检测策略,因为它关注邻域内的相似性而不是全局分布。其基本原理如图 7 所示,检测过程如下:

1)使用近邻算法计算待测样本在嵌入空间中的邻居,并统计其分布;

2)比较待测样本和邻居的类别差异,或者对比待测样本邻居与正常样本邻居的分布,如果存在较大差距,则将待测样本判定为对抗样本。

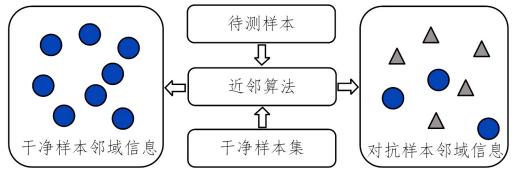


图 7 基于邻域对比的检测方法的原理流程

Fig. 7 Principle flow of detection method based on neighborhood comparison

3.3.1 基于深度 k 近邻

最近,有研究[31]发现输入样本的合法性与其在嵌入空间中的邻域信息存在显著的相关性。这为研究人员提供了一个新的思路:提取待测样本的邻域信息并整合到检测方法中,将待测样本的邻域信息与正常样本进行对比,以检测对抗样本。

深度 k 近邻(Deep k-Nearest Neighbors, DkNN)检测器[42]是应用该思路检测对抗样本的方法之一。该方法的核心思想是计算神经网络的每一层输入的近邻嵌入。这里的“嵌入”是指样本在神经网络中的特征表示。比较每个待测样本与其在嵌入空间中的最近邻居,如果存在过多邻居与待测样本的类别不同,或者邻居属于多个不同的类别,则可判定输入为对抗样本。

该方法在对抗样本检测中表现出色,特别是对于 FGSM[6]攻击算法,其检测性能非常出色,并且误检率极低。然而,深度 k 近邻方法在实际应用中仍然面临一些挑战。首先,阈值的选择问题是一个难题,因为在不同的应用场景中,最佳阈值可能会有所不同。算法设计者通常难以预先确定最适合特定情况的阈值,这可能需要在实际部署时进行精细调整。其次,该方法的性能受到中间层特征选择和模型结构的影响,这意味着它可能不具备很强的通用性。如果更换模型

或引入新的数据,可能需要重新训练模型并重新选择适合的中间层和阈值,这会增加实施和维护的复杂性。最后,深度 k 近邻检测器根据待测样本的近邻进行检测,本质上属于从数据分布角度出发的统计方法,其虽然在一定程度上能够捕获对抗样本的特征,但可解释性相对较差,难以提供深入的对抗样本分析。

3.3.2　基于邻域图对比

Abusnaina 等[43]使用了潜在邻域图(Latent Neighborhood Graph,LNG)和图神经网络(Graph Neural Networks,GNN)来解决对抗样本检测问题。在该方法中,首先为输入样本生成一个潜在邻域图,其中节点代表样本,边代表样本之间的关系;然后利用 GNN 模型从图节点的局部流形中找到用于对抗样本检测的高阶模式特征。该方法可以分析邻域图中节点之间的连接方式,以确定输入是否属于对抗样本。

这种方法的优势在于不依赖于对数据分布的具体假设,因此对各种类型的数据都具有普适性。图神经网络能有效地从邻域图节点的局部流形中找到用于对抗样本检测的高阶模式特征,能够捕获复杂的对抗样本特征,不仅限于传统的低级特征。但是,构建和处理潜在邻域图以及运行图神经网络需要大量计算资源和时间,在一些大型数据集或需要实时分析的应用中适用性较差。此外,由于不同的数据集和任务可能需要不同的设置,图神经网络的超参数通常需要进行一定的实验和调整。最后,该方法的性能高度依赖于潜在邻域图的构建和样本关系的建模,不正确的图构建或关系建模可能导致性能下降。因此,在实际应用中需要综合考虑这些优势和局限性。

3.4　基于特征变换的检测方法

对抗性扰动微小且对模型的预测结果影响极大,因此对图像进行一些处理以削弱扰动可能会导致模型输出发生变化。基于特征变换的检测方法旨在通过比较模型对变换前后样本的预测结果来判别对抗样本,其基本检测过程如图 8 所示,分为以下几步:

1)对待测样本进行特征变换,例如降噪、特征降维、图像增强等操作,得到变换样本;

2)将待测样本和变换样本分别输入深度学习模型,比较其输出的预测结果,如果结果不同,则将待测样本判定为对抗样本。

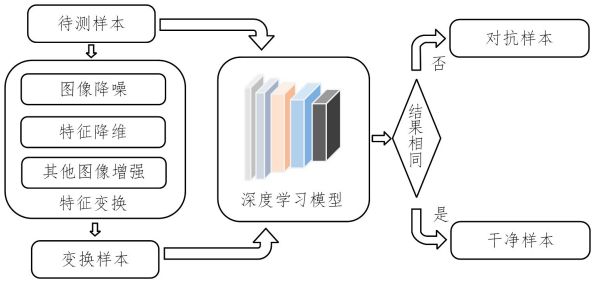


图 8　基于特征变换的检测方法的原理流程

Fig. 8　Principle flow of detection method based on feature transformation

3.4.1　基于降噪

在对抗样本中,对抗性扰动可被视为一种附加噪音。基于此,Liang 等[44]提出了一种自适应降噪方法,结合标量量化和平滑空间滤波来消除对抗性扰动,通过比较输入图像 x 和经过降噪处理的图像 x' 的分类结果来判断 x 是否为对抗样本。如果分类发生变化,即 $f(x)\neq f(x')$,那么 x 被标记为对抗样本。然而,在处理良性图像时,直接降噪可能损害其有用信息,从而造成降噪后的误分类,导致高的误检率。为了解决这一问题,Wang 等[45]采取了一种策略,首先在图像 x 上添加一些额外的高斯噪音,然后进行去噪,以减轻对有效信息的破坏。

这种方法具有较高的普适性,因为它不依赖于特定的数据分布或假设,因此适用于各种不同类型的数据。但它存在一些显著的限制。其中之一是需要仔细选择适当的高斯噪音水平和去噪方法,以确保方法的有效性,这需要进行大量的调试和参数设置,因为不同的数据集和任务可能需要不同的配置。此外,该方法的性能高度依赖于高效的去噪技术,如果去噪技术不够先进或不适用于处理特定类型的对抗性扰动,那么其检测性能可能会下降。

3.4.2　基于特征压缩

Xu 等[33]观察到图像特征的输入空间通常包含许多对于模型预测不必要的特征,这些特征的存在提高了对抗样本攻击的自由度。为了应对这一问题,他们提出了一种新策略,并称之为特征压缩,通过"压缩"那些不必要的输入特征,减少了对抗样本攻击可能利用的特征。接着,模型对原始输入和经过特征压缩处理后的输入进行评估。如果模型对压缩后输入的预测与其对原始输入的预测之间的差异超过了预先设定的阈值,那么该输入就被识别为对抗性。这一策略有助于提高模型的鲁棒性,降低对抗样本攻击的成功率。

特征压缩策略是一种创新思路,旨在减少深度学习模型所用的特征维度,无需显著改变模型结构,就可以降低对抗样本攻击的风险。该方法的性能高度依赖于特征压缩算法的有效特征保留能力以及差异阈值的设置,如果特征压缩算法或阈值不合适,则可能导致误报或漏报。因此,需要谨慎选择特征压缩算法和阈值,以实现最佳性能。

3.4.3　基于特征映射

Drenkow 等[18]提出了一种技术,利用随机映射特征的不一致性来体现不同子空间集合中正常样本和对抗样本的区别。首先通过随机投影将图像特征降维,并映射到一系列随机子空间中;然后比较这些特征映射与类原型之间的一致性(在每个子空间中)。

该方法的随机特征映射仅需要在正常样本上进行训练,因此被归类于无监督检测,且不对攻击策略或目标做出假设,具有真正的不可知性。随机投影尽管是一种计算高效的降维方法,但在高维空间中进行计算仍较为复杂,特别是在大规模数据集上。

4　监督检测

监督检测方法的核心思想是设计和训练检测器。不同于无监督方法,监督检测方法需要使用已知标签的对抗样本或者其他形式的先验知识,其优势在于,当有足够的带标签对抗样本可用时,它们可以实现相对较优的检测性能。但是对于

新的、未知的对抗攻击，需要重新收集并标记大量的对抗样本，这是一项昂贵且耗时的任务。此外，监督方法也存在过拟合的风险，因为它们可能会在模型上过度优化已知的攻击样本。

本章将介绍两种主要的监督检测方法，即基于网络特征和基于统计分布的方法，对它们的原理进行介绍，并分析其优点和限制，以更好地理解它们在对抗样本检测中的作用。

4.1 基于网络特征的检测方法

对于正常样本和对抗样本，它们在模型特征上通常表现出不同的模式。因此，可以通过分析输入图像特征或模型中间特征来检测对抗样本。该类方法利用模型特征的差异性来进行对抗样本检测，如输入图像及相关特征、激活函数、梯度信息等。

4.1.1 基于自然场景统计特征

自然场景统计(Natural Scene Statistics，NSS)在图像处理领域中被广泛应用于图像质量估计。有研究[46]证明，自然图像的统计特性与经过处理的图像不同。Kherchouche 等[47]遵循了这一假设，构建了一个二元分类器，该分类器以一些特征参数作为输入，这些参数是由正常图像和对抗样本的均值减去对比度归一化的 MSCN 系数[48]计算得出的。这些系数具有广义高斯分布(Generalized Gaussian Distribution，GGD)和非对称广义分布(Asymmetric Generalized Gaussian Distribution，AGGD)的统计特性。具体来说，该分类器使用了一些特征参数，这些参数是从图像的亮度差异数据中计算得出的，表示图像的亮度分布情况。然后，将这些特征参数输入到一个二元分类器中，分类器将判断输入图像是正常样本还是对抗样本。

自然场景统计方法在某些情境下能够有效检测对抗样本，但它们存在一些限制。首先，它们依赖于已知的攻击样本和特定领域的特征，在面对新的对抗攻击时，需要耗费相当多的时间和资源来收集足够的带标签数据用于训练分类器。其次，这些方法在设计检测器时通常是特定于某一领域或模型的，因为不同领域的图像特征和统计特性可能存在差异。因此，需要更多的研究来提高该方法的通用性和鲁棒性，使其更好地应对不断演进的对抗攻击。

4.1.2 基于对抗样本的分类器

Gong 等[49]训练了一个二元分类器，将正常图像和对抗性图像作为输入，训练该分类器，它与基线分类器相互独立，因此不会对基线模型的性能产生负面影响。Hosseini 等[50]重新训练了基线分类器，并添加了一个新的类别，即对抗样本类别，用于拒绝对抗样本。这些方法侧重于构建对抗样本的分类器，采用不同的策略来区分正常样本和对抗样本，从而提高检测性能。

该方法的优势在于通过引入独立的检测器，确保了对抗样本的有效检测，同时避免了对基线模型性能的不良影响。但是，对基线分类器的重新训练和引入新的类别可能需要额外的计算资源和时间。此外，其检测性能是否能够在更广泛的数据集和实际应用场景中得到验证，仍然需要进一步的实证研究。

4.1.3 基于梯度的分类器

Lust 等[51]提出了一种名为 GraN 的检测器。首先，对输入进行平滑处理，使得对梯度的计算更加稳定，在每一层，GraN 方法计算了平滑输入相对于基线分类器预测类别的梯度范数。然后，使用这些梯度范数训练一个二值分类器用于检测对抗样本。GraN 方法通过分析梯度信息来检测对抗样本，其高效性是一个显著优势。然而，它需要对输入进行平滑处理，并且在一定程度上受到基线模型的影响。这种方法的创新点在于充分考虑了梯度信息，使得对抗样本检测更加准确和高效。

通过分析梯度信息，该类方法可以高效检测对抗样本，这种高效性使其在实际应用中具有可行性。然而，GraN 方法也存在一些限制。首先，对输入进行平滑处理会增加预处理的复杂性。其次，因为它依赖于基线模型的预测类别和梯度信息，所以会在一定程度上受到基线模型的影响。

4.1.4 基于影响函数

影响函数(Influence Function)可以用于衡量样本对模型参数的影响程度，也就是样本的重要性。通过计算影响函数，可以评估每个训练样本对模型预测结果的贡献，从而找出对验证集数据影响最大的训练样本，这其中包括对训练有利和有害的样本。影响函数与最近邻模型相结合，构成了一种新的策略来检测对抗样本。正常输入的 k-最近邻训练样本(在嵌入空间中最接近的邻居)和通过影响函数找到的最有帮助的训练样本应该是相关的。然而，对于对抗样本，这种相关性相对较弱，因此 Cohen 等[52]使用有帮助的训练样本和 k-最近邻训练样本在嵌入空间中的距离以及它们与输入样本的 L2 距离训练了一个对抗样本检测器，用于检测对抗样本。

该方法将通过影响函数得到的有助样本和 k-最近邻样本进行比较，根据其相关性检测对抗样本，能够同时关注模型行为和样本特征。然而，影响函数及相关样本之间的距离计算复杂度较高，需要大量的计算资源，且效果受所使用的模型影响，不同的模型可能会导致不同的检测结果。此外，对抗样本可能会操纵影响函数的计算结果，从而降低了影响函数对模型行为的准确估计。

4.2 基于统计分布的检测方法

正常样本通常是从真实世界中收集的数据，因此它们的统计属性反映了自然界的规律。这意味着正常样本的特征和标签之间存在一定的内在关系，这些关系在数据分布中得到体现。对抗样本是有意制作的，目的是欺骗机器学习模型，使其产生错误的预测，这会改变样本的统计分布，因此可以使用各种统计方法来区分对抗样本和正常样本。例如，比较样本的特征、分布、距离度量或其他统计属性。这些方法依赖于对数据分布的统计分析，旨在捕获对抗样本与正常样本之间的分布差异。

4.2.1 基于分布多样性

Ma 等[53]发现对抗样本的局部固有维数(LID)不同于正常数据。因此，根据样本与其相邻样本之间的距离分布，可以很容易地将对抗样本与正常样本区分开来。Lorenz 等[54]评估了基于 LID 分布的对抗样本检测方法，发现 LID 分布具有相当长的尾部，在处理分布的尾部时，检测任务较为困难，

因此他们将聚合的 LID 估计展开,分别考虑样本与其邻居之间的归一化对数距离,形成一个特征向量。该特征向量被称为 multiLID,可以更好地捕捉样本之间的特征相似性,提高对抗样本和正常样本的区分度。此方法简单且轻量级,但在一些可绕过攻击方法上的有效性未得到验证。

此外,Lee 等[55]在高斯判别分析的基础上,尝试利用输入图像在模型特征空间上的概率密度来区分对抗样本和正常样本。上述方法的计算时间复杂度较低,能够有效识别大多数现有攻击算法生成的对抗样本。然而,当应对如 C&W[8] 等扰动较小的攻击方法时,其性能将受到限制。为了解决这个问题,Chen 等[56]的研究发现对抗性扰动和预测置信度之间存在一致性,这对于检测对抗样本非常有用,特别是对于轻微的扰动。他们开发了一个对抗样本检测框架,即像素伪像和置信度伪像(Pixel Artifacts and Confidence Artifacts,PACA)双流检测框架。利用该框架,可以同时从像素特征和置信度特征中提取有用信息用于检测对抗样本。尽管 PACA 的性能优越,但其难以处理对预测置信度不敏感的场景。

4.2.2　基于最大平均差异

Grosse 等[57]采用了一种统计测试方法,称为最大均值差异(Maximum Mean Discrepancy,MMD),用于区分对抗样本和正常样本。这是一种基于核的双样本检验方法,不影响目标模型。首先,检测器 D 计算了正常样本 x 和待测样本 x'之间的最大平均差异,用符号 a 表示,即 $a=MMD(x,x')$。MMD 是一种用于测量两个样本集之间相似性的统计方法,可以检测不同样本是否来自相同的分布。其次,将 x 和 x'的元素重新排列成两个新的集合 y_1 和 y_2。然后,计算 y_1 和 y_2 之间的 MMD,用符号 b 表示,即 $b=MMD(y_1,y_2)$。如果 $a<b$,表示 x 和 x'来自不同的分布,从而认定输入数据 x'可能是对抗样本。然而,仅使用 MMD 进行对抗样本检测忽略了一些关键因素,如高斯核的有限表示能力以及对抗样本的非独立性。Gao 等[58]提出了一个简单而有效的改进,将 MMD 与语义感知内核结合(SAMMD),这样可以有效提高对抗样本的检测精度。

MMD 方法采用基于核的统计检验,这使得它对于数据分布的敏感性相对较低,有助于提高方法的泛化性能,使其在不同数据集上更具鲁棒性。然而,需要注意的是,MMD 的计算涉及核方法和对整个数据集的操作,因此在大规模数据集上的计算成本较高,并且需要谨慎选择 MMD 的核函数以适应不同的数据分布,这可能会影响方法的实际可行性。

4.2.3　基于核密度估计

Feinman 等[59]的研究发现对抗样本的子空间通常比正常样本的密度低,特别是在输入样本远离类别流形(Class Manifold)时。基于这一特性,他们提出了针对训练数据中每个类别的核密度估计(Kernel Density Estimation,KDE)。对抗样本通常会位于数据分布的稀疏区域,而不是类别的主要区域(类别流形),这与正常样本存在区别。因此,他们将这些密度估计值作为训练数据训练一个二元分类器 D,该分类器可以有效识别对抗样本。

这种方法的优点在于它可以有效地利用数据的密度信息来进行检测,从而更好地区分对抗样本和正常样本。然而,它的性能也受到核函数的选择和参数的设置,以及训练数据的质量和数量等因素的影响。因此,在应用时需要仔细选择和调整参数,以适应特定的数据集和问题。

4.2.4　基于傅里叶变换

利用傅里叶频谱来提取人眼无法感知的特征已经被证明是成功的,例如,检测 Deepfakes[60-61]。因此,Harder 等[17]在输入图像和特征图的傅里叶域中进行分析,以区分正常样本和对抗样本。首先,对输入样本和特征图进行傅里叶变换,得到频谱特征;然后,分别提取频谱特征的幅度部分和相位部分,训练分类器以检测对抗样本。Lorenz 等[62]在该项工作上进行补充,提出了两种检测算法:黑盒检测和白盒检测。在黑盒设置下,提取并连接每个颜色通道的傅里叶功率谱作为输入图像的特征表示,训练分类器以检测对抗样本。白盒设置下,在黑盒设置所用特征的基础上将目标网络的特征映射加入到输入特征来训练分类器,以捕获更多特征。

该类方法可以有效检测 AutoAttack[63] 方法产生的对抗样本且具有较强的泛化性,但是需要对输入图像和特征映射进行额外的傅里叶变换,增加了计算复杂度。

4.2.5　基于 k-近邻

Neighbor Context Encoder(NCE)[64] 利用 k 个最近邻来帮助检测样本是正常还是对抗的。首先,对于一个待测样本,NCE 会在特征空间中搜索其 k 个最近邻。每个邻近样本被视为一个 token,根据它们到待测样本的距离降序排列并连接,形成一个长度为 k 的序列。被检测的样本被添加在序列的起始位置作为一个特殊的分类 token。接着,对序列进行编码,从邻近点的拓扑结构中学习特征,并推断被检测样本是正常还是对抗的。

NCE 能够利用被检测样本周围的邻近样本信息,从而更准确地判断样本的性质。但 NCE 的性能受到样本空间的影响,如果样本的分布不均匀或存在离群点,其效果可能受到影响。

4.3　无监督与监督方法对比

在对抗样本检测方法中,无监督方法和监督方法各具优势,适用于不同的攻击方法和使用场景。为了更全面地理解对抗样本检测方法的性能和适用性,本文对无监督和监督检测方法进行了全面对比,包括数据需求、检测精度、模型需求、检测速度、通用性和鲁棒性等指标的对比,部分指标在第 2 章的相关定义中有介绍。对比结果如表 3 所列。

表 3　无监督和监督检测方法的全面对比

Table 3　Comprehensive comparison of unsupervised and supervised detection methods

指标	无监督检测	监督检测
数据需求	仅需要正常样本	需要带标签正常样本和对抗样本
检测精度	相对较低	对于特定类型对抗样本精度高
模型需求	部分需要额外模型	需要额外模型
检测速度	快	较慢
通用性	有一定通用性,适用于多种领域或模型	通常特定于某一领域或模型
鲁棒性	具有较高的鲁棒性	在某些情况下可能不够鲁棒

总体来看,无监督方法通常在通用性和鲁棒性上具有一定优势,而监督方法则在特定类型对抗样本的检测上表现更出色。

5 融合跨领域先进技术的检测方法

除了前面介绍的检测方法，一些研究试图结合其他领域的先进技术，例如生成模型、语义分割等，以提高检测的效率和准确性。探讨这些方法，可以为对抗样本检测领域提供新的思路。本章将详细介绍一些融合其他领域先进技术的检测方法，探讨它们的原理和应用，以及它们在增强对抗样本检测性能方面的潜力。

5.1 模型显著图

早期的对抗样本检测方法通常假设对抗数据集和原始数据集在本质上是不同的，因此设计并训练分类器来实现对抗样本的检测。然而，Zhang 等[65]认为这些检测方法容易被一些攻击方法绕过。因此，他们从模型的角度出发，试图找到对模型分类影响较大的图像区域，并使用显著图来展示这些区域，然后使用原始数据和显著性数据训练一个二分类器来检测对抗样本。显著图体现了模型做出决策所需的各个像素的重要性。最初的显著性水平通过式(3)来计算：

$$\theta=\frac{\partial f(x)}{\partial x} \tag{3}$$

其中，x 为输入图像，$f()$为模型。因为显著性水平体现了模型对图像不同区域的关注程度，这种方法不仅局限于模型或者图像，因此具有较好的泛化性和可转移性，在面对多种迭代性攻击[8,27]时表现出良好的检测效果。

除此之外，Prakash 等[66]通过 Grad-CAM(Gradient-weighted Class Activation Mapping)[67]展示了模型对正常样本和对抗样本关注区域的不同，如图 9 所示。基于此，研究人员开始研究如何利用显著图的不同之处来检测对抗样本。Wang 等[68]采用了类判别的可解释性方法，自动获取在模型决策中相对重要的特征，通过分析结果发现对抗样本和正常样本在模型决策方面逐渐显现出差异，尤其是在模型的隐层中。为了全面综合这些显著性特征，他们提出了一种基于多层显著性特征的对抗样本检测方法，该方法通过比较正常样本和对抗样本在模型不同层次上的显著性特征来进行检测。

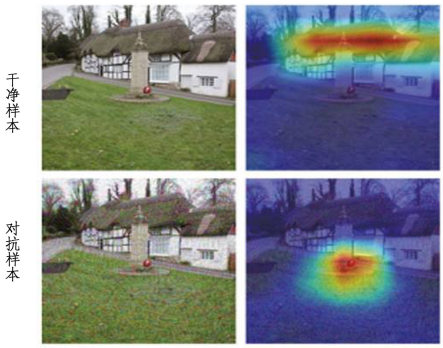


图 9 对抗样本和正常样本显著图对比

Fig. 9 Comparison of saliency maps between adversarial samples and normal samples

这些研究表明了可解释性技术和对抗样本检测之间的紧密联系，为对抗样本检测方法提供了独特的思路。

5.2 生成模型

生成模型能够生成与训练数据类似的数据样本。例如自回归模型(PixelCNN)[69]，生成模型的目标是捕获数据的内在结构和规律，使其能够生成具有相似统计特性的新数据。因此有研究方法将输入数据与生成模型生成的数据进行比较，如果输入数据在某种程度上偏离了正常数据的分布，就可能被识别为对抗样本。PixelDefend[70]使用 PixelCNN 对正常的训练数据进行了重构，使其接近真实图像。然后，计算每幅图像在训练数据分布中的概率，重构后的图像在训练数据的分布中有更高的概率。接着，计算待测输入在训练数据分布中的概率密度，并与训练数据的概率密度进行比较，如果待测输入的概率较低，则其被认为是对抗样本。

Kiani 等[71]将对抗样本中的特征分为两种：对应于真实标签的特征和对应于误分类标签的扰动特征。这两种特征具有较强的不一致性，为此他们提出了一种方法，首先获得目标模型对待测图像的预测标签，然后使用生成模型根据该标签生成图像，如果标签是对抗性的，则生成的图像将与待测图像显著不同。Wang 等[72]通过辅助分类生成器(Auxiliary Classifier GAN，AC-GAN)对预测类标签下的干净数据进行建模。给定一个测试样本及其预测类，基于 AC-GAN 生成器和 SVM 计算 3 个检测统计量用于检测对抗样本。

5.3 语义分割

Freitas 等[73]提出了一种独特的方法用于检测对抗样本，称之为 UnMask。他们认为分类模型的脆弱性是由一些非鲁棒性特征引起的。UnMask 方法首先从图像样本(比如一张可能包含“鸟”的图像)中提取一些鲁棒性特征(如喙、翅膀、眼睛等)，然后将这些特征与模型预期的分类特征进行比较。例如，从一张被模型分类为“鸟”的图像中提取出的鲁棒性特征包括车轮、踏板和车架等与鸟类无关的特征，那么 UnMask 会判断这个样本很可能是一个对抗样本。这种方法利用了鲁棒性特征的不匹配来检测潜在的对抗样本，具有较高的可解释性。

然而这种检测方法依赖于鲁棒性特征提取模型，如果提取方法存在误差或者对于不同数据集的提取表现不佳，可能会导致误判或漏检。对此，Gong 等[74]使用语义分割模型替换鲁棒性特征提取模型，引入了一种名为 SeMatch 的对抗样本检测方法。其核心思想是通过构建输入图像语义图并将其与预测类别语义图原型进行比对，关注语义图中的冲突，并以此检测对抗样本，原理如图 10 所示。首先将图像输入到语义分割模型和目标分类模型，得到语义特征和预测类别；然后比较图像语义特征和预测类别的语义特征，如果差异过大，则将待测样本判为对抗样本。

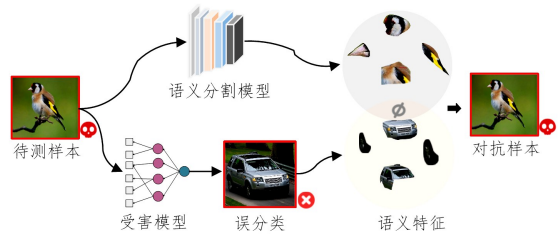


图 10 结合语义分割的对抗样本检测原理

Fig. 10 Principle of adversarial sample detection combined with semantic segmentation

SeMatch 方法的一大优势在于它结合了较为成熟的语义分割技术，可以检测各种攻击方法产生的对抗样本。此外，该方法具有良好的可解释性。通过对语义图的部分分割结果进行分析，可以直观地了解模型作出决策的依据，并根据决策的合理性识别对抗样本。

5.4　模型路径

文献[75]将深度学习模型内部的神经元分成两部分：一部分是与人类可察觉的图像属性强相关的“见证神经元”，另一部分是与人类不可察觉的图像属性强相关的“非见证神经元”。通常情况下，对抗样本更容易激活“非见证神经元”，因为这些神经元更加关注人类不可察觉的图像属性，且同样会对模型的输出产生重要影响。基于这个观察，Qiu 等[76]提出了一种方法，通过提取一组关键神经元，形成一组“有效路径”。研究人员发现，模型在不同类别的图像上做出决策时，会依赖于不同的有效路径，如果输入样本的有效路径与其预测类别对应的路径相差较大，则该样本很可能为对抗样本。Nwaigwe 等[77]受到图论视角启发，在样本输入到目标模型后，使用了层次相关传播算法(Layer-wise Relevance Propagation)为目标模型的每个神经元分配一个量，这些量可以解释为神经元对输出的影响大小。基于这些神经元的量和连接方式构建一个稀疏图，并从中提取了 3 个特征量进行比较：节点的度、Wasserstein Sums Ratio(WSR)和最后两层节点的邻接矩阵。通过比较这 3 个特征量来识别对抗样本。

该方法只使用了来自训练数据和输入图像本身的信息，不需要关于具体攻击方法的先验知识，除了用于对抗样本的防御外，还有助于更好地理解深度神经网络的内部工作原理。这种方法为深度学习模型的可解释性和对抗性防御提供了新的途径。

6　数据集与评估

6.1　常用数据集

本节介绍在对抗样本检测领域常用的几个数据集，它们在计算机视觉和深度学习模型研究中起到了重要的作用。这些数据集不仅广泛用于测试机器学习和深度学习模型的性能，还在对抗样本的生成和检测方面提供了有力的支持，主要包括 MNIST[78]，CIFAR-10[79]，ImageNet[80]，Tiny ImageNet[81]和 SVHN[82]这几个代表性的数据集。

6.1.1　MNIST

MNIST(Modified National Institute of Standards and Technology database)[78]是一个经典的数据集，常用于测试机器学习和深度学习模型。该数据集包含 0－9 的手写数字的灰度图像，每个数字由 28×28 像素的图像组成。MNIST 总共包含 60 000 张用于训练的图像和 10 000 张用于测试的图像。

最初，MNIST 被设计用于测试和验证机器学习算法。它通常用于入门级的图像分类问题，旨在帮助研究人员和学生迅速熟悉图像处理和分类任务。因为 MNIST 图像相对较小且清晰，因此大多数标准的深度学习模型可以在 MNIST 上实现高精度的分类性能。这也使得对抗样本的生成和检测在 MNIST 上更加容易，实验和测试可以在较短的时间内完成，几乎所有涉及图像分类的攻击、防御、检测方法都会使用该数据集进行验证。

然而，正因为 MNIST 图像相对简单，一些对抗样本检测算法在这个数据集上表现良好，但在处理更复杂的图像数据集时可能失效。在实际应用中，图像数据通常更加复杂和多样化，与 MNIST 相比有很大不同。因此，在评估算法性能时，需要考虑更具挑战性的数据集和更复杂的应用场景。

6.1.2　CIFAR-10

CIFAR-10(Canadian Institute For Advanced Research-10)[79]是另一个常用于测试机器学习和深度学习模型的重要数据集。它包含了 10 个不同类别的物体图像，每个类别有 6 000 张图像，总计包含了 60 000 张图像。这些图像以 32×32 像素的大小呈现，具有彩色通道。

与 MNIST 不同，CIFAR-10 数据集更具挑战性，因为它包含了更复杂的图像，涵盖了 10 个不同种类的物体，分别是飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车。每个类别的图像都相对多样化，并且存在不同的视角、光照条件和背景。

正因为 CIFAR-10 的图像更具多样性和复杂性，它更适用于评估对抗样本检测算法在实际场景中的性能。与 MNIST 相比，CIFAR-10 更接近真实世界中的图像数据，因此在这个数据集上测试算法可以得到更有挑战性的结果。

然而，与挑战性相伴随的是复杂性。由于图像更大且包含多个通道，因此处理 CIFAR-10 数据集需要更多的计算资源和模型复杂性。这也使得对抗样本的生成和检测在 CIFAR-10 上变得更加复杂。因此，在选择适当的数据集时，需要根据特定应用的需求和算法的性质来权衡使用 MNIST 或 CIFAR-10 等数据集。

6.1.3　ImageNet

ImageNet(Image Database for Object Recognition)[80]是一个大规模的图像数据库，有 ImageNet-1K 和 ImageNet-21K 两个主要版本，它们之间的主要区别在于类别数量和图像数量。ImageNet-1K 包含 1 000 个类别；ImageNet-21K 更加庞大，包含了约 21 000 个不同的类别。比较常用的 ImageNet-1K 即 ILSVRC2012 数据集，它包含了 1 000 个不同的图像分类类别，每个类别约有 1 000 张图片用于训练。总计约有 120 万张用于训练的图片，以及 5 万张验证集和 10 万张测试集图片。

ImageNet 的主要目标是促进计算机视觉和模式识别研究的发展，特别是对大规模图像数据进行分类和对象识别的任务。正因为 ImageNet 的规模大且具有多样性，它被广泛用于深度学习模型的训练和评估中。许多最先进的计算机视觉模型都是在 ImageNet 上进行训练的，并且这些模型在各种视觉任务中表现出色，包括图像分类、对象检测和图像分割等。

然而，ImageNet 的数据量过于庞大，对抗样本的生成和检测在 ImageNet 上变得更加复杂。因此，在使用它时需要谨慎考虑计算资源和算法复杂性。

6.1.4　Tiny ImageNet

Tiny ImageNet[81]是一个基于 ImageNet 的小型图像

数据集，通常用于计算机视觉和深度学习研究。它是 ImageNet 的一个精简版本，包含 200 类，每个类有 500 张训练图片、50 张验证图片和 50 张测试图片，旨在提供一个更加轻量级的数据集，以便研究人员和开发者可以更容易地进行实验和原型设计。

相比 ImageNet，Tiny ImageNet 规模较小，处理和训练模型相对来说更加快速和容易，这对于对抗样本检测算法的快速原型设计和实验非常有利。同时，其保留了数据的多样性分布特征，使得在该样本上进行的检测结果具有较高的代表性。

6.1.5 SVHN

SVHN(The Street View House Numbers)[82]是一个真实世界的图像数据集，来源于谷歌街景图片中的门牌号，用于解决现实世界问题。数据集中的每张图片是带有字符级边界框的彩色门牌号图像，类似于 MNIST 的 32×32 以单个字符为中心的图像，总共包含 0－9 这 10 个类别。与 MNIST 的单通道灰度图像不同，SVHN 包含彩色图像，每个图像都有3个通道(红、绿、蓝)，并且图像是从 Google 街景图像中提取的，因此它们包含了来自真实世界的各种噪音，例如模糊、光照变化、角度变化等。它的训练集包含 73257 个样本，测试集包含 26032 个样本，另有 531131 个难度稍低的样本用作额外的训练数据。

表 4 中介绍了上述数据集和其他一些使用较少但仍具有一定代表性的数据集。

表 4 检测性能评估数据集

Table 4 Detection performance evaluation datasets

名称	用途	数据量	类别	来源
Flower[83]	花分类	8189 张图片	102 类产自英国的花卉	https://download.csdn.net/my
Caltech101[84]	物体分类	每个类别约有 40～800 张图片	101 个类别的物体图片	https://data.caltech.edu/records/mzrjq-6wc02
Caltech256[85]	物体分类	30607 张图片	256 个物体类别	https://data.caltech.edu/records/nyy15-4j048
PubFig83[86]	人脸识别	8300 张图片	83 位公众人物每人 100 张图像	http://vision.seas.harvard.edu/pubfig83/
Dogs vs. Cats[87]	猫狗分类	25000 张图片	猫、狗	www.kaggle.com/c/dogs-vs-cats/data
Cardiac Arrhythmia Database[88]	区分是否存在心律失常，并分类为16 种状态	452 条数据，每条数据 279 个属性	01 级指正常心电图，02－15 级指不同等级的心律失常，16 级指其余未分类的心律失常	https://archive.ics.uci.edu/ml/datasets/Arrhythmia
Wine[89]	酒分类	178 条数据，每条数据 13 个属性，代表葡萄酒中 13 种成分的含量	意大利同一地区葡萄所酿的 3 种不同葡萄酒	https://archive.ics.uci.edu/dataset/109/wine
fashion-mnist[90]	服装商品种类识别	60000/10000 的训练/测试数据划分，28×28 的灰度图片	10 种不同商品的正面图片	https://github.com/zalandoresearch/fashion-mnist

6.2 检测器评价指标

检测率：用于衡量检测器的准确性，即检测器成功找到的对抗样本数量与所有实际对抗样本的数量之比。数值越大，表示检测器性能越好。

误报率：检测器将无害的清洁样本错误地识别为对抗样本的程度，即清洁样本被误判为对抗样本的数量与总的清洁样本数量之比。数值越小，表示检测器的误报率越低，性能越好。

复杂性：训练检测器所需的时间和计算资源。在实际应用中，复杂性是一个重要因素，因为在有些情况下，训练一个复杂的检测器可能不切实际。

额外开销：检测器的架构和部署所需的额外参数大小。较小的额外开销适合在资源有限的环境中部署，如移动设备。

检测时间：检测器识别输入是否为对抗性所需的响应时间。较短的检测时间适用于需要实时响应的应用。这些指标有助于更好地理解不同检测方法的性能和适用性。

通用性：检测器是否具有适用于多种不同数据集和模型生成的对抗样本的能力。

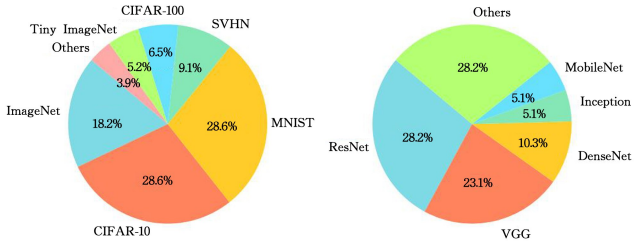
可解释性：对检测器输出结果的理解和解释的难易程度。

鲁棒性：检测器在面对对抗样本攻击者明确知晓检测器存在且攻击者特意添加了绕过检测器的机制时的表现。

6.3 实验配置统计

对抗样本的多样性意味着它们的特性在不同模型、方法和数据集下都可能不同。因此，在评估检测器性能时，需要使用具有代表性的对抗样本。为此，对本文引用的 32 项研究的实验配置进行了调查统计，主要包括数据集和模型使用占比情况，以及测试检测性能所用的攻击方法统计。

数据集和模型占比的统计结果如图 11 所示。一项研究可能会使用多个数据集以及目标模型来测试其检测方法。由统计结果可知，数据集使用最多的是 MINIST，ImageaNet 和 CIFAR-10；模型使用最多的是 ResNet 和 VGG，以及一些自行设计的模型(统一归类为“Others”)。



(a)数据集占比情况 (b)模型占比情况

图 11 评估检测器所用数据集和模型占比情况

Fig. 11 Proportion of datasets and models used to evaluate the detector

测试检测器所用的攻击方法统计结果如图 12 所示，每项研究可能使用多种攻击方法产生的对抗样本测试检测器性能和泛用性。

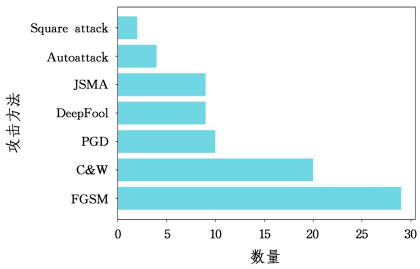


图 12　检测器评估中所用攻击方法数量的统计

Fig. 12　Statistics of the number of attack methods used in detector evaluation

7　未探索的检测方法

人工智能及相关领域发展迅速，涌现出许多新技术，其中有些在对抗样本检测领域非常有潜力。这些方法或技术尚未广泛应用，但它们可以为对抗样本检测带来新的可能性。

本章将探讨对抗样本检测领域中的一些前沿，即尚未在相关论文中广泛讨论的技术或原理。虽然这些技术或者方法尚未被广泛应用于对抗样本检测，但它们具有潜在的应用前景和引人瞩目的性能特征，有望推动对抗样本检测领域的创新发展。

7.1　非鲁棒性特征

Ilyas 等[91]的最新研究提出了一个引人深思的观点：对抗样本并不是机器学习中的错误，而是一种有价值的特征。该研究将对抗样本的出现解释为标准机器学习数据集中存在具有高度预测性但缺乏鲁棒性的特征。他们的研究通过实验将数据集中的特征分为两类：鲁棒特征和非鲁棒特征。鲁棒特征指对模型的输入变化不敏感的特征，而非鲁棒特征则容易受到输入的微小改变的影响而产生显著变化。

正如前文所提到的，对抗性扰动通常会作用于非鲁棒性特征，这为对抗样本检测提供了一个有潜力的方向。如果能够有效地提取出图像中的非鲁棒性特征，就有可能更好地检测对抗样本。但其可行性和性能还需要更深入的研究来验证。

7.2　图像特征分解

人脸伪造检测是一种重要的计算机视觉任务，旨在检测处理后的图像中是否存在伪造的人脸，通常包括合成、修改或替换真实人脸的图像。与对抗样本检测相似，两者都涉及检测输入数据是否经过操纵或变换，以欺骗机器学习模型，因此在人脸伪造检测领域有许多值得借鉴的方法。

例如，有研究[92]将人脸图像看作是三维几何和光照环境干预的产物，将人脸图像分解为三维形状、普通纹理、身份纹理、环境光和直射光，且发现直射光和身份纹理中存在伪造线索。基于这些伪造线索可以有效地检测人脸伪造，这类人脸伪造检测方法为对抗样本检测提供了有价值的启发。在对抗样本攻击中，攻击者可能会通过修改图像的颜色、纹理或形状等特征来欺骗模型，因此对这些特征进行分解并进行分析可能有助于检测对抗样本。最近一项研究[93]提出了一种可以将图像隐式地分解成形状、纹理和颜色特征的方法，通过对这些分解得到的特征进行分析，可以更有效地检测对抗样本，提高模型的鲁棒性。该方法为对抗样本检测提供了一种新颖而有前景的研究方向。

7.3　距离相关性

部分距离相关性（Partial Distance Correlation）用于测量两个变量或不同维度特征空间的相关性，其不仅在统计学中应用广泛，在计算机视觉和对抗样本防御中也很有应用潜力。Zhen 等[94]探讨了部分距离相关性在计算机视觉领域的一些应用，其中包括训练多个神经网络，并通过距离相关性限制它们在损失函数中的表现，以确保这些网络学习到不同的特征。这种方法旨在增强不同子网络的独立性，从而提高模型的鲁棒性。距离相关性首先计算分布中的不同样本间的距离，得到距离矩阵；然后通过计算距离矩阵的相似性来衡量两个分布的相关性。

部分距离相关性在对抗样本检测中也具有潜在应用前景。例如，它可以用来描述一个特征分布的性质，对于一个给定类别的正常数据分布，在引入对抗样本后，这一分布可能会发生显著的变化，这些变化可以通过距离相关性来量化和反映，因为距离相关性能够衡量分布之间的相似性或差异性。与基于统计的无监督方法相比，距离相关性更能体现分布的相对变化，更适合用来描述分布的特性，因此，它可能成为对抗样本检测中一个强大的工具，用于捕获对抗样本引入的分布变化，为对抗样本检测领域提供新的视角和方法。

7.4　对比学习

对比学习是一种机器学习方法，旨在通过比较数据样本之间的相似性来学习有意义的表示。该方法通过对比样本，学习到更加鲁棒和通用的表示。例如，OpenAI 开发的一种深度学习模型 CLIP（Contrastive Language-Image Pretraining）[95]可以同时理解文本和图像。它采用对比学习的方法，将图像和文本嵌入映射到相同的空间，使得 CLIP 可以将文本和图像结合在一起。其具有广泛的应用前景，不仅可用于解决各种涉及文本和图像的任务，而且为对抗样本检测方法的研究提供了新的视角。CLIP 的原理如图 13 所示，首先使用文本图像编码器得到文本和图像特征，然后计算文本图像的相似度，并且在训练过程中最大化相匹配的图像文本的相似度，最小化不匹配的图像文本的相似度。

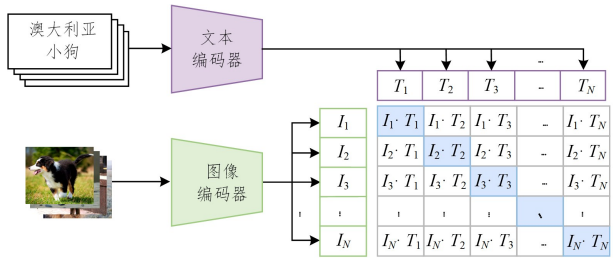


图 13　CLIP 原理示意图

Fig. 13　Schematic diagram of CLIP principle

CLIP 这一原理可以用于检测对抗样本，通过将图像和标签信息进行对比，来判断图像是否为对抗样本，具体步骤如下。

1）特征提取：首先，使用卷积神经网络（CNN）或其他

深度学习模型对输入图像进行特征提取。这将生成包含图像的低级特征和高级特征的向量。

2)标签信息嵌入:从数据集中的标签信息中提取特征,这些特征可以表示类别、语义信息等。可以使用一种编码方法将标签信息转化为固定长度的向量表示。

3)相似性比较:通过计算图像特征和标签信息特征之间的相似性分数,如余弦相似度或欧氏距离,来衡量它们之间的相似程度。

4)阈值判定:设定一个相似性阈值,如果图像特征和标签信息特征的相似性得分低于此阈值,则该图像可能是对抗样本。

这种方法与结合语义分割的检测方法[73-74]类似。结合语义分割的检测依赖于对图像及其语义分割特征的对比,需要提取复杂的语义分割特征;而这种方法则更加高效,因为它依赖于图像和标签信息的对比,由于提取标签信息的特征通常比提取复杂的语义分割特征更加高效,因此计算成本更低。然而,这种方法仍然需要考虑标签信息的准确性和数据质量,因为不准确的标签信息可能导致错误的判定。因此,在选择使用图像和标签信息对比的方式时,需要在计算效率和准确性之间进行权衡。

8 研究挑战

在回顾本文所述的对抗样本检测领域的工作时,明显可见这一领域的研究依然面临着多方面的挑战和问题,本章将详细探讨这些问题。

1)对抗样本的多样性。对抗样本的多样性主要体现在攻击方式和目标模型的多样性上,为对抗样本检测方法的研发带来了巨大挑战。这种多样性源于攻击者采用各种生成对抗样本的方法,例如 FGSM[6],C&W[8]和 PGD[24]等,同时也对不同类型的深度学习模型进行攻击,如 VGG[2]和 ResNet[3]等。对抗样本的多样性对检测方法造成了两个主要影响。首先,由于对抗样本具有不同的特征和攻击原理,因此开发能够泛化到所有情况的特征检测变得更加复杂。其次,多样性导致检测方法很可能对某一种攻击或特定模型过于依赖,难以适应各种对抗样本。对于无监督方法,对抗样本多样性导致找到使得检测方法能够更好泛化的阈值等参数变得困难。在监督方法中,多样性导致检测方法本身的泛用性难以保证。因此,未来的研究应当聚焦于对抗样本的本质,寻找其广泛存在的共同特征。

2)数据不平衡。对抗样本检测面临的一个主要挑战是正常样本和对抗样本之间的数据不平衡问题。尽管现在有许多数据集可供不同任务直接下载和使用,但是对抗样本仍然需要通过正常数据生成。这一过程需要耗费大量的算力和时间。在对抗样本检测的研究中,由于生成对抗样本的成本较高,通常只能生成数量较少的对抗样本用于模型的训练和性能评估。因此,为了更好地推动对抗样本检测的研究,有必要收集更为多样化且具有代表性的对抗样本,以构建更为完备的对抗样本数据集。

3)实时对抗样本检测。实时对抗样本检测是一项充满挑战的任务,主要受深度学习模型复杂性的影响。复杂的模型虽然更能捕捉对抗样本微小的变化,但在推理时却需要更多的计算资源,从而影响实时性。相较于离线批处理,实时系统必须立即处理数据,这要求对抗样本检测模型具备迅速适应新数据的能力,以确保系统高效运行。未来的研究应该注重提高深度学习模型的计算效率,通过优化对抗样本检测方法,在算法和架构设计中找到平衡点,以更好地适应实时在线场景。

4)跨模态对抗样本检测。现实世界中的数据通常呈现多模态特性,许多深度学习应用需要有效处理来自多个模态的信息。攻击者可能会充分利用跨模态数据的特点,采用更为复杂的对抗样本生成方法。因此,跨模态对抗样本检测的一个关键问题在于如何构建具有鲁棒性的检测方法,以适应不同模态的数据分布和攻击方式。

对抗样本检测是一个不断发展的领域,有许多机会和挑战等待着研究者们去探索和解决。未来研究可侧重于对抗样本本质、多样性特征的理解,解决数据不平衡问题,提高实时检测速度,以及探索跨模态环境中的鲁棒性检测方法。

参 考 文 献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Annual Conference on Neural Information Processing Systems. 2012.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[3] 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.

[4] REN S Q, HE K M, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. arXiv: 1506.01497, 2015.

[5] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv: 1312.6199, 2013.

[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.

[7] ZHANG J, HUANG Y, WU W, et al. Transferable adversarial attacks on vision transformers with token gradient regularization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16415-16424.

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

[9] XU H, LI Y, LIU X, et al. Yet meta learning can adapt fast, it can also break easily[C]// Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 2021: 540-548.

[10] KARIMI M P, AMIRKHANI A, SHOKOUHI S B. Robust object detection against adversarial perturbations withgabor filter [C]// 2021 29th Iranian Conference on Electrical Engineering

(ICEE). IEEE,2021:187-192.

[11] WANG L,YOON K J. Psat-gan:Efficient adversarial attacks against holistic scene understanding[J]. IEEE Transactions on Image Processing,2021,30:7541-7553.

[12] ABDULLAH H,RAHMAN M S,GARCIA W,et al. Hear "no evil",see"kenansville":Efficient and transferable black-box attacks on speech recognition and voice identification systems [C]// 2021 IEEE Symposium on Security and Privacy (SP). IEEE,2021:712-729.

[13] CHEN G,CHENB S,FAN L,et al. Who is real bob? adversarial attacks on speaker recognition systems[C]// 2021 IEEE Symposium on Security and Privacy(SP). IEEE,2021:694-711.

[14] HU Z,HUANG S,ZHU X,et al. Adversarial texture for fooling person detectors in the physical world[C]// Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition. 2022:13307-13316.

[15] WANG D,JIANG T,SUN J,et al. Fca:Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022:2414-2422.

[16] LIU J,LAU C P,SOURI H,et al. Mutual adversarial training:Learning together is better than going alone[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2364-2377.

[17] HARDER P,PFREUNDT F J,KEUPER M,et al. Spectraldefense:Detecting adversarial attacks on cnns in the fourier domain[C]// 2021 International Joint Conference on Neural Networks(IJCNN). IEEE,2021:1-8.

[18] DRENKOW N,FENDLEY N,BURLINA P. Attack agnostic detection of adversarial examples via random subspace analysis [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022:472-482.

[19] LIU Z,CAO C,TAO F,et al. From Spatial to Spectral Domain, a New Perspective for Detecting Adversarial Examples[J/OL]. Security and Communication Networks,2022. [2022-09-05]. https://doi.org/10.1155/2022.5501035.

[20] NADERI H,NOORBAKHSH K,ETEMADI A,et al. Lpf-defense:3d adversarial defense based on frequency analysis[J]. Plos one,2023,18(2):e0271388.

[21] ZHANG T,YANG K W,WEI J H,et al. A review of adversarial sample detection and defense technology for image data [J]. Computer Research and Development,2022,59(6):1315-1328.

[22] ALDAHDOOH A,HAMIDOUCHE W,FEZZA S A,et al. Adversarial example detection for DNN models:A review and experimental comparison[J]. Artificial Intelligence Review,2022, 55(6):4403-4462.

[23] ZHOU T,GAN R,XU D W,et al. A review of image adversarial example detection [J/OL]. Journal of Software:1-35. [2023-10-23]. https://doi.org/10.13328/j.cnki.jos.006834.

[24] MADRY A,MAKELOV A,SCHMIDT L,et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv:1706.06083,2017.

[25] 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.

[26] XIAO C,LI B,ZHU J Y,et al. Generating adversarial examples with adversarial networks[J]. arXiv:1801.02610,2018.

[27] 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.

[28] 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.

[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] MAO X,CHEN Y,LI Y,et al. Gap++:Learning to generate target-conditioned adversarial examples[J]. arXiv:2006.05097, 2020.

[31] CARRARA F,FALCHI F,CALDELLI R,et al. Detecting adversarial example attacks to deep neural networks[C]// Proceedings of the 15th International Workshop on Content-based Multimedia Indexing. 2017:1-7.

[32] SHI C,HOLTZ C,MISHNE G. Online adversarial purification based on self-supervision[J]. arXiv:2101.09387,2021.

[33] XU W,EVANS D,QI Y. Feature squeezing:Detecting adversarial examples in deep neural networks[J]. arXiv:1704.01155, 2017.

[34] MA S,LIU Y,TAO G,et al. Nic:Detecting adversarial samples with neural network invariant checking[C]// 26th Annual Network and Distributed System Security Symposium(NDSS 2019). Internet Soc,2019.

[35] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.

[36] ALDAHDOOH A,HAMIDOUCHE W,DéFORGES O. Revisiting model's uncertainty and confidences for adversarial example detection[J]. Applied Intelligence,2023,53(1):509-531.

[37] GEIFMAN Y,EL-YANIV R. Selectivenet:A deep neural network with an integrated reject option[C]// International Conference on Machine Learning. PMLR,2019:2151-2159.

[38] HENDRYCKS D,GIMPEL K. A baseline for detecting misclassified and out-of-distribution examples in neural networks[J]. arXiv:1610.02136,2016.

[39] KULLBACK S,LEIBLER R A. On information and sufficiency [J]. The Annals of Mathematical Statistics,1951,22(1):79-86.

[40] WIYATNO RR,XU A,DIA O,et al. Adversarial examples in modern machine learning:A review[J]. arXiv:1911.05268, 2019.

[41] ZHENG Z H,HONG P Y. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks[J]. Neural Information Processing Systems, 2018, 31:

7924-7933.
[42] PAPERNOT N, MCDANIEL P. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning[J]. arXiv:1803.04765,2018.
[43] ABUSNAINA A, WU Y, ARORA S, et al. Adversarial example detection using latent neighborhood graph[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:7687-7696.
[44] LIANG B, LI H, SU M, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2018, 18(1):72-85.
[45] WANG Y, LI X, YANG L, et al. ADDITION: Detecting Adversarial Examples With Image-Dependent Noise Reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 21(3):1139-1154.
[46] MOORTHY A K, BOVIK A C. Blind image quality assessment: From natural scene statistics to perceptual quality[J]. IEEE Transactions on Image Processing, 2011, 20(12):3350-3364.
[47] KHERCHOUCHE A, FEZZA S A, HAMIDOUCHE W, et al. Detection of adversarial examples in deep neural networks with natural scene statistics[C]// 2020 International Joint Conference on Neural Networks(IJCNN). IEEE, 2020:1-7.
[48] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on image processing, 2012, 21(12):4695-4708.
[49] GONG Z, WANG W. Adversarial and clean data are not twins [C]// Proceedings of the Sixth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management. 2023:1-5.
[50] HOSSEINI H, CHEN Y, KANNAN S, et al. Blocking transferability of adversarial examples in black-box learning systems[J]. arXiv:1703.04318,2017.
[51] LUST J, CONDURACHE A P. Gran: An efficient gradient-norm based detector for adversarial and misclassified examples [J]. arXiv:2004.09179,2020.
[52] COHEN G, SAPIRO G, GIRYES R. Detecting adversarial samples using influence functions and nearest neighbors[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:14453-14462.
[53] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv: 1801.02613,2018.
[54] LORENZ P, KEUPER M, KEUPER J. Unfolding local growth rate estimates for(almost) perfect adversarial detection[J]. arXiv:2212.06776,2022.
[55] LEE K, LEE K, LEE H, et al. A simple unified framework for detecting out-of-distribution samples and adversarial attacks[J]. arXiv:1807.03888,2018.
[56] CHEN K, CHEN Y, ZHOU H, et al. Adversarial examples detection beyond image space[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2021:3850-3854.
[57] GROSSE K, MANOHARAN P, PAPERNOT N, et al. On the (statistical) detection of adversarial examples[J]. arXiv:1702.06280,2017.
[58] GAO R, LIU F, ZHANG J, et al. Maximum mean discrepancy test is aware of adversarial attacks[C]// International Conference on Machine Learning. PMLR, 2021:3564-3575.
[59] FEINMAN R, CURTIN RR, SHINTRE S, et al. Detecting adversarial samples from artifacts[J]. arXiv:1703.00410,2017.
[60] DONG C, KUMAR A, LIU E. Think twice before detectinggan-generated fake images from their spectral domain imprints[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:7865-7874.
[61] JUNG S, KEUPER M. Spectral distribution aware image generation[C]// Proceedings of the AAAIConference on Artificial Intelligence. 2021:1734-1742.
[62] LORENZ P, HARDER P, STRABEL D, et al. Detecting autoattack perturbations in the frequency domain[J]. arXiv: 2111.08785,2021.
[63] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]// International Conference on Machine Learning. PMLR, 2020: 2206-2216.
[64] MAO X, CHEN Y, LI Y, et al. Learning to characterize adversarial subspaces[C]// ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020:2438-2442.
[65] ZHANG C, YANG Z, YE Z. Detecting Adversarial Perturbations with Salieny[C]// Proceedings of the 6th International Conference on Information Technology: IoT and Smart City. 2018:25-30.
[66] PRAKASH A, MORAN N, GARBER S, et al. Deflecting adversarial attacks with pixel deflection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8571-8580.
[67] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.
[68] WANG S, GONG Y. Adversarial example detection based on saliency map features[J]. Applied Intelligence, 2022(6): 6262-6275.
[69] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Conditional image generation with pixelcnn decoders[J]. arXiv:1606.05328,2016.
[70] SONG Y, KIM T, NOWOZIN S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[J]. arXiv:1710.10766,2017.
[71] KIANI S, AWAN S, LAN C, et al. Two souls in an adversarial image: Towards universal adversarial example detection using multi-view inconsistency[C]// Proceedings of the 37th Annual Computer Security Applications Conference. 2021:31-44.
[72] WANG H, MILLER D J, KESIDIS G. Anomaly detection of adversarial examples using class-conditional generative adversarial

networks[J]. Computers & Security,2023,124:102956.
[73] FREITAS S,CHEN S T,WANG Z J,et al. Unmask:Adversarial detection and defense through robust feature alignment[C]// 2020 IEEE International Conference on Big Data(Big Data). IEEE,2020:1081-1088.
[74] GONG Y,WANG S,JIANG X,et al. Adversarial example detection using semantic graph matching[J]. Applied Soft Computing,2023,141:110317.
[75] TAO G H,MA S Q,LIU Y Q,et al. Attacks meet interpretability:Attribute-steered detection of adversarial samples[J]. arXiv:1810.11580,2018.
[76] QIU Y,LENG J,GUO C,et al. Adversarial defense through network profiling based path extraction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4777-4786.
[77] NWAIGWE D,CARBONI L,MERMILLOD M,et al. Graph-based methods coupled with specific distributional distances for adversarial attack detection[J]. Neural Networks,2024,169:11-19.
[78] DENG L. The mnist database of handwritten digit images for machine learning research[J]. IEEE Signal Processing Magazine,2012,29(6):141-142.
[79] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images[D]. Toronto:University of Toronto,2009.
[80] DENG J,DONG W,SOCHER R,et al. ImageNet:A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2009:248-255.
[81] LE Y,YANG X. Tiny imagenet visual recognition challenge[J]. CS 231N,2015,7(7):3.
[82] NETZER Y,WANG T,COATES A,et al. Reading digits in natural images with unsupervised feature learning[C]// NIPS Workshop on Deep Lear-ning and Unsupervised Feature Learning. 2011:7.
[83] NILSBACK M E,ZISSERMAN A. Automated flower classification over a large number of classes[C]// 2008 Sixth Indian Conference on Computer Vision,Graphics & Image Processing. IEEE,2008:722-729.
[84] LI F F,FERGUS R,PERONA P. Learning generative visual models from few training examples:An incremental bayesian approach tested on 101 object categories[C]// 2004 Conference on Computer Vision and Pattern Recognition. IEEE,2004:178.
[85] GRIFFIN G,HOLUB A,PERONA P. Caltech-256 object category dataset[R]. Pasadena:Technical Report 7694,California Institute of Technology,2007.
[86] PINTO N,STONE Z,ZICKLER T,et al. Scaling up biologically-inspired computer vision:A case study in unconstrained face recognition on facebook[C]// CVPR 2011. IEEE,2011:35-42.
[87] CUKIERSKI W. Dogs vs. cats,2013[J/OL]. https://kaggle.com/competitions/dogs-vs-cats.
[88] GUVENIR H A,ACAR B,DEMIROZ G,et al. A supervised machine learning algorithm for arrhythmia analysis[C]// Computers in Cardiology 1997. IEEE,1997:433-436.
[89] AEBERHARD S,COOMANS D,DE VEL O. Comparative analysis of statistical pattern recognition methods in high dimensional settings[J]. Pattern Recognition,1994,27(8):1065-1077.
[90] XIAO H,RASUL K,VOLLGRAF R. Fashion-mnist:a novel image dataset for benchmarking machine learning algorithms[J]. arXiv:1708.07747,2017.
[91] ILYAS A,SANTURKAR S,TSIPRAS D,et al. Adversarial examples are not bugs,they are features[J]. arXiv:1905.02175,2019.
[92] ZHU X,WANG H,FEI H,et al. Face forgery detection by 3d decomposition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:2929-2939.
[93] GE Y,XIAO Y,XU Z,et al. Contributions of shape,texture,and color in visual recognition[C]// European Conference on Computer Vision. Cham:Springer Nature Switzerland,2022:369-386.
[94] ZHEN X,MENG Z,CHAKRABORTY R,et al. On the versatile uses of partial distance correlation in deep learning[C]// European Conference on Computer Vision. Cham:Springer Nature Switzerland,2022:327-346.
[95] RADFORD A,KIM J W,HALLACY C,et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR,2021:8748-8763.

ZHANG Xin, born in 1999, postgraduate. His main research interests include image processing and information security.

JI Lixia, born in 1979, associate professor. Her main research interests include multi-modal learning and information security.

(责任编辑:柯颖)