引用格式:黄知涛,柯达,王翔. 电磁信号对抗样本攻击与防御发展研究[J]. 信息对抗技术,2023, 2(4/5):37-52. [HUANG Zhitao, KE Da, WANG Xiang. Survey of electromagnetic signal adversarial example attack and defense [J]. Information Countermeasure Technology, 2023, 2(4/5):37-52. (in Chinese)]

电磁信号对抗样本攻击与防御发展研究

黄知涛[1],柯　达[2*],王　翔[2]

(1. 国防科技大学电子对抗学院,安徽合肥 230037; 2. 国防科技大学电子科学学院,湖南长沙 410073)

摘　要　以深度学习为代表的智能化技术在提升电磁频谱控制与利用系统性能水平的同时,也暴露出其脆弱性,催生出一批以对抗样本为代表的智能电磁攻防技术。随着智能化的快速应用和发展,该领域势必成为电磁频谱竞争的又一个“制高点”。首次尝试着明确了电磁对抗样本攻防的概念内涵,为规范后续的关键技术研究和具体应用提供参考。分析了智能模型脆弱性机理,认为智能模型脆弱性与可解释性存在一定的关系,将专家知识嵌入到模型学习中是下一步改善模型鲁棒性的研究方向。系统梳理了电磁信号对抗样本攻击和对抗样本防御的研究脉络,总结了通用对抗样本领域的共性研究规律,可以直接为电磁信号对抗样本研究提供借鉴。通过总结电磁信号对抗样本的研究规律,提炼出电磁信号对抗样本特有的问题。在此基础上,结合团队近年在该领域的研究积累,提出下一步的发展趋势,对抗攻击下一步的研究趋势是适应跨模型、跨任务的场景,应更加注重领域知识的应用,目标是要对抗多源综合的传感器体系;对抗防御的研究趋势是寻找鲁棒性与泛化性的权衡,通过利用信号处理知识优化处理流程,提高模型的对抗防御性能。同时关注鲁棒性评估,这可能是下一代智能化系统可靠性评估的关键技术之一。

关键词　对抗样本攻击;对抗样本防御;电磁频谱控制与利用;深度学习

中图分类号　TN 97　　**文章编号**　2097-163X(2023)04/05-0037-16

文献标志码　A　　**DOI**　10.12399/j.issn.2097-163x.2023.04-05.003

Survey of electromagnetic signal adversarial example attack and defense

HUANG Zhitao[1], KE Da[2*], WANG Xiang[2]

(1. College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China;
2. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract　The intelligent technology represented by deep learning has exposed vulnerabilities while improving the performance of electromagnetic spectrum control and utilization system. However it has given rise to a number of intelligent electromagnetic attack and defense technologies represented by adversarial examples. With the rapid application and development of intelligence, this field is bound to become another “high point” in the competition of electromagnetic spectrum. This paper attempted to clarify the content of electromagnetic adversarial-example attack and defense, and to provide reference for standardizing the subsequent research and applications, analyzed the vulnerability mechanism of intelligent

收稿日期:2023-05-17　　**修回日期**:2023-05-30

通信作者:柯达,E-mail:1747884404@qq.com

基金项目:国家自然科学基金资助项目(62271494)

models and concluded that there was a relationship between the vulnerability and interpretability of intelligent models. Embedding expert knowledge into model learning is the next research direction to improve the robustness of models. The research lineage of electromagnetic signal adversarial example attack and defense was systematically sorted out, and the common laws in the field of adversarial examples were summarized, which could directly referred by electromagnetic signal research. By summarizing the research laws of electromagnetic signal adversarial examples, some the specific problems were refined. On this basis, combining the accumulation in this field in recent years, the next development trend was proposed: adapt to cross-model and cross-task scenarios should be paid more attention, more domain knowledge should be embedded in the adversarial example, the goal was fighting against multi-source integrated sensor systems. The research trend of adversarial defense was to find the trade-off between robustness and generalization, and optimize the processing flow by using signal processing knowledge. Besides, attention should be paid to robustness assessment, which is likely to be one of the key techniques for reliability assessment of next-generation intelligent systems.

Keywords adversarial-example attack; adversarial-example defense; electromagnetic spectrum control and utilization; deep learning

0 引言

近年来,以深度学习为代表的人工智能(artificial intelligence, AI)技术在陆、海、空和网电等多个作战域均得到了广泛的应用。其中,在频谱感知、频谱利用和无人空战等领域已取得了优越的性能水平。2017 年,美国哈佛大学肯尼迪政治学院发布研究报告,认为未来 AI 将会像核、航空航天、网络、生物技术一样,成为深刻影响国家安全的革命性技术[1]。但是 AI 的安全性也存在重大隐患,通过对输入样本添加人为精心设计的微小扰动,尽管这些扰动难以察觉,却能使 AI 的识别性能出现严重下降,这类样本被称为对抗样本[2]。2018 年,美国国防高级计划研究局(DARPA)宣布将投资超过 20 亿美元发展“AI Next”运动[3],重点部署了 5 大方向,其中,对抗性 AI 和强健的 AI 便是从对抗样本攻击和防御的角度研究人工智能。时任 DARPA 副局长彼得·海拉姆强调“AI Next”的 3 大目标之一就是增强人工智能技术的稳健性。美国新成立的 AI 顶层管理机构联合人工智能中心主任、海军陆战队中将迈克·格伦在 2020 年宣布“美未来将重点发展对抗性人工智能技术,确保人工智能算法的安全受到保护”。美国人工智能安全委员会在 2021 年提交的报告中指出,鲁棒性、弹性以及抗欺骗和干扰的人工智能技术是大国军事竞争对抗的关键点。2022 年,兰德公司发布研究报告《对抗攻击如何影响美国防部军事人工智能系统》,分析了光电系统、合成孔径雷达、信号情报等 3 个军事人工智能系统实例中的人工智能对抗攻击,以及对美国防部人工智能系统和作战的影响,并给出了评估风险、追踪研究进展、开发鲁棒模型、提供支持等应对建议[4]。

电磁频谱空间也不例外。近年来,以美国为代表的军事强国一直致力于围绕智能化电磁频谱作战创新作战概念、研发先进技术,试图始终保持其在电磁频谱作战中的绝对优势。2021 年出版的首部关于认知电子战的国际性论著《认知电子战:人工智能方法》指出[1]:“现代电子战面临的挑战远远超出了传统电子战系统的能力范畴,将 AI 嵌入到电子战系统是应对当前挑战的唯一方法。”但是,AI 为电子战系统带来新的增量的同时,也会导致电子战系统对 AI 的高度依赖。现有研究揭示了 AI 存在明显的脆弱性[2,5],以对抗样本为代表的攻击技术将会制约未来智能化手段部署到电子战、雷达、通信和导航等装备中的安全性。因此,AI 在电磁频谱作战领域的鲁棒性和安全性不容忽视。特别是近年来,以美国为首的世界各主要军事强国在智能化电磁频谱对抗样本攻防领域的研究飞速发展[6-15],充分说明开

展面向电磁频谱作战的对抗样本攻防技术具有迫切的现实意义。

对抗样本的概念自 2014 年提出至今，已出现了大量的学术成果和应用案例，但是关于对抗样本在电磁频谱控制与利用方面的研究还处于起步阶段，尚没有形成明确的概念。本文认为对抗样本在未来的电磁频谱控制与利用行动中具有深远的前景，为了更好地指导后续的技术研究和作战运用，本文尝试着给出电磁对抗样本攻击和防御的概念，即围绕电磁频谱智能控制与利用，以对抗样本为主要实现手段开展的攻防对抗行动，主要包括电磁对抗样本攻击和电磁对抗样本防御。电磁对抗样本攻防是电子对抗（或电磁战）的一种特殊形式，以电磁频谱智能控制与利用系统中的智能模型为作用对象，采用对抗样本等攻击方法以降低对方系统效能，同时采取针对性措施确保己方系统的效能而采取的军事行动。

其中，电磁频谱控制与利用这一概念源于美国参谋长联席会议于 2020 年发布的《JP3-85：联合电磁频谱作战》条令。电磁频谱控制主要指传统的电子战和电磁频谱管理行动，而电磁频谱利用则指雷达、通信、导航等利用电磁频谱的行动。

电磁对抗样本攻击是为影响敌方智能电磁频谱控制和利用系统的主动攻击行动，通过有意识地发射对抗样本波形，扰乱和欺骗敌方军事电子信息系统和武器控制系统，使其不能正常工作。区别于传统的电子干扰，电磁对抗样本攻击的作用对象是敌方的智能化系统，且攻击所需的功率远小于电子干扰，有效降低了被发现被摧毁的概率。

电磁对抗样本防御的主要任务是在受到敌方电磁对抗样本攻击威胁和己方实施电磁对抗样本攻击时，尽量减少己方智能电磁频谱控制和利用系统作战效能受到的影响。

本文重点对基于深度学习模型的电磁对抗样本攻防技术研究情况进行分析总结。随着技术的发展，未来还会涌现出更多反制智能模型的手段和方法，需要研究者保持持续的关注。

作者团队在文献[16]中简要梳理了电磁领域智能攻防对抗的研究现状，针对典型应用，仿真分析了对抗样本攻击的效果，并总结了智能攻击技术的主要特点。在此基础上，进一步聚焦概念内涵、智能模型脆弱性数学机理，对研究进展和发展趋势进行分析、研判，系统性更强，以期对后续该领域的研究提供参考。

1　典型智能模型的脆弱性分析

1.1　对抗样本定义

为便于理解，本文从一个典型的最小化智能单元——智能电子侦察系统调制识别模块入手，详细分析 AI 模型的脆弱性。首先，从理论层面就智能调制识别模型的脆弱性进行详细的分析。

调制识别本质上是一个分类问题，也是目前深度学习研究最广泛的问题，本节将主要以调制识别模型为研究对象，详细阐述深度学习模型的脆弱性及背后的机理，类似的结论可以推广到回归模型[17]、生成模型[18]等其他模型中。

设模型输入为信号时域波形，则一个基于深度学习的调制识别模型可以描述为一个映射：$f_{\boldsymbol{\theta}}:X\to Y$，将输入空间中的信号 $\boldsymbol{x}\in X=\mathbf{R}^D$ 映射到输出空间 $y\in Y=\{1,\cdots,C\}$，其中，$\boldsymbol{\theta}\in W$ 为神经网络的参数，X，Y 和 W 分别构成模型的输入空间、输出空间和参数空间。模型对输入 $\boldsymbol{x}$ 的分类结果可以表示为：

$$f_{\boldsymbol{\theta}}(\boldsymbol{x})=\underset{k\in\{1,\cdots,C\}}{\operatorname{argmax}}\left[\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{x})\right]_k \tag{1}$$

式中，$\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{x})$ 为输入 $\boldsymbol{x}$ 属于每一类的概率组成的向量，一般为神经网络的 softmax 输出结果。在有监督的条件下，找到上述最优分类器 $f_{\boldsymbol{\theta}}$ 的学习算法的目标可以描述为，对于给定数据和标签 $(\boldsymbol{x},y)$ 的分布 D，最小化将输入 $\boldsymbol{x}$ 映射到标签 y 的期望风险为：

$$\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},y)\sim D}\left[L(\boldsymbol{x},y;\boldsymbol{\theta})\right] \tag{2}$$

式中，$L(\cdot)$ 为损失函数的一般表达式。

一般深度学习模型对输入的随机噪声具备一定的鲁棒性，但是对精心设计的对抗性扰动却表现出明显的脆弱性[2]。研究表明，对于任意输入 $\boldsymbol{x}$ 和模型 f_θ，总是存在对抗扰动，因此在这些方向上添加微小的扰动总是能改变分类器的输出结果。求解对抗扰动 $\boldsymbol{\delta}$ 的过程可以定义为最优化问题：

$$\begin{cases}\min\limits_{\boldsymbol{\delta}\in\mathbf{R}^D} Q(\boldsymbol{\delta})\\ \text{s.t.}\quad f_\theta(\boldsymbol{x}+\boldsymbol{\delta})\neq f_{\boldsymbol{\theta}}(\boldsymbol{x})\\ \qquad\boldsymbol{\delta}\in C\end{cases} \tag{3}$$

式中，$Q(\boldsymbol{\delta})$ 是优化目标的一般形式，C 是描述扰动 $\boldsymbol{\delta}$ 特性的一组约束集合。例如：

1) 在最小化扰动 $\boldsymbol{\delta}$ 的 l_p 范数的约束下可以定义为：

$$Q(\boldsymbol{\delta})=\|\boldsymbol{\delta}\|_p=\left(\sum_{k=1}^{D}\delta_k^p\right)^{1/p} \tag{4}$$

式中，$\boldsymbol{\delta}=[\delta_1,\delta_2,\cdots,\delta_D]^{\mathrm{T}}$，此时 C 为空集，即 $C=\varnothing$。整个优化问题可以理解为在使分类器 $f_{\boldsymbol{\theta}}$ 识别错误的前提下，扰动 $\boldsymbol{\delta}$ 的 l_p 范数需要尽可能小[19-20]。

2) 在 ε 邻域约束下，$Q(\boldsymbol{\delta})$和 C 可以定义为：

$$\begin{cases} Q(\boldsymbol{\delta})=-L(\boldsymbol{x}+\boldsymbol{\delta},y;\boldsymbol{\theta}) \\ C=\{\boldsymbol{\delta}\in\mathbf{R}^D:\|\boldsymbol{\delta}\|_p\leqslant\varepsilon\} \end{cases} \tag{5}$$

式中，$L(\cdot)$为损失函数，在该定义下整个优化问题可以理解为在给定的扰动范围 ε 内，找到满足识别错误条件的最坏扰动[5,21]。对抗样本还可以用其他距离测度的定义来描述，如数据流形测地线距离[22-24]、感知度量[25-26]和 Wasserstein 距离[27-28]等距离测度，l_p 范数和 ε 约束是目前使用最广泛的。一般来说，大部分关于对抗样本攻击的研究，均是围绕如何约束和求解公式(3)展开的。然而，由于深度学习模型的非凸性和输入空间的高维性[29]，精确求解公式(3)往往是困难的。因此，大多数的对抗攻击方法仅仅是对公式(3)的近似求解[5,19-21]。

1.2 对抗脆弱性机理

关于深度学习模型对抗脆弱性的成因，最开始 SZEGEDY[2]认为是由于深度神经网络的高度线性化和正则化不充分导致的。作者团队认为，当前的全连接神经网络和卷积神经网络都属于线性运算，故均满足 Lipschitz 性，即：

$$\|\varphi_k(\boldsymbol{x};\boldsymbol{W}_k)-\varphi_k(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{W}_k)\|\leqslant L_k\|\boldsymbol{\delta}\|,\forall\,\boldsymbol{x},\boldsymbol{\delta} \tag{6}$$

那么可以用 Lipschitz 常数 L_k 来描述扰动 $\boldsymbol{\delta}$ 给每一层网络带来的不稳定性上界。相应的，最终的网络输出结果满足：

$$\|\varphi(\boldsymbol{x})-\varphi(\boldsymbol{x}+\boldsymbol{\delta})\|\leqslant L\|\boldsymbol{\delta}\| \tag{7}$$

式中，$L=\prod_{k=1}^{K}L_k$。

尽管实际网络中，每一层线性运算后还会经过非线性的激活函数，如 ReLU 函数，但是，这些函数均满足收缩性(contractive)，即对于一个函数 $\|\rho(\boldsymbol{x})-\rho(\boldsymbol{x}+\boldsymbol{\delta})\|\leqslant\|\boldsymbol{\delta}\|$。所以，非线性激活函数不影响神经网络满足式(6)所表示的 Lipschitz 性。

文献[2]以典型深度神经网络 ImageNet[30]为例，分析了该神经网络每一层的 Lipschitz 上界，结果如表 1 所列。根据式(6)，表 1 中不稳定性上界的含义是，扰动 $\boldsymbol{\delta}$ 经过该层神经网络后，其造成的偏差最多会被放大多少。例如，经过第一层后，扰动的影响最多会被放大 2.75 倍。这些结果可以解释神经网络存在对抗脆弱性，而这并不能解释为什么对抗样本可以泛化到不同的网络或训练集。计算出不稳定上界的意义在于：尽管大的界限不代表对抗样本的广泛存在，但是小的界限可以尽可能保证不会出现对抗样本。这表明可以对参数进行简单的正则化，包括对每个 Lipschitz 边界进行惩罚，这可能有助于改善网络的泛化误差。SZEGEDY 的研究解释了对抗样本的存在性，即由于神经网络的线性性导致了对抗样本的存在。但是，上述分析过程只能说明线性性是导致对抗样本存在的一种充分不必要条件，并不能解释为什么对抗样本能在不同网络结构、不同训练数据中均具有泛化性和普遍性。

表 1 ImageNet 网络每一层的不稳定性上界

Tab. 1 The upper bound on the instability of each layer of the ImageNet

网络层	参数规模	不稳定性上界
Conv1	3×11×11×96	2.75
Conv2	96×5×5×256	10
Conv3	256×3×3×384	7
Conv4	384×3×3×384	7.5
Conv5	384×3×3×256	11
FC1	9 216×4 096	3.12
FC2	4 096×4 096	4
FC3	4 096×1 000	4

GOODFELLOW 等[5]认为对抗脆弱性的成因是由于参数空间的高维特性，微小的扰动和高维的参数向量进行点积运算，经过多层的误差积累，最终形成了巨大的偏差。具体而言，作者以简单的线性分类器 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}$ 为例进行了详细阐述。对带有扰动的输入 $\boldsymbol{x}+\boldsymbol{\delta}$，即使扰动 $\boldsymbol{\delta}$ 的范数非常小，由于模型参数的维数和深度都很大，扰动经过神经网络的运算后 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\delta}$ 依然有可能很大。对于线性分类器 $f(\boldsymbol{x}+\boldsymbol{\delta})=\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}+\boldsymbol{\delta})+\boldsymbol{b}$，不同模型之间的对抗样本的泛化性可以解释为对抗性扰动与模型的权重向量高度一致，而不同的模型在训练执行相同的任务时学习相似

的函数。GOODFELLOW 的观点可以看作是对上文 SZEGEDY 观点的补充，为对抗样本泛化性给出了方向性的解释。

关于对抗样本的解释，TSIPRAS 等[31-32]给出了一种新颖的观点，认为对抗样本不是深度学习的故障，而是一种特征。他们认为模型会尽可能学习一切有助于识别样本的特征，这些特征分为鲁棒性特征和非鲁棒性特征，并且发现鲁棒性特征往往包含容易理解的语义信息，而非鲁棒特征有利于模型的标准泛化性。其中，对抗样本便是反应了非鲁棒的这部分特征。他们将对抗样本的现象归结为标准数据集中存在高度预测性但非鲁棒性特征的自然结果，通过明确区分标准数据集中的鲁棒性和非鲁棒性特征，并证明仅非鲁棒性特征足以实现良好的泛化，为这一假设提供了支持；最后，建立了一套理论框架用以更详细地研究这些现象，据此可以严格研究对抗脆弱性和鲁棒的训练：一方面，作者的发现表明在标准数据集上的高准确率和鲁棒性是存在矛盾的，因为非鲁棒的特征往往具有良好的可预测性；另一方面，从可解释性的角度来看，只要模型依赖于这些非鲁棒的特征，就不能指望对模型本身既有可解释性又忠实于模型自己的解释。总体而言，要获得可靠且可解释的模型，就需要在训练过程中明确编码专家先验知识。

1.3　小结

通过梳理深度学习模型脆弱性机理的研究历程，可以得到以下 2 点关于深度学习脆弱性的结论：

1）输入与状态空间的高维特性导致了深度学习的脆弱性。复杂的深度识别模型包含数千万量级参数（最新的超大模型甚至达到了千亿量级的参数规模），且由于模型的高度非线性，输入数据的微小扰动可能会产生积累，导致巨大的输出差异，该现象通常也称为“维度灾难”。

2）深度学习模型的脆弱性与难解释性存在关联。鲁棒/非鲁棒表征的观点揭示了当前的深度学习模型普遍是不鲁棒的，而鲁棒的模型往往能提取可解释性强的特征。为了达到更高的对抗鲁棒性，下一步应该研究如何将专家知识编码到深度学习模型中。

2　研究进展

本节将系统梳理关于对抗样本攻击和对抗样本防御的研究进展，并总结其发展规律。由于对抗样本这一概念起源于图像领域的研究，且前期的大多数研究成果也来源于图像领域，所以，本文将首先介绍图像领域对抗样本的研究历程，梳理出其中共性的研究规律，然后结合电磁信号领域的对抗样本研究历程，总结出电磁信号领域独有的研究特点，进而服务于下一步电磁信号对抗样本的研究。

2.1　电磁对抗样本攻击

2.1.1　图像领域的对抗样本攻击

2014 年 SZEGEDY 等[2]在国际学习表征大会（international conference learning representation，ICLR）发表文章称，首次发现深度神经网络存在一种现象，即通过在图像中添加精心设计的微小扰动，尽管这些扰动对于人类视觉系统来说几乎无法察觉，也往往会导致深度学习模型以很高的置信度对扰动后的图像进行错误分类。这些扰动后的数据称为对抗样本。

作者团队将对抗样本形式化地描述为一个基本的优化问题。设神经网络模型为 $f_\theta: X \to Y$，将输入空间中的样本 $\boldsymbol{x} \in X = \mathbf{R}^D$ 映射到输出空间 $y \in Y = \{1, \cdots, C\}$，其中，$\boldsymbol{\theta} \in W$ 为神经网络的参数。对抗扰动 $\boldsymbol{\delta} \in \mathbf{R}^D$ 可以定义为：

$$\begin{cases} \min \|\boldsymbol{\delta}\|_2 \\ \text{s.t.} \quad f_\theta(\boldsymbol{x}+\boldsymbol{\delta}) = l \cdot \boldsymbol{x} + \boldsymbol{\delta} \in [0,1]^D \end{cases} \tag{8}$$

由于目标函数难以优化，SZEGEDY 提出使用有限内存的 BFGS（limited memory BFGS，L-BFGS）算法将此目标函数近似为盒约束（box-constrained）优化的形式：

$$\begin{cases} \min\limits_{\boldsymbol{\delta}} c\|\boldsymbol{\delta}\|_2 + L(\boldsymbol{x}+\boldsymbol{\delta}, l) \\ \text{s.t.} \quad \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^D \end{cases} \tag{9}$$

式中，输入 $\boldsymbol{x}$ 的每个元素都被正则化到区间 $[0,1]$，l 是目标被错误分类的标签，c 为用于控制损失函数 $L(\cdot)$ 和扰动 δ 的相对权重，其中，$c>0$。该方法一经提出，就使得当时最先进的 Alexnet[30]模型和 QuocNet[33]模型对图片分类的识别率大大下降，且生成的对抗样本在视觉上与干净的图片几乎一致。

尽管 L-BFGS 方法取得了显著的攻击成果，但是该方法是一种基于优化的方法，生成对抗样本的过程需要多轮迭代，算法效率较低。随后，GOODFELLOW 等[5]又提出了效率更高的快速梯度符号法（fast gradient sign method，FGSM），

该算法能够在给定输入的情况下快速找到扰动方向，从而使目标模型的训练损失增加，增加类间混淆的可能性。GOODFELLOW 等给出了 FGSM 算法的表达式：

$$\boldsymbol{\delta}=\varepsilon\cdot\operatorname{sign}(\nabla_x L(\boldsymbol{x},y;\boldsymbol{\theta})) \tag{10}$$

式中，$L(\cdot)$为损失函数的一般表达式，ε 为约束扰动强度的参数。根据式(9)可知，FGSM 算法仅需计算一次神经网络的反向梯度即可生成对抗扰动，避免了 L-BFGS 方法的多步迭代。西交利物浦大学的 LYU 等[34]从梯度正则化的角度出发，将 FGSM 算法推广为基于“梯度正则化族”的对抗样本。这类对抗样本可以统一定义为满足以下优化目标：

$$\min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_p\leqslant\varepsilon} L(\boldsymbol{x}+\boldsymbol{\delta},y;\boldsymbol{\theta}) \tag{11}$$

即在 l_p 正则化约束下的优化问题(其中，p 取不同的值)，可以得到不同形式的对抗扰动求解方法。若 $p=\infty$，则得到式(10)的表达式；若 $p=2$，则 $\boldsymbol{\delta}=\varepsilon\dfrac{\nabla_x L(\boldsymbol{x},y;\boldsymbol{\theta})}{\|\nabla_x L(\boldsymbol{x},y;\boldsymbol{\theta})\|_2}$，该方法也称为快速梯度法(fast gradient method，FGM)[34-35]。上述方法统一被称为“单步法”，即求解对抗扰动的过程只计算了一次梯度。基本迭代法(basic iterative method，BIM)[36]是 FGSM 的另一变种，也被称为迭代的 FGSM，其通过迭代计算 FGSM 并限制总体扰动在输入空间的 ε 范围内实现更强的对抗攻击，其表达式为：

$$\boldsymbol{x}^{i+1}=\operatorname{Clip}_{\varepsilon}\{\boldsymbol{x}^i+\alpha\cdot\operatorname{sign}(\nabla L(\boldsymbol{x}^i,y;\boldsymbol{\theta}))\} \tag{12}$$

式中，$\boldsymbol{x}^{i+1}$ 表示第 i 轮迭代后被扰动的对抗样本，$\operatorname{Clip}_{\varepsilon}\{\cdot\}$表示将数据约束在 $\boldsymbol{x}$ 的 ε 邻域内，在 BIM 算法的框架下，扰动的强弱由 α 和 ε 共同决定。BIM 算法是 l_∞ 约束下的投影梯度下降法(projected gradient descent，PGD)[21]。

上述的 FGSM 算法、BIM 算法等均属于非定向攻击方法，即攻击目标是使得分类器“犯错”，而不限定于错判到哪一类。通过简单修改，即可实现定向的对抗攻击，即可以指定分类器将样本错判到任意特定的类别。迭代的最小似然分类方法(iterative least-likely class method，ILLCM)[34,36]便是 BIM 算法的一种定向攻击变种，其定向攻击的类别是原始分类器中可能性最小的一类，所以称为极小似然分类器，其表达式为：

$$\boldsymbol{x}^{i+1}=\operatorname{Clip}_{\varepsilon}\{\boldsymbol{x}^i-\alpha\cdot\operatorname{sign}(\nabla L(\boldsymbol{x}^i,t;\boldsymbol{\theta}))\} \tag{13}$$

式(13)与式(12)非常相似，区别在于梯度更新的符号相反，并把真实标签 y 改为了目标标签 $t=\operatorname{argmin} f(\boldsymbol{x})$。两者的物理意义是，式(12)的优化目标是使得分类器对扰动后的样本识别在损失函数空间上尽可能远离真实标签，而式(13)的优化目标则是尽可能靠近目标标签。

此外，还可以通过修改目标函数实现对抗样本的生成，如 C W 算法[20]。

其优化目标可以写为：

$$\begin{cases}\min\ D(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{\delta})\\ \text{s.t.}\ \ f(\boldsymbol{x}+\boldsymbol{\delta})=t\\ \quad\ \boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n\end{cases} \tag{14}$$

式中，$D(\cdot)$是衡量原始样本 $\boldsymbol{x}$ 和对抗样本 $\boldsymbol{x}+\boldsymbol{\delta}$ 差异的相似性度量，可以是 l_0、l_2、l_∞ 等不同的距离测度。t 表示定向攻击的目标标签，$f(\cdot)$代表分类器。由于 $f(\cdot)$是一个高度非线性的映射，文献[20]采用了更适合优化的形式来表示 $f(\boldsymbol{x}+\boldsymbol{\delta})=t$。本文定义了目标函数 L_{CW}，L_{CW} 满足当且仅当 $L_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta})\leqslant 0$ 时，$f(\boldsymbol{x}+\boldsymbol{\delta})=t$。此时式(14)可以转换为：

$$\begin{cases}\min\ D(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{\delta})+cL_{\mathrm{CW}}(\boldsymbol{x}+\boldsymbol{\delta})\\ \text{s.t.}\ \ \boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^n\end{cases} \tag{15}$$

本文给出了 F 的多种表达式[20]。例如：

$L_{\mathrm{CW}_1}(\boldsymbol{x}')=-\operatorname{loss}_{f,t}(\boldsymbol{x}')+1$；

$L_{\mathrm{CW}_2}(\boldsymbol{x}')=(\max_{i\neq t}\{Z(\boldsymbol{x}')_{(i)}\}-Z(\boldsymbol{x}')_{(t)})^+$；

$L_{\mathrm{CW}_3}(\boldsymbol{x}')=(0.5-f(\boldsymbol{x}')_t)^+$。

其中，$\operatorname{loss}_{f,t}(\boldsymbol{x}')$表示交叉熵函数，$(e)^+=\max(e,0)$，$Z(\boldsymbol{x}')_{(i)}$表示分类器 softmax 之前输出的第 i 个分量，即 $f(\boldsymbol{x})=\operatorname{softmax}(Z(\boldsymbol{x}))$。

可以发现，其核心思想与公式(9)表示的 L-BFGS 方法类似，均是将一般约束优化策略转化为无约束优化。本文进一步给出了若干种经验选择的损失函数 L_{CW}。

通用领域对抗样本的研究已基本成熟。在理论方面，已有大量学者在不同的假设前提和约束下提出了不同的对抗样本生成方法，在现有的各类深度学习模型上均取得了显著的攻击成果；在应用层面，对抗样本在人脸识别、自动驾驶等实际应用场景中均得到了成功的尝试[51-59]。完备的理论研究和成功的应用案例也使得对抗样本在电磁信号处理场景中的应用具

有广阔的前景。

2.1.2　电磁信号对抗样本攻击

鉴于对抗样本在图像、语音等领域的研究已取得了大量的成果和应用，电磁信号对抗样本的研究也得到了大量的关注。关于电磁信号对抗样本的公开研究大致可以分为 2 个阶段。

第一个阶段自 2019 年始，由于电磁信号处理的问题模型与图像领域的很多场景是一致的，即均可表示为分类问题，通用图像领域的对抗样本研究成果可以直接借鉴到电磁信号领域。所以该时期的研究主要表现为简单的将现有成熟的对抗样本算法应用到智能电磁频谱控制与利用的场景中，证明对抗样本在电磁信号领域中的可行性。SADEGHI 等[46]首次在对基于深度学习的无线信号调制识别任务中提出了一种白盒对抗攻击方法和通用对抗攻击方法。结果表明，对抗攻击可以在极小的扰动下大大降低智能调制识别的分类性能，这给无线电物理层使用基于深度学习的算法带来了重大的安全性和鲁棒性问题。同年，KE 等[47]也针对通信信号调制识别模型进行了对抗样本攻击的验证。作者采用了 FGSM 算法对基于深度学习的调制识别模型进行攻击，并验证了对抗训练对调制识别模型标准鲁棒性和对抗鲁棒性的改善作用。LIN 等[48]验证了多种主流对抗攻击在调制识别场景中的有效性和可行性，分析了多种攻击方法在调制识别场景中的表现。上述工作属于早期研究对抗样本在电磁信号场景中应用的尝试，为对抗样本在电磁信号场景中的可行性验证奠定了基础。

随着研究的深入，研究人员开始将对抗样本应用在智能电磁频谱控制与利用更具体的场景中。智能自动化公司（Intelligent Automation Inc.，IAI）的 SHI 等[12]提出了一种对抗性机器学习方法。作者针对基于深度神经网络的通信发射机构造了一个对抗性发射机，其中，通信发射机能实时感知周围的频谱环境，以最小的感知误差预测空闲信道用于数据传输；对抗性发射机通过推断通信发射机的行为并伪造频谱感知数据来发动频谱数据投毒攻击，其目的是在发射机感知频谱环境时将信道占用状态从空闲变为繁忙，从而使发射器做出错误的通信。实验结果表明，这种攻击手段可以大大降低发射机的吞吐量，且与直接干扰数据传输过程相比，这种攻击更节能，也更难被发现。IAI 的 HOU 等[11]结合生成对抗网络（generative adversarial network，GAN）技术[49]，针对基于机器学习的物联网设备认证提出了一种物联网生成对抗网络（Internet of thing generative adversarial network，IoTGAN）攻击技术，该技术不仅可以攻击深度神经网络模型，也可以攻击随机森林、决策树和支持向量机等传统机器学习模型，极大地拓展了电磁信号对抗攻击的适用场景。

第二个阶段自 2021 年始，研究人员开始结合真实物理场景，结合领域特有的知识，针对性地优化对抗样本的生成算法，以期增强对抗样本对真实物理场景的适应能力。HAMMAD 等[50]结合无线信号的特点，提出了一种新的对抗样本生成的理论框架，首先确定了如图 1 所示的对抗样本在通信环路中的具体实现方式。作者将对抗样本看作一种特殊的调制方式，对抗性扰动应该在基带调制的环节添加到原信号中。

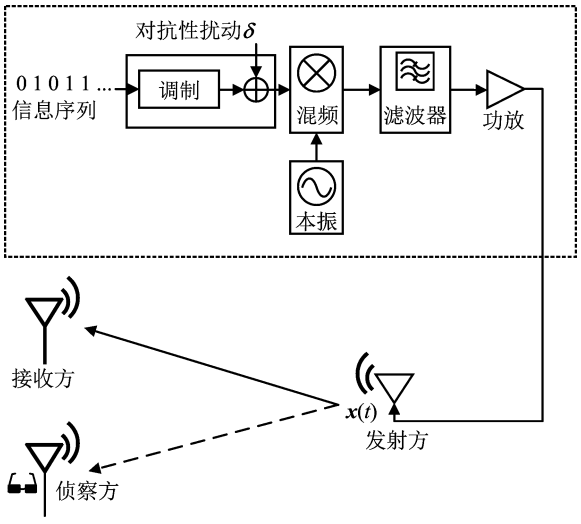


图 1　电磁信号对抗样本的实现模型Ⅰ

Fig. 1　Implementation model Ⅰ of electromagnetic signal adversarial example

文献[50]还认为，对抗样本的两大目标分别是使模型犯错和不易被察觉。其中，不易被察觉在电磁信号中应该具有新的内涵，而非传统认为的仅仅是能量上的微弱；在通信系统中，对抗样本的不易被察觉应该表现为添加了对抗扰动的信号不会影响原本的合作通信过程，而只破坏非合作的“窃听者”的调制识别过程。基于上述观点，作者将误码率作为优化目标补充到了式（3）的对抗样本求解框架中，具体可以表达为：

$$\begin{cases}\min\limits_{\boldsymbol{\delta}\in\mathbf{R}^D} Q(\boldsymbol{\delta})+\lambda e(\boldsymbol{x}+\boldsymbol{\delta}) \\ \text{s.t.}\quad f_{\theta}(\boldsymbol{x}+\boldsymbol{\delta})\neq f_{\theta}(\boldsymbol{x}) \\ \qquad \boldsymbol{\delta}\in C\end{cases} \tag{16}$$

式中，$e(\boldsymbol{x}+\boldsymbol{\delta})$为解调后的误码率。

文献[50]将误码率与交叉熵函数进行联合优化，即同时实现“难以察觉”和使分类器犯错的目标；由于通信码流的离散化，导致优化函数中的误码率约束项无法反向计算梯度，为此，采用同步扰动随机近似法(simultaneous perturbation stochastic approximation，SPSA)解决了这一问题。具体为，通过误码率 $e(\boldsymbol{x}+\boldsymbol{\delta})$对神经网络输入 $\boldsymbol{x}$ 梯度的求解：

$$\nabla_x e(\boldsymbol{x})=\frac{1}{K}\sum_{k=1}^{K}\frac{e(\boldsymbol{x}+\eta\boldsymbol{r}_k)-e(\boldsymbol{x}-\eta\boldsymbol{r}_k)}{2\eta}\boldsymbol{r}_k^{\mathrm{T}} \tag{17}$$

式中，$\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_K$ 为与输入 $\boldsymbol{x}$ 具有相同维度且满足均匀分布的随机向量。实验结果表明，所提对抗攻击算法在实现同样攻击性能的前提下，具有更小的误码率损失。

ZHANG 等[37]研究发现了通用领域的对抗攻击方法直接应用在电磁信号领域存在频谱泄露的问题，即大部分的对抗扰动能量分布在信号的频带之外，这在电磁信号对抗样本领域很容易被带通滤波器滤除。因此，本文提出了一种频谱聚焦的对抗攻击方法。同时，文献[37]考虑到实际场景中往往难以获取对手的模型信息，所以结合元学习的思想，提出了适应黑盒场景的对抗攻击方法。

KIM 等[6-10]提出了另一种对抗攻击的实现方式，设计了独立的对抗性扰动发射机，如图 2 所示，实现了在传输信号中添加对抗性扰动。

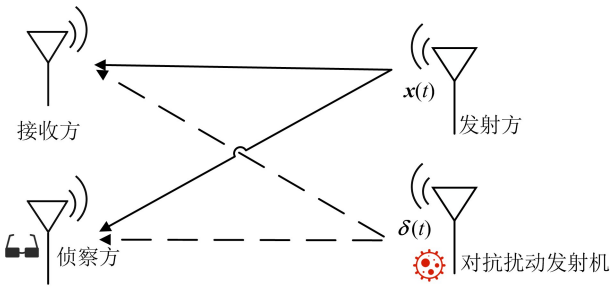


图 2 电磁信号对抗样本的实现模型Ⅱ

Fig. 2 Implementation model Ⅱ of electromagnetic signal adversarial example

研究发现，在实施电磁信号伪装攻击时，由于不能破坏自身的工作过程，伪装攻击信号的功率往往很微弱，传输过程中无线信道的非理想效应的功率足以破坏其扰动信号的结构，致使攻击失效，KIM 等[6,9-10]通过考虑从对抗样本发射机到接收机的信道效应，提出了一种基于信道反演的物理世界电磁信号伪装攻击方法，能够改善信道畸变对电磁信号伪装攻击的破坏，并将该技术推广到了 5G、物联网、毫米波等多种应用场景；进一步的研究发现，信道感知攻击是有选择性的(即它只影响其信道在扰动设计中被考虑的接收者)之后，提出了一个广播式对抗性攻击，通过精心设计一个通用的对抗性扰动来同时“愚弄”不同接收者的分类器。

总的来说，现有的案例表明将对抗样本应用到智能电磁频谱控制与利用场景中，民用上，可以利用对抗样本技术增强无线通信的反侦察能力，防止用户的通信内容被智能化的“窃听者”解译；军用上，基于对抗样本技术的电子对抗手段将会对智能化电子侦察系统带来新的威胁，通过对抗样本“愚弄”对手侦察系统，可以起到“四两拨千斤”的效果。

通过梳理整个对抗样本研究脉络，可以看出，对抗样本的研究主要是围绕式(3)的优化框架设计不同的优化目标和约束条件，以期增强对抗样本的性能，或是满足不同的应用需求，这是电磁信号对抗样本和通用领域对抗样本共通的地方。具体到电磁信号对抗样本的个性问题，主要表现为当应用到真实物理世界中后，需要结合领域知识来具体设计对抗样本的求解目标，例如，2.1.2 节中介绍的将解调误码率最小作为目标函数，是为了适应通信场景下对抗样本在攻击窃听者的同时，不能影响己方正常通信的需求；又如 KIM 等[6-10]则是关注到了现实物理世界中信道效应的影响，在求解式时加入了信道补偿。从文献[6-10]的研究成果也可以得到启发，虽然克服了信道对对抗攻击的影响，但是真实的无线信道环境要复杂得多，除了信道的影响，发射机和接收机器件的影响同样不可忽视。例如，对抗样本被非合作的窃听者截获后，由于对抗性扰动的频谱未进行带宽约束，使得典型的对抗样本频谱覆盖范围大大超出被干扰信号带宽，因此很容易被接收系统滤波器滤除，从而无法发挥正常的攻击效能。所以真实物理世界中的对抗攻击将会是下一步需要重点研究的方向。

2.2　电磁对抗样本防御

由于大量的对抗样本生成方法被相继提出，严重威胁到了深度学习模型的安全性。相应地，也催生了对抗样本防御技术的发展。目前针对电磁信号对抗样本的防御方法研究较少，主要是采用滤波、对抗样本检测等方法[78]。一方面是因为对抗防御在该领域还处于起步阶段，另一方面是因为通用机器学习领域的对抗样本防御方法均可迁移到电磁信号防御领域。基于此，本文重点总结了通用机器学习领域的对抗防御方法研究。AKHTAR 等[51]于 2018 年将对抗防御的策略总结为 3 类，即修改数据类、修改模型类和增加辅助模型。修改数据类方法是通过在训练或测试阶段修改数据及特征实现防御；修改模型类方法的是通过修改从数据学习得到的模型结构或参数信息实现防御；增加辅助模型方法通过引入额外的网络模型增强鲁棒性。AKHTAR 等又于 2021 年将认证防御（certified defenses，CD）总结为第 4 种防御策略。但是，基于认证的防御策略其方法大多来源于第一种策略，所以本文仍然按照 3 种分类的方法来梳理对抗防御的发展情况。

2.2.1　基于修改数据的防御

伴随着对抗样本的产生，对抗训练也作为一种防御对抗攻击的手段被提出[5]。其核心思想是在每一轮训练迭代中加入对抗样本构成新的训练集，重新训练以得到适应对抗样本的模型。要解决的关键问题是设计合适的目标函数，使得对抗样本能够更好地表征数据的分布，保证训练出的模型更鲁棒。设 $L(\boldsymbol{x},y)$ 为训练的损失函数，则对抗训练的损失函数可修改为：

$$L(\boldsymbol{x},y)+(1-\alpha)L(\boldsymbol{x}',y) \tag{18}$$

式中，$\boldsymbol{x}'$ 代表 $\boldsymbol{x}$ 对应的对抗样本，α 是用来控制正常训练和对抗训练之间损失权重的超参数，通常设置为 0.5。

用于产生对抗训练所需对抗样本的方法很多，例如，最常用的是可以直接用 FGSM 算法产生对抗样本，此时可以称为 FGSM 对抗训练。一般来说，每提出一种新的对抗扰动生成方法，便可得到一种对应的对抗训练方法[2,5,19]。研究发现，FGSM 对抗训练的结果是使得神经网络进一步正则化，降低了过拟合，反过来提高了对对抗样本攻击的鲁棒性。受到启发，MIYATO 等提出了“虚拟对抗训练”方法[36,52]，通过平滑神经网络输出特征的分布，可以在半监督条件下实现对神经网络的正则化。

虽然对抗训练是一种最直接、最简单，且性能较好的对抗样本防御方法，但对抗训练会成倍地增加训练的时间和空间复杂度。尽管神经网络经过了对抗训练，KANBAK 等[24]的研究发现依然能够找到新的对抗样本攻击神经网络。例如，对使用 FGSM 进行对抗训练的模型对迭代梯度攻击（如 BIM、PGD）的鲁棒性较低。另外，采用梯度攻击进行对抗训练的模型还有可能出现标签泄露的问题[35]，即采用特定的单步对抗攻击对模型进行对抗训练，与干净的样本相比，再采用同一方法生成的对抗样本上反而会使模型的准确率更高。这说明单步方法生成的对抗样本过于简单，容易导致模型过拟合。集成对抗训练（ensemble adversarial training，EAT）[53]是为解决对抗训练过拟合而产生的变种，其核心思想是使用在不同的预训练模型上生成的对抗样本重新训练目标模型。得到的目标模型和对抗样本不存在紧密的耦合关系，克服了原始对抗训练所观察到的过拟合现象。

ZHANG 等[37]在研究中发现经过 PGD 对抗训练后的模型分类界面会随着攻击步数逐渐增大而变得模糊。这导致对抗训练后自然数据和对抗数据严重的交叉混合问题（cross-over mixture problem），即模型要正确识别对抗样本就无法正确识别原始样本，所以鲁棒性比较强的模型会对正常样本分类错误，从而导致正确率下降。文献[37]认为传统对抗训练的极小极大公式，即 $\min\limits_{\boldsymbol{\theta}}\max\limits_{\boldsymbol{x}'}L(\boldsymbol{x}',y;\theta)$ 不适用于对抗训练，因此提出了友好的对抗训练（friendly adversarial training，FAT）算法，其核心思想是找到既使模型分类错误，又不破坏分类边界的“友好对抗样本”用于训练。FAT 算法的基础是极小-极小公式，表示为：

$$\begin{cases}\boldsymbol{x}'_i=\underset{\boldsymbol{x}'\in B_\varepsilon[\boldsymbol{x}_i]}{\arg\min}\,L(\boldsymbol{x}',y_i)\\ \text{s.t.}\quad L(\boldsymbol{x}',y_i)-\min\limits_{y\in Y}L(\boldsymbol{x}',y)\geqslant\rho\end{cases} \tag{19}$$

式中，$\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ 为给定的数据-标签对，$B_\varepsilon[\boldsymbol{x}]=\{\boldsymbol{x}'\in X\mid d_{\inf}(\boldsymbol{x},\boldsymbol{x}')\leqslant\varepsilon\}$ 表示输入空间 X 内以样本 $\boldsymbol{x}$ 为中心，ε 为半径的邻域，$\boldsymbol{x}'$ 为邻域内的对抗样本。公式表示的含义为首先确保对抗样本 $\boldsymbol{x}'$ 会被分错，同时确保 $\boldsymbol{x}'$ 的损失函数比预期最坏的损失函数值要提升 r。

ZHANG 等[38]提出了一种基于特征散射的对抗训练(feature scattering-based adversarial training, FSAT)方法。传统对抗训练方法在生成训练所用的对抗样本时利用了有监督信息,会造成标签泄露的问题。文献[38]基于上述观点,提出通过隐空间中的特征散射来生成用于对抗训练的对抗样本,由于这个过程没有利用标签信息,本质上是无监督的,因此避免了标签泄露。同时,传统的对抗训练方法只是独立的对待每一个样本,没有很好地利用样本之间的关系。在生成用于训练的对抗样本时,扰动样本的方向完全基于从当前数据点到决策边界的方向,而不考虑其他样本。虽然有效,但是忽略了不同特征点之间的相互关系,也忽略了集体分布特性,导致生成的扰动高度偏向决策边界。文献[38]将最优运输距离(optimal transport, OT)定义为:

$$D(\mu,\upsilon)=\inf_{\gamma\in\prod(\mu,\upsilon)}E_{(x,y)\sim\gamma}c(x,y) \tag{20}$$

式中,$D(\mu,\upsilon)$用来描述 2 个概率分布 μ 和 υ 之间的分布情况,表示从 μ 转向 υ 的最小代价。在 FSAT 中,μ 可以看作干净数据的经验分布,υ 可以看作是对抗样本的经验分布。对抗样本的目的是最大化 $D(\mu,\upsilon)$来产生对抗扰动。如图 3 所示,基于修改数据的对抗防御的研究脉络主要围绕对抗训练进行推广。首先,将不同的对抗攻击方法融入到对抗训练框架内,便诞生了如 PGD 方法、集成对抗训练等;将小样本学习的思想融入到对抗训练中,解决了标注困难的问题,便衍生出了虚拟对抗训练方法;经典的对抗训练是利用了模型的信息,容易造成标签泄露的问题,而将数据分布的信息引入对抗样本的生成中,便产生了基于特征散射的对抗训练。

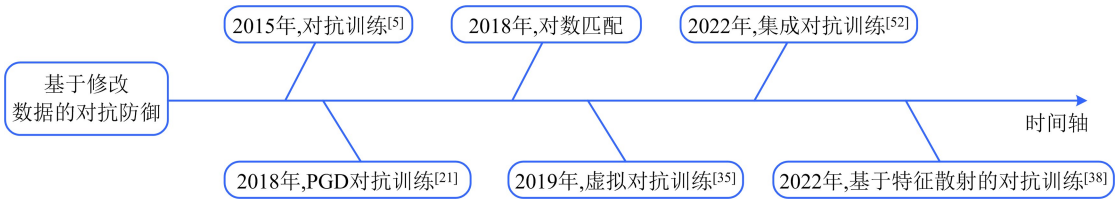


图 3 基于修改数据的对抗防御方法研究脉络

Fig. 3 Research lineage of adversarial defense methods based on modifying data

2.2.2 基于修改模型的防御

最早尝试通过改进模型来提高对抗鲁棒性的工作是 GU 等[39]提出的深度收缩网络(deep contractive networks, DCN),通过在对抗样本中加入噪声来破坏对抗样本,再利用降噪自编码器[38]进行预处理,发现降噪自编码能够去除大量的对抗噪声,但是如果把去噪自编码和原来的深度网络堆叠起来,新的深度网络更容易受到对抗样本的攻击,然后又对压缩自编码(contractive auto encoders, CAE)[41]进行分析和实验。CAE 是在自编码损失函数的基础上引入分层收缩惩罚,使得输入的微小变化不会给隐层激活值带来太大改变,从而使输出变量对一定范围内输入变量的变化不敏感。由于损失函数的平滑度惩罚项增强了 CAE 的泛化能力,因此对于对抗样本这种输入变化不明显的样本具有较好的鲁棒性。实验证明该网络可提高神经网络对对抗样本的鲁棒性。

PAPERNOT 等[44-45]基于蒸馏网络[45]提出了一种提升神经网络对抗鲁棒性的“防御性蒸馏”思想。其基本思想是利用网络的知识来提高自身鲁棒性。首先,根据原始的训练样本 X 和标签 Y 训练一个初始的深度神经网络,网络会输出用于分类判决的概率类向量$\boldsymbol{F}(X)$;然后,再利用样本 X 和概率向量 $\boldsymbol{F}(X)$作为软标签训练一个结构完全相同的蒸馏网络,得到新的概率向量 $\boldsymbol{F}^d(X)$,利用新的蒸馏网络进行分类识别。研究表明,这种方法可以提高网络对图像中的小扰动的适应性。蒸馏的过程使得蒸馏网络学到的样本知识不仅被编码到学习到的权重中,也被编码到网络输出的概率向量中。

2.2.3 增加辅助模型的防御

AKHTAR 等[54]提出了一种针对通用对抗扰动的防御框架。该框架在目标网络上附加了额外的“预输入”层,并训练它们对扰动的图像进行矫正,使分类器的预测与它对同一图像的干净版本的预测相同。预输入层被称为扰动矫正网络(perturbation rectifying network, PRN),它们的训练不需要更新目标网络的参数。通过从训练图像的 PRN 的输入输出差异中提取特征来训

练一个单独的检测器。测试数据首先通过 PRN，然后使用其特征来检测扰动。若检测到对抗性扰动，PRN 的输出则被用来对测试图像进行分类。

KINGMA 等[55]使用生成对抗网络的流行框架来训练一个对 FGSM 类似攻击具有鲁棒性的网络，提议沿着一个试图为该网络产生扰动的生成器网络直接训练该网络；在其训练过程中，分类器不断尝试对干净和扰动的数据进行正确分类。这种技术可以被归类为“附加”方法。在另一个基于 GAN 的防御中，SHEN 等[56]使用网络的生成器部分来校正一个扰动的数据。

MENG 等[57]提出了一个框架，该框架使用 1 个或多个外部检测器将输入数据分类为对抗性或清洁性。在训练期间，该框架旨在学习干净数据的流形。在测试阶段，出现远离流形的数据被视为对抗性数据而被拒绝的情况 。接近流形的数据（但不完全在流形上）总是被改造成位于流形上，分类器被送入改造后的数据。将附近的数据吸引到干净的数据流形上，而放弃远处的数据，该框架因这一概念而被命名为 MagNet。

3　发展趋势

通过对电磁对抗样本攻击和对抗样本防御的发展脉络进行梳理，可以总结规律，指导下一步的研究。本节将结合前文总结的规律以及电磁频谱控制与利用的特殊需求，从攻击和防御两方面讨论下一步发展趋势。

3.1　电磁对抗攻击发展趋势

未来，电磁对抗攻击面临的问题主要有攻击模型未知条件的对抗样本、物理世界中的对抗样本以及多传感器综合条件下的对抗样本等。

3.1.1　适应跨模型和跨任务的对抗样本

未来的频谱控制与利用场景，将会面临敌方系统模型未知和任务未知的场景。模型未知指敌方用于实现电磁频谱控制与利用的深度学习模型的结构和参数未知；任务未知指敌方的系统可能是多环节的智能化，如信号检测、信号识别等，因此需要设计针对不同任务的对抗攻击方法。上述 2 个问题归纳为科学问题就是要研究适应跨模型、跨任务的对抗攻击方法。其中，跨模型的对抗攻击可以借鉴基于迁移的黑盒对抗攻击方法，即虽然对手的模型未知，但是己方可以在已知模型上设计对抗样本，然后对未知模型也能生效。对抗样本的可迁移性已被证明是普遍存在的，关键是研究不依赖于模型信息的对抗攻击方法。跨任务的对抗攻击，首先要设计不依赖于模型信息的对抗攻击方法，其次应解决用于不同任务的模型对数据的采样率、细粒度需求不同的问题。

3.1.2　物理世界中的对抗样本

2.1.2 节中梳理了电磁信号对抗攻击的研究情况，大部分的研究都是在计算机仿真等理想条件下进行的。对抗样本与被干扰样本相比，往往非常“微小”。在现实世界中，当微小的对抗样本叠加在其他相对大功率的样本中时（即大小信号叠加），电磁发射或接收系统中各类物理器件（如放大器等）、信道等的非线性效应会对功率相对很小的对抗样本产生极大的抑制作用。文献[8-9]初步探索了电磁对抗攻击如何适应无线信道的影响。下一步，还需要考虑如接收机、发射机等更多的真实物理环境影响下的电磁对抗攻击。从 2.1 节总结的对抗样本攻击的发展规律来看，下一步研究真实世界下电磁信号对抗样本的关键是对真实物理世界的因素进行准确的建模。比如，发射机、接收机中的滤波器也是制约电磁信号对抗样本性能的重要因素，那么则需要将合适的滤波器模型约束到公式的对抗样本优化框架中。同时，约束后的优化目标仍然要求是可微的模型，便于采用梯度下降法求解最终的对抗样本表达式。

3.1.3　对抗综合传感器体系的对抗样本

在频谱控制与利用场景中，往往同时存在雷达、电子侦察，甚至光电等多类传感器，且未来的单一传感器也可能综合利用多种特征实现电磁频谱的控制与利用任务。现有的工作大多只关注对单一传感器的对抗攻击技术，无法攻击利用多特征的多种传感器。已有研究发现，深度学习在多源数据处理任务结构上的相似性会导致耦合攻击的风险，该成果可以将光学识别模型的正确率由 92.36％降低到 30.98％，同时将 SAR 识别模型的正确率由 81.24％降低到 42.07％[58]。体系智能将会成为未来智能化的趋势，美国海军曾发布了对抗综合传感器的多元素信号特征网络仿真（netted emulation of multi-element signature against integrated sensors，NEMESIS）项

目，旨在将多种前沿的电子战概念融合，实现对广阔区域内敌方多种传感器的迷惑、欺骗或致盲。对抗样本将有望作为对抗综合传感器的有效手段之一。

3.2 电磁对抗防御发展趋势

提升未来智能化电磁频谱控制与利用的对抗防御能力，应从优化流程和优化模型 2 个方面入手。

3.2.1 流程优化

从优化流程的角度来说，可以从数据层面进行优化，进而增强智能系统的对抗防御能力。通过在推理阶段建立对抗样本检测环节，判断接收信号是否含有恶意的对抗样本数据，从而设计专门的对抗样本处理环节进行后续处理或者选择拒绝处理对抗样本。对抗样本检测的核心思想是在推理阶段检测可靠泛化区域外的异常样本。常用的方法有集成检测、度量检测、一致性检测和生成检测。

3.2.2 模型优化

从提升模型的鲁棒性的角度来说，应该关注以下 2 个问题。

3.2.2.1 鲁棒性与泛化性的权衡

尽管已有大量关于对抗防御的研究，深度学习模型仍然面临泛化性和鲁棒性之间权衡的挑战，即模型的泛化性和鲁棒性是矛盾的[32]，希望以较低的泛化性成本实现相当水平的鲁棒性。即使在最强的攻击环境下，是否还能获得准确性和鲁棒性兼顾的模型，仍然是一个开放的问题[59-64]。聚焦到电磁对抗样本防御方面，本文在 2.2 节中总结了对抗防御和提升对抗鲁棒性的方法。为了进一步提高模型的对抗防御能力，研究者尝试了不同方法。对于电磁信号处理而言，已有很多传统的信号处理手段对复杂环境中的信号进行提取、分析等。将专家知识嵌入到对抗防御中有 2 种思路，分别是信号变换和将知识作为目标函数的约束条件。

1）信号变换。研究表明，一些简单的变换能够提高模型的对抗防御性能。WONG 等[65]基于神经网络的线性导致对抗脆弱性的假设，提出对输入数据进行量化和离散化，有效地减小了对抗样本的微小扰动所产生的影响。DHILLON 等[66]在测试时采用随机裁剪模型的某些激活层来减少模型对对抗样本所带来的在输入上的微弱扰动的响应。这些防御与模型无关，意味着不需要重新训练或微调模型。考虑到电磁信号处理领域已经积累了大量的信号变换手段用于抑制各类干扰和噪声，可以将信号变换的防御方法与其他防御方法结合起来。

2）将知识作为约束。深度学习面临可解释性差的挑战，也制约了其鲁棒性。已有研究指出，鲁棒的深度学习模型，其提取的特征更具有可解释性[31]。通过将专家知识作为深度学习的目标函数，有望推动深度学习模型的可解释性研究，增强模型鲁棒性。GUO 等[67]采用低通滤波器作为约束，缩小了模型学习的搜索空间，不仅能够降低计算复杂度，还能使学到的特征具有更多的低频成分。LONG 等[68]也从频域的角度分析了模型的鲁棒性，提出结合专家知识进行有针对性的数据增强可以改善模型的鲁棒性。在未来，深入探究结合专家知识和智能学习的鲁棒性模型将很好地帮助人们解决鲁棒泛化性问题。

3.2.2.2 鲁棒性评估

一般认为，深度神经网络模型对随机噪声的鲁棒性被称为标准鲁棒性。而对抗鲁棒性 ρ_p^{ϵ} 则是描述模型对对抗攻击的适应性能[2]。对抗鲁棒性评价也是一个比较重要的问题，如何正确地评判鲁棒性可以为模型的选择和优化提供准确的指导。首先是通过模型在特定对抗攻击下的性能损失来评价模型的对抗鲁棒性，定义为：

$$\rho_p^{\epsilon}(f_{\boldsymbol{\theta}})=P_{(\boldsymbol{x},y)\sim D}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}+\boldsymbol{r}_p^*(\boldsymbol{x}))=y\right] \tag{21}$$

这种定义方式突出了深度神经网络对某些对抗性攻击的脆弱性。

但是，很多研究者认为单凭模型在某些特定对抗样本的分类效果不足以评判该模型的鲁棒性，因为还可能存在一些潜在的攻击使模型失效[65,69-71]。事实上，衡量一个分类器的“真正”对抗鲁棒性是很有挑战性的，目前的对抗性扰动 $\boldsymbol{r}_p^*(\boldsymbol{x})$ 的计算均是近似解，因此，对抗鲁棒性的评估也是如此。尽管如此，仍然可以通过检查一个分类器的决策边界和所有数据样本之间是否存在安全距离来验证其鲁棒性。也就是说，如果一个分类器总是在任何典型数据样本周围半径为 ϵ 的 l_p 范数球中输出一个恒定的标签，那么就可以认为其在 l_p 范数意义上是鲁棒的。WENG 等[69]提出 CLEVER 准则去评估模型的鲁棒性，理论上证明了 CLEVER 评分是生成对抗样本的最小

半径的 l_p 范数的上界，其物理意义指的是在最小特定半径内不会受到对抗样本的攻击、CLEVER 评分越大，最小攻击半径越大，则鲁棒性越强，因此，可以作为鲁棒性的评价标准，且与攻击类型无关，一并提出了极限值理论方法来估计 CLEVER 评分。FRANCESCO 等[70]提出了针对所有 l_p 范数（$p \geqslant 1$）扰动的鲁棒性评价准则 MMR-Universal，并证明了此准则能得到对于 1、2 和无穷范数的扰动更紧致的鲁棒性上界和下界。此外，一些研究者也提出了针对于 Relu 激活函数的神经网络的鲁棒性评价。不过，在高维空间下获得这样的评估需要相当大的计算负担，而且为了便于操作，鲁棒性评估通常被限制在特定类型的分类器中。该领域有待进一步的探索。

4　结束语

基于深度神经网络模型的新一代智能化识别与处理系统已逐步在电磁频谱控制与利用的多个环节广泛部署。然而现有深度识别模型缺乏可解释性，容易受到对抗攻击的威胁，只能提供有限的可靠性能保证，给模型在复杂电磁环境下强对抗场景中的实际应用带来严重的安全隐患。首先，从对抗样本攻防的角度总结了深度学习模型对抗脆弱性成因：一方面是由于输入与状态空间的高维特性导致了深度学习的脆弱性；另一方面是深度学习模型缺乏可解释性。下一步，研究方向是将专家知识编码到模型学习的过程中。其次，总结了对抗样本攻防以及对抗鲁棒性评估等方面的研究进展。通过梳理通用领域对抗样本的研究规律，总结出电磁信号对抗样本研究可以借鉴的共性规律。通过分析电磁信号对抗样本的研究进程，指出电磁信号对抗样本研究需要关注的个性问题：即在真实物理世界中研究对抗样本需要考虑发射/接收系统和无线信道的效应，以及无论是在攻击还是防御中，均需要结合领域知识。最后，给出了下一步的研究建议，为建立鲁棒可信的高性能智能化电磁频谱控制与利用系统提供参考。

参 考 文 献

[1] HAIGH K Z, ANDRUSENKO J. Cognitive electronic warfare: an artificial intelligence approach[M]. Boston: Artech House, 2021.

[2] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the 2nd International Conference on Learning Representations. [S. l. : s. n.], 2014: 1-10.

[3] PIKNER L C S. Leveraging multi-domain military deception to expose the enemy in 2035[R]. [S. l. : s. n.], 2021: 81-87.

[4] ZHANG L A, HARTNETT G S, AGUIRRE J, et al. Operational feasibility of adversarial attacks against artificial intelligence[R]. RAND Corporation, 2022.

[5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2015: 1-11.

[6] KIM B, SAGDUYU Y, ERPEK T, et al. Adversarial attacks on deep learning based mmwave beam prediction in 5G and beyond[C]//Proceedings of 2021 IEEE Statistical Signal Processing Workshop. [S. l. : s. n.], 2021: 590-594.

[7] KIM B, SAGDUYU Y E, ERPEK T, et al. Channel effects on surrogate models of adversarial attacks against wireless signal classifiers[C]//Proceedings of 2021 IEEE International Conference on Communications. [S. l.]: IEEE, 2021: 1-6.

[8] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers[J]. IEEE Transactions on Wireless Communications, 2021, 21(6): 3868-3880.

[9] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels[C]//Proceedings of the 54th Annual Conference on Information Sciences and Systems. [S. l.]: IEEE, 2020: 1-6.

[10] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. How to make 5G communications "Invisible": adversarial machine learning for wireless privacy[C]//Proceedings of the 54th Asilomar Conference on Signals, Systems, and Computers. [S. l.]: IEEE, 2020: 763-767.

[11] HOU T, WANG T, LU Z, et al. IoTGAN: GAN powered camouflage against machine learning based IoT device identification[C]//Proceedings of 2021 IEEE International Symposium on Dynamic Spectrum Access Networks. [S. l.]: IEEE, 2021: 280-287.

[12] SHI Y, ERPEK T, SAGDUYU Y E, et al. Spectrum data poisoning with adversarial deep learning[C]// Proceedings of 2018 IEEE Military Communications Conference. [S. l.]: IEEE, 2018: 407-412.

[13] SHI Y, DAVASLIOGLU K, SAGDUYU Y E. Generative adversarial network in the air: deep adversarial learning for wireless signal spoofing[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 7(1): 294-303.

[14] KOKALJ-FILIPOVIC S, MILLER R. Adversarial examples in RF deep learning: detection of the attack and its physical robustness[J]. IEEE Wireless Communications Letters, 2019, 8(1): 213-216.

[15] KOKALJ-FILIPOVIC S, MILLER R, CHANG N, et al. Mitigation of adversarial examples in RF deep classifiers utilizing autoencoder pre-training[C]//Proceedings of 2019 International Conference on Military Communications and Information Systems. [S. l.]: IEEE, 2019: 1-6.

[16] 黄知涛, 柯达, 王翔. 先进电磁频谱智能攻击与防御发展及启示[J]. 国防科技, 2023, 44(1): 5-11.
HUANG Zhitao, KE Da, WANG Xiang. Advanced electromagnetic spectrum intelligent attack and defense development and insights[J]. National Defense Technology, 2023, 44(1): 5-11. (in chinese)

[17] NGUYEN A T, RAFF E. Adversarial attacks, regression, and numerical stability regularization[C]//Proceedings of Engineering Dependable and Secure Machine Learning Systems. [S. l.: s. n.], 2019:1-8.

[18] KOS J, FISCHER I, SONG D. Adversarial examples for generative models[C]//Proceedings of 2018 IEEE Security and Privacy Workshops. [S. l.: s. n.], 2018: 36-42.

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

[20] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. [S. l.: s. n.], 2017: 39-57.

[21] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of International Conference on Learning Representations. [S. l.: s. n.], 2018: 1-23.

[22] FAWZI A, FROSSARD P. Manitest: are classifiers really invariant? [C]//Proceedings of the British Machine Vision Conference. [S. l.: s. n.], 2015:1-13.

[23] ENGSTROM L, TRAN B, TSIPRAS D, et al. Exploring the landscape of spatial robustness[C]//Proceedings of the 36th International Conference on Machine Learning. [S. l.: s. n.], 2019: 1802-1811.

[24] KANBAK C, MOOSAVI-DEZFOOLI S M, FROSSARD P. Geometric robustness of deep networks: analysis and improvement[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.: s. n.], 2018: 4441-4449.

[25] LAIDLAW C, FEIZI S. Functional adversarial attacks [C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.: s. n.], 2019:1-11.

[26] SHARIF M, BAUER L, REITER M K. On the suitability of Lp-Norms for creating and preventing adversarial examples[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S. l.: s. n.], 2018: 1605-1613.

[27] WU K, WANG A, YU Y. Stronger and faster wasserstein adversarial attacks[C]//Proceedings of the 37th International Conference on Machine Learning. [S. l.: s. n.], 2020: 10377-10387.

[28] WONG E, SCHMIDT F, KOLTER Z. Wasserstein adversarial examples via projected Sinkhorn iterations [C]//Proceedings of the 36th International Conference on Machine Learning. [S. l.: s. n.], 2019: 6808-6817.

[29] KATZ G, BARRETT C, DILL D, et al. Reluplex: an efficient SMT solver for verifying deep neural networks[C]//Proceedings of the 29th International Conference on Computer Aided Verification. [S. l.: s. n.], 2017:1-31.

[30] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.: s. n.], 2012: 1-7.

[31] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features [C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.: s. n.], 2019:1-37.

[32] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[C]//Proceedings of International Conference on Learning Representations. [S. l.: s. n.], 2019:1-24.

[33] LE Q V. Building high-level features using large scale unsupervised learning[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.: s. n.], 2013: 8595-8598.

[34] LYU C, HUANG K, LIANG H N. A unified gradient regularization family for adversarial examples [C]//Proceedings of 2015 IEEE International Conference on Data Mining. [S. l.]: IEEE, 2015: 301-309.

[35] KURAKIN A, GOODFELLOW I J, BENGIO S. Ad-

versarial machine learning at scale[C]//Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2016: 1-17.

[36] MIYATO T, MAEDA S I, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.

[37] ZHANG J, XU X, HAN B, et al. Attacks which do not kill training make adversarial learning stronger [C]//Proceedings of International Conference on Machine Learning. [S. l. : s. n.], 2020: 11278-11287.

[38] ZHANG H, WANG J. Defense against adversarial attacks using feature scattering-based adversarial training[J]. Advances in Neural Information Processing Systems, 2019, 32: 1-11.

[39] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[C]//Proceedings of 2015 International Conference on Learning Representations Workshop. [S. l. : s. n.], 2015: 1-9.

[40] BENGIO Y. Learning deep architectures[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127.

[41] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: explicit invariance during feature extraction[C]//Proceedings of the 28th International Conference on Machine Learning. [S. l. : s. n.], 2011: 833-840.

[42] ROS A S, DOSHI-VELEZ F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]// Proceedings of the 32th Conference on Artificial Intelligence. [S. l. : s. n.], 2018: 1-10.

[43] DRUCKER H, LE CUN Y. Improving generalization performance using double backpropagation[J]. IEEE Transactions on Neural Networks, 1992, 3(6): 991-997.

[44] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[C]//Proceedings of 2014 Conference and Workshop on Neural Information Processing Systems. [S. l. : s. n.], 2014: 1-9.

[45] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of 2016 IEEE Symposium on Security and Privacy. [S. l. : s. n.], 2016: 582-597.

[46] SADEGHI M, LARSSON E G. Adversarial attacks on deep-learning based radio signal classification[J]. IEEE Wireless Communications Letters, 2019, 8(1): 213-216.

[47] KE D, HUANG Z T, WANG X, et al. Application of adversarial examples in communication modulation classification[C]//Proceedings of 2019 International Conference on Data Mining Workshops. Beijing, China: IEEE, 2019: 877-882.

[48] LIN Y, ZHAO H, MA X, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2021, 70(1): 389-401.

[49] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[C]//Proceedings of Advances in Neural Information Processing Systems. [S. l. : s. n.], 2014: 1-9.

[50] HAMEED M Z, GYORGY A, GUNDUZ D. The best defense is a good offense: adversarial attacks to avoid modulation detection[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1074-1087.

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

[52] MIYATO T, DAI A M, GOODFELLOW I. Adversarial training methods for semi-supervised text classification[C]//Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2017: 1-17

[53] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses [C]//Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2022: 1-22.

[54] AKHTAR N, LIU J, MIAN A. Defense against universal adversarial perturbations[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l. : s. n.], 2018: 3389-3398.

[55] KINGMA D P , WELLING M . Auto-encoding variational Bayes[C]//Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2013: 1-9.

[56] JIN G, SHEN S, ZHANG D, et al. APE-GAN: adversarial perturbation elimination with GAN[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l. : s. n.], 2019: 3842-3846.

[57] HE W, WEI J, CHEN X, et al. Adversarial example defense: ensembles of weak defenses are not strong [C]//Proceedings of the 11th USENIX Workshop on Offensive Technologies. [S. l. : s. n.], 2017: 1-11.

[58] 孙浩，陈进，雷琳，等. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述[J]. 雷达学报，2021，10

(4): 571-594.
SUN Hao, CHEN Jin, LEI Lin, et al. Adversarial robustness of deep convolutional neural network-based image recognition models: a review[J]. Journal of Radars, 2021, 10(4): 571-594. (in chinese)

[59] RAGHUNATHAN A, XIE S M, YANG F, et al. Understanding and mitigating the tradeoff between robustness and accuracy[C]//Proceedings of the 37th International Conference on Machine Learning. [S. l.: s. n.], 2020: 7909-7919.

[60] ALAYRAC J B, UESATO J, HUANG P S, et al. Are labels required for improving adversarial robustness? [C]//Proceedings of Conference and Workshop on Advances in Neural Information Processing Systems. [S. l.: s. n.], 2019: 1-10.

[61] CARMON Y, RAGHUNATHAN A, SCHMIDT L, et al. Unlabeled data improves adversarial robustness [C]//Proceedings of Conference and Workshop on Advances in Neural Information Processing Systems. [S. l.: s. n.], 2019: 25-37.

[62] TAORI R, DAVE A, SHANKAR V, et al. Measuring robustness to natural distribution shifts in image classification [C]//Proceedings of Conference and Workshop on Advances in Neural Information Processing Systems. [S. l.: s. n.], 2020: 18583-18599.

[63] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proceedings of International Conference on Machine Learning. [S. l.: s. n.], 2019: 7472-7482.

[64] LAMB A, VERMA V, KAWAGUCHI K, et al. Interpolated adversarial training: achieving robust neural networks without sacrificing too much accuracy[J]. Neural Networks, 2022, 154: 218-233.

[65] WONG E, KOLTER Z. Provable defenses against adversarial examples via the convex outer adversarial polytope[C]//Proceedings of International Conference on Machine Learning. [S. l.: s. n.], 2018: 5286-5295.

[66] DHILLON G S, AZIZZADENESHELI K, LIPTON Z C, et al. Stochastic activation pruning for robust adversarial defense [C]//Proceedings of International Conference on Learning Representations. [S. l.: s. n.], 2018: 1-13.

[67] GUO C, FRANK J S, WEINBERGER K Q. Low frequency adversarial perturbation [C]//Proceedings of Conference on Uncertainty in Artificial Intelligence. [S. l.: s. n.], 2020: 1127-1137.

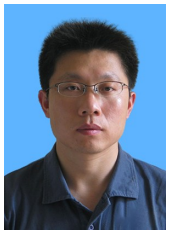
[68] LONG Y, ZHANG Q, ZENG B, et al. Frequency domain model augmentation for adversarial attack[C]//Proceedings of European Conference on Computer Vision. [S. l.: s. n.], 2022: 549-566.

[69] WENG T W, ZHANG H, CHEN P Y, et al. Evaluating the robustness of neural networks: an extreme value theory approach [C]//Proceedings of International Conference on Learning Representations. [S. l.: s. n.], 2018: 1-18.

[70] CROCE F, HEIN M. Provable robustness against all adversarial l_p-perturbations for $p>1$ [C]//Proceedings of International Conference on Learning Representations. [S. l.: s. n.], 2020: 1-20.

[71] JORDAN M, LEWIS J, DIMAKIS A G. Provable certificates for adversarial examples: fitting a ball in the union of polytopes[C]//Proceedings of Conference and Workshop on Advances in Neural Information Processing Systems. [S. l.: s. n.], 2019, 32: 1-11.

作 者 简 介



黄知涛
男,1976 年生,博士,教授,博士研究生导师,研究方向为电子对抗
E-mail:huangzhitao@nudt.edu.cn

柯 达
男,1994 年生,博士研究生,研究方向为电子侦察
E-mail:1747884404@qq.com

王 翔
男,1985 年生,博士,副教授,研究方向为电子对抗
E-mail:christopherwx@163.com

责任编辑 安 蓓