计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(7): 1713−1737, 2025

面向信号调制识别的对抗攻击与防御综述

徐冬月[1] 田蕴哲[1] 陈 康[1] 李轶珂[1] 吴亚伦[1] 童恩栋[1] 牛温佳[1] 刘吉强[1] 史忠植[2]

[1](智能交通数据安全与隐私保护北京市重点实验室(北京交通大学) 北京 100044)

[2](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

(dongyuexu@bjtu.edu.cn)

Survey on Adversarial Attack and Defense for Signal Modulation Recognition

Xu Dongyue[1], Tian Yunzhe[1], Chen Kang[1], Li Yike[1], Wu Yalun[1], Tong Endong[1], Niu Wenjia[1], Liu Jiqiang[1], and Shi Zhongzhi[2]

[1](*Beijing Key Laboratory of Security and Privacy in Intelligent Transportation* (*Beijing Jiaotong University*), *Beijing* 100044)

[2](*Key Laboratory of Intelligent Information Processing*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract With the rapid development of deep learning, signal modulation recognition based on deep neural networks has gained popularity in wireless communications research. However, it has been observed that the deep neural network model is vulnerable to adversarial perturbations, rendering the modulation identification task ineffective. Currently, there are theoretical gaps and bottlenecks in wireless communication security research. Due to the multidimensional nature of wireless communication, including factors such as experimental environments, data structures, and signal characteristics, it is not feasible to transfer the established attack and defense methods from other domains to signal countermeasures. In this paper, we comprehensively summarize the research on adversarial attack and defense technology in the field of signal modulation recognition. We propose a generic classification framework and threat model for adversarial attacks in this field. And we classify the research in this field into two categories: physical self-defense attacks and digital direct access attacks. Then, we systematically integrate and visualize the research as two-dimensional diagrams to demonstratively showcase the methods, models, and techniques of adversarial attack. Additionally, we provide details on the methods and models of adversarial attack. We present the existing research on adversarial attack methods, adversarial examples generation techniques, theoretical formulas, and adversarial detection and defense techniques. We systematically refine the characteristics of the three dimensions of adversarial attacks on wireless communications and summarize the corresponding processing methods. Finally, we summarize the future research and development direction of the attack and defense security field oriented towards signal modulation recognition.

Key words deep neural network; wireless communication; signal modulation recognition; adversarial attack; defense

摘 要 随着深度学习的飞速发展,基于深度神经网络的信号调制识别任务成为无线通信领域热门研究.研究发现,深度神经网络模型极易受到对抗性扰动的影响,使调制识别任务失效.目前面向无线通信安全性的研究工作仍存在些许瓶颈问题与理论空白,源于无线通信固有的实验环境、数据结构与信号特征等多维度特性,不能将其他领域较为成熟的攻防方法简单迁移到信号对抗攻击中.面向信号调制识别领域

收稿日期: 2023-10-19; **修回日期**: 2025-01-13

基金项目: 国家自然科学基金项目(62372021)

The work was supported by the National Natural Science Foundation of China (62372021).

通信作者: 牛温佳(niuwj@bjtu.edu.cn)

的对抗攻防,总结该领域对抗攻击与防御技术研究工作,提出信号调制识别领域的通用对抗攻击分类框架与威胁模型,将该领域研究工作分类为物理自我防御式攻击和数字直接访问式攻击,并以2维图形式进行系统化整合与可视化展示,详细阐述对抗攻击方法、对抗样本生成技术与理论公式、对抗检测与防御技术最新的研究工作,系统提炼无线通信对抗攻击研究的3个维度特性并归纳相应的处理方法,最后总结面向信号调制识别的攻防安全领域在未来的研究发展方向.

关键词 深度神经网络;无线通信;信号调制识别;对抗攻击;防御

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330826 **CSTR:** 32373.14.issn1000-1239.202330826

信号调制识别是无线通信系统的关键技术,能够在无先验知识的情况下通过对接收信号的处理判断出信号的调制方式. 调制识别在包括认知无线电、频谱感知等民用或军用通信场景中至关重要[1-3].

传统的信号调制识别方法包括基于似然理论和基于特征的调制识别方法. 前者追求最佳识别精度,但计算较为复杂. 后者依赖于信号代表性特征和训练模型进行调制识别. 随着深度学习(deep learning, DL)在如计算机视觉(computer vision, CV)和自然语言处理(natural language processing, NLP)等领域取得突破性进展,其在信号调制识别领域的应用也逐渐受到关注[4-12]. 相较于传统方法,深度学习方法利用神经网络自动提取特征,对未知信号具有较强的泛化能力,性能优于传统方法,适应多变信号环境,在信号识别领域取得突破性进展.

然而研究发现,深度神经网络(deep neural networks, DNNs)极易受到对抗性扰动影响,这在CV领域已得到证实[13],即在输入中添加微小的扰动使DNNs识别分类错误. 对抗样本的存在揭示了神经网络在高维空间中的线性性质[14]. 在无线通信领域,由于无线媒介的共享性与广播性,进一步降低了基于深度学习的无线通信任务的安全性[15]. Sadeghi等人[16]将对抗攻击引入无线通信领域,验证了对抗样本能够降低信号识别的性能,随后研究人员在此基础上进一步深入探究[17-23]. 对抗深度学习在无线通信领域处于初步阶段,针对调制识别的对抗攻击主要集中在提高攻击性能上[24],不能直接应用CV领域的成熟方法开展研究,且对抗样本解释大多局限于假设场景中,仍存在理论研究空白.

为应对对抗攻击,部分研究人员进一步提出了多种对抗防御方法,如Kim等人[15,25-26]利用对抗训练防御机制迭代生成对抗样本,以增强模型鲁棒性;再如Kokalj-Filipovic等人[27]利用统计学防御方法,通过接收信号的统计量特征进行对抗检测. 尽管这些防御策略初显成效,但该领域的对抗防御技术尚处于早期阶段,防御机制的泛化能力和可验证鲁棒性仍需深入研究.

虽有多篇综述文献讨论对抗攻击与防御,但面向信号识别任务的对抗攻防研究仍缺乏系统性总结梳理工作. 文献[28]是侧重于整体无线通信系统的对抗机器学习英文综述,本文作为侧重于信号调制识别的对抗攻防研究的中文综述,与此篇并不重复. 本文工作包括提出调制识别领域的通用对抗攻击威胁模型,总结3个维度通信特性问题以及相对应的对抗研究处理方法,详细梳理9种对抗样本生成技术和3种对抗防御方法,讨论探究该领域研究的3个未来发展方向. 本文的主要贡献体现在3个方面:

1)总结面向信号识别领域的对抗攻击与防御技术已有的研究工作,提出了全新的对抗攻击威胁模型框架,将该领域研究工作分类为物理自我防御式攻击和数字直接访问式攻击,并以2维图形式进行系统化整合与可视化展示来提高整体直观性与方法间的相关性,利于发掘对抗研究空白与痛点.

2)详细阐述面向调制识别的对抗样本生成技术方法与理论公式,相比于文献[28]仅仅给出无线通信领域对抗攻击粗略框架,本文提供更具细粒度的对抗攻击方法.

3)总结调制识别与其他领域对抗研究的特性问题,从3个角度(环境、数据、特征)提出通信特性问题,并归纳通信特性应用于对抗攻击的详细处理方法. 本文整体框架图如图1所示.

1 预备知识

1.1 面向信号调制识别的深度学习

1.1.1 调制识别的任务描述

Kim等人[15,19,29]在对信号调制识别进行对抗攻击研究中考虑到真实物理场景下的无线通信系统,该

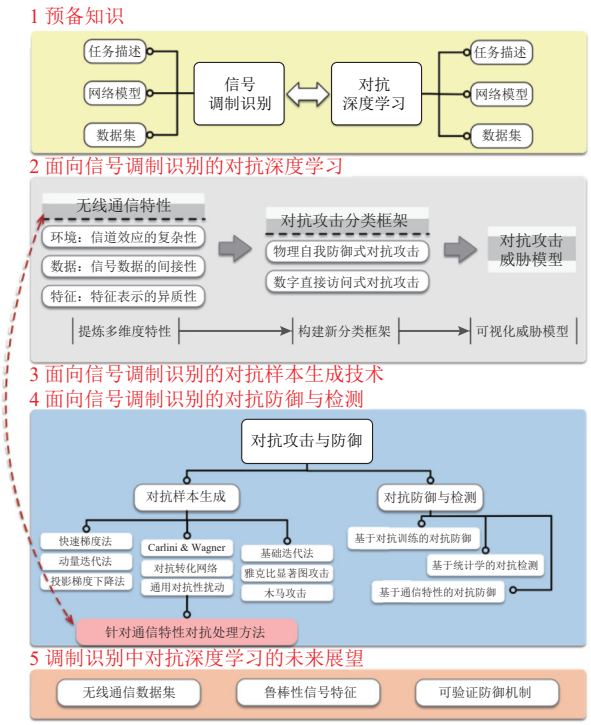
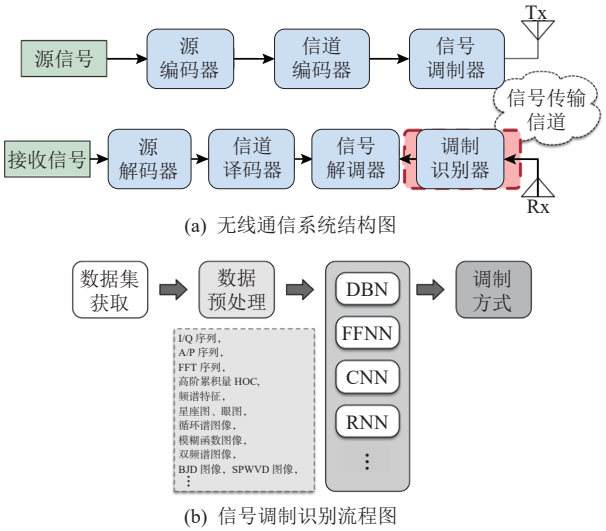


Fig. 1 Overall structural diagram

图 1 整体架构图



(a) 无线通信系统结构图

(b) 信号调制识别流程图

Fig. 2 Structure diagram of wireless communication system and signal modulation recognition based on DL

图 2 无线通信系统与基于深度学习的信号调制识别结构图

系统由发射器、信道、接收器组成，如图 2(a)所示. 无线通信系统原始信号通过采样和量化处理进行数字化，再对数字信号进行编码，以提高数据的安全性，减少错误传输；调制器根据预先设定的调制方式将数字基带信号转换成适于信道传输的无线信号；在信道上，传输信号受到偏移、衰减等影响，同时由于接收器的硬件缺陷，噪声也随传输信号一起被接收器接收；对于非合作信号应用场景下，接收器对于接收信号的先验知识有限，故在解调信号之前需要对接收信号进行调制识别，再通过解码器对信号进行解调以重构发射器发送的信息[30].

信号调制识别作为无线通信系统中信号解调的一个先导过程[31]，旨在识别检测到的信号的调制方式，以帮助正确解调收到的信号，在频谱监测、频谱管理和安全通信等通信应用场景下发挥着关键作用，同时也成为通信信号领域研究的热点. Erpek 等人[32]侧重于介绍深度学习应用于物理层通信方面的研究进展，Peng 等人[33]侧重于信号表示方式和数据预处理方法对基于深度学习的调制识别研究进行总结，Zhou 等人[34]侧重于各类深度学习的算法模型，对信号调制识别进行介绍.

具体来说，基于深度学习的调制识别通常包含 3 个步骤，即接收信号的预处理、特征提取和调制分类，如图 2(b)所示. 接收到的信号经过预处理调整为合适的数据格式，以便进行后续的特征提取和模型训练，特征提取和特征识别可以由深度学习模型通过端到端的方式完成，最后得到信号调制方式的预测结果.

信号通过信道传输到接收器的过程如式(1)所示：

$$r_l = A_l \mathrm{e}^{\mathrm{j}(\omega l+\varphi)} x_l + n_l, \tag{1}$$

其中，x_l 为发射器经过某种调制方式调制后发射的信号值，n_l 为加性高斯白噪声(additive white Gaussian noise, AWGN)，A_l 为信道增益，ω 为频率偏移量，φ 为相位偏移量，r_l 为接收器接收到的第 l 个信号值，发射信号 $\boldsymbol{x}$ 和接收信号 $\boldsymbol{r}$ 可以表示为 $\boldsymbol{x}=(x_1, x_2, \cdots, x_L)$, $\boldsymbol{r}=(r_1, r_2, \cdots, r_L)$.

信号调制识别任务，可以看作是一个多分类任务，即利用基于深度学习的调制识别模型来求解目标函数，如式(2)：

$$\arg\min_{\boldsymbol{\theta}} \mathcal{L}(f(\boldsymbol{\theta}, \boldsymbol{x}), \boldsymbol{y}), \tag{2}$$

从而学习得到输入信号样本的制式映射. 其中 $f(\cdot)$ 为调制识别模型的网络架构，对于输入信号 $\boldsymbol{x}$, $\boldsymbol{y}$ 为其对应的调制类别，$\mathcal{L}(\cdot)$ 为损失函数，如交叉熵损失函数，其通常与优化器(如随机梯度下降或 Adam[35])配合使用来训练网络模型，从而学习得到模型参数 $\boldsymbol{\theta}$.

1.1.2 调制识别的开源数据集

深度学习的信号调制识别方法主要依赖于数据集，其中全面且高质量的数据集是关键前提. 表 1 总结了当前通信领域主流的开源调制识别数据集并比较了它们的属性，其中 SNR 为信噪比.

Table 1 Summary of Main Open-Source Datasets for Signal Modulation Recognition

表 1 信号调制识别主要的开源数据集汇总

数据集	调制方式	样本大小	样本总数	SNR/dB
RML2016.10A	11 类（8PSK, BPSK, CPFSK, GFSK, PAM4, 16QAM, AM-DSB, AM-SSB, 64QAM, QPSK, WBFM）	2 × 128	220 000	−20~18
RML2016.10B	10 类（8PSK, BPSK, CPFSK, GFSK, PAM4, AM-DSB, 16QAM, 64QAM, QPSK, WBFM）	2 × 128	1 200 000	−20~18
RML2016.04C	11 类（8PSK, BPSK, CPFSK, GFSK, PAM4, AM-DSB, AM-SSB, 16QAM, 64QAM, QPSK, WBFM）	2 × 128	162 060	−20~18
RML2018.01A	24 类（OOK, 4ASK, 8ASK, BPSK, QPSK, 8PSK, 16PSK, 32PSK, 16APSK, 32APSK, 64APSK, 128APSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, AM-SSB-WC, AM-SSB-SC, AM-DSB-WC, AM-DSB-SC, FM, GMASK, OQPSK）	2 × 1 024	2 555 904	−20~10
HisarMod2019.1	26 类（AM-DSB, AM-SC, AM-USB, AM-LSB, FM, PM, 2FSK, 4FSK, 8FSK, 16FSK, 4PAM, 8PAM, 16PAM, BPSK, QPSK, 8PSK, 16PSK, 32PSK, 64PSK, 4QAM, 8QAM, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM）	2 × 1 024	780 000	−20~18

RML 数据集是 O'Shea 等人[4,36]通过 GNU Radio[37]生成的开源数据集，已被广泛用于基准测试. 该数据集模拟了实际情况下无线系统中常见的时变随机信道效应，包括中心频率偏移、加性高斯白噪声、多路径衰落等. RML2016.10A 和 RML2016.10B 是通过模拟恶劣环境下的传播特性产生的，而 RML2018.01A 是在相对良好的真实实验室环境下产生的. 为引入更全面的数据集，Tekbiyik 等人[38]利用仿真软件 MATLAB 创建名为 HisarMod2019.1 的新数据集，包含 26 个调制类别信号. 该数据集提供了在理想静态服从 Rayleigh 分布、Rician(k = 3)分布和 Nakagami-m(m = 2)分布的信道衰落模型条件下的无线信号.

目前，RML2016.10A 因数据量适中且包含常见调制类型，如 QAM, AM 等而被广泛应用于信号调制识别的研究. RML2016.10B 数据集的数据量更大，需要更多的计算资源. RML2018.01A 数据集中的调制类型增加到 24 种，数据长度从 128 维增加到 1 024 维，在训练模型过程中提高了对算力的要求，但也促进了更先进模型的开发. HisarMod2019.1 包含 26 种调制类型，但该数据集在较理想的环境下生成，使调制方式更易识别.

基于深度学习的信号数据集采样与收集面临资源消耗大等困难. 部分开源数据集往往未考虑真实物理通信环境，限制了其在复杂多变的信道条件下的适用性. 同时，考虑在真实物理环境下进行面向信号调制识别的对抗深度学习研究也将遇到实验严谨性的问题.

1.1.3 调制识别的深度学习模型

目前针对于深度学习的调制识别模型的研究中，具代表性深度学习模型框架如图 3 所示.

1)基于无监督学习的模型

无监督模型如深度信念网络(deep belief network, DBN)和自编码器被应用于信号调制识别. DBN 模型是由多层受限玻尔兹曼机(restricted Boltzmann machine, RBM)组成的概率生成式模型，Mendis 等人[39,42]结合光谱相关函数(SCF)以更有效地学习复杂模型，但计算复杂性限制了其应用于大规模问题. Ali 等人[43]提出的低复杂度 k 稀疏自编码器通过激活 k 个最大的隐藏节点以实现稀疏性，有效进行信号特征提取和调制分类. Dai 等人[44]中采用模糊函数(AF)作为输入，通过稀疏自编码器中进行调制识别. 然而，大多数无监督的深度学习方法只被用来测试少数调制方式的分类任务，未广泛应用于具有普适性的实际场景中.

2)基于前馈神经网络的模型

前馈神经网络(feedforward neural network, FFNN)含多个隐藏层，信息单向从输入节点传递到输出节点，无需反向. 前馈神经网络结构特点使得 DNN 模型在高维空间中提取复杂特征的能力较为出众. Lee 等人[40]采用 4 层 DNN 模型，通过多种统计量特征进行调制分类. Xie 等人[45]使用相似的 DNN 结构和 6 个高阶累积量特征进行调制识别. 为实现自动选择并优化隐藏层节点数，Shi 等人[46]引入粒子群优化(particle swarm optimization, PSO)算法，显著提高分类精度.

3)基于 CNN 的模型

卷积神经网络(convolutional neural network, CNN)在处理空间特征数据(如图像分割、物体检测)具有突出优势，同样 CNN 模型也被成功引入到信号调制识别应用中，通过利用其空间特征提取能力来识别信号制式. 根据输入数据的类型，现有的基于 CNN 的信号调制识别方法可以被大体分为 2 类：原始 I/Q 数据或经过预处理后的信号数据作为输入的 CNN 模型，以及高效 CNN 模型架构. 这 2 类模型满足通信系统的延迟和复杂性要求.

①原始 I/Q 数据作为输入的 CNN 模型. O'Shea 等人[4]使用简单的 4 层 CNN 模型并将 I/Q 数据作为

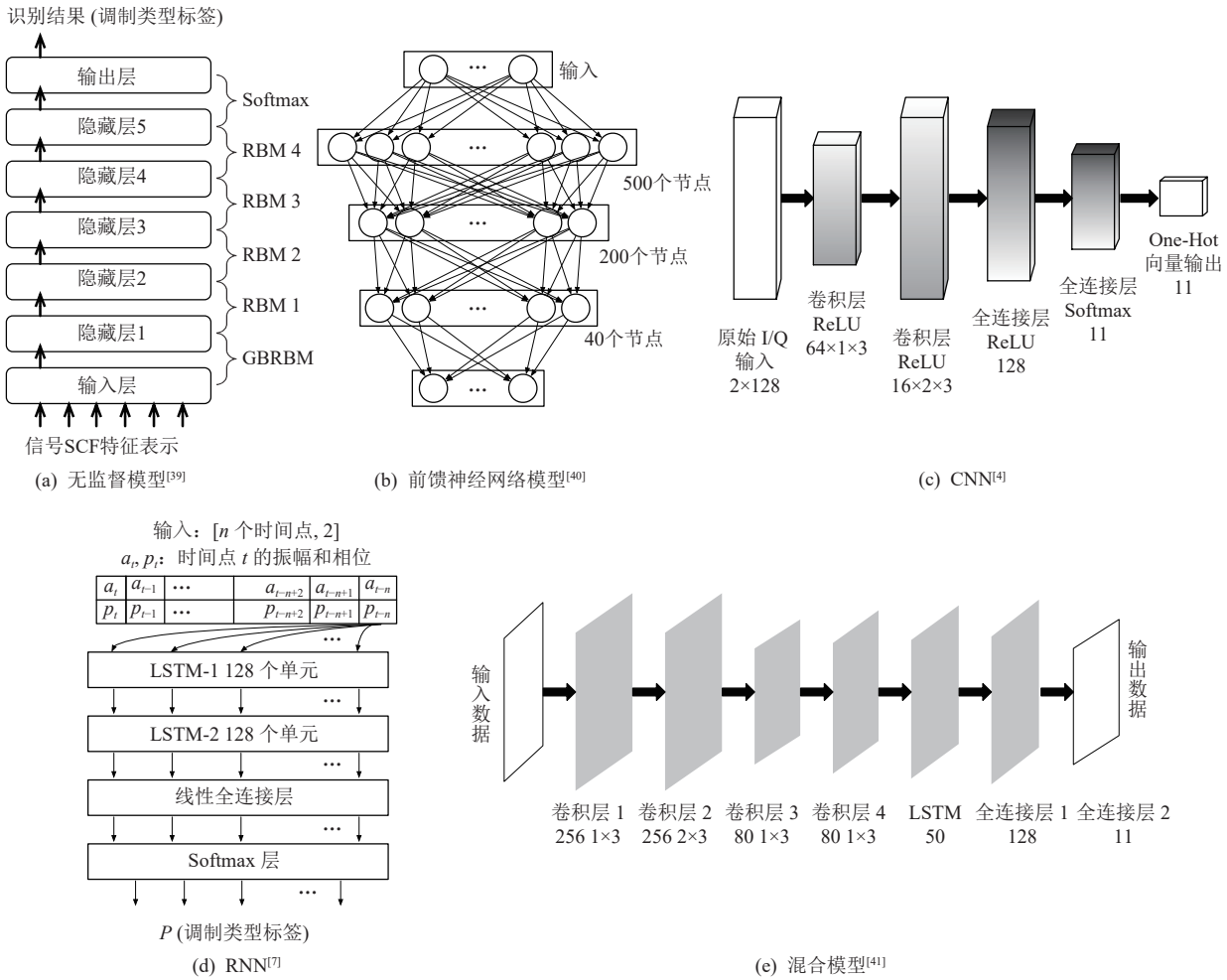


Fig. 3 Representative DL models for modulation recognition

图 3 面向调制识别的代表性深度学习模型

输入，实现了比传统方法更高的识别准确性. 为增强性能，研究者们对 CNN 模型进行改进. Liu 等人[41] 基于 ImageNet 2015 的获胜架构，提出了结合残差神经网络（residual neural network, ResNet）和稠密连接网络（densely connected network, DenseNet）的深度学习模型，高效传输多层特征至识别模块. 但 Liu 等人[41] 在识别准确性方面的提升是以显著增加计算复杂度为代价的，并且忽略了通信系统中固有的信号特征. 为解决这些限制，Yashashwi 等人[47] 提出预估接收信号的载波频率偏移和相位噪声，通过可训练的函数校正接收信号.

②经过预处理后的信号数据作为输入的 CNN 模型. 将基于传统特征（feature-based, FB）的信号调制识别方法与 CNN 模型相结合，解决直接使用原始 I/Q 信号作为输入可能导致的重要特征丢失的问题，包括高阶累积量特征、频谱图像特征和星座图特征等. Zeng 等人[48] 通过短时离散傅里叶变换和高斯滤波器降噪，将 1 维信号数据转换为频谱图像，提升了识别准确率. Peng 等人[49] 将星座图转换为 3 通道图像，利用 AlexNet 和 GoogLeNet 模型进行调制识别，以发挥 CNN 模型的彩色图像处理能力. Wang 等人[50-52] 通过特征融合，结合原始 I/Q 数据、星座图、循环光谱图像等多种特征，以提高调制识别的准确率. 其他信号表示方法，如眼图[53]、特征点（feature point, FP）图像[54] 和方形特征矩阵[55] 也被用来作为信号调制识别模型的输入进行研究.

③高效 CNN 模型架构. 为满足超 5 代（B5G）通信系统的低时延高可靠要求，Hermawan 等人[56] 通过添加滤波器和减少 CNN 模型的可训练参数，实现小于 0.01 ms 的处理时间并保持高的识别精度. Huynh-The 等人[10] 则采用非对称卷积核和跳跃连接的方法，开发低成本高性能 CNN 模型.

4）基于 RNN 的模型

无线通信信号的时间相关性特征可循环神经网络（recurrent neural network, RNN）学习可用于调制识别. Hong 等人[6] 提出的基于 RNN 模型，利用门控循

环单元(GRU)比部分 CNN 模型实现更高的识别精度. Rajendran 等人[7]将 I/Q 数据转化为信号的振幅和相位输入到 LSTM 模型中, 同样得到较高的识别精度. Ke 等人[57]设计基于 LSTM 的去噪自编码器, 该编码器能够在低成本的计算平台上实现, 性能超过之前的模型.

5)混合模型

为提高信号调制识别性能, 研究者们结合 CNN 模型与 RNN 模型的优势, 构建混合模型. West 等人[5]提出卷积长短期深度神经网络(convolutional long-short-term deep neural network, CLDNN)模型融合 LSTM 层和 CNN 层, 通过跳跃连接提供更长时的上下文, 优化时间特征提取, 实现更高的识别精度与更稳定的梯度下降过程. Liu 等人[41]优化 West 等人[5]提出的 CLDNN 模型, 在增加参数和网络层数的代价下, 获得更高的识别精度. Xu 等人[9]提出新型多通道深度学习模型, 从时间和空间 2 个角度提取特征. Chang 等人[58]基于双向门控循环单元(bidirectional gated recurrent unit, BiGRU)结构, 引入步长注意力融合网络(step attention fusion network, SAFN), 优化 I/Q 和 A/P (振幅/相位)信号的融合过程. 除 I/Q 和 A/P 数据之外, 高阶统计量特征也被融合到调制识别任务中[59]. 混合模型虽然提高了识别精度, 但相比于单一模型结构, 成本也有显著提高.

表 2 总结用于信号调制识别的深度学习模型的优缺点.

Table 2 Signal Modulation Recognition for Deep Learning Models

表 2 信号调制识别的深度学习模型

模型类型	模型名称	优点	缺点
无监督学习模型	SCF-DBN[39,42]、自编码器 1[43]、自编码器 2[44]	·处理未标记数据; ·特征提取表现优异.	·计算复杂性较高, 限制大规模问题的应用; ·只用来测试少数调制方式分类, 应用受限.
前馈神经网络模型	FFNN1[40], FFNN2[45], FFNN3[46]	·擅长提取高维特征; ·结构简单, 优化灵活.	·可能不适合处理如时间序列数据; ·需要大量训练数据, 以避免过拟合.
CNN 模型	CNN1[4], CNN2[50], CNN3[51], ResNet[41], DenseNet[41], CM+CNN[47], SCNN[48], AlexNet[49], GoogLeNet[49], DrCNN[52], IC-AMCNet[56], MCNET[10]	·强大的空间特征提取能力; ·适用于不同输入数据类型; ·高效架构满足低时延高可靠.	·高计算复杂度, 增加训练成本; ·可能忽略通信系统中固有的信号特征; ·原始 I/Q 信号直接输入模型导致信息损失.
RNN 模型	GRU[6], LSTM[7], DAE[57]	·处理时间序列数据的能力; ·适用于低成本计算平台.	·计算复杂度相对较高, 训练难度较大; ·可能存在梯度消失或爆炸问题.
混合模型	CLDNN1[5], CLDNN2[41], MCLDNN[9], MLDNN[58], DBN+SNN[59]	·结合 CNN 和 RNN 的优势; ·提高识别精度和泛化能力.	·模型复杂性更高, 计算和维护成本增加; ·计算资源需求更高.

1.2 对抗深度学习

尽管深度学习在各个领域都取得了显著的成功, 但已被证实深度神经网络极易受到对抗性扰动的影响, 即在 DNNs 的输入数据中加入很小的扰动, 导致其对被扰动的输入数据识别分类错误. 形式化地表达为, 对于输入 $\boldsymbol{x} \in \boldsymbol{X}$, 对抗攻击旨在生成不可察觉的对抗性扰动 $\boldsymbol{\delta}$ 来更改输入, 使得模型 $f(\cdot)$ 出现决策失误, 即 $f(\boldsymbol{\theta}, \boldsymbol{x}) \neq f(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta})$. 目前普遍认为对抗样本存在的原因在于神经网络在高维空间中的线性性质[14].

1.2.1 研究现状

对抗深度学习已经在 CV 和 NLP 等领域得到了广泛研究. 例如, 在 CV 领域, 随着无人驾驶汽车和外科医疗手术等应用的发展, 研究人员探究了大量不可察觉的对抗样本产生的影响[60-66]. 基于潜在安全漏洞的认知, 以减轻对抗扰动的影响同时确保分类模型的鲁棒性为目标, 研究人员设计和开发了多种防御算法, 如对抗性训练, 即在训练阶段用对抗样本增加训练数据[13-14,67-69]. 在 NLP 领域, 攻击者可以通过扰乱语音来欺骗基于深度学习的语音识别系统, 如苹果 Siri 等. Carlini 等人[70]通过添加不可察觉的声音扰动, 在最先进的语音识别系统 Deep Speech[71]中制作了音频对抗样本, 攻击成功率达 100%. 郑海斌等人[72]从 NLP 模型、攻防鲁棒分析、平台和工具包的逻辑关系 3 方面对 NLP 对抗深度学习进行系统地总结, 将攻击扰动分为字符级攻击、单词级攻击和句子级攻击, 并分别对 3 种不同粒度的攻击方法展开介绍. 在 NLP 领域的研究工作中也提出了多种防御策略[73-76], 如随机平滑法[74]、有监督的对比性对抗学习(SCAL)[76]和对抗性去伪存真法[77].

相比之下, 对抗深度学习于无线通信领域的研究较为滞后. Sadeghi 等人[16]将对抗攻击引入该领域, 验证对抗样本能够降低信号识别性能, 随后研究人员在此基础上进一步深入探究[17-23,78-96]. 无线通信领域的研究仍处于初步阶段, 对抗攻击集中在提高性能的研究, 对抗样本解释多限于假设场景, 存在理论研究空白.

1.2.2 分类方法

本文参考对抗攻击研究的综述[97]提出了对抗深度学习分类方法, 如图 4 所示.

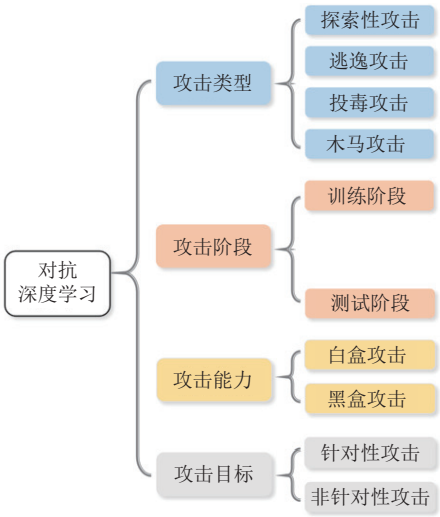


Fig. 4 Categorization framework diagram based on adversarial deep learning

图 4 基于对抗深度学习的分类框架图

1)攻击类型

探索性攻击，也称为推理攻击，其试图通过收集训练数据以及用类似的输入和输出类型模仿目标模型的功能，即建立一个代理模型，来探索目标模型的内部工作原理[98]. 探索性攻击通常是逃逸攻击、投毒攻击、木马攻击的攻击方式之前的领先步骤，因为其旨在运用主动学习[99-100]等技术探索被攻击者模型，或利用生成对抗网络(generative adversarial network, GAN)来增加有限信息[101]. 逃逸攻击，其攻击目的是通过操纵输入的测试数据来骗取目标模型做出错误的决策[13-14]. 投毒攻击，其攻击目的是通过向目标模型注入漏洞(如虚假训练数据)来操纵模型的训练过程[102]. 木马攻击，也称为后门攻击，是逃逸攻击与投毒攻击的结合，攻击者在训练数据中注入触发器(后门)，在测试阶段利用输入样本进行激活[103].

2)攻击阶段

对抗攻击通常发生在目标模型的训练阶段或测试阶段. 具体而言，投毒攻击发生在目标模型的训练阶段；探索性攻击、逃逸攻击发生在目标模型的测试阶段；而木马攻击，既发生在目标模型的训练阶段，又发生在目标模型的测试阶段.

3)攻击能力

根据攻击者对目标模型所掌握知识的不同，对抗攻击可以分为白盒攻击和黑盒攻击. 在白盒攻击中，攻击者知道训练数据、模型结构、算法和优化方法，即其可以完全访问训练好的目标模型，并知道分类模型的输入数据. 黑盒攻击作为许多安全威胁场景中更现实、更严格的目标模型[104]，由于黑盒攻击者既不了解也无法访问训练数据或训练过的目标模型，攻击者只能试图从模型返回的结果中推断出信息来指导攻击. 黑盒攻击通常利用代理模型，被训练为在推理攻击期间执行与目标网络相同的任务.

4)攻击目标

针对性攻击在模型决策方面有一个特定的目标，即生成扰动 $\boldsymbol{\delta}$，使输入样本 $\boldsymbol{x}\in X$(真实标签为 $\boldsymbol{y}^{\text{true}}$)被错误分类为特定的目标标签 $\boldsymbol{y}^{\text{target}}$，$f(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta})=\boldsymbol{y}^{\text{target}}$，$\boldsymbol{y}^{\text{target}}\neq\boldsymbol{y}^{\text{true}}$. 非针对性攻击的目的是最大化模型决策错误率，即生成扰动 $\boldsymbol{\delta}$，使算法的输出出现错误，即 $f(\boldsymbol{\theta}, \boldsymbol{x})\neq f(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta})$. 非针对性攻击与标签类别无关，通过准确性下降来降低算法的可信度.

2 面向信号调制识别的对抗深度学习

对抗攻击在无线通信领域的应用与其他应用领域(如 CV 和 NLP)有很大的不同，因为对原始信号数据进行适当的数据表示、预处理流程等都需无线通信领域的专业知识作为支撑，与其他领域数据结构(图像、文本等)有着显著差异；此外，信道、干扰和流量效应等复杂通信环境因素不仅对于信号通信产生影响，对于对抗攻击的有效性也会带来不确定性阻碍. 因此面向信号调制识别领域的对抗深度学习与其他领域相比，存在特殊性，本节将基于无线通信环境下信号调制识别任务总结 3 个维度的独特属性，并提出新的对抗攻击分类框架与威胁模型，对目前相关研究工作进行系统整合.

2.1 信号对抗攻击的特性

面向信号调制识别的对抗攻击任务具有 3 个维度特性，分别是环境特性、数据特性、特征特性.

2.1.1 环境特性：信道效应的复杂性

无线通信信道模型对对抗攻击有显著影响，因为传输信号与扰动信号均需要通过信道才能抵达接收器并对目标 DNNs 模型进行对抗攻击. 首先，信道模型会对精心制作的对抗性扰动造成路径损耗和相位偏移等影响，还可能在对抗性信号样本抵达目标接收器之前对其进行削弱或方向改变；其次，在攻击者无法获得目标接收器 DNNs 分类模型的先验知识时，用于训练代理模型的信号数据的收集过程也是在信道中进行的，则训练数据并非完美，攻击者训练的代理模型的有效性依赖于信道效应[105]；此外，对于复杂通信环境也可以利用信道效应的差异进行多目标问题研究，如实现目标接收器正确分类同时避免被窃听接收器正确识别.

2.1.2 数据特性：信号数据的间接性

在真实无线通信环境下，由于信道的存在以及接收器在进行调制识别之前的信号处理操作，攻击者不能直接操纵输入到分类器模型的训练或测试数据. 因此，攻击者需要监控无线通信中的行为状态，并试图间接地操纵或影响 DNNs 模型的结果.

2.1.3 特征特性：特征表示的异质性

各种通信系统的共存使得无线电信号数据具有更加多样化和复杂的特征表示[89]，这也在很大程度上影响了精心制作的对抗性扰动的有效性. Flowers 等人[29]指出对抗性扰动信号并不能很好地保持与原始信号相同的光谱形状，信号频谱形状的改变降低了对抗攻击的成功率，接收器可以利用信号的预处理环节(如滤波器)直接排除扰动信号，增加了对抗样本检测的可能性；此外，Sahay 等人[88]也通过实验表明，在通信信道存在的情况下，以基于信号时域特征训练的 DNNs 作为目标模型而设计的对抗攻击，在基于信号频域特征训练的 DNNs 目标模型上并不一定具有相似的攻击效果.

2.2 面向信号调制识别的对抗攻击分类框架

本文对于对抗攻击的研究着眼于无线通信领域下的信号调制识别任务，1.2.2 节中提到的 CV 领域对抗攻击分类方法源于对抗攻击的共性特征，本文在此基础上充分考虑 2.1 节中提到的无线电信号对抗攻击的特性特征，参考 Flowers 等人[19]提出的根据攻击的发起地点进行分类，提出新的分类类别：物理自我防御式对抗攻击和数字直接访问式对抗攻击(下文简称物理对抗攻击和数字对抗攻击)，物理对抗攻击和数字对抗攻击模型对比图分别如图 5 所示.

基于框架图可以从以下 3 方面深入分析物理对抗攻击与数字对抗攻击的不同点.

1)攻击发生地点不同

数字攻击主要是在发射器与接收器之外的攻击者发射器处进行的. 传统的对抗性机器学习，如 CV 中一般考虑的攻击或 Sadeghi 等人[16]考虑的攻击，都属于数字直接访问式攻击的范畴. 部分研究工作没有将信道效应等环境影响考虑在内，将攻击近似为信号处理链的一部分，对于分析分类器模型在最坏情况下的脆弱性有极大的意义. 但由于其假定信号处理链已被破坏，在复杂的无线通信背景下不太现实. 进一步的研究工作中，考虑到真实物理环境条件通过添加约束规范，实现对抗攻击理论方法.

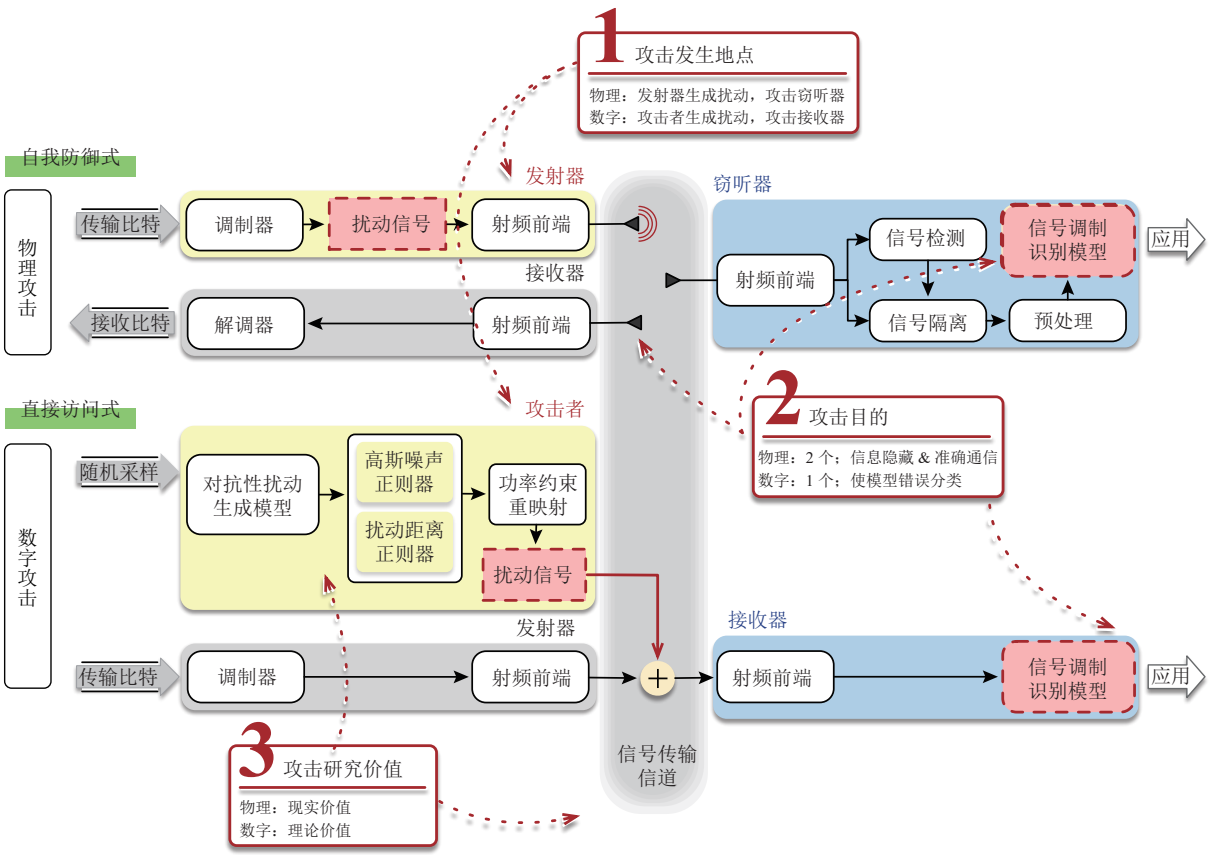


Fig. 5 Comparison of physical adversarial attack model and digital adversarial attack model

图 5 物理对抗攻击模型与数字对抗攻击模型对比

2)攻击应用目的不同

数字攻击只有 1 个目的就是攻击目标接收器中的信号调制识别模型，使其分类错误. 物理攻击有 2 个目的，分别是信息隐藏(欺骗窃听器中的信号调制识别模型分类错误)和准确通信(保证目标接收器中的信号调制识别模型分类正确)，最终目标是二者的相互博弈达到最优结果.

3)攻击研究价值不同

数字攻击更具理论价值，以近似理想化的攻击场景假设为基础，或通过添加约束条件来模拟，聚焦于研究提高对抗攻击性能的方法，为调制识别领域提供基线模型和科学理论支撑. 文献 [20, 81, 86, 89] 的研究工作将 CV 领域的对抗攻击方法应用于调制识别领域，并证明其有效性.

物理攻击更具现实价值，因其考虑到现实复杂的通信场景，即除发射器和接收器外，还存在第三方窃听者. 窃听者依次通过信号检测、信号隔离、信号预处理、信号解调以获得二进制数据流，从而实现无限通信链路窃听. 为加强安全性，可以进行信号加密，保护解调后的比特流. 而物理层安全[106-107]旨在防止窃听者对信号的正确解调，以减少可用信息的获取. 文献 [19, 29, 90] 的研究着眼于通信物理层安全问题，降低窃听者分类器对信号调制识别的准确率. 同时需要与无线通信的主要目标(保证经过调制的信号向期望接受器的可靠性传输)相平衡.

2.3 面向信号调制识别的对抗攻击威胁模型

本文扩展 Papernot 等人[108]和 Chakraborty 等人[97]提出的 CV 领域对抗攻击威胁模型分类方法，提出在无线通信领域下信号调制识别任务的对抗攻击威胁模型，如图 6 所示. 以图解的方式构建对抗攻击威胁模型，并全面总结了目前信号调制识别对抗攻击领域的研究工作，旨在更清晰直观地提供信号调制识别对抗威胁模型分类方法的展示，同时强调该领域特有的信号对抗攻击特性与攻击目标，与 2.2 节中的分类方法相结合，细化信号调制识别的对抗目标和对抗能力.

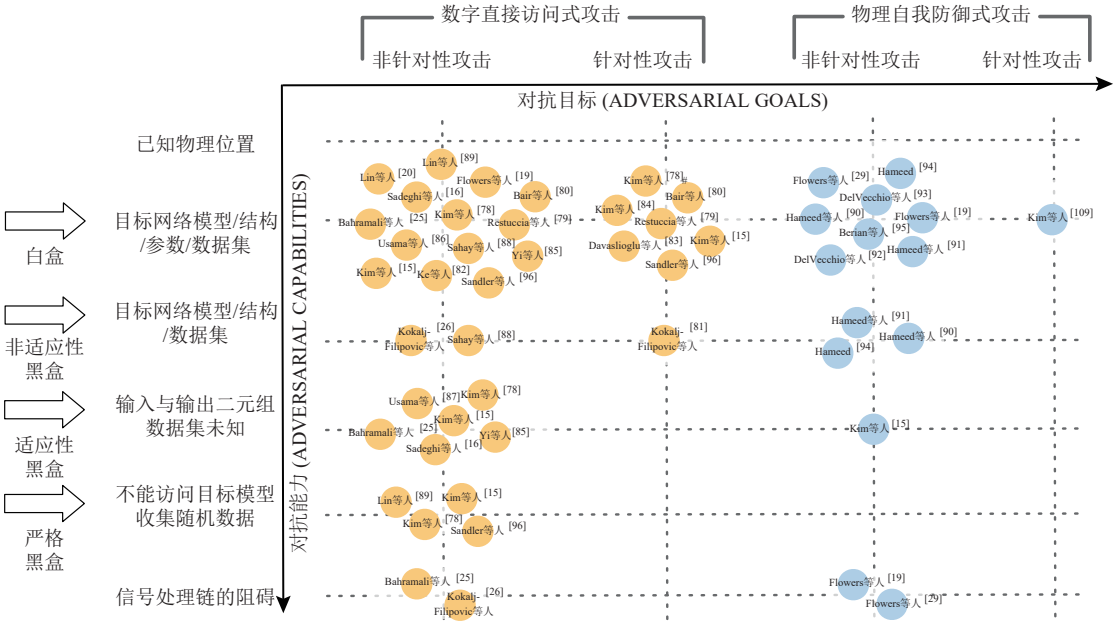


Fig. 6 Threat model of adversarial attack based on signal modulation recognition

图 6 基于信号调制识别的对抗攻击威胁模型

对抗目标的复杂性沿着横轴延伸，依次递增，对抗能力沿着纵轴延伸，依次递减，以对抗目标和对抗能力 2 个影响因子共同作为对抗攻击复杂性的衡量指标，攻击复杂性沿着对角线延伸，“最容易”的攻击是在左上角，“最难”的攻击是在右下角.

1)对抗目标

传统对抗攻击有 2 个主要目标：非针对性对抗攻击和针对性对抗攻击. 非针对性对抗攻击旨在使信号被错误分类为任何非真实调制类别. 针对性对抗攻击更复杂，目标是让分类器将信号分类为特定目标类别，其难度取决于真实与目标类别的信号格式[19].

2)对抗能力

对抗攻击的能力指攻击者利用目标 DNN 信息进行攻击的能力[108]. 对抗能力可分为白盒攻击和黑盒攻击，其中黑盒攻击可以进一步分类为非适应性黑盒攻击、适应性黑盒攻击、严格黑盒攻击[97].

① 白盒攻击. 在最理想的情况下, 攻击者可以完全了解目标模型的结构以及学习参数, 并且可以访问训练数据的分布, 这种攻击下目标模型被认为具有较高的错误率. 在无线通信中, 当数字直接访问式攻击(数字攻击)为白盒攻击时, 假设信道等外界物理环境对攻击不产生可变性影响, 通过直接操纵接收器的调制分类器模型得以实现[16,25,89]; 当物理自我防御式攻击(物理攻击)为白盒攻击时, 如 Flowers 等人[29]在发射器上使用对抗性残余网络生成扰动, 以规避窃听者调制分类器的正确识别.

② 非适应性黑盒攻击. 攻击者对目标模型结构与学习参数未知, 但可以利用其训练数据集或数据分布来训练一个代理模型, 以接近目标分类器, 再利用白盒攻击策略将精心制作的对抗样本输入到目标模型, 致使其错误分类. 在无线通信中, 数字攻击与前述步骤相同; 而物理攻击通过与窃听者调制分类模型结构相同/不同的网络模型训练生成对抗扰动, 添加到其传输信号中以规避窃听接收器正确识别[90-91,94].

③ 适应性黑盒攻击. 攻击者对目标模型的训练数据集或数据分布未知, 仅可以通过访问目标模型, 即对于任意的输入, 攻击者通过查询目标模型获得其标签, 通过得到的输入输出二元组训练代理模型, 再利用白盒攻击策略致使目标模型错误分类. 在无线通信中, 数字攻击未知训练数据集, 如 Sadeghi 等人[16]构造与目标模型输入无关的通用对抗性扰动, 并通过实验证明亦能有效实现对抗攻击目标; 物理攻击的过程是将扰动添加到发射器上, 当充分考虑信道效应与接收器的信号处理误差, 此时的对抗攻击并非直接操纵在输入信号样本上. Kim 等人[15]考虑了信道效应, 并由实验证明信道状态信息对于对抗攻击的成功有较大影响.

④ 严格黑盒攻击. 攻击者不能从目标模型的数据集中获得信息, 需要收集少量数据或部分随机数据, 再将替代数据集输入到目标模型中进行数据集扩展[108].

面向信号调制识别的对抗能力还有 2 类特殊能力. 其一, 已知物理位置的能力, 主要用于物理对抗攻击中, 攻击者通过智能天线精确地将对抗样本信号传输到窃听接收器, 从而最大限度降低窃听者的识别准确率, 同时减少对期望接收器的干扰. 其二, 信号处理链的阻碍, 信号处理链可以通过射频前端过滤外频干扰和功率放大器的非线性作用, 降低对抗攻击成功率, 不仅对高频扰动进行屏蔽还可以使扰动信号发生扭曲. 此外, 信号处理链中的信号检测和信号隔离阶段对影响带宽的扰动特别敏感, 可能导致信号调制识别的级联影响.

3 面向信号调制识别的对抗样本生成技术

本节将结合现有的研究工作详细阐述信号调制识别任务中的对抗样本生成技术, 并通过考虑无线通信环境下信号调制识别任务的特性, 总结针对该场景下对抗样本生成的特殊处理方法.

3.1 对抗样本生成技术

对抗样本是指在对抗攻击中产生的样本, 即在原始输入数据集中的样本中通过有目的地加入少量的扰动信息, 使得 DNNs 对输入出现误判的样本[13]. 对抗攻击的函数表示如表 3 所示. 对抗攻击的理论方法如式(3)所示, 在原始输入 $\boldsymbol{x}\in X$ 中加入精心设计且不可察觉的扰动 $\boldsymbol{\delta}$, 更改输入为 $\boldsymbol{x}^*=\boldsymbol{x}+\boldsymbol{\delta}$, 从而使模型错误分类. 由于对抗攻击的隐蔽性要求, 扰动 $\boldsymbol{\delta}$ 应尽可能小, 通常这一要求被近似为约束 $\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$, 其中 ε 为一个很小的正数, $\|\cdot\|_p$ 为 l_p 范式, $p\in\{1,2,\cdots\}$ 为范式类型.

Table 3 Function Expression of Adversarial Attack Methods

表 3 对抗攻击方法的函数表示

方法	函数表示	
对抗攻击	$\min\limits_{\boldsymbol{\theta}}\|\boldsymbol{\delta}\|_p$ s.t. $f(\boldsymbol{\theta},\boldsymbol{x})\neq f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta})$	(3)
非针对性对抗攻击	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$	(4)
针对性对抗攻击	$\min\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{target}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$	(5)

在无线通信的信号调制识别任务中, 除了添加的扰动信号样本之外, 对抗攻击还会受到信道噪声和其他信道效应的影响. 由于 DNNs 的非线性特性, 直接解决式(3)的优化问题是很困难的, 通常利用分类模型 $f(\cdot)$ 训练过程中使用的损失函数 $\mathcal{L}(\cdot)$ 来实现.

非针对性对抗攻击旨在最大化误分类损失 $\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{true}})$, 限制条件为扰动 $\boldsymbol{\delta}$ 在规定的约束范围内, 即可由目标函数式(4)求解得到最佳扰动, 其中 ε 为对抗性扰动的约束上限. 针对性对抗攻击, 为了使原始信号被识别为目标标签, 则试图最小化目标标签的损失 $\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{target}})$, 限制条件仍为扰动 $\boldsymbol{\delta}$ 在规范约束范围内, 即可由目标函数式(5)求解得到最佳扰动.

目前已经提出多种方法来求解上述优化问题, 从而产生对抗样本, 见表 4. 例如, 作为经典方法的快

Table 4 Summary of Adversarial Attack Function Experssion for Modulation Recognition

表 4 面向调制识别的对抗攻击函数表达汇总

方法	具体技术	对抗样本函数表示	优化目标函数表示
FGM	FGSM [26]	$\boldsymbol{x}^*=\boldsymbol{x}+\varepsilon\times\mathrm{sgn}(\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y}))$	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\mathrm{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$
	FGM [16]	$\boldsymbol{x}^*=\boldsymbol{x}+\varepsilon\times\dfrac{\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})}{\|\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})\|_2}$	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\mathrm{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$
	FGM 考虑信道效应[15]	$r_{\mathrm{ar}}=\boldsymbol{H}_{\mathrm{tr}}\boldsymbol{x}+\boldsymbol{H}_{\mathrm{ar}}\boldsymbol{\delta}+\boldsymbol{n}$	$\mathcal{L}(f(\boldsymbol{\theta},r_{\mathrm{ar}}),\boldsymbol{y}^{\mathrm{target}})\approx\mathcal{L}(f(\boldsymbol{\theta},r_{\mathrm{tr}}),\boldsymbol{y}^{\mathrm{target}})+(\boldsymbol{H}_{\mathrm{ar}}\boldsymbol{\delta})^{\mathrm{T}}\nabla_x\mathcal{L}(f(\boldsymbol{\theta},r_{\mathrm{tr}}),\boldsymbol{y}^{\mathrm{target}})$
MI-FGSM	MI-FGSM [80]	$\boldsymbol{x}_{t+1}^*=\boldsymbol{x}_t^*+\left(\dfrac{\varepsilon}{T}\right)\times\mathrm{sgn}(\boldsymbol{g}_{t+1})$ $\boldsymbol{g}_{t+1}=\mu\times\boldsymbol{g}_t+\dfrac{\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}^*),\boldsymbol{y})}{\|\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}^*),\boldsymbol{y})\|_1}$	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\mathrm{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$
PGD	PGD [91]	$\boldsymbol{x}_{t+1}^*=\Pi_{\mathcal{B}_\varepsilon(x_2)}(\boldsymbol{x}_t^*+\beta\times\mathrm{sgn}(\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}_t^*),\boldsymbol{y})))$	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\mathrm{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$
C&W	Carlini & Wagner [81]	$\boldsymbol{x}^*=\boldsymbol{x}+\boldsymbol{\delta}$	$\min\limits_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_2^2+c\times f_t(\boldsymbol{x}+\boldsymbol{\delta})$ s.t. $\boldsymbol{x}+\boldsymbol{\delta}\in[\lceil\min(x_i)\rceil,\lceil\max(x_i)\rceil]^k$ $f_t(\boldsymbol{x}')=\max(\max\{Z(\boldsymbol{x}')_i:i\neq t\}-Z(\boldsymbol{x}')_t,0)$
ATN	P-ATN 考虑通信损失[29]	$\boldsymbol{x}^*=\boldsymbol{x}+g(\boldsymbol{\theta},\boldsymbol{x})$	$\mathcal{L}(\cdot)=(1-\alpha)\times\mathcal{L}_{\mathrm{adv}}(\cdot)+\alpha\times\left[\beta\times\mathcal{L}_{\mathrm{comm}}(\cdot)+(1-\beta)\times\mathcal{L}_{\mathrm{pwr}}(\cdot)\right]$
	AAE 考虑通信损失[92]	$\boldsymbol{x}^*=g(\boldsymbol{\theta},\boldsymbol{x})$	$\mathcal{L}_{\mathrm{total}}=\alpha\times\mathcal{L}_{\mathrm{adv}}+\beta\times\mathcal{L}_{\mathrm{comm}}+\gamma\times\mathcal{L}_{\mathrm{pwr}}$
BIM	BIM [89]	$\boldsymbol{x}_{t+1}=Clip_{x,\varepsilon}\{\boldsymbol{x}_t+\varepsilon\times\mathrm{sgn}(\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}_t),\boldsymbol{y}_t))\}$ $Clip_{x,\varepsilon}\{z\}:z\in[\boldsymbol{x}-\varepsilon,\boldsymbol{x}+\varepsilon]$	$\max\limits_{\boldsymbol{\theta}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\mathrm{true}})$ s.t. $\min\|\boldsymbol{\delta}\|_p\leqslant\varepsilon$

速梯度法(fast gradient method, FGM)[16]和快速梯度符号法(fast gradient sign method, FGSM)[26], 再如以其为基础引入迭代思想的衍生方法(基本迭代法(basic iterative method, BIM)[89]、动量迭代法(momentum iterative- FGSM, MI-FGSM)[80]等), 以及引入随机性的衍生方法(投影梯度下降法(projected gradient descent, PGD)[91]), 并且随着神经网络的广泛应用, 也出现了基于对抗生成网络方法, 如对抗转化网络(adversarial transformation network, ATN)[110]等.

根据第 2 节提出的面向调制识别的对抗攻击分类框架与威胁模型, 将现有的对抗攻击研究工作按照对抗样本生成技术的不同进行了分类汇总, 如表 5 所示. 并在 3.1.1~3.1.7 节中对现有的针对信号调制识别任务中的对抗样本生成技术的研究工作进行详细阐述.

Table 5 Summary of Research Work on Digital/Physical Attacks for Different Adversarial Examples Generation Methods

表 5 针对不同对抗样本生成方法的数字攻击/物理攻击研究工作汇总

对抗攻击方法	数字攻击的相关文献	物理攻击的相关文献
快速梯度法（FGM）	[15−16, 20, 26, 78−79, 82, 84,	[19, 95]
动量迭代法（MI-FGSM）	[20, 80, 89]	
投影梯度下降法（PGD）	[20, 89, 96]	[90−91, 94]
Carlini & Wagner	[81, 85−87]	
对抗转化网络（ATN）		[29, 92−93]
通用对抗扰动（UAP）	[15−16, 25, 78, 96]	
基础迭代法（BIM）	[20, 88−89]	
雅克比显著图攻击（JSMA）	[89]	
木马攻击	[83]	

3.1.1 基于 FGM

FGSM[14]作为一种主流的对抗样本生成技术, 基于 DNNs 在高维空间中的线性性质是对抗样本成因这一假设, 围绕模型参数的当前值对损失函数实现线性化, 得到一个最佳的符合最大规范约束的扰动, 即

$$\boldsymbol{x}^*=\boldsymbol{x}+\varepsilon\times\mathrm{sgn}(\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})),\tag{6}$$

其中$\boldsymbol{x}^*$为对抗样本, $\boldsymbol{\theta}$为模型的参数, $\boldsymbol{x}$为模型输入, $\boldsymbol{y}$为与$\boldsymbol{x}$相关的目标标签, $\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})$为用于训练 DNNs 的损失函数, ε为限制扰动变化的一个极小值.

在无线通信领域, 更常用的方法是 FGSM 方法的泛化形式, 即 FGM, 该方法满足 l_2 规范约束 $\|\boldsymbol{x}^*-\boldsymbol{x}\|_2<\varepsilon$, 即

$$\boldsymbol{x}^*=\boldsymbol{x}+\varepsilon\times\frac{\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})}{\|\nabla_x\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}),\boldsymbol{y})\|_2}.\tag{7}$$

在无线通信数字攻击的研究中, Sadeghi 等人[16]针对 FGM 方法生成粗粒度扰动及低攻击成功率问题, 提出了新算法, 依靠计算上高效的 FGM 作为算法的核心, 生成细粒度的对抗性扰动. 首先, FGM 通过人为设定ε来找到使模型分类准确率最低的、符合扰动约束的最佳值; 而 Sadeghi 等人[16]使用二分法确定能够使模型错误分类的扰动值ε(在扰动规范的约束范围内), 减少人为误差. 其次, 非针对性攻击中,

FGM 试图最大化$\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{true}})$, 针对性攻击中, FGM 试图最小化特定类别$\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y}^{\text{target}})$, 而 Sadeghi 等人[16]在所有可能的针对性攻击中进行搜索, 选择具有最小扰动的攻击来执行错误分类. 除此之外, Sadeghi 等人[16]还重新定义了信号扰动约束, 由图像特征空间的距离约束类推测信号的功率比的约束.

Kim 等人[15]的研究着眼于信道效应对于对抗攻击的影响, 探讨如何在现实的信道效应和不同接收器的多个分类模型存在的情况下实现物理的对抗攻击, 如信道信息已知的前提下, 针对对抗攻击者通过使用 FGM 方法线性化损失函数:

$$\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{ar}}),\boldsymbol{y}^{\text{target}})\approx\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{tr}}),\boldsymbol{y}^{\text{target}})+(\boldsymbol{H}_{\text{ar}}\boldsymbol{\delta})^{\text{T}}\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{tr}}),\boldsymbol{y}^{\text{target}}),\tag{8}$$

其中$\boldsymbol{y}^{\text{target}}$为攻击者期望的目标类别, $\boldsymbol{\delta}$为对抗性扰动, $\boldsymbol{H}_{\text{ar}}$为从攻击者到接收器之间的信道效应, $\boldsymbol{r}_{\text{ar}}$为发生对抗攻击的接收器接收到的信号, $\boldsymbol{r}_{\text{tr}}$为未发生对抗攻击的接收器接收到的信号, $\boldsymbol{x}$为发射器发射的信号. 为实现针对性对抗攻击, 通过设置$\boldsymbol{H}_{\text{ar}}\boldsymbol{\delta}=-\alpha\times\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{tr}}),\boldsymbol{y}^{\text{target}})$来最小化$\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{ar}}),\boldsymbol{y}^{\text{target}})$, 其中$\alpha$为比例系数, 用于将攻击者添加的扰动限制在最大功率内. 同理对于非针对性对抗攻击者的目标是最大化损失函数$\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{ar}}),\boldsymbol{y}^{\text{true}})$, 其中$\boldsymbol{y}^{\text{true}}$为$\boldsymbol{x}$的真实类别, 设置$\boldsymbol{H}_{\text{ar}}\boldsymbol{\delta}=-\alpha\times\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{r}_{\text{tr}}),\boldsymbol{y}^{\text{true}})$得以实现.

Kim 等人[15]还通过对比指出 Sadeghi 等人[16]未考虑信道效应而导致攻击性能下降, 原因是无线信道改变了接收器感知到的扰动信号的相位与大小, 并提出了充分最大接收扰动功率(maximum received perturbation power, MRPP)法. 该方法将对抗性扰动$\boldsymbol{\delta}$与信道效应的共轭数$\boldsymbol{H}_{\text{ar}}^{*}$相乘, 得到通过信道后接收器接收到的扰动为$\|\boldsymbol{H}_{\text{ar}}\|^2\boldsymbol{\delta}$, 使得在相位不变情况下, 有效提升扰动功率和误导分类器的能力.

Kim 等人[84] 2020 年的一个研究利用多根天线在不同的信道效应生成多个并发扰动(受总功率预算的影响), 其使用与 Kim 等人[15] 2022 年相同的 MRPP 攻击方法. Kim 等人[84]于 2020 年提出了元素最大信道增益(elementwise maximum channel gain, EMCG)法优化天线功率分配, 提升攻击性能. 该方法通过选择在符号级具有最佳信道增益的天线来发射扰动, 并表明增加天线数量可显著增强攻击.

无线通信物理攻击的研究中, Flowers 等人[19]与 Sadeghi 等人[16]同样重新定义了信号扰动约束为相对功率, 但其考虑了真实物理场景, 分析了 FGSM 方法的有效性, 并提出将误码率作为评估指标.

Berian 等人[95]提出 2 种基于滤波器的方法生成对抗性扰动. 第 1 种为梯度上升滤波法(gradient ascent filter, GAF), 利用迭代优化方法使滤波器系数被有效训练; 第 2 种为快速梯度滤波法(fast gradient filter method , FGFM), 即结合 FGM 思想, 最大限度提高与滤波器系数相关的损失. 基于滤波器的方法有别于 FGSM 加性扰动生成方法, 该方法使用滤波器来创建有限脉冲响应滤波器, 以最小化扰动功率并使分类错误. Berian 等人[95]通过实验表明, 基于滤波器的方法攻击性能更好, 且计算效率更高, 功耗更低.

3.1.2 基于 MI-FGSM

由于 FGSM 生成的对抗样本只是 1 阶近似最优解, 且其阈值是人为选择的[111], 在 CV 领域也提出了很多基于 FGSM 的衍生方法, 如引入迭代思想的动量迭代法[112]. MI-FGSM 并非沿梯度方向只调整 1 步, 而是采取较小的步长, 即ε/T, 其中T为迭代次数, 调整多步, 旨在对扰动进行更精细的调整.

目前 MI-FGSM 攻击方法多用于无线通信数字攻击的场景中. Bair 等人[80]使用 MI-FGSM[112]来生成用于信号调制识别的对抗样本, 该方法在 FGSM 的基础上引入了迭代思想和动量参数, 在找到最佳扰动之前, 扰动信号的当前迭代梯度表示为

$$\boldsymbol{g}_{t+1}=\mu\times\boldsymbol{g}_t+\frac{\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}^{*}),\boldsymbol{y})}{\|\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}^{*}),\boldsymbol{y})\|_1},\tag{9}$$

其中t为当前迭代轮次, $\boldsymbol{g}$为梯度, μ为某一动量值, 则通过 MI-FGSM 方法求得的对抗样本可表示为

$$\boldsymbol{x}_{t+1}^{*}=\boldsymbol{x}_t^{*}+\left(\frac{\varepsilon}{T}\right)\times\text{sgn}(\boldsymbol{g}_{t+1}),\tag{10}$$

其中T为迭代的总次数. 迭代$T-1$次后, $\boldsymbol{x}^{*}$将是最终得到的对抗信号样本. Dong 等人[112]通过实验发现超参数最佳值$\mu=0.5$和$T=6$, 在调制识别的对抗攻击研究中, 还未涉及超参数最佳值的讨论工作.

3.1.3 基于 PGD

在迭代方法基础上, FGSM 的衍生方法中还将引入随机性的思想, 如 PGD[67], 该方法以均匀的随机噪声作为初始化, 并作为最强大的 1 阶攻击方法之一, 同样在无线通信环境下信号调制识别任务中得到广泛的应用.

在无线通信数字攻击的研究中, Lin 等人[89]利用 PGD[67]对调制识别模型进行白盒对抗攻击. 并通过对比实验表明, PGD 基于迭代方法的攻击效果明显优于 FGSM 单步攻击效果. Lin 等人[89]还对信号不同调制方式的鲁棒性问题进行了讨论研究, 并通过实验说明 AM-SSB 和 PAM4 的安全性更高, 而如 BPSK, QAM16, QPSK 等安全性较低, 攻击成功率相对较高.

PGD 方法在求解最佳扰动时需多轮反向传播，因此该方法的运行迭代步数依赖于计算资源[91]，而无线通信信号发射器的吞吐率通较高，PGD 方法的应用显然面临计算成本较高、能耗较大的问题. Sandler 等人[96]提出利用通用对抗扰动[113]来解决这一问题，通过随机选择训练集输入并迭代生成扰动，同时在每次迭代之前对信号添加了随机相位偏移以模拟真实信道.

在无线通信物理攻击的研究中，Hameed 等人[91]利用随机噪声作为初始信号样本值，第 t 次迭代信号样本表示为

$$\boldsymbol{x}_{t+1}^{*}=\Pi_{\mathcal{B}_{\epsilon}(\boldsymbol{x}_2)}(\boldsymbol{x}_t^{*}+\beta\times\operatorname{sgn}(\nabla_{\boldsymbol{x}}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}_t^{*}),\boldsymbol{y}))),\tag{11}$$

其中 $\beta>0$ 为步长，$\operatorname{sgn}(\cdot)$ 为符号运算，$\Pi_{\mathcal{B}_{\epsilon}(\boldsymbol{x}_2)}$表示以 ϵ 为半径、$\boldsymbol{x}_2$ 为中心的 l_2 领域内的欧几里德投影算子，算法通过多次迭代找到最佳扰动. Hameed 等人[91]还考虑了真实物理场景需求，保证期望接收器误码率最小，同时最大化窃听接收器号识别误差，通过 PGD 方法计算局部最优解.

3.1.4 基于 Carlini & Wagner 方法

Carlini 和 Wagner[114]针对 Hinton 等人[115]提出的防御蒸馏网络提出 Carlini & Wagner 攻击方法，与前期的对抗样本生成技术不同的是，Carlini & Wagner 将扰动最小化和损失函数最大化这 2 个优化问题结合为一个目标函数.

目前 Carlini & Wagner 攻击方法，在无线通信对抗攻击研究中均为数字攻击. 由于 Carlini & Wagner 方法的原始定义中规定了对抗样本 $\boldsymbol{x}^{*}=\boldsymbol{x}+\boldsymbol{\delta}\in[0,1]^k$ 的约束条件，因此该条件只适用于图像（像素），不适用于信号样本. Kokalj-Filipovic 等人[81]对 CleverHans 库[116]中的 Carlini & Wagner-l_2 方法[114]进行调整，以适应信号调制识别任务，即对于信号输入样本 $\boldsymbol{x}$，Carlini & Wagner 通过解决以下目标优化问题来生成对抗扰动 $\boldsymbol{\delta}$：

$$\begin{aligned}&\min_{\boldsymbol{\delta}}\|\boldsymbol{\delta}\|_2^2+c\times f_t(\boldsymbol{x}+\boldsymbol{\delta}),\\&\text{s.t. }\boldsymbol{x}+\boldsymbol{\delta}\in[\lceil\min(x_i)\rceil,\lceil\max(x_i)\rceil]^k,\end{aligned}\tag{12}$$

其中 t 为 $\boldsymbol{x}$ 的目标标签，f_t被定义为

$$f_t(\boldsymbol{x}')=\max(\max\{Z(\boldsymbol{x}')_i:i\neq t\}-Z(\boldsymbol{x}')_t,0),\tag{13}$$

其中 $Z(\cdot)$ 为 Softmax 函数，f_t为原始非线性优化问题的近似目标函数，即当且仅当 $f_t(\boldsymbol{x}+\boldsymbol{\delta})\leqslant 0$ 时，对抗样本 $\boldsymbol{x}^{*}=\boldsymbol{x}+\boldsymbol{\delta}$被识别为目标类别 t. 对于非针对性攻击，在除真实标签外的所有标签中找到满足最小化式（13）的 t，同时限制扰动的约束规范，使扰动功率与信号的噪声功率相同. Kokalj-Filipovic 等人[81]为了充分分析修改后的 Carlini & Wagner 攻击方法的性能，在模拟数据集与真实世界数据集中共同进行训练，并由实验表明 Carlini & Wagner 攻击方法具有与 FGSM 攻击方法相似的效果.

Yi 等人[85]利用 Carlini & Wagner 方法来生成对抗样本，应用于无线通信环境下，将微小的扰动约束在攻击者所需的最小发射功率之下，致使分类器模型的性能显著下降. 但 DNNs 模型的漫长训练时间对其在无线通信中的应用造成了严重的瓶颈，需要频繁地重新训练以适应实时变化的环境，因此，Yi 等人[85]参考 Ramjee 等人[117]提出的数据驱动子采样策略，采用 DNNs 模型架构模拟从每组训练样本中删除候选样本的过程，该过程实际为有监督的封装式特征选择算法的实现，即子采样网络，通过对输入进行向下采样，有效地减少网络模型大小和训练时间. Yi 等人[85]通过实验验证在对抗攻击过程中，通过信噪比选择性训练，在不牺牲攻击性能的情况下可以有效节省计算资源，为数据驱动的子采样策略应用于无线通信领域下对抗攻击任务奠定了基础.

3.1.5 基于 ATN

Baluja 等人[110]提出了 ATN，与前面所述的攻击方法不同，ATN 作为独立的 DNN 模型，可以将任何输入转化为对抗样本，并且 Baluja 等人[110]提出 2 种 ATN 的变体：其一，扰动-对抗转化网络（perturbation ATN, P-ATN），即基于 ATN 训练只生成输入相关扰动的网络模型；其二，对抗自编码器（adversarial auto-encoder, AAE），即基于 ATN 训练生成关于输入的对抗自编码器模型.

基于 ATN 的方法多用于无线通信物理攻击的场景中. Flowers 等人[29]利用 ATN 的第 1 个变体 P-ATN 在真实物理场景下对信号调制识别的对抗攻击进行研究. P-ATN 作为 ATN 的变体，通过将残差模块结构加入 ATN 模型中，生成的对抗样本可以表示为

$$\boldsymbol{x}^{*}=\boldsymbol{x}+g(\boldsymbol{\theta},\boldsymbol{x}),\tag{14}$$

其中 $\boldsymbol{\theta}$ 为 P-ATN 在训练过程中学习到的参数矩阵，$g(\cdot)$ 为 P-ATN（Flowers 等人[29]以 ARN 作为简称），其损失函数被定义为

$$\begin{aligned}\mathcal{L}(\cdot)=&(1-\alpha)\times\mathcal{L}_{\text{adv}}(\cdot)+\alpha\times\\&\left[\beta\times\mathcal{L}_{\text{comm}}(\cdot)+(1-\beta)\times\mathcal{L}_{\text{pwr}}(\cdot)\right],\end{aligned}\tag{15}$$

其中 $\mathcal{L}_{\text{adv}}$ 为导致错误分类的对抗性损失，$\mathcal{L}_{\text{comm}}$ 为最小化误码率的通信损失，$\mathcal{L}_{\text{pwr}}$ 为扰动功率约束，三者之间的权衡通过调整 2 个超参数的值来实现. 由于 P-ATN 中残差模块的特性，P-ATN 可以更容易地学习

到满足要求的最佳扰动.

由于 Flowers 等人[29]考虑的是物理对抗攻击场景, 其 P-ATN 的目标函数具有多个目标, 最小化期望接收器的误码率, 同时最小化窃听接收器的识别精度, 即在窃听器的对抗损失与接收器的通信损失之间寻求一种平衡, 并且扰动信号功率在发射之前也受到一定的功率限制, Flowers 等人[29]将其以正则器的形式作为通信损失的一部分. 此外, Flowers 等人[29]在计算通信损失的过程中, 出于接近更真实的物理场景的考虑, 假设对信道模型不可知, 没有直接使用符号空间中的距离来推导通信传输中的误码率, 提出了计算通信损失的新的模型, 即通信合页损失(communications hinge loss), 可用于在训练期间发生误码时根据经验对其进行惩罚. 由实验结果显示, P-ATN 在实现与 FGSM 同等的对抗攻击效果的同时, 还可以保证接收器的传输误码率处于较低的水平. 并且, 由于 P-ATN 不需要计算梯度就可以生成对抗样本, 使得计算复杂度较低, 说明 P-ATN 较 FGSM 更适合于实时的无线通信系统.

DelVecchio 等人[92]则运用的是 ATN 的第 2 个变体 AAE, AAE 与 P-ATN 相类似, 其通过以下方式生成对抗样本:

$$\boldsymbol{x}^* = g(\boldsymbol{\theta}, \boldsymbol{x}), \tag{16}$$

其中 $g(\cdot)$ 为 AAE 模型, $\boldsymbol{\theta}$ 为在训练 AAE 时学到的参数矩阵. AAE 与 P-ATN 的明显差别在于, P-ATN 生成的扰动相对于原始输入信号的缩放是在网络模型之外完成的, 而 AAE 是将扰动与原始输入信号囊括在内一起进行训练完成的. 这个区别使得 AAE 更好地考虑到扰动功率与原始信号功率之间的平衡问题, 因此 DelVecchio 等人[92]利用 AAE 方法构造应用于信号调制识别的对抗样本, 并对其额外添加了功率限制, 将其称之为对抗性突变网络(adversarial mutation network, AMN). 此外, DelVecchio 等人[92]还考虑了前向纠错编码(forward error correction, FEC)的存在对于对抗攻击性能的影响, 将对抗攻击模型扩展到含有 FEC 的信号训练的通信感知攻击框架.

3.1.6 基于通用对抗性扰动(UAP)

在 CV 领域, 研究者还考虑到对抗样本的普遍适用性, 生成与输入无关的通用扰动, 实现跨类别的扰动迁移, 即通用对抗性扰动(universal adversarial perturbation, UAP)[113]. 该方法的输入包括: 目标模型、所需的 UAP 规范以及数据输入的随机子集, 该方法的核心是迭代方法, 即在每次迭代中需要为 N 个数据点中的每一个数据点生成一个对抗性扰动, 计算成本较高.

基于 UAP 的方法多用于无线通信数字攻击的场景中. Sadeghi 等人[16]在 UAP 方法的基础上, 提出了一种新的与输入无关的对抗样本生成方法, 利用主成分分析方法来生成 UAP, 与 UAP 方法相比, 该方法的计算复杂度有显著的降低, 并且通过实验证明该方法同样能达到良好的攻击效果.

与 Sadeghi 等人[16]的研究方法相类似, Bahramali 等人[25]也同样考虑到输入不可知这一实际应用上的约束条件, 但 Bahramali 等人[25]指出单一的 UAP 向量很容易被目标接收器识别并进行过滤, 对抗攻击成功率不高, 因此对 Sadeghi 等人[16]提出的仅生成单一的 UAP 向量的方法进行了改进. Bahramali 等人[25]将生成对抗性扰动的过程建模为一个优化问题, 通过构建扰动生成模型(perturbation generator model, PGM)进行解决. 该模型能够针对目标模型生成大量与输入无关的对抗性扰动向量(即 UAP), 在对抗攻击中, 攻击者从大量可用的 UAP 集合中挑选并生成一个随机的对抗样本对目标模型进行攻击. Bahramali 等人[25]通过实验证明, PGM 扰动生成模型在白盒与黑盒场景下均有效, 并且该方法的攻击效果优于 Sadeghi 等人[16]的方法. 此外, Bahramali 等人[25]还对 PGM 生成的 UAP 添加了不可检测性和鲁棒性 2 个约束. 具体而言, 在生成 UAP 的同时, 约束它们遵循加性高斯白噪声(AWGN)无线信道的预期噪声分布, 使其不可检测; 并对不同的 UAP 之间的距离进行最大化, 使其对防御技术表现为高鲁棒性. 通过实验证明, 在调试识别任务中, 尽管防御机制的存在, 该方法仍能使识别准确率由 69% 降低到 23%.

3.1.7 其他技术

Lin 等人[20,88-89]利用基础迭代法(basic iterative method, BIM)[118]生成面向调制识别模型的对抗样本. BIM 作为 FGSM 的衍生方法, 在 FGSM 基础上通过多次迭代生成对抗样本, 每次迭代的步长保持较小长度, 每一步迭代后截取中间值以确保其位于原始输入的扰动附近. 并且, Sahay 等人[88]还通过实验表明, 以基于信号时域特征训练的 DNNs 作为目标模型而设计的对抗攻击, 在基于信号频域特征训练的 DNNs 目标模型上并不一定具有相似的攻击效果.

Lin 等人[89]利用对抗样本的可迁移性, 基于雅克比显著图攻击(Jacobian-based saliency map attack, JSMA)方法[108]对目标模型实现黑盒攻击. 在没有任何模型先验知识的情况下, 使用代理 DNNs 模型来模拟近似目标模型的决策边界.

Davaslioglu 等人[83]利用木马攻击方法，即在目标模型的部分训练数据集中嵌入特洛伊木马作为相位偏移触发器，稍后在模型测试阶段触发木马以欺骗分类模型. 该方法不同于操纵训练数据的投毒攻击[119-120]，在木马攻击中，数据投毒的过程不是随机应用于输入样本，只有被选定的样本才会进行投毒，并且该样本携带着特定触发器. Davaslioglu 等人[83]通过实验表明，木马攻击只需要对少量的训练数据样本进行木马嵌入，其攻击效果就可以在信噪比全域范围内达到较好的水平.

综上所述，表 6 对上述对抗样本生成方法进行了归纳，包括各个方法的关键技术、攻击分类、攻击目标、攻击能力、实验数据集以及攻击效果等.

3.2 针对无线通信特性的攻击方法提炼

3.2.1 环境特性：信道效应的复杂性

调制识别任务的对抗攻击研究中，考虑更为真实的黑盒场景与复杂信道环境具有现实价值. 针对信道效应，信道模型可分为 2 种场景：攻击者与目标模型具有相同和不同信道分布，如图 7 所示.

场景 1：相同信道分布（攻击发起之前）. 攻击者具有和接收器相同的训练数据集或数据分布，信道效应只会影响对抗样本的攻击效果. 按照假设条件，该场景细分为 3 种情况.

1）不考虑模拟信道模型：在对抗性扰动中添加

Table 6 Summary of Adversarial Samples Generation Methods for Modulation Recognition

表 6 面向调制识别的对抗样本生成方法总览

攻击方法	相关研究工作	关键技术	攻击分类	攻击目标	攻击能力	数据集	攻击后最低准确率/%(SNR/dB)
FGM	Sadeghi 等人[16]	使用二分法改进 FGM	数字	非针对性	白盒	RML2016.10A	0 (10)
	Kim 等人[15,78]	最大接收扰动功率算法 MRPP 改进 FGM	数字	针对性	白盒	RML2016.10A	13 (10)
				非针对性	白盒	RML2016.10A	17 (10)
	Restuccia 等人[79]	在一组连续的输入信号切片上计算扰动	数字	针对性	白盒	RML 2018.01A	-
				非针对性	白盒	RML 2018.01A	31 (0)
	Kokalj-Filipovic[26]	通过自编码器对调制识别模型进行预训练，提高模型鲁棒性	数字	非针对性	非适应性黑盒	RML 2018.01A (BPSK, QPSK, 9-PSK)	60 (14)
	Ke 等人[82]	FGSM 和 L-BFGS 对抗攻击	数字	非针对性	白盒	RML2016.04C	21 (8)
	Lin 等人[20]	比较多种方法性能，讨论波形变化	数字	非针对性	白盒	RML2016.10A	17 (10)
	Kim 等人[84]	采用元素最大信道增益 EMCG 法来优化天线信号功率分配，生成并发对抗性扰动	数字	针对性	白盒	RML2016.10A	10 (10)
	Sahay 等人[88]	基于频域特征分类，削弱非适应性黑盒对抗样本的可迁移性	数字	非针对性	白盒	RML2016.10B	23 (18)
				非针对性	非适应性黑盒	RML2016.10B	78 (18)
	Lin 等人[89]	比较多种方法攻击性能，讨论调制类型鲁棒性	数字	非针对性	白盒	RML2016.10A	14 (10)
	Flowers 等人[19]	FGSM 实现对抗攻击，并考虑接收器误码率	物理	非针对性	白盒	RML2016.10A (BPSK)	58 (10)
	Berian 等人[95]	基于滤波器方法实现对抗攻击	物理	非针对性	白盒	RML2016.10A	25 (0)
MI-FGSM	Bair 等人[80]	利用 MI-FGSM 实现针对性 (QPSK 被误分为 8PSK) 与非针对性对抗攻击	数字	针对性	白盒	RML2016.10A	0 (18)
				非针对性	白盒	RML2016.10A	7 (18)
	Lin 等人[20]	比较多种方法性能，讨论波形变化	数字	非针对性	白盒	RML2016.10A	10 (10)
	Lin 等人[89]	比较多种方法攻击性能，讨论调制类型鲁棒性	数字	非针对性	白盒	RML2016.10A	10 (10)
PGD	Lin 等人[20]	比较多种方法性能，讨论波形变化	数字	非针对性	白盒	RML2016.10A	13 (10)
	Lin 等人[89]	比较多种方法攻击性能，讨论调制类型鲁棒性	数字	非针对性	白盒	RML2016.10A	13 (10)
	Sandler 等人[96]	利用 PGD 方法实现对抗攻击	数字	非针对性	白盒	RML 2018.01A, 自建数据集	-
	Hameed 等人[90]	PGD 方法实现对抗攻击，考虑接收器误码率的影响，将 l_2 范数随机噪声作为对照组	物理	非针对性	白盒	自建数据 (QAM64)	55 (10)
				非针对性	非适应性黑盒	自建数据 (QAM64)	55 (10)
	Hameed 等人[91]	改进 PGD 方法，并通过降低传输速率来权衡误码率与分类精度	物理	非针对性	白盒	自建数据 (QAM64)	55 (10)
				非针对性	非适应性黑盒	自建数据 (QAM64)	72 (10)
	Hameed [94]	改进 PGD 方法，考虑误码率的优化	物理	非针对性	白盒	自建数据 (QAM64)	55 (10)
				非针对性	非适应性黑盒	自建数据 (QAM64)	72 (10)

表 6（续）

攻击方法	相关研究工作	关键技术	攻击分类	攻击目标	攻击能力	数据集	攻击后最低准确率/%(SNR/dB)
Carlini & Wagner	Kokalj-Filipovic 等人[81]	改进 Carlini&Wagner-l_2 方法以适应调制识别对抗攻击	数字	针对性	非适应性黑盒	RML 2018.01A 自建数据集	-
	Yi 等人[85]	使用数据驱动的子采样策略，对未知子采样策略实现适应性黑盒攻击	数字	非针对性	白盒	RML2016.10B	48 (10)
				非针对性	适应性黑盒	RML2016.10B	70 (10)
	Usama 等人[86]	实现白盒对抗攻击	数字	非针对性	白盒	RML2016.10A	-
	Usama 等人[87]	实现适应性黑盒对抗攻击	数字	非针对性	适应性黑盒	RML2016.10A	33 (10)
ATN	Flowers 等人[29]	利用扰动-对抗转化网络 ARN 实现对抗攻击，引入通信合页损失减少期望接收器误码率	物理	非针对性	白盒	RML2016.10A (BPSK)	25 (10)
	DelVecchio 等人[92]	利用 AMN 实现对抗攻击，引入通信合页损失减少期望接收器误码率，并考虑前向纠错编码 FEC 影响	物理	非针对性	白盒	自建数据 (QPSK)	0 (10)
	DelVecchio 等人[93]	改进 AMN 网络，考虑频谱欺骗损失，同时减少期望接收器的误码率	物理	非针对性	白盒	自建数据 (QPSK)	0 (10)
UAP	Sadeghi 等人[16]	利用主成分分析方法生成 UAP	数字	非针对性	适应性黑盒	RML2016.10A	37 (10)
	Bahramali 等人[25]	构建扰动生成模型 PGM 生成大量 UAP，基于 GAN 构建噪声正则器，同时最大化 UAP 之间的距离	数字	非针对性	适应性黑盒	RML2016.10A	25 (10)
	Kim 等人[15,78]	利用主成分分析和自编码器 2 种方法生成 UAP，并考虑信道信息有限	数字	非针对性	适应性黑盒	RML2016.10A	46 (10)
				非针对性	严格黑盒	RML2016.10A	65 (10)
	Sandler 等人[96]	生成 UAP 和特定类别的 UAP，并添加随机相位偏移模拟真实信道	数字	非针对性	适应性黑盒	RML 2018.01A, 自建数据集	-
				针对性	适应性黑盒	RML 2018.01A, 自建数据集	-
BIM	Lin 等人[20]	比较多种方法性能，讨论波形变化	数字	非针对性	白盒	RML2016.10A	12 (10)
	Sahay 等人[88]	基于频域特征分类，削弱非适应性黑盒对抗样本的可迁移性	数字	非针对性	白盒	RML2016.10B	20 (18)
				非针对性	非适应性黑盒	RML2016.10B	74 (18)
	Lin 等人[89]	比较多种方法攻击性能，讨论调制类型鲁棒性	数字	非针对性	白盒	RML2016.10A	10 (10)
JSMA	Lin 等人[89]	基于 JSMA 方法实现严格黑盒攻击	数字	非针对性	严格黑盒	RML2016.10A	25 (10)
木马攻击	Davaslioglu 等人[83]	在一部分训练数据集中嵌入木马，在测试时触发木马以欺骗分类器	数字	针对性	白盒	RML2016.10A	攻击成功率 88 (12)

注：“-”为当前文献未给出较为明确的攻击后最低准确率.

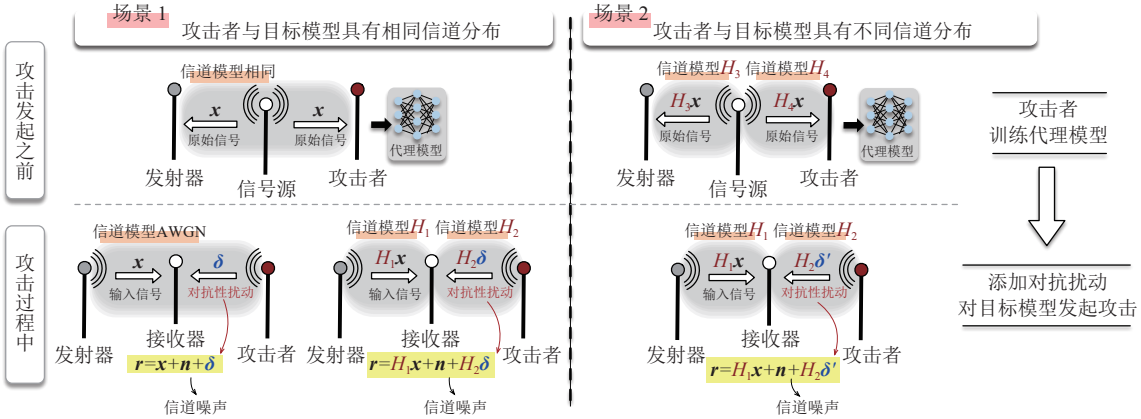


Fig. 7 Two adversarial attack scenarios considering channel effect

图 7 考虑信道效应的 2 种对抗攻击场景

相对随机相位偏移[25,96].

2)完整信道模型：假设已知信道模型且恒定不变，利用信道效应“升级”对抗样本生成算法，解决扰动信号相位偏移与大小变化等问题[15,78].

3)有限信道模型：假设未知信道模型，仅可知信道分布，利用主成分分析(PCA)对有限信道信息进行降维[15,78]；或者考虑到信道效应的时变性，对连续信号切片计算扰动[79].

在物理对抗攻击场景下，还可以利用信道差异构建多目标优化，如通过 DNNs 训练模型参数，解决正常通信的误码率与窃听接收器的分类精度的最优化问题[90-91,94-95]. 其中基于 GAN 的 ATN 在该优化问题中有较好的实验效果[29,92-93].

场景 2: 不同信道分布(攻击发起之前). 攻击者具有和接收器不同的训练数据集或数据分布，信道效应不仅会影响对抗攻击效果，而且攻击者在训练数据的收集过程中已经默认是不完美的.

该场景下的对抗攻击效果与攻击者和发射器的位置距离、扰动信号的信号功率等有关，需进行深入的探讨研究. Kim 等人[105]通过实验表明，在黑盒场景下，选择适当的扰动信号发射功率，以及与信号源的位置更接近(与发射器间的距离固定)，可以提高对抗攻击效果. 此外，通过生成 UAP，可实现跨类别扰动迁移，增强对抗攻击的普遍适用性.

3.2.2 数据特性：信号数据的间接性

当不考虑无线通信环境影响，假设攻击者生成的对抗性扰动直接作用于接收器的分类模型上[25,89]，通过理想实验证明 DNNs 调制识别模型的脆弱性. 当考虑真实无线通信环境影响，对抗性扰动信号的发射必须通过信道模型以及接收器的信号处理环节，因此攻击者只能间接影响接收器的 DNNs 调制识别模型. 很多研究工作[15,29,78-79,92-93]尝试近似模拟信道模型以及讨论分析接收器对信号数据的影响.

3.2.3 特征特性：特征表示的异质性

大部分研究工作集中在基于时频特征训练的 DNNs 上，对于信号的其他特征表示方法的研究较少. DelVecchio 等人[93]引入新定义的损失量，即光谱欺骗损失量，以保持扰动信号频谱的完整性，同时实现有效攻击与稳定通信；Lin 等人[20,88]关注到对抗攻击前后的输入信号波形变化，探讨如何保持攻击效果的同时，实现对抗攻击的隐蔽性，寻找最佳扰动方案.

综上所述，表 7 与表 8 对无线通信特性的对抗攻击方法进行了总结与归纳，包括针对各通信特性的假设/应用场景与处理方法.

Table 7 Summary of Adversarial Attack Methods Considering Wireless Communication Characteristics

表 7 针对于无线通信特性的对抗攻击方法汇总

通信特性		相关研究工作	处理方法
信道效应的复杂性	场景 1：不考虑模拟信道模型	Sadeghi 等人[16]	UAP 具有移位不变性
		Bahramali 等人[25]	UAP 中添加相位旋转
	场景 1：完整信道模型	Kim 等人[15,78,95]	利用信道效应改进对抗算法
		Kim 等人[84]	利用信道效应改进对抗算法，并合理分配多天线信号功率
		Flowers 等人[19]	提出特定信道与动态信道场景实验
	场景 1：有限信道模型	Flowers 等人[29]	引入通信合页损失
		Kim 等人[15,78]	采用 PCA 方法对信道信息降维，生成 UAP
		Restuccia 等人[79]	在一组连续的输入信号切片上计算扰动
		Sandler 等人[96]	真实信道上使用 SDR 广播对抗性扰动信号
		Flowers 等人[19]	提出特定信道与动态信道场景的实验
	场景 1：物理对抗攻击中利用信道差异	Flowers 等人[29]	利用 ATN 的变体 ARN 进行对抗攻击
		DelVecchio 等人[92-93]	利用 ATN 的变体 AMN 进行对抗攻击
		Restuccia 等人[90,94]	基于 DNN 模型以多目标的线性组合为目标函数
		Hameed 等人[91]	多目标优化，并降低传输速率来权衡误码率与分类精度
	场景 2：训练数据集不同	Sadeghi 等人[15-16,25,78,96]	生成与输入无关的 UAP
特征表示的异质性		Lin 等人[20]	保证对抗攻击前后信号波形的一致，同时不减弱攻击性能
		DelVecchio 等人[93]	引入频谱欺骗损失，保持对抗性扰动信号的频谱完整性
		Sahay 等人[88]	基于频域特征分类，削弱黑盒对抗样本的可迁移性
信号数据的间接性		Flowers 等人[15,19,29,78-79,91-93]	模拟信道模型，讨论分析前向纠错编码 FEC 等影响

4 面向信号调制识别的对抗检测与防御

4.1 基于对抗训练的防御

对抗训练通过使用对抗样本训练 DNNs 模型来增强鲁棒性，减轻对抗攻击的影响，是较为主流的防御方法. 每次迭代生成新的对抗样本重新训练模型，扩大训练集，提高模型抵抗力，防止攻击者破坏通信系统.

实际上，基于对抗训练的防御方法即是经验风险最小化(empirical risk minimization, ERM)的对抗攻击方法的变体，旨在解决对抗样本风险最小化问题：

$$\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\delta}\|_p\leqslant\varepsilon}\mathcal{L}(f(\boldsymbol{\theta},\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y})\right], \tag{17}$$

其中输入数据 $\boldsymbol{x}$ 与对应的真实标签 $\boldsymbol{y}$ 采样于分布 $\mathcal{D}$，E 为期望值. 式(17)可以解释为近似解决内部损失函数的最大化问题和外部经验风险的最小化问题，即通过梯度上升找到最坏情况以模拟对抗攻击，并通过梯度下降在模型参数上找到最具鲁棒性的参数，实现防御目的.

对抗训练防御机制已被应用于保护信号调制识别模型免受对抗攻击[25-26]. Bahramali 等人[25]在单一

Table 8 Summary of Related Terms of Three Communication Characteristics on Adversarial Attack Research for Modulation Recognition

表 8 调制识别的对抗攻击研究工作对 3 种通信特性涉及项汇总

相关研究工作	信道效应的复杂性	信号数据的间接性	特征表示的异质性
Kim 等人[15]	✓	✓	
Sadeghi 等人[16]	✓		
Flowers 等人[19]	✓	✓	
Lin 等人[20]			✓
Bahramali 等人[25]	✓		
Kokalj-Filipovic 等人[26]			
Flowers 等人[29]	✓	✓	
Kim 等人[78]	✓	✓	
Restuccia 等人[79]	✓	✓	
Bair 等人[80]			
Kokalj-Filipovic 等人[81]			
Ke 等人[82]			
Davaslioglu 等人[83]			
Kim 等人[84]			
Yi 等人[85]			
Usama 等人[86]			
Usama 等人[87]			
Sahay 等人[88]			✓
Lin 等人[89]			
Hameed 等人[90]	✓	✓	
Hameed 等人[91]	✓	✓	
DelVecchio 等人[92]	✓	✓	
DelVecchio 等人[93]	✓	✓	✓
Hameed [94]	✓	✓	
Berian 等人[95]	✓	✓	
Sandler 等人[96]	✓	✓	

UAP 攻击场景下, 使用单一扰动向量生成对抗样本训练模型; 在扰动生成模型 PGM 攻击场景下, 使用 PGM 生成对抗样本以提高模型鲁棒性.

对抗训练防御机制存在些许问题, 如防御者需要预知攻击者使用的对抗攻击方法[15], 而攻击者也可以对经过对抗训练后的模型生成新的对抗性扰动, 并且对抗训练后通常会降低干净信号样本的识别精度. Madry 等人[67] 提出使用多种对抗样本生成方法训练模型; 为减轻对抗攻击影响, Kokalj-Filipovic 等人[26] 通过自编码器对 DNNs 调制识别模型进行预训练, 从而增强模型鲁棒性[26]; Hameed 等人[91] 引入课程学习[121] 的思想, 使用按信噪比降序的新训练样本, 对分类模型进行训练, 并逐回合增加低信噪样本. Hameed 等人[91] 通过实验证明增加训练样本的数量与类别以及使用基于课程学习的训练策略能够提高分类器模型精度及鲁棒性.

4.2 基于随机平滑的可验证防御

与需预知具体攻击方法的基于对抗训练的防御方法不同, 可验证防御方法对各类攻击方法具有普适性, 即若分类器模型对于输入样本 $\boldsymbol{x}$ 的 l_2 或 l_∞ 领域内的所有样本预测值恒定, 则该模型被认为是具有可靠性的[122]. 如随机平滑, 通过高斯噪声对训练集进行数据增强, 提高模型对多个梯度方向的鲁棒性.

Kim 等人[15,25] 已经在调制识别中实现基于随机平滑的可验证防御方法的应用. 作为对单一 UAP 攻击的防御机制, Kim 等人[15] 将随机平滑作为防御机制, 在训练阶段, 将高斯噪声作为扰动对模型进行训练, 以重新构造平滑分类器 $g(\cdot)$. 对于被扰动的输入样本, $g(\cdot)$ 输出原分类器模型 $f(\cdot)$ 最有可能输出的类别的概率:

$$g(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(f(\boldsymbol{x}+\boldsymbol{n}) = c), \tag{18}$$

其中 $\boldsymbol{n}$ 为高斯噪声, 即 $\boldsymbol{n}\sim\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. 式(18)表示 $g(\boldsymbol{x})$ 可以输出标签 c, 其对应的输入信号样本 $\{\boldsymbol{x}^*\in\mathbb{R}^d: f(\boldsymbol{x}^*) = c\}$ 在高斯分布 $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ 上关于标签 c 具有较高的概率测度. σ 为高斯噪声的标准差, 另一个超参数为添加到每个训练样本的噪声样本数 k, 即 x_i+n_1, x_i+n_2, …, x_i+n_k, 其中 n_j 为标准差 σ 的高斯噪声, 防御者通过调整 2 个超参数以平衡分类准确率及模型对扰动的鲁棒性. 由于在无线通信传输过程中, 接收信号中已包含信道噪声, 随机平滑防御机制会一定程度上增加噪声水平. 在测试阶段, 使用随机平滑方法验证模型分类结果的可信度, 即通过双侧假设检验法来判断分类模型对扰动信号的分类结果的置信度.

4.3 基于降低噪声的防御

基于降低噪声的思想, 防御者接收器根据对攻击者的了解程度, 对受到扰动的接收信号进行信号处理, 以消除扰动信号的影响, 试图重建原始传输信号, 从而提高分类器模型的准确率.

Bahramali 等人[25] 基于对攻击者对抗性扰动的先验知识, 在接收信号中消除扰动信号的影响, 并通过实验表明, 对于单一 UAP 向量攻击, 降低噪声可以完全破坏攻击效果, 而对于扰动生成模型 PGM 生成大量 UAP 的攻击场景下无效.

在真实物理场景下, 降噪处理如噪声检测和滤波是无线通信关键步骤之一, 有助于消除一部分对抗扰动信号的影响, 如 Hameed 等人[91] 讨论了纠错码

率的提高可以有效提高对抗攻击的防御能力，使得被攻击后的信号在接收端能被发现或纠正.

4.4 基于通信特性的防御

在无线通信领域，对抗攻击防御方法也同样存在不同于其他领域的特性，如 Sahay 等人[88] 利用信号特征异质性，提出了基于频域特征的信号调制识别模型. 该模型能削弱黑盒攻击下对抗样本的可迁移性，有效减轻基于时域特征训练的代理模型发起的对抗攻击.

4.5 基于统计学的对抗检测

4.5.1 射频信号样本的峰值平均功率

射频信号对抗样本检测可以利用射频信号的统计学特征，如信号的数字化射频样本数据的峰值平均功率(peak to average power ratio, PAPR)[27]. PAPR 的统计学分布是无线通信中的重要指标，被用作调制信号的签名. 如果信号的 PAPR 与其调制方式的预期统计分布不符，则需进一步分析以确保准确性. 在未知信道条件下，由于多路径衰减以及无线环境影响，相比于其他统计量如信号振幅，PAPR 统计量指标对于线性时间不变的信道而言表现出更高的可靠性[27]. 使用双样本 KS 检验[123]，可计算和评估样本的 PAPR 分布，来判断 PAPR 是否与对抗样本或合法输入样本的统计量相似[27]，该有效性依赖于样本数量.

4.5.2 DNNs 分类模型的 Softmax 输出

Softmax 分布方法是通用的对抗样本检测方法之一，其依赖于 DNNs 模型最后一层的统计数据来检测由对抗样本引起的分布变化，从而判断输入是否具有对抗性. 与 4.5.1 节相同，Kokalj-Filipovic 等人[27] 同样利用双样本 KS 检验来确定训练输入和其对抗样本的 Softmax 分布是否相似，该方法有效性取决于信号波形和信道模型的影响.

4.5.3 基于统计学方法的对抗触发器检测

对于木马攻击，即通过对训练数据嵌入木马等对抗性触发器进行的对抗攻击，可以利用中位数绝对偏差(median absolute deviation, MAD)算法和聚类算法等统计学方法进行检测[83]. MAD 算法通过计算数据点与中位数绝对偏差的中位数，即 $median(|x_i-\tilde{\boldsymbol{X}}|)$，其中 x_i 是 $\boldsymbol{X}$ 的元素，$\tilde{\boldsymbol{X}}=median(|\boldsymbol{X}|)$，从而进行异常点检测，Davaslioglu 等人[83] 的实验指出基于 MAD 算法的对抗触发器检测方法的有效性依赖于训练数据集中毒样本的数量；Davaslioglu 等人[83] 还提出了基于聚类的 2 步异常点检测方法，第 1 步利用 t-SNE 算法对样本数据进行降维，第 2 步再利用聚类方法进行异常点检测，并通过实验指出聚类方法的有效性不受中毒样本数量影响.

综上所述，表 9 对面向调制识别的对抗防御与检测方法进行了比较. 对抗检测方法更偏向数据驱动，依赖于对抗实验先决条件. 防御方法主要围绕提升模型鲁棒性与降低扰动影响 2 方面展开，可以着重考虑无线通信特性在防御方法研究中的应用.

5 调制识别中对抗深度学习的未来展望

随着深度学习成为当前新兴通信系统(如 5G 通信)的核心，其自身面对对抗攻击展现出的脆弱性成为研究焦点. 为了实现更为智能与高效的信息通信与资源管理，应当充分考虑无线通信信号的独特属性，开发安全、灵活、高鲁棒性的深度学习信号调制识别模型，支撑无线安全深入研究和快速发展. 本文将对无线通信中对抗深度学习的未来研究工作进行讨论与展望.

5.1 无线通信数据集

标准化的真实世界信号数据集可以更充分展现无线通信领域的真实物理场景，为研究工作提供更具应用价值的数据支撑. 与 CV 和 NLP 领域相比，无线领域缺少公开的深度学习数据集，现有数据集如 RML 等[4,36,38] 往往未考虑对抗攻击以及信号处理链的影响，仅包含有限的中心频率偏移等硬件与信道的自然影响. 因此，开发更多公开可用的信号数据集对于研究工作非常有价值，这些数据集可以代表不同的真实物理场景，不仅包括信道变化、多重干扰以及波形变化，还可以包括对抗攻击的影响. 虽然有更多的数据集涌现[124-126]，对抗攻击与相应防御技术的实现仍然是亟待解决的应用研究需求，为更好地评估真实信道与无线电硬件对于调制识别任务的影响，满足真实物理环境下基于深度学习模型的对抗攻击与防御的延迟、功率和计算复杂性等研究需求.

5.2 鲁棒性信号特征

在构建无线通信中的深度学习模型过程中，信号特征的鲁棒性也是未来研究工作中的关键. 目前，关于信号特征鲁棒性的研究与讨论较少. Lin 等人[89] 关注信号不同调制方式的鲁棒性差异，通过实验说明 AM-SSB 和 PAM4 的安全性更高，而 BPSK, QAM16 和 QPSK 等易受攻击，但未对信号特征的鲁棒性进行讨论；Ilyas 等人[127] 通过实验表明，目标模型为抵御对抗攻击应将信号的鲁棒性特征纳入构建模型策略之内. 相比之下，在 CV 和 NLP 领域对于特征工程和识别鲁棒性特征的研究较为深入. 因此应当探索新

Table 9 Comparison of Adversarial Defense and Detection Methods for Modulation Recognition

表 9 面向调制识别的对抗防御与检测方法比较

类别	方法	详细技术	主要思想	优点	缺点
防御	对抗训练	UAP 对抗训练[25] 自编码器预训练[26] PGD 对抗训练[91]	最大化对抗损失、最小化模型参数损失，实现对抗信号防御	提升模型鲁棒性，操作相对容易	对训练数据与攻击方法具有依赖性，容易出现过拟合现象
	随机平滑	基于随机平滑的可验证防御[15,25]	通过高斯噪声对训练集进行数据增强，以提高分类器模型对多个梯度方向的鲁棒性	提升模型的泛化能力，对 C&W 攻击比其他防御机制的防御效果好	训练时间较长
	降低噪声	扰动减法[25]	根据对攻击者的了解程度，对受到扰动的接收信号进行信号处理，以消除扰动信号的影响	降低扰动影响，减小攻击成功率	需要扰动信号先验知识，防御效果较差
		纠错编码[91]	提高编码效率	降低扰动影响，提升信息传输速率	防御效果较差
	通信特性	利用信号特征的异质性[88]	提出基于频域特征的信号调制识别模型，以抵御攻击者通过时域特征训练的代理模型发起的对抗攻击	频域特征训练的模型识别性能优于时域，并且对抗攻击防御效果较为优秀	依赖于信号数据集的特征表示形式
检测	统计学方法	PAPR[27]	利用接收信号的数字化射频样本数据的 PAPR 进行统计测试	对抗检测精度较高	依赖于收集的输入样本的数量
		Softmax 输出[27]	利用 DNNs 模型最后一层的统计数据来检测由对抗样本引起的分布变化	对抗检测精度较高	依赖于信号波形和信道模型的影响
		对抗触发器检测[83]	利用 MAD 算法和聚类算法等统计学方法对触发器进行检测	对抗检测精度较高	依赖于训练数据集被投毒样本的数量

技术来识别对于对抗攻击有鲁棒性的重要信号特征，并将非鲁棒性特征与鲁棒性特征相隔离，加入到训练模型中进行特征学习，从而提升模型的泛化能力和对抗攻击的防御能力.

5.3 可验证防御机制

为无线通信中的对抗攻击设计并开发可验证的防御机制也是一项亟待探索的研究方向. 尽管研究者们提出了多种防御技术，包括对抗训练和使用统计学方法等，这些防御技术大多是在攻击方法的启发下而提出的后验防御机制，并且只在受限的实验环境下才具有有效性，缺乏可靠的泛化能力和充足的理论依据支撑，且面对未知的新型攻击可能无效，或将面对更强大的对抗攻击时，防御方法往往是无效的. 已有研究工作提出了使用随机平滑的可验证防御技术[15]显示了良好的防御性能，但仍需进一步严谨地评估与分析，尝试寻找更多能用理论证明的可验证的防御技术，并对调制识别模型及性能进行深入的安全性与鲁棒性分析.

除以上讨论的未来可考虑研究的重点工作内容为将深度学习模型应用于未来无线安全通信中，还可以考虑将不可解释的黑箱模型转向可解释模型，以促进无线安全解决方法的广泛采用，包括但不限于考虑将物理学的基本规律注入到 DNNs 模型的学习过程中，以提高通用性与可解释性，并确保对攻击的鲁棒性.

6 总 结

本文重点关注于信号调制识别领域的对抗深度学习研究现状，系统全面地总结了该领域的对抗攻击与防御技术研究工作，同时充分考虑无线通信环境下信号对抗攻击的独特性质以及模型鲁棒性、技术灵活性、无线通信安全性. 提出了信号调制识别领域的通用对抗攻击分类框架与威胁模型，将该领域研究工作系统地分类为物理自我防御式攻击和数字直接访问式攻击. 同时，详细阐述了对抗攻击方法、对抗样本生成技术与理论公式、对抗检测与防御技术的研究工作，系统提炼了无线通信对抗攻击研究的 3 个维度特性并归纳相应的处理方法. 最后对该领域的对抗深度学习的未来前景进了讨论. 面向信号调制识别领域的对抗深度学习研究属于一个待探索的研究领域，迫切寻求研究者们进一步的探索，以量化对抗攻击的影响，更好地进行对抗检测与防御工作，为无线通信安全性提供更有潜力的深度学习解决方案.

作者贡献声明：徐冬月负责调研并完成论文的撰写；田蕴哲负责论文审阅，并给出详细修改指导意见；陈康负责论文部分格式修订；李轶珂、吴亚伦、童恩栋、牛温佳、刘吉强、史忠植负责论文审阅.

参 考 文 献

[1] Bhatti F A, Khan M J, Selim A, et al. Shared spectrum monitoring using deep learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(4): 1171−1185

[2] Dobre O A, Abdi A, Bar-Ness Y, et al. Survey of automatic

modulation classification techniques: Classical approaches and new trends[J]. IET Communications, 2007, 1(2): 137−156

[3] Zhang Wenhan, Feng Mingjie, Krunz M, et al. Signal detection and classification in shared spectrum: A deep learning approach[C/OL]//Proc of the 40th IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2021[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9488834/

[4] O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognition networks[C]//Proc of the 17th Int Conf on Engineering Applications of Neural Networks. Cham: Springer, 2016: 213−226

[5] West N E, O'Shea T. Deep architectures for modulation recognition[C/OL]//Proc of the 2017 IEEE Int Symp on Dynamic Spectrum Access Networks. Piscataway, NJ: IEEE, 2017[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/7920754/

[6] Hong Dehua, Zhang Zilong, Xu Xiaodong. Automatic modulation classification using recurrent neural networks[C]//Proc of the 3rd IEEE Int Conf on Computer and Communications. Piscataway, NJ: IEEE, 2017: 695−700

[7] Rajendran S, Meert W, Giustiniano D, et al. Deep learning models for wireless signal classification with distributed low-cost spectrum sensors[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(3): 433−445

[8] Hu Shisheng, Pei Yiyang, Liang P P, et al. Deep neural network for robust modulation classification under uncertain noise conditions[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1): 564−577

[9] Xu Jialang, Luo Chunbo, Parr G, et al. A spatiotemporal multi-channel learning framework for automatic modulation recognition[J]. IEEE Wireless Communications Letters, 2020, 9(10): 1629−1632

[10] Huynh-The T, Hua C H, Pham Q V, et al. MCNet: An efficient CNN architecture for robust automatic modulation classification[J]. IEEE Communications Letters, 2020, 24(4): 811−815

[11] Zhang Fuxin, Luo Chunbo, Xu Jialang, et al. An efficient deep learning model for automatic modulation recognition based on parameter estimation and transformation[J]. IEEE Communications Letters, 2021, 25(10): 3287−3290

[12] Shi Fengyuan, Hu Zeming, Yue Chunsheng, et al. Combining neural networks for modulation recognition[J]. Digital Signal Processing, 2022, 120: 103264

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

[14] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2015

[15] 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, 2022, 21(6): 3868−3880

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

[17] Tu Ya, Lin Yun, Wang Jin, et al. Semi-supervised learning with generative adversarial networks on digital signal modulation classification[J]. Computers, Materials & Continua, 2018, 55(2): 243−254

[18] Zhao Haojun, Lin Yun, Gao Song, et al. Evaluating and improving adversarial attacks on DNN-based modulation recognition[C/OL]//Proc of the 63rd IEEE Global Communications Conf. Piscataway, NJ: IEEE, 2020[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9322088/

[19] Flowers B, Buehrer R M, Headley W C. Evaluating adversarial evasion attacks in the context of wireless communications[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1102−1113

[20] Lin Yun, Zhao Haojun, Ma Xuefei, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2021, 70(1): 389−401

[21] Tu Ya, Lin Yun, Zha Haoran, et al. Large-scale real-world radio signal recognition with deep learning[J]. Chinese Journal of Aeronautics, 2022, 35(9): 35−48

[22] Bao Zhida, Lin Yun, Zhang Sicheng, et al. Threat of adversarial attacks on DL-based IoT device identification[J]. IEEE Internet of Things Journal, 2022, 9(11): 9012−9024

[23] Zhang Sicheng, Lin Yun, Bao Zhida, et al. A lightweight modulation classification network resisting white box gradient attacks[J]. Security and Communication Networks, 2021(1): 8921485

[24] Hou Changbo, Liu Guowei, Tian Qiao, et al. Multisignal modulation classification using sliding window detection and complex convolutional network in frequency domain[J]. IEEE Internet of Things Journal, 2022, 9(19): 19438−19449

[25] Bahramali A, Nasr M, Houmansadr A, et al. Robust adversarial attacks against DNN-based wireless communication systems[C]//Proc of the 28th ACM SIGSAC Conf on Computer and Communications Security(CCS'21). New York: ACM, 2021: 126−140

[26] Kokalj-Filipovic S, Miller R, Chang N, et al. Mitigation of adversarial examples in RF deep classifiers utilizing autoencoder pre-training[C/OL]//Proc of the 2019 Int Conf on Military Communications and Information Systems. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/8842663/

[27] Kokalj-Filipovic S, Miller R, Vanhoy G. Adversarial examples in RF deep learning: Detection and physical robustness[C/OL]//Proc of the 2019 IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/8969138/

[28] Adesina D, Hsieh C C, Sagduyu Y E, et al. Adversarial machine learning in wireless communications using RF data: A review[J]. IEEE Communications Surveys & Tutorials, 2022, 25(1): 77−100

[29] Flowers B, Buehrer R M, Headley W C. Communications aware adversarial residual networks for over the air evasion attacks[C]//Proc of the 2019 IEEE Military Communications Conf. Piscataway, NJ: IEEE, 2019: 133−140

[30] Huynh-The T, Pham Q V, Nguyen T V, et al. Automatic modulation classification: A deep architecture survey[J]. IEEE Access, 2021, 9: 142950−142971

[31] O'Shea T, Hoydis J. An introduction to deep learning for the physical layer[J]. IEEE Transactions on Cognitive Communications and Networking, 2017, 3(4): 563−575

[32] Erpek T, O'Shea T J, Sagduyu Y E, et al. Deep learning for wireless communications[J]. Development and Analysis of Deep Learning

Architectures, 2020, 867: 223−266

[33] Peng Shengliang, Sun Shujun, Yao Y D. A survey of modulation classification using deep learning: Signal representation and data preprocessing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7020−7038

[34] Zhou Ruolin, Liu Fugang, Gravelle C W. Deep learning for modulation recognition: A survey with a demonstration[J]. IEEE Access, 2020, 8: 67366−67376

[35] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014

[36] O'Shea T J, Roy T, Clancy T C. Over-the-air deep learning based radio signal classification[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 168−179

[37] Blossom E. GNU Radio: Tools for exploring the radio frequency spectrum[J/OL]. Linux Journal, 2004[2023-02-02]. https://dl.acm.org/doi/fullHtml/10.5555/993247.993251

[38] Tekbiyik K, Ekti A R, Gorcin A, et al. Robust and fast automatic modulation classification with CNN under multipath fading channels[C/OL]//Proc of the 91st IEEE Vehicular Technology Conf. Piscataway, NJ: IEEE, 2020[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9128408/

[39] Mendis G J, Wei-Kocsis J, Madanayake A. Deep learning based radio-signal identification with hardware design[J]. IEEE Transactions on Aerospace and Electronic Systems, 2019, 55(5): 2516−2531

[40] Lee J, Kim B, Kim J, et al. Deep neural network-based blind modulation classification for fading channels[C]//Proc of the 8th Int Conf on Information and Communication Technology Convergence. Piscataway, NJ: IEEE, 2017: 551−554

[41] Liu Xiaoyu, Yang Diyu, Gamal A E. Deep neural network architectures for modulation classification[C]//Proc of the 51st Asilomar Conf on Signals, Systems, and Computers. Piscataway, NJ: IEEE, 2017: 915−919

[42] Mendis G J, Wei Jin, Madanayake A. Deep learning-based automated modulation classification for cognitive radio[C/OL]//Proc of the 2016 IEEE Int Conf on Communication Systems. Piscataway, NJ: IEEE, 2016[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/7833571/

[43] Ali A, Yangyu F. *k*-sparse autoencoder-based automatic modulation classification with low complexity[J]. IEEE Communications Letters, 2017, 21(10): 2162−2165

[44] Dai Ao, Zhang Haijian, Sun Hong. Automatic modulation classification using stacked sparse auto-encoders[C]//Proc of the 2016 IEEE Int Conf on Signal Processing. Piscataway, NJ: IEEE, 2016: 248−252

[45] Xie Wenwu, Hu Sheng, Yu Chao, et al. Deep learning in digital modulation recognition using high order cumulants[J]. IEEE Access, 2019, 7: 63760−63766

[46] Shi Wenzhe, Liu Dejun, Cheng Xing, et al. Particle swarm optimization-based deep neural network for digital modulation recognition[J]. IEEE Access, 2019, 7: 104591−104600

[47] Yashashwi K, Sethi A, Chaporkar P. A learnable distortion correction module for modulation recognition[J]. IEEE Wireless Communications Letters, 2019, 8(1): 77−80

[48] Zeng Yuan, Zhang Meng, Han Fei, et al. Spectrum analysis and convolutional neural network for automatic modulation recognition[J]. IEEE Wireless Communications Letters, 2019, 8(3): 929−932

[49] Peng Shengliang, Jiang Hanyu, Wang Huaxia, et al. Modulation classification based on signal constellation diagrams and deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(3): 718−727

[50] Zhang Zufan, Wang Chun, Gan Chenquan, et al. Automatic modulation classification using convolutional neural network with features fusion of SPWVD and BJD[J]. IEEE Transactions on Signal and Information Processing over Networks, 2019, 5(3): 469−478

[51] Wu Hao, Li Yaxing, Zhou Liang, et al. Convolutional neural network and multi - feature fusion for automatic modulation classification[J]. Electronics Letters, 2019, 55(16): 895−897

[52] Wang Yu, Liu Miao, Yang Jie, et al. Data-driven deep learning for automatic modulation recognition in cognitive radios[J]. IEEE Transactions on Vehicular Technology, 2019, 68(4): 4074−4077

[53] Wang Danshi, Zhang Min, Li Ze, et al. Modulation format recognition and OSNR estimation using CNN-based deep learning[J]. IEEE Photonics Technology Letters, 2017, 29(19): 1667−1670

[54] Lee J H, Kim K Y, Shin Y. Feature image-based automatic modulation classification method using CNN algorithm[C/OL]//Proc of the 2019 Int Conf on Artificial Intelligence in Information and Communication. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexplore.iee e.org/abstract/document/8669002/

[55] Chen Zhuangzhi, Cui Hui, Xiang Jingyang, et al. SigNet: A novel deep learning framework for radio signal classification[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 8(2): 529−541

[56] Hermawan A P, Ginanjar R R, Kim D S, et al. CNN-based automatic modulation classification for beyond 5G communications[J]. IEEE Communications Letters, 2020, 24(5): 1038−1041

[57] Ke Ziqi, Vikalo H. Real-time radio technology and modulation classification via an LSTM auto-encoder[J]. IEEE Transactions on Wireless Communications, 2022, 21(1): 370−382

[58] Chang Shuo, Huang Sai, Zhang Ruiyun, et al. Multitask-learning-based deep neural network for automatic modulation classification[J]. IEEE Internet of Things Journal, 2022, 9(3): 2192−2206

[59] Ghasemzadeh P, Banerjee S, Hempel M, et al. A novel deep learning and polar transformation framework for an adaptive automatic modulation classification[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13243−13258

[60] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410−14430

[61] Cao Yangjie, Jia Lili, Chen Yongxia, et al. Recent advances of generative adversarial networks in computer vision[J]. IEEE Access, 2019, 7: 14985−15006

[62] Akhtar N, Mian A, Kardan N, et al. Advances in adversarial attacks and defenses in computer vision: A survey[J]. IEEE Access, 2021, 9: 155161−155196

[63] Zou Zhengxia, Lei Sen, Shi Tianyang, et al. Deep adversarial decomposition: A unified framework for separating superimposed images[C]//Proc of the 2020 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 12803−12813

[64] Sarafianos N, Xu Xiang, Kakadiaris I. Adversarial representation

learning for text-to-image matching[C]//Proc of the 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 5814−5824

[65] Wang Hao, Gong Dihong, Li Zhifeng, et al. Decorrelated adversarial learning for age-invariant face recognition[C]//Proc of the 2019 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3522−3531

[66] Zhong Yaoyao, Deng Weihong. Adversarial learning with margin-based triplet embedding regularization[C]//Proc of the 2019 IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6548−6557

[67] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[68] Bai Tao, Luo Jinqi, Zhao Jun, et al. Recent advances in adversarial training for adversarial robustness[C]//Proc of the 21st Int Joint Conf on Artificial Intelligence. Piscataway, NJ: IEEE, 2021: 4312−4321

[69] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv preprint, arXiv: 1705.07204, 2017

[70] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C/OL]//Proc of the 2018 IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018[2023-02-02]. https://ieeexplor e.ieee.org/abstract/document/8424625/

[71] Hannun A, Case C, Casper J, et al. Deep speech: Scaling up end-to-end speech recognition[J]. arXiv preprint, arXiv: 1412.5567, 2014

[72] Zheng Haibin, Chen Jinyin, Zhang Yan. Survey of adversarial attack, defense and robustness analysis for natural language processing[J]. Journal of Computer Research and Development, 2021, 58(8): 1727−1750 (in Chinese)
(郑海斌, 陈晋音, 章燕. 面向自然语言处理的对抗攻防与鲁棒性分析综述[J]. 计算机研究与发展, 2021, 58(8): 1727−1750)

[73] Zhang W E, Sheng Q Z, Alhazmi A, et al. Adversarial attacks on deep-learning models in natural language processing: A survey[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(3): 1−41

[74] Zhou Yi, Zheng Xiaoqing, Hsieh C J, et al. Defense against adversarial attacks in NLP via dirichlet neighborhood ensemble[J]. arXiv preprint, arXiv: 2006.11627, 2020

[75] Araujo V, Carvallo A, Aspillaga C, et al. On adversarial examples for biomedical NLP tasks[J]. arXiv preprint, arXiv: 2004.11157, 2020

[76] Miao Deshui, Zhang Jiaqi, Xie Wenbo, et al. Simple contrastive representation adversarial learning for NLP tasks[J]. arXiv preprint, arXiv: 2111.13301, 2021

[77] Han Xudong, Baldwin T, Cohn T. Decoupling adversarial training for fair NLP[C/OL]//Proc of the Joint Conf of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing(ACL-IJCNLP 2021). 2021: 471−477. [2023-02-02]. https://aclanthology.org/2021.findings-acl.41.pdf

[78] Kim B, Sagduyu Y E, Davaslioglu K, et al. Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels[C/OL]//Proc of the 54th Annual Conf on Information Sciences and Systems(CISS). Piscataway, NJ: IEEE, 2020[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9086166/

[79] Restuccia F, D'Oro S, Al-Shawabka A, et al. Generalized wireless adversarial deep learning[C]//Proc of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2020: 49−54

[80] Bair S, DelVecchio M, Flowers B, et al. On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition[C/OL]//Proc of the ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2019: 25−30

[81] Kokalj-Filipovic S, Miller R, Morman J. Targeted adversarial examples against RF deep classifiers[C]//Proc of the ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2019: 6−11

[82] Ke Da, Huang Zhitao, Wang Xiang, et al. Application of adversarial examples in communication modulation classification[C]//Proc of the 19th Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2019: 877−882

[83] Davaslioglu K, Sagduyu Y E. Trojan attacks on wireless signal classification with adversarial machine learning[C/OL]//Proc of the 12th IEEE Int Symp on Dynamic Spectrum Access Networks. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/8935782/

[84] Kim B, Sagduyu Y E, Erpek T, et al. Adversarial attacks with multiple antennas against deep learning-based modulation classifiers[C/OL]//Proc of the 2020 IEEE Globecom Workshops. Piscataway, NJ: IEEE, 2020[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9367473/

[85] Yi Jinho, Gamal A E. Gradient-based adversarial deep modulation classification with data-driven subsampling[J]. arXiv preprint, arXiv: 2104.06375, 2021

[86] Usama M, Asim M, Qadir J, et al. Adversarial machine learning attack on modulation classification[C/OL]//Proc of the 2019 UK/China Emerging Technologies. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexpl ore.ieee.org/abstract/document/8881843

[87] Usama M, Qadir J, Al-Fuqaha A. Black-box adversarial ML attack on modulation classification[J]. arXiv preprint, arXiv: 1908.00635, 2019

[88] Sahay R, Brinton C G, Love D J. Frequency-based automated modulation classification in the presence of adversaries[C/OL]//Proc of the 56th IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2021[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9500583/

[89] Lin Yun, Zhao Haojun, Tu Ya, et al. Threats of adversarial attacks in DNN-based modulation recognition[C]//Proc of the 39th IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2020: 2469−2478

[90] Hameed M Z, Gyorgy A, Gunduz D. Communication without interception: Defense against modulation detection[C/OL]//Proc of the 7th IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2019[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/8969541/

[91] 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

[92] DelVecchio M, Flowers B, Headley W C. Effects of forward error

correction on communications aware evasion attacks[C/OL]//Proc of the 31st Annual Int Symp on Personal, Indoor and Mobile Radio Communications. Piscataway, NJ: IEEE, 2020[2023-02-02]. https://ieeexplore.ieee.org/abst ract/document/9217343/

[93] DelVecchio M, Arndorfer V, Headley W C. Investigating a spectral deception loss metric for training machine learning-based evasion attacks[C]//Proc of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2020: 43−48

[94] Hameed M Z. New quality measures for adversarial attacks with applications to secure communication[D]. London: Imperial College London, 2020

[95] Berian A, Staab K, Ditzler G, et al. Adversarial filters for secure modulation classification[C]//Proc of the 55th Asilomar Conf on Signals, Systems, and Computers. Piscataway, NJ: IEEE, 2021: 361−367

[96] Sandler R A, Relich P K, Cho C, et al. Real-time over-the-air adversarial perturbations for digital communications using deep neural networks[J]. arXiv preprint, arXiv: 2202.11197, 2022

[97] Chakraborty A, Alam M, Dey V, et al. Adversarial attacks and defences: A survey[J]. arXiv preprint, arXiv: 1810.00069, 2018

[98] Yi Shi, Sagduyu Y, Grushin A. How to steal a machine learning classifier with deep learning[C/OL]//Proc of the 2017 IEEE Int Symp on Technologies for Homeland Security. Piscataway, NJ: IEEE, 2017[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/7943475/

[99] Shi Y, Sagduyu Y E, Davaslioglu K, et al. Active deep learning attacks under strict rate limitations for online API calls[C/OL]//Proc of the 2018 IEEE Int Symp on Technologies for Homeland Security. Piscataway, NJ: IEEE, 2018[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/8574124/

[100] Pi Lei, Lu Zhuo, Sagduyu Y, et al. Defending active learning against adversarial inputs in automated document classification[C]//Proc of the 4th IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2016: 257−261

[101] Shi Yi, Sagduyu Y E, Davaslioglu K, et al. Generative adversarial networks for black-box API attacks with limited training data[C]//Proc of the 2018 IEEE Int Symp on Signal Processing and Information Technology. Piscataway, NJ: IEEE, 2018: 453−458

[102] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines[J]. arXiv preprint, arXiv: 1206.6389, 2012

[103] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]// Proc of the 40th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 707−723

[104] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[J]. arXiv preprint, arXiv: 1607.02533

[105] Kim B, Sagduyu Y E, Erpek T, et al. Channel effects on surrogate models of adversarial attacks against wireless signal classifiers[C/OL]//Proc of the 57th IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2021[2023-02-02]. https://ieeexplore.ieee.org/abstract/document/9500374/

[106] Wyner A D. The wire-tap channel[J]. Bell System Technical Journal, 1975, 54(8): 1355−1387

[107] Gunduz D, Brown D R, Poor H V. Secret communication with feedback[C/OL]//Proc of the Int Symp on Information Theory and Its Applications. Piscataway, NJ: IEEE, 2008[2023-02-02]. https://ieeexplor e.ieee.org/abstract/document/4895417/

[108] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//Proc of the 1st IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 372−387

[109] Kim B, Sagduyu Y E, Davaslioglu K, et al. How to make 5G communications "Invisible": Adversarial machine learning for wireless privacy[C]//Proc of the 54th Asilomar Conf on Signals, Systems, and Computers. Piscataway, NJ: IEEE, 2020: 763−767

[110] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks[C/OL]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018[2023-02-02]. https://ojs.aaai.org/index.php/AAAI/article/view/11672

[111] Zhang Tian, Yang Kuiwu, Wei Jianghong. Survey on detecting and defending adversarial examples for image data[J]. Journal of Computer Research and Development, 2022, 59(6): 1315−1328 (in Chinese)
(张田, 杨奎武, 魏江宏. 面向图像数据的对抗样本检测与防御技术综述[J]. 计算机研究与发展, 2022, 59(6): 1315−1328)

[112] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9185−9193

[113] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 86−94

[114] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//Proc of the 38th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39−57

[115] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[116] Papernot N, Faghri F, Carlini N, et al. Technical report on the CleverHans v2.1. 0 adversarial examples library[J]. arXiv preprint, arXiv: 1610.00768, 2016

[117] Ramjee S, Ju Shengtai, Yang Diyu, et al. Ensemble wrapper subsampling for deep modulation classification[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(4): 1156−1170

[118] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale[C/OL]//Proc of Int Conf on Learning Representations.[2024-05-26]. https://openreview.net/forum?id=BJm4T4Kgx

[119] Sagduyu Y E, Shi Yi, Erpek T. Adversarial deep learning for over-the-air spectrum poisoning attacks[J]. IEEE Transactions on Mobile Computing, 2021, 20(2): 306−319

[120] Shi Yi, Erpek T, Sagduyu Y E, et al. Spectrum data poisoning with adversarial deep learning[C]//Proc of the 37th IEEE Military Communications Conf. Piscataway, NJ: IEEE, 2018: 407−412

[121] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning[C]//Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 41−48

[122] Cohen J, Rosenfeld E, Kolter J Z. Certified adversarial robustness via randomized smoothing[C]//Proc of the 36th Int Conf on Machine Learning. New York: PMLR. 2019: 1310−1320

[123] Dodge Y. Kolmogorov–Smirnov test[J/OL]. The Concise

Encyclopedia of Statistics, 2008[2023-03-08]. http://link.springer.com/10.1007/978-0-387-32 833-1_214

[124] Al-shawabka A, Restuccia F, D'Oro S, et al. Massive-scale I/Q datasets for WiFi radio fingerprinting[J]. Computer Networks, 2020, 182: 107566

[125] Alkhateeb A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications[J]. arXiv preprint, arXiv: 1902.06435, 2019

[126] Ezuma M, Erden F, Anjinappa C, et al. Drone remote controller RF signal dataset [DB/OL]. IEEE DataPort, 2020[2023-03-08]. https://ieee-dataport.org/open-access/drone-remote-controller-rf-signal-dataset

[127] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features[J/OL]. Advances in Neural Information Processing Systems, 2019[2023-03-08]. https://proceedings.neurips.cc/paper/2019/has h/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html



Xu Dongyue, born in 1999. Master. Her main research interest includes adversarial machine learning.

徐冬月, 1999 年生. 硕士. 主要研究方向为对抗机器学习.



Tian Yunzhe, born in 1999. PhD candidate. His main research interest includes AI security.

田蕴哲, 1999 年生. 博士研究生. 主要研究方向为人工智能安全.



Chen Kang, born in 1999. Master. His main research interest includes signal modulation recognition.

陈　康, 1999 年生. 硕士. 主要研究方向为信号调制识别.



Li Yike, born in 1995. PhD. Her main research interest includes AI security.

李轶珂, 1995 年生. 博士. 主要研究方向为人工智能安全.



Wu Yalun, born in 1994. PhD candidate. His main research interest includes AI security.

吴亚伦, 1994 年生. 博士研究生. 主要研究方向为人工智能安全.



Tong Endong, born in 1986. PhD, lecturer. Member of CCF. His main research interests include AI security, services computing, and data mining.

童恩栋, 1986 年生. 博士, 讲师. CCF 会员. 主要研究方向为人工智能安全、服务计算、数据挖掘.



Niu Wenjia, born in 1982. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include AI security and data mining.

牛温佳, 1982 年生. 博士, 教授, 博士生导师. CCF 杰出会员. 主要研究方向为人工智能安全、数据挖掘.



Liu Jiqiang, born in 1973. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include trusted computing, cryptographic protocols, privacy preserving, and network security.

刘吉强, 1973 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为可信计算、加密协议、隐私保护、网络安全.



Shi Zhongzhi, born in 1941. Professor, PhD supervisor. Fellow of CCF. His main research interests include artificial intelligence, machine learning, and neural computing and cognitive science.

史忠植, 1941 年生. 教授, 博士生导师. CCF 会士. 主要研究方向为人工智能、机器学习、神经计算与认知科学.