ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 6,Nov. 2023,pp. 1235－1256
DOI:10. 16337/j. 1004-9037. 2023. 06. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

信号调制识别的对抗样本攻防技术研究进展

江　汉[1]，胡　林[2]，李　文[1]，焦雨涛[1]，徐煜华[1]，徐逸凡[1]

(1. 陆军工程大学通信工程学院,南京 210007;2. 中国电子科技集团公司第二十九研究所,成都 610036)

摘　要：对调制识别的对抗样本攻击这一研究热点进行了综述,首先给出调制识别中对抗样本的的相关概述和专业术语,将对抗样本攻击和防御方法的相关研究成果进行梳理回顾,并对现有对抗攻击方法进行分类,阐述了其生成机理。针对在无线通信场景下对抗攻击所面临的挑战,梳理了调制识别任务下的对抗样本攻击现有关键技术,深入分析了调制识别对抗样本攻击的防御技术。最后,依据现有研究与潜在的机遇和挑战,结合人工智能算法的各项优势,对下一代智能无线通信中对抗攻击的技术方向和发展前景进行了展望。

关键词：深度学习;自动调制识别;对抗攻击;对抗攻击防御;无线安全

中图分类号：TN972　　**文献标志码**:A

Research Progress of Adversarial Attack and Defense for Signal Modulation Recognition

JIANG Han[1],HU Lin[2], LI Wen[1], JIAO Yutao[1], XU Yuhua[1], XU Yifan[1]

(1. College of Communication Engineering, Army Engineering University of PLA, Nanjing 210007, China;2. China Electronics Technology Corporation 29th Research Institute, Chengdu 610036, China)

Abstract: The hot research topic of adversarial sample attacks on modulation recognition is reviewed. Firstly, we introduce the concepts and terms related to modulation recognition adversarial samples. Then we review and sort out the related research results on adversarial sample attacks and defense methods, and classify the existing adversarial attack methods and explain their generation mechanisms. Finally, based on the existing research, potential opportunities and challenges, and the advantages of artificial intelligence algorithms, the technical directions and development prospects of adversarial attacks in next-generation intelligent wireless communications are presented.

Key words: deep learning; automatic modulation recognition; adversarial attack; defense against attacks; wireless security

引　　言

机器学习(Machine learning, ML)包括传统的机器学习方法、深度学习、强化学习和联邦学习等。机器学习具有特征自提取等优势,减少了人为操作。机器学习用于无线通信网络有以下优势:(1)由于

基金项目:国家自然联合基金(U22B2002);国家自然科学基金(62101594);江苏省基础研究计划(自然科学基金)前沿引领技术基础研究专项(BK20212001)。
收稿日期:2022-04-16;**修订日期**:2022-08-29

无线通信中海量的通信数据，ML可以识别无法通过传统分析方法的相关性和异常，便于更好地提取无线网络的特性。(2)ML具有处理海量的、结构化的和非结构数据化的能力。(3)ML能容忍在数据缺失或部分数据异常情况下，通过机器学习的方式仍然可以获得比较好的识别效果和决策判断。这也是传统方法所不具有的。(4)ML具有强大的控制决策能力，有助于实现无线网络的智能化。因此，业界以射频数据为依托，开发人工智能驱动的无线通信系统解决方案备受关注[1-3]。例如，卷积神经网络(Convolutional neural network, CNN)已用于调制识别[4]和信道译码[5]。前馈神经网络(Feedforward neural network, FNN)和长短时间记忆(Long and short-term memory, LSTM)模型已被用于设备指纹识别[6-9]。各种深度神经网络(Deep neural network, DNN)模型已被提出用于频谱感知频谱预测[10]、无线资源管理[11]、波束预测[12]以及智能反射面通信[13]。文献[14]指出，通信与计算融合是无线通信系统可持续发展的一个有前景的解决方案，因此在高动态的通信环境中，海量通信数据的计算和资源分配越来越依赖机器学习方法。

尽管基于深度学习(Deep learning, DL)的技术在解决无线通信中复杂任务有很大的优势和潜力，但在现实世界部署中，与传统的信号处理方法相比，这种数据驱动方法的可靠性和鲁棒性不能得到保证。其中一个主要原因就是在DL训练和测试过程来自对抗样本的恶意干扰。研究对抗攻击的目标是对其攻击过程和原理进行全面系统理解，并设计防御算法避免此类攻击，从而设计更鲁棒、更安全的DL模型。对抗样本的产生具体为，在特征空间上精心设计一个矢量扰动从而误导已经训练好的DL模型。这类攻击在模型的梯度方向上相对于输入加入极小的增量值，或者通过解决一个约束优化问题，在输入特征空间上产生一个可能误导目标DL模型的向量。

对抗攻击已经在计算机视觉(Computer vision, CV)和自然语言处理(Natural language processing, NLP)等领域得到了广泛的研究应用[15-17]。一个典型的例子就是当对一张熊猫图像加入精心设计的扰动后，DL分类器将其误分类为一张长臂猿[18]。在调制识别领域，已有研究表明，如需构造出信号领域中的对抗样本，只需在干净信号样本上添加精心设计的微弱扰动。如对原始二进制相移键控(Binary phase shift keying, BPSK)、正交相移键控(Quadrature phase shift keying, QPSK)等调制方式的信号添加一个小的对抗扰动，形成对抗样本，对抗样本输入到分类器，这就可能导致预先训练好的分类器将信号误分类为相正交振幅调制(Quadrature amplitude modulation, 64QAM)或者其他调制方式的信号。在通信任务中，如果调制识别任务出错，这就会给智能无线通信系统、认知无线电网络、电磁侦察和卫星导航等应用带来严重的危害。

在军事领域，信号识别技术是在拦截敌方信号之后获取敌方机密信息的关键技术，识别敌方信号的调制方式和参数，最后完成解调解密等其他后续工作。此外，调制识别在频谱监测、频谱管理和认知无线电等领域也扮演着重要角色。研究面向信号调制识别的对抗攻击技术意义颇丰。基于DNN的自动调制识别(Automatic modulation recognition, AMR)技术具有特征自提取、识别精度高、人工干预少的优势，故基于人工智能的调制识别手段也成为了如今的研究热点。此外，随着物联网(Internet of things, IoT)的普及，未来的第6代移动蜂窝网络(6G)除了能力比5G提高10～100倍外，还应实现智能化和开放性，以适应物联网中不断变化复杂的服务[19]。基于机器学习的智能通信是未来6G的重要方向。因此设计更安全、更高效和更鲁棒的人工智能算法也为“万物互联的时代”提供了坚实的基石。

本文通过分析无线通信的独特特性及其对抗攻击对AMR的攻击威胁，重点讨论了在无线通信领域中产生、检测和防御对抗攻击的方法。已有多项在CV[20]和NLP[21]领域的对抗攻击综述工作。但本文是首次聚焦调制识别领域对抗攻击的攻击和防御的综述研究工作。本文从人工智能通信的优势入手，首先介绍了通信领域的对抗机器学习研究的重要现实意义，分析了通信领域对抗攻击技术研究的特点与挑战。其次对调制识别中的对抗攻击现有研究进行总结。从对抗样本概念、分类、产生机制以

及对抗攻击防御技术层层递进展开阐述。最后,依据现有研究基础与潜在的机遇和挑战,结合人工智能算法的各项优势,探讨了下一代智能无线通信中对抗机器学习的技术方向和发展前景。本文研究结构如图1所示。

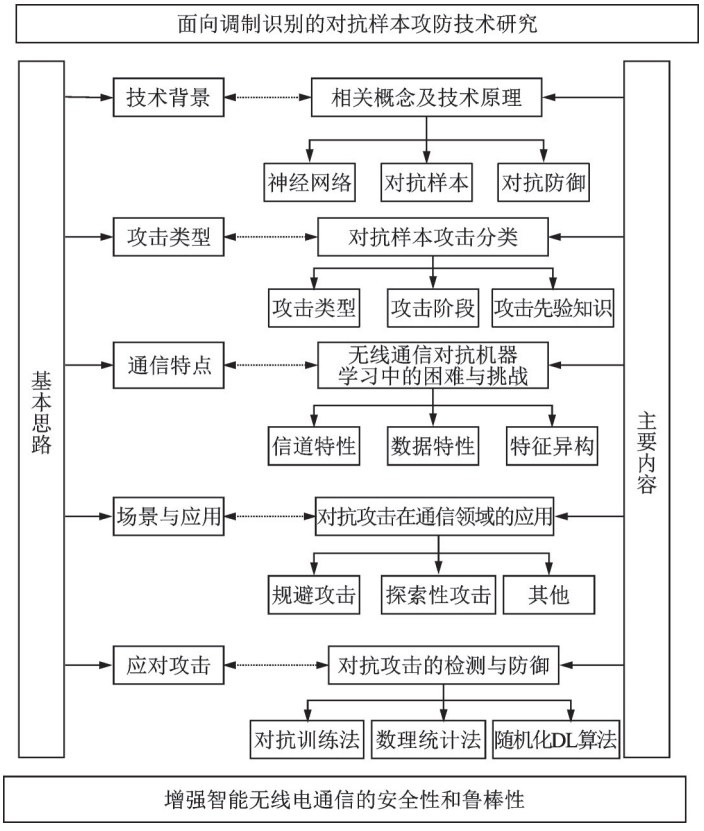

图1　通信领域的对抗机器学习特点与问题研究

Fig.1　Study of characteristics and problems of adversarial machine learning in communications

1　对抗样本攻防

1.1　深度神经网络基本概念

DNN起源于生物神经网络,是用神经元节点按特定的层次结构连接而成的网络。在一个人工智能系统中,使用DNN的有监督学习通常表示为一个映射f,f从输入数据到类(标签)的映射建模。为了学习这个输入和输出的关系,需要估计出权重矩阵$\boldsymbol{w}$的值。损失函数$l(x,y,\boldsymbol{w})$用来进行逐点计算模型的预测值$\hat{y}=f(x)$和真实标签值y之间的误差。X是数据集中的样本点,y是标签类别里某一类别,为了估算权重矩阵$\boldsymbol{w}$,代价函数$J(\boldsymbol{w})$为X中n个训练数据样本的平均损失,计算为

$$J(\boldsymbol{w})=J(X,y,\boldsymbol{w})=\frac{1}{n}\sum_{(x_i,y_i)}l(x_i,y_i,\boldsymbol{w}) \tag{1}$$

将代价函数$J(\boldsymbol{w})$求最小值为

$$\arg\min_{\boldsymbol{w}\in\mathbf{R}^n}J(\boldsymbol{w}) \tag{2}$$

然后,通过DNN权值$\boldsymbol{w}$索引的DNN分类器被确定为$f_{\boldsymbol{w}}:X\to\mathbf{R}^C$,$C$是总的标签数。对于每个$x\in X$,$f_{\boldsymbol{w}}(x)$分配一个标签表,$\hat{y}=\arg\max_{y\in C}f_{\boldsymbol{w}}(x,y)$,$f_{\boldsymbol{w}}(x)$对应在$C$中类$y$的输出。

1.2　对抗样本攻击

对抗样本是指在输入样本数据中添加精心设计且人眼难以察觉的细微扰动的样本,对抗样本攻击DL分类器后,错误的分类结果将被输出为高置信度。无目标对抗样本攻击产生的核心技术是使损失函数$L(x,y)$最大化,致使模型预测值与真实标签结果之间的差异最大化。

当前的对抗样本攻击研究工作大多是数据攻击,即假设对抗样本能直接进入基于人工智能的分类器中。具体而言,DL中的对抗攻击是指对干净数据添加精心设计的扰动后(此扰动依托于设计的算法),故意扰乱、愚弄机器学习分类、识别等决策的技术。用数学语言表述即为对深度学习分类器进行对抗攻击,方法就是将扰动r_x与输入样本x一起送至训练好的分类器$f_{\boldsymbol{w}}$中去,从而让分类器分类出错,即$f_{\boldsymbol{w}}(x+r_x)\neq f_{\boldsymbol{w}}(x)$。因为攻击还需要具有隐蔽性,扰动$r_x$的功率要小,即$x+r_x$信号与未受扰动的干净信号保持细微的差距。一般用范数对r_x进行功率约束,即$\|r_x\|_p\leqslant\varepsilon$,$\varepsilon$是一个很小的正数值,$\|\cdot\|_p$表示$l_p$范数,$p\in\{1,2,\cdots\}$表示范数的类型。故$x$和$f_{\boldsymbol{w}}$的对抗扰动$r_x$通过求解以下优化问题来确定,即

$$\min_{r_x} \| r_x \|_p \quad \text{s.t. } f_w(x) \neq f_w(x + r_x) \tag{3}$$

在无线通信场景下，r_x最好选择l_2范数来进行功率约束[19]，而在CV领域，使用l_1范数来隐蔽人眼的感知[22]。对抗样本有“四两拨千斤”的作用，对抗样本大多都是在深度网络模型梯度上设计的，只在智能模型决策阶段影响分类器，所以比普通噪声所需功率更小、更具隐蔽性，故对抗样本对干净信号的影响不仅比噪声干扰小，而且更容易被分类器误分。另外，在相同的对抗性样本下也会对不同的网络结构模型产生误分。即对抗攻击具有迁移性。这表明以对抗攻击为代表的反智能技术会影响学习算法的鲁棒性和安全性。神经网络内部主要基于线性块构建，这会导致DL模型过度线化。在实施基于DL的通信任务时，DL模型整体函数被证明是高度线性的，这些线性函数很容易优化，但如果一个线性函数有许多输入，那么它的值可以非常迅速地改变。所以在神经网络高维度中，只要输入一点细微的增量，由于高维的积累，输出结果将会有巨大变化，更直观的表述如图2所示，扰乱后分类边界会发生改变，这些不重叠区域的样本会导致目标模型做出错误的预测，将A类样本误判为B类。在信号调制识别领域对抗扰动添加到原始信号上使得目标方的识别模型分类出错，例如将QPSK信号误判为最小频移键控（Minimum shift keying，MSK）信号。因此，对抗攻击的目的是寻找一种在对抗区域内生成恶意样本的有效方法，从而达到欺骗模型的目的。为此许多学者提出了不同的对抗样本生成方法。

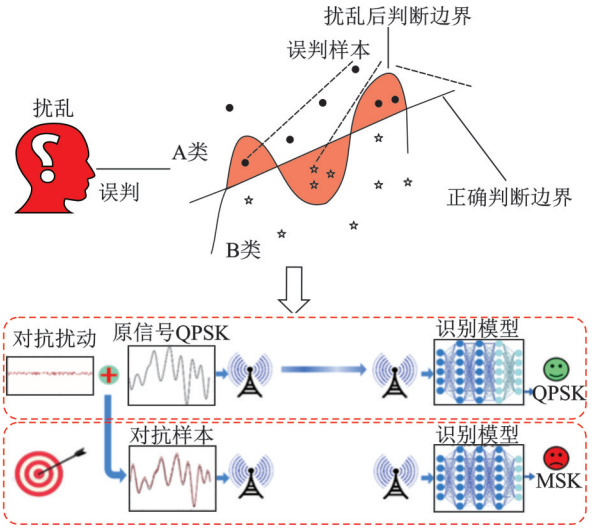

图2 对抗样本攻击基本概念和信号调制识别分类的具体应用

Fig.2 Specific application of the basic concept of adversarial attack and signal modulation recognition

1.3 对抗样本防御

对抗样本的存在不仅严重威胁DL模型的安全性，甚至使人一致怀疑基于DL模型通信任务的实用性和可靠性。攻击与防御是一个不断博弈的过程，攻击者不断提出新的攻击方法，防御者则不断推出更加鲁棒的算法。对抗样本的防御可以分为完全防御和检测防御。完全防御就是去提高模型的鲁棒性，目前有对抗训练、梯度掩盖、输入转换和随机化DL算法等4种主要思想解决此类问题。对抗训练是指把某种攻击算法生成的对抗样本也作为训练集去训练模型，此类方法耗费时间长，且攻击算法层出不穷，每出现一个新的对抗样本算法都要重新训练。梯度掩盖就是利用目前生成对抗样本的方式都是基于梯度的，故可以把梯度的作用削弱，常用的方法是防御蒸馏、深度压缩网络和输入梯度正则化。输入转换即是指用一些图像预处理和转换算法来削弱对抗样本的影响，在信号识别领域，常见的方式就是把频域和时域数据互相转换，以此来减弱对抗样本的影响。随机化是指在DL模型中引入随机层数或随机变量，使得该模型具有一定的随机性，从而增强了该模型的健壮性，提高了该模型的抗噪性。对抗样本的检测防御就是在分类模型进行分类识别之前先进行恶意样本和干净样本的检测。即通过细粒度的手段探究干净样本和对抗样本的数字特征信息差异，如提取两者的均值、标准差、能量、相关性和熵，通过这些值的差异筛选出恶意样本，恶意样本不再参与识别过程。

为了便于介绍对抗样本攻防技术研究成果，表1列出了对抗样本领域常见的相关符号，表2列出了对抗攻击中常用术语。

表1　对抗攻击中常用数学符号

Table 1　Commonly used mathematical symbols in adversarial attacks

数学符号	释义
X	全部数据
x	一个数据点
x^*	一个对抗样本
Y	数据集的标签
y'	特定目标标签
r_x	对抗扰动
w	模型参数

表2　对抗攻击中常用术语

Table 2　Commonly used terms in adversarial attacks

术语	定义
对手	特指制作对抗样本欺骗ML模型的攻击者
对抗扰动	对手产生的扰动
对抗样本	将精心设计的扰动加入原始样本后的样本
对抗攻击	对抗样本攻击的过程
对抗训练	将对抗样本添加到训练数据中训练以提高模型抗攻击的鲁棒性

2　信号调制识别对抗攻击分类

如图3所示,根据攻击方法不同可分为模型窃取攻击、对抗样本攻击、数据投毒攻击和木马攻击,根据攻击阶段可分为训练阶段、测试阶段攻击。根据攻击目的可分为有目标攻击和无目标攻击,根据攻击对先验知识的了解程度可分为白盒、灰盒和黑盒攻击。表3列出了这些类别下针对调制识别的对抗攻击工作。

图3　对抗攻击方法分类

Fig.3　Classification of adversarial attack methods

2.1　基于攻击方式的分类

根据攻击策略可分为不同类别,如模型窃取攻击、对抗样本攻击、数据投毒攻击和木马攻击,图4展示了不同攻击的工作原理以及形成攻击的一般过程。

2.1.1　模型窃取攻击

模型窃取攻击也被称为探索性攻击,模型窃取攻击探索了解模型内部工作原理,攻击方

表3　对抗攻击方法的应用

Table 3　Application of adversarial attack methods

分类	攻击类型	原理	相关文献
攻击类型	模型窃取攻击	探索模型的内部工作原理,构建替代模型	[22-25]
	对抗样本攻击	输入数据加噪声后攻击愚弄DL模型	[26-28]
	木马攻击	破坏模型训练过程	[29-30]
攻击目标	定向攻击	特定类别的攻击	[26,31-32]
	非定向攻击	非特定类别的攻击	[33]
先验知识	白盒攻击	知道所有先验信息	[22]
	黑盒攻击	不知道先验信息	[34]
	灰盒攻击	知道有限的先验信息	[35]

通过收集训练数据摸清DL模型和算法的内部运作机理,并通过仿真相似的输入和输出数据来训练一个网络模型,作为替代模型[36-37]。模型窃取攻击也可探索在无线信道决策中使用的深度强化学习(Deep reinforcement learning, DRL)机制并干扰底层通信。攻击者在一段时间内观察被攻击者与其环境的交互,并学习其活动模式。攻击者采用动态策略的方式对对抗扰动进行动态调整,避免对方破解防御我方的对抗攻击算法。模型窃取攻击是典型的训练过程中的攻击,为了获取更多输入样本、权重值、激活函数、体系结构以及训练方法等先验信息,使用主动学习等技术探索目标模型[38]或使用生成式对抗网络(Generative adversarial nets, GANs)增加有限的信息[39]。文献[40]指出用替代模型设计的对抗攻击具有一定的迁移性,对抗攻击会转移到目标模型。例如在无线通信场景中,可以通过探索空中频谱信息来了解通信系统的传输模式。另外,文献[23]研究了无线信道对替代模型的对抗攻击迁移性的影响。

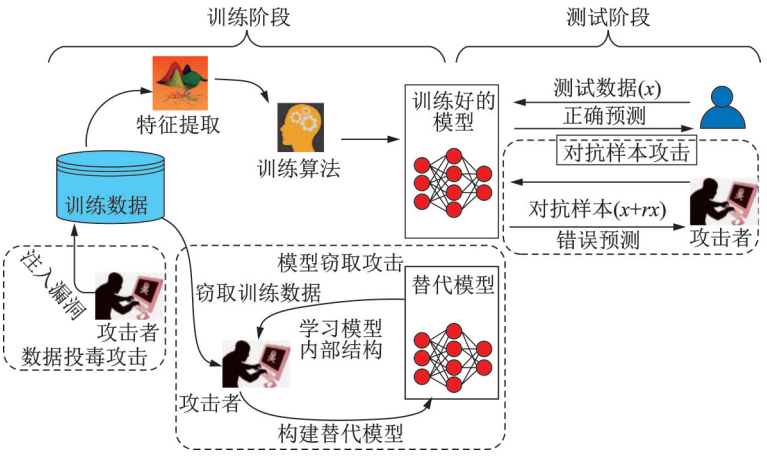

图4 对抗机器学习中的攻击方式

Fig.4 Adversarial attacks in machine learning

2.1.2 对抗样本攻击

对抗样本攻击也被称为规避攻击,目标是通过操纵输入数据来欺骗ML模型做出错误的分类[41-42]。规避攻击已应用到无线通信中,用来欺骗频谱感知的分类器[24]、调制识别的分类器[22]、基于自动编码器的端到端通信系统[43]、用于大规模MIMO的信道状态信息(Channel state information, CSI)反馈[33]以及定向通信中的初始接入[44],无线通信领域的对抗攻击挑战是目标分类器接收到的对抗扰动会受到信道效应的影响,因此需要通过考虑信道效应来精心设计这些扰动[45-46],这些攻击通常比传统的干扰攻击更隐蔽、更节能。传统的干扰攻击目的是对数据传输造成干扰[47],而对抗攻击只需要在短时间内传输低功率信号,在对手决策过程中去混淆ML模型。

2.1.3 数据投毒攻击

数据投毒攻击也叫诱导攻击,这种攻击假定攻击者可以进入DL模型的训练中,在训练集中插入伪造的恶意样本。训练集的恶意样本会扰乱分类器的训练过程[48-49],导致网络在进行测试时得到错误的预测结果,导致分类精度降低。诱导攻击对ML模型的误导效果很好。如频谱中毒攻击就是典型的诱导攻击方式,此过程使ML模型训练过程出错。最后导致ML模型失效。现有研究如频谱感知攻击[24]、合作式频谱感知[50]和物联网系统[51]都是利用了此类攻击原理。

2.1.4 木马攻击

木马攻击也称为后门攻击,是规避攻击和诱导攻击的结合,在这种攻击中,对手向训练数据中注入触发器,然后在测试过程中为某些输入样本激活触发器[52-53]。文献[54]研究了针对无线信号分类器的木马攻击问题,将受控相移构造为后门,并将其添加到发射和接收的射频数据样本中。

2.2 基于攻击方式的分类

一般而言,基于DL的目标系统分为模型训练阶段和测试验证阶段。攻击方可选择在不同的阶段

实施对抗攻击。在训练阶段实施攻击主要选择数据投毒和木马攻击来污染目标系统的训练数据,扰乱目标分类器的训练收敛过程,甚至迫使分类器以高置信度输出到既定的错误类别。在测试阶段实施攻击主要选择模型窃取、对抗样本攻击和木马攻击等方式来篡改输入测试样本数据,由于DL模型的高度线性,对数据设计细微的改动就能导致分类器以高置信度错误分类。

由于无线信道的开放性,攻击方要获得大量的无线电数据从而进行网络训练与验证工作,所以由于训练阶段数据侦获难、与真实信号数据有误差等原因,导致训练阶段实施攻击更难。

2.3　基于先验知识的分类

基于先验知识的分类,即是对目标模型的构造、目标模型的参数和数据集等先验信息的了解程度,可分为白盒攻击、灰盒攻击和黑盒攻击。

2.3.1　白盒攻击

白盒攻击需要了解一些先验信息,如输入样本、权重值、激活函数、体系结构和训练方法等,即可以直接访问攻击目标的分类模型f,是一种最理想攻击方式。此类攻击通过不断地访问模型来计算梯度,从而产生信号对抗样本。在白盒攻击下,对抗样本的生成较为容易和迅速。由于白盒攻击对目标网络模型和参数有充分的认知,攻击方能清晰地理解目标方分类模型的弱点,所以白盒攻击在众多场景和领域有广泛的应用。而对于信号调制识别中的白盒攻击,攻击方除了掌握上述分类模型的先验信息之外,还需知道影响信号传播的信道参数和特性,不同的信道特性会对对抗扰动造成较大的影响[45,55]。

2.3.2　灰盒攻击

灰盒攻击中,预先知道数据或算法的部分先验知识以及对模型f有限的访问权限。攻击者对被攻击模型的结构、参数信息、损失函数和梯度信息的了解程度介于白盒和黑盒之间。如无线通信领域中不知道通信过程中准确的信道信息,但可以通过信道估计、信道数据的分布预测等方式获得部分信道信息。还可以通过瑞利衰落、路径损失指数的值或信噪比(Signal to noise ratio, SNR)的值来对信道信息进行估计和预测。依据推理和预测的模型结构及参数,采用外推的方法来生成对抗样本。

2.3.3　黑盒攻击

黑盒攻击是一种更符合现实、更贴近实际场景的攻击方式[56],此类攻击假定不能直接访问目标方训练数据和训练模型f,只能通过感知目标方输入和输出之间的映射关系来获取目标方训练数据和训练模型的部分信息,因此常利用白盒模型中对抗样本的迁移性或黑盒模型的输出结果训练替代模型等来获得对抗样本。黑盒攻击通常构建替代模型作为目标攻击模型,在信号调制识别应用中,替代模型的有效性还取决于信道信息[23]。

2.4　基于攻击目的性的分类

针对攻击目标的不同,可以将其划分为无目标攻击和有目标攻击。无目标攻击不会使目标分类器错误地输出具体的类别,即攻击者产生导致目标分类器随机出错的对抗样本,其目的是降低原始类别的预测置信度。具体解释如下,非定向攻击产生的对抗扰动为r_x,添加到目标分类器的输入中,即$f_w(x+r_x)=y$,模型输出$y=f(x)$为任意的调制信号类别,即只需要以高置信度让模型分类随机出错,不考虑特定类别。如BPSK、QPSK、QAM16和QAM64的4类调制分类问题,攻击者欺骗调制分类器,将不同类别的调制信号分类为其他任意类别,非目标攻击具有随机性。

目标攻击使目标分类器错误输出特定的调制类别,即确保当前的攻击可以使得分类器朝着己方所期望的方向进行学习和分类。定向攻击产生的对抗扰动为r_x,添加到目标分类器的输入中,即$f_w(x+r_x)=y_0$, y_0为特定的调制信号类别。例如,考虑BPSK、QPSK、QAM16和QAM64等4类调制分类问题,欺骗误导调制分类器将这4类调制信号皆分类为特定的BPSK调制类别。

非目标攻击更易执行并且所需要的代价小，因此其实用范围更为广阔。目标攻击是持续性的，攻击者需要通过一定的代价来保持攻击的持续性，使得目标分类器的梯度值和损失值向攻击方预期的方向逐渐接近，最后将其全部修改为期望的特定类别，从而实现目标攻击的效果。

3 无线通信中对抗样本攻击技术面临的挑战

无线通信领域中DL算法的应用不同于计算机视觉和自然语言处理领域，其数据与处理以及编码译码技术需要一定的专业知识支撑。例如本文聚焦的基于DL的调制识别过程，根据调制信号对抗样本是否经过空中传播，可以将攻击方法分为数据攻击和物理攻击，数据攻击是指对抗样本以数据信号的形式直接进入被攻击者分类器，而物理攻击是指对抗样本信号由发射机发射、经空口传播、被攻击者接收并进入其分类器识别系统。显然，在物理攻击过程中，信号经过硬件设备的处理会出现失真，调制信号在空中传播不可避免会受到各种干扰，攻击难度比数据攻击难度更大。即使模拟仿真真实环境实施攻击，也存在与真实环境差异大、计算量大等问题。

对抗攻击应用到无线通信系统中时，应从通信实际出发，即考虑动态的信道信息、不完美的通信数据、特征表示的异构、信号层面的评估指标等特点。这些动态不确定信息给对抗样本攻击在无线通信中的应用和发展带来了一定的挑战，同时也由于通信智能化的发展，越来越多的通信任务依赖于DL算法。从人工智能的安全角度来研究对抗样本攻击，并通过防御对抗样本攻击的手段，设计更鲁棒、更智能的人工智能算法。

（1）无线信道的动态特性

首先无线通信信道对空口对抗攻击有主要影响。在物理攻击中，攻击者产生的对抗攻击信号需要通过无线信道传输才能到达通信接收方的模型分类器。在此过程中，信道衰落、信道噪声、信道延迟和频谱偏移等效应很容易削弱对抗样本攻击的效果[23,45]。为使对抗样本攻击经过空口物理传播依然有效，在构建对抗样本时需要将动态的信道信息纳入考量。由于无线信道的动态特征影响[57]，在时域进行特征训练的DL模型产生的对抗攻击在频域特征训练DL模型中将会失效。研究表明，由于信道特征的差异性，对抗攻击可以成功欺骗某一个目标接收机的分类器，但无法在其他接收机上产生同样的攻击效果[31,58-59]。

（2）不完美的训练数据

在图像识别或自然语言处理领域，通过API（Application programming interface）函数的调用可以很容易地获取目标网络的结构信息和训练数据，但在无线通信环境下，攻击者一般不具有直接访问目标模型的能力，一般需要持续感知被攻击者的通信过程，以此来获取并预测目标模型参数。同时需要从空口进行数据截获形成训练数据来训练替代网络，而截获的数据受信道特性的影响与原始数据存在误差不具备完美性，利用不完美的数据来训练替代网络必然会影响其有效性。

文献[27]中攻击方式考虑到了物理信道的影响，其基本思想是持续观察通信方的动作，在频谱感知和数据传输期间发射对抗扰动信号[28]，并间接地操纵DL模型并影响其分类过程。

（3）信号特征的异构性

各种通信系统的共存引入了无线电数据特征表示的异构性，使无线电数据呈现更加多样化和复杂的特征[60]。这对精心设计的有效对抗扰动带来了巨大挑战。无线信道一般具有全向传输的广播特性，攻击方发送的对抗扰动可能会同时影响多个接收机的信号分类器，而在不同通信系统中接收机的接收效应和信道特性具有异构性，因此在设计对抗扰动时，必须考虑不同接收机的不同信道影响，这将大大增加对抗样本的生成复杂性[46]。

（4）攻击效果评估受限

面向无线通信独有的数据格式和信号属性的对抗样本研究正处于蓬勃发展阶段,也有很多研究关注用于处理这些非结构化无线通信信号深度学习的模型,取得了一系列丰富的成果,但是,目前在无线通信领域中,针对对抗样本攻击的评测还比较复杂,缺乏统一的规范,无法对真实通信场景下的对抗攻击进行科学的评估。攻击效果是判断攻击算法是否有意义的重要指标,考虑到通信领域信号非结构化的独特属性,传统的图像、自然语言处理对抗样本攻击的分析指标无法完全适用于面向通信领域的效果评测。比如误分类率是评估攻击方法的一项重要指标。然而众多研究表明,误分类率不足以全面地评估信号的识别和分类领域里对抗样本攻击的有效性。另一方面,区别于图像的可视性,无线通信信号的特性更加抽象化,具有相位、幅度、功率和信噪比等特有的属性,因此必须针对这些独有的特性建立信号特性指标,用于描述攻击前后信号特性的变化,为后续的解调译码等工作提供量化评估的依据[61]。

4　信号调制识别对抗样本攻击技术

4.1　无线信道的动态特性

最近几年,一些国际人工智能与机器学习领域的顶级学术会议设立了有关对抗样本的专题研讨会,谷歌公司也在NeurIPS上组织了对抗攻防算法博弈比赛,此举引起了众多学者的高度关注,也在对抗样本攻击、防御和应用研究领域取得了丰硕的成果。许多学者对信号调制识别中的对抗攻击进行了深入的研究,研究成果总结为表4。

传统的通信系统的攻击如主用户仿真、干扰、窃听攻击和频谱感知数据伪造等已被广泛研究[62-64]。为了进一步保障通信系统的安全,ML/DL技术本身可以用来检测和缓解这种传统的攻击[65-67]。对抗样本攻击已经应用于CV[20],自动驾驶技术[68]和网络安全[69]。文献[70-71]也详细讨论了对抗样本攻击算法和防御策略从而设计更安全的DL模型。文献[70]重点介绍对抗攻击的分类问题,而文献[71]则专注于对抗样本的生成技术及其防御机制。

Sadeghi等[22]首次将对抗样本应用到信号调制识别领域。研究表明基于DL的无线通信信号识别也易受到对抗样本的攻击,文章研究了如何在白盒、黑盒场景下产生信号对抗样本。研究结果表明,此类精心设计的灵巧干扰攻击比传统的噪声干扰攻击效果更显著,所需功率代价也更小。因此,在智能无线通信领域中,基于DL模型通信任务的安全性和鲁棒性问题引起了关注。

当对抗样本应用到通信领域后,Flowers等[72]进一步设计更贴近真实物理世界的对抗样本算法,设计了考虑信道信息的调制识别过程中对抗样本攻击的工作,主要根据执行攻击的位置分别展开论述,包括直接访问数据攻击,空口传播攻击(Over the air, OTA)。作者设定一个包含合作通信双方,一个窃听者的通信系统,使用快速梯度符号法(Fast gradient sign method, FGSM)方法,以合作通信方的误码率为主要优化目标,以窃听者的误识别率为次要目标。研究表明,在OTA攻击时,分类模型仍然很容易被对抗样本攻击。但是由于信道效应、接收机效应等原因,OTA攻击的有效性将会大幅度下降。在认知无线电系统中,当发射机探测到空闲信道时,Shi等[73]提出一种对抗攻击的办法,以发起频谱数据中毒攻击,这种攻击会在运行时操控发送者的输入数据并导致其作出错误的传输决策。文献[46]提出的攻击模式可以应用在具有功率约束的通信场景中,攻击者仅需要很短的一段时间就能控制发射端的空口信号,因为传输时间很短,所以不易被探测到。

在信号调制识别中,除研究攻击策略外,将对抗样本技术作为一种灵活的防御策略也具有重要意义。Hameed等[74]设定了一个包含合作通信方、窃听者的通信系统。窃听者拦截己方通信信号,然后采用DL模型对截获的信号进行调制识别工作。此工作就是从通信防御方出发。将信号和对抗样本一起传输至接收机,在避免窃听者识别传输的信号调制样式的同时,也要考虑合作通信接收方能正确获取接收机传输信息,从而达到隐蔽通信的目的。文献[43]提出一种新颖的方法使优化合作通信方的误码

表4 对抗样本攻击在调制识别中的应用

Table 4 Application of adversarial attacks in modulation recognition

攻击类型	攻击方法	文献研究目标	所采用数据集	文献
白盒攻击 黑盒攻击 非/定向攻击	FGM 通用对抗扰动(Universal adversarial perturbations, UAP)	考虑信道效应和攻击者的范数限制来发起真实的对抗样本攻击	RML2016.10A	[39]
对抗样本攻击	FGSM L-BFGS	测试比较两种对抗样本的攻击效率	RML2016.04C	[75]
对抗样本 白盒攻击	FGSM 投影梯度下降(Projected gradient descent, PGD) BIM MIM	分析比较4种对抗攻击算法的可行性和有效性	RML2016.10A	[60]
木马攻击	在训练数据中嵌入木马,在测试时触发	通过在训练数据样本中插入木马(触发器)来隐蔽操纵训练数据,在测试阶段激活触发器	RML2016.10A	[54]
对抗样本攻击	FGSM	评估基于原始I/Q信号的调制分类漏洞	RML2016.10A	[26]
对抗样本攻击	PGD	通过干扰信道输入数据,使入侵者在确定发射机使用的调制方案时的识别准确率降到最低	在本地生成的数据	[55,58,76]
灰盒攻击	FGSM	在射频分类器中使用自动编码器预处理来缓解对抗样本的攻击	RML2018.01A	[35]
定向白盒 对抗样本攻击	FGM	利用合作干扰机发送对抗扰动来欺骗基于DL的窃听者,从而使5G通信不被窃听者发现	使用MATLAB 5G工具箱生成的数据	[77]
对抗样本攻击	通过敌对变异网络造成扰动	通过使用前向纠错(Forword error-correction, FEC)扩展具有通信感知的对抗样本攻击	使用Liquid DSP生成合成数据	[78]
白盒非定向攻击	对抗攻击波形干扰、合成	提出了一种广义的无线对抗机器学习问题(Generalized wireless adversarial machine learning problem, GWAP),并针对无线DL系统的对抗机器学习攻击进行评估	RML2018.01A	[79]
对抗样本攻击	敌对的突变网络(Adversarial mutation network, AMN)	在训练过程中引入光谱欺骗损失值,使光谱形状更符合原始信号	在本地生成的数据	[80]
对抗样本攻击	PGD 均匀随机噪声	将窃听者的识别准确性降到最低,同时确保预期接收者成功解码信号	在本地生成的数据	[81]
对抗样本攻击	C&W	采用不同的数据驱动下的采样策略,研究对抗攻击对基于DL的调制识别模型的影响	RML2016.10B	[82]
白盒攻击 黑盒攻击	UAP	提出一种输入不可知的对抗性攻击,降低了该攻击的检测概率	RML2016.10A	[83]

率最低,窃听者的误识别率最高。具体而言,通过预先确定的调制模式,协作通信方的接收端可靠地进行解调,忽视干扰带来的对调制识别的影响。窃听者却因为这些扰动的存在导致模型将截获的信号识别出错、不能可靠解码,从而不能破译截获的信号。实验结果也证明该方法以最小的通信性能损失保护己方信号防侦察、防破译等优势。

4.2　对抗样本攻击技术

针对无线通信中对抗样本攻击技术面临无线信道高动态性、通信数据不完美、信号特征异构和攻击效果受限等挑战。本节先对调制识别领域内对抗样本攻击、模型窃取攻击等方法进行展开描述,再结合通信信号幅值相位等独有特性对其进行全面科学评估。

信号调制识别对抗样本攻击流程如图5所示。首先目标方产生通信信号的原始波形,并对信号数据进行一系列数据封装处理再空口传播该信号。由于无线信道的开放性,攻击者窃听到当前信道上传输的通信信号,并对调制信号进行特征提取和分析,完成网络的梯度计算,选择不同的攻击方法来生成对抗样本,这些经过精心设计的对抗扰动会随着原始信号一同进入到接收端的接收机。接收端采集到发送端所传输的信号并利用DL模型对无线信号进行调制识别工作,由于受到对抗样本的攻击,导致目标DL模型分类器出错。

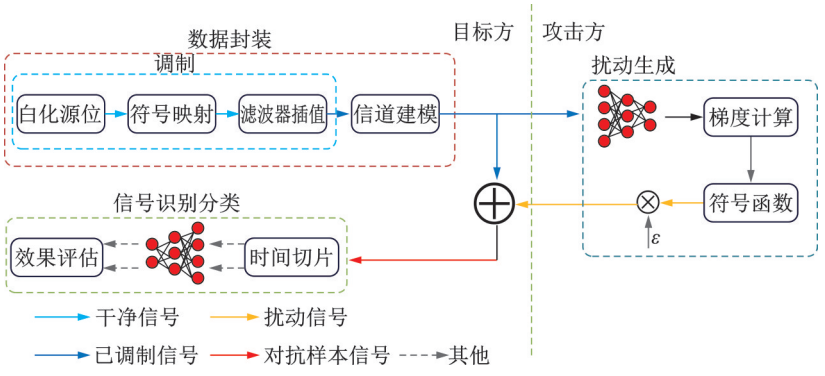

图5　调制识别中对抗攻击的流程图

Fig.5　Flow chart for adversarial attack in modulation classification

具体而言,对抗样本攻击就是在DL模型的原始输入信号样本x上添加微妙的、人眼难以察觉的扰动r_x,这样分类器的输入是$x'=x+r_x$,求解r_x的优化问题见式(3)。在无线通信领域,因为信道的存在有所改变,设置

$$x'=h_{t,r}x+h_{t,a}r_x+n \tag{4}$$

式中:x'为接收的信号,x为传输的干净信号,r_x为传输的扰动信号,$h_{t,r}$为从发射机到接收机的信道增益,$h_{t,a}$为从发射机到对手的信道增益,n为信道中的噪声。

由于DNN的映射是非线性的,求解式(3)比较困难,故可以求其近似解。比如通过最大化损失函数来实现,$L(w,x,y)$用于分类器f_w的训练过程,如交叉熵损失函数为$L(w,x,y)=-\log_2(1+\exp(-f_w(x,y)))$,因此,根据扰动上界的条件。可以通过求解式(5)得到对抗扰动r_x为

$$\max\quad L(w,x|+r_x,y)\quad \text{s.t. } \min\|r_x\|_p\leqslant\varepsilon \tag{5}$$

式中:ε为对抗性扰动的上界,研究者们提出了各种近似求解式(5)的方法,如FGSM[42]及其迭代变体。如最大优化法(Large BFGS, L-BFGS)[41]、C&W,基本迭代法(Basic iterative method, BIM)[56],投影梯度下降(Projected gradient descent, PGD)[84],动量迭代方法(Momentum iteration method, MIM)[85]等,研究者已经用上述方法来产生无线通信领域的对抗样本。

4.2.1　快速梯度符号法

在无线通信领域中,FGSM是一种快速且有效的对抗样本生成方法,FGSM在损失函数梯度方向

上寻求最小增量,其原理是通过线性化模型的代价函数$J(w,x,y)$来实现,即FGSM生成一个与原始样本非常相似的对抗样本,再将对抗样本输入模型,最后导致模型误分类。具体为

$$x^*=x+\varepsilon\cdot\mathrm{sign}(\nabla_x J(x,y,w)) \tag{6}$$

式中:x^*为产生的对抗样本,x为原始信号样本,$J(x,y,w)$为参数为w的损失函数,y为x对应的标签,ε为用来限制扰动的超参数值。快速梯度法(Fast gradient method, FGM)是FGSM满足L_2范数边界的一种推广方法,$\|x^*-x\|_2<\varepsilon$。

$$x^*=x+\varepsilon\cdot\frac{\nabla_x J(x,y,w)}{\|\nabla_x J(x,y,w)\|_2} \tag{7}$$

FGSM也用于如下的迭代方法中。

(1)BIM:BIM为FGSM的变体,将每次迭代后的结果进行裁剪,以确定结果都在原始样本的邻域内。BIM具体为

$$x_i^*=\mathrm{clip}_{x,\in}(x_{i-1}^*+\varepsilon\cdot\mathrm{sign}(\nabla_x J(x_{i-1}^*,y,w))) \tag{8}$$

(2)PGD:PGD为FGSM在损失函数上的多步变体,有助于发起更有效的对抗性攻击。PGD具体为

$$x_t^*=\prod_{B\in(x)}(x_{t-1}^*+\varepsilon\cdot\mathrm{sign}(\nabla_x J(x_{t-1}^*,y,w))) \tag{9}$$

(3)MIM:MIM的发展是为了解决与过拟合和局部极小值相关的问题,它有助于加速梯度下降迭代。在FGSM由于强干扰和失真而性能受到抑制的情况下,MIM对迭代攻击施加动量,以稳定和保持更新的正确方向,并对对抗样本进行泛化。MIM的梯度计算如下

$$g_{n+1}=\mu g_n+\frac{\nabla_x J_\theta(x_n',y)}{\|\nabla_x J_\theta(x_n',y)\|_1} \tag{10}$$

首先输入x_n'到目标函数f,求出梯度$\nabla_x J_\theta(x_n',y)$,并在梯度方向上累积速度矢量,从而对$g_{n+1}$参数进行更新,再利用符号梯度进行迭代更新生成的样本x_{n+1}'为

$$x_{n+1}'=x_n'+\varepsilon\cdot\mathrm{sign}(g_{n+1}) \tag{11}$$

4.2.2 L-BFGS方法

Box-constrained L-BFGS是基于优化方法来生成对抗样本。如一个样本x,L-BFGS方法通过找到一个x^*,x^*的L_2范数距离和x非常相近,但是其分类器的标签不一致,可以建模为下面的约束最小化问题

$$\min\|x-x^*\|_2^2 \quad \text{s.t. } f_w(x+r_x)=y',(x+r_x)\in[0,1]^n \tag{12}$$

$$\min\lambda\cdot\|x-x^*\|_2^2+J(x,y',w) \quad \text{s.t. } x^*\in[0,1]^n \tag{13}$$

式中:y'为目标类别标签;$J(x,y',w)$为一种损失函数,它会将x映射为一个正实数。在式(13)中,损失函数J设定为典型分类损失函数交叉熵,λ为用来约束生成最小距离对抗样本的超参数。

4.3 模型窃取攻击技术

针对信号特性的异构性,攻击者采用动态策略的方式对对抗扰动进行动态调整,避免对方破解防御我方的对抗攻击算法。在模型窃取攻击中,攻击者试图探索理解DL模型的内部结构,并利用这些探索的结构知识来攻击目标DL模型,最终破坏DL模型系统的有效性。该过程往往是通过训练一个可以模拟目标DL模型的替代模型,先窃取训练数据,再构建具有与原始模型相同或类似输出的模型。探索了目标模型工作的信息特征后,攻击者就可以做出精妙的攻击决策,从而选择如何攻击和改变基于DL

系统的整体性能。例如,文献[86]认知发射机使用一个预先训练的分类器,根据最新的感知结果预测当前信道状态,并决定是否发送信号。而干扰机则收集信道状态和确认字符(Acknowledge character, ACK),并建立一个深度学习分类器,可靠地预测下一个信号成功传输时间,并有效地干扰其传输,这种干扰方法大大降低了发射机的功率。文献[29]中,对手观察频谱并用DNN来推断物联网发射机使用的信道接入算法,并在干扰它之前预测传输的结果。模型窃取攻击也可探索在无线信道决策中使用的深度强化学习(Deep reinforcement learning, DRL)机制并干扰底层通信。文献[87-88]设计了一种基于DRL的对抗攻击,该攻击采用动态策略避免对方破解防御我方的对抗攻击算法,目标是最小化基于DRL的用户访问动态信道的精度。DRL攻击者不知道被攻击者行动策略上的信道切换模式等先验信息,攻击者在一段时间内观察被攻击者与其环境的交互,并学习其活动模式,然后两个基于DRL的系统再相互交互,并重新训练它们的模型并适应对手的策略。文献[89-90]介绍了强化学习(Reinforcement learning, RL)算法,通过观察频谱和使用基于RL的替代模型来中断5G网络切片,该模型的目标是选择性阻塞资源块,以获得大量失败网络切片的请求,从而阻塞资源块,导致资源块数量减少。RL算法的奖励会影响模型的性能,奖励的引入可以动态改变RL算法。

4.4　对抗攻击效果评估技术

针对在无线通信中对抗攻击效果评估受限等挑战,Sadeghi等[22]在对信号攻击的研究中,提出了干扰信号功率比(Perturbation signal rate, PSR),PSR是扰动信号和原始信号功率的比率,结合扰动噪声功率比(Perturbation noise rate, PNR),可以更直观反映出信号采样攻击前后的信噪比变化。不过单以此来作为衡量指标还远远不够。故文献[61]从信号特性、隐蔽性和误分类等3个方面提出面向信号识别领域对抗样本攻击方法的有效性评估指标与方法,指标体系的构成如图6所示。考虑到误分类、隐蔽性、计算代价的指标体系及计算代价都有成熟的研究方案,此处不在赘述,本文将重点针对信号识别领域的信号特征指标展开论述。

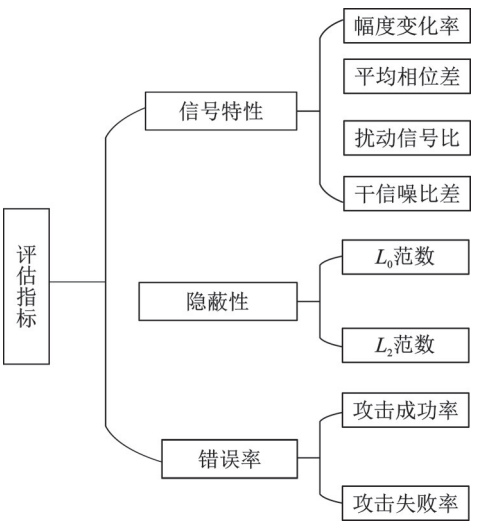

图6　面向信号识别的对抗攻击指标体系

Fig.6　Adversarial indicator system for signal recognition

(1) 信号幅值变化率(Amplitude change rate, ACR)

为了描述攻击前后的信号幅值变化情况,引入信号幅值变化率这一指标。

$$\mathrm{ACR}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_{oi}-A_{ai}}{A_{oi}}\right| \tag{14}$$

式中:A_{oi}为第i个原始信号的有效幅度,A_{ai}为攻击后信号第i个采样点的有效幅度,n为一个信号样本中采样点总数。

与图像中各象素点相互独立的情况不同,信号数据集中的I/Q两路存在一一对应的关系,分别为复信号实部与虚部的采样值,在计算过程中要考虑I/Q两路的相关性。

(2) 平均相位差(Average phase difference, APD)为

$$\mathrm{APD}=\frac{1}{n}\sum_{i=1}^{n}\left|\arctan\frac{Q_{oi}}{I_{oi}}-\arctan\frac{Q_{ai}}{I_{ai}}\right| \tag{15}$$

式中APD为一个信号样本中每个采样点的相位差平均值。信号波形有无变化可以很好地判断信号波

形是否有扰动攻击存在，如果发生相位延迟，那该信号可能就会失去有效真实信息。I_{oi}为干净信号第i个采样点的实部系数，Q_{oi}为干净信号第i个采样点的虚部系数。I_{ai}为攻击后信号第i个采样点的实部系数，Q_{oi}为原始信号第i个采样点的虚部系数。

（3）扰动信号比（Perturbation signal rate，PSR）为

$$\begin{cases} P=\dfrac{\sum_{i=1}^{n}A_i^2}{n} \\ \mathrm{PSR}=\dfrac{P_P}{P_s} \end{cases} \tag{16}$$

式中：PSR为扰动信号功率和原始信号功率之比，P为信号功率，A_i为信号第i个采样点的有效幅度。

（4）信噪比差（Signal noise rate difference，SNRD）为

$$\begin{cases} \mathrm{SNR}=\dfrac{P_s}{P_n} \\ \mathrm{SNRD}=\mathrm{SNR}_{\mathrm{adv}}-\mathrm{SNR}_{\mathrm{ori}} \end{cases} \tag{17}$$

SNRD计算的是原信号样本与对抗样本信号的信噪比差值，$\mathrm{SNR}_{\mathrm{ori}}$、$\mathrm{SNR}_{\mathrm{adv}}$分别为攻击前后信号样本的信噪比。

5 信号调制识别中对抗样本攻击防御技术

无线通信DL算法的设计必须考虑新的安全问题，如对抗样本信号的识别和防御[55]。由于射频数据的动态特性，设计射频数据中的对抗样本有一定的难度和挑战，DL方法依赖于训练数据生成过程的分布，这样训练的模型可以很好地泛化测试数据。然而，这一前提不一定适用于存在对抗性的情况，因为射频数据的分布会因对抗样本的存在而发生极大的变化，而对抗性的影响不能直接预测[41]。在这样的对抗样本攻击下，5G和6G等无线通信系统中基于DL设计的系统组件将容易受到干扰、使性能损失、并最终导致通信过程受阻等重大影响。因此，分析对抗样本攻击所带来的威胁，并应用对抗攻击防御技术来缓解并最终消除对抗样本对人工智能系统的影响至关重要。

本节集中介绍在无线通信系统中对抗样本攻击的检测和防御算法，检测是指在进行网络识别和决策之前就先把对抗样本筛选出来，不进行识别，而防御则是靠自身算法的鲁棒性来抵御对抗样本的愚弄，最终依然达到预期的分类效果。常用的对抗样本防御策略为对抗样本训练技术、对抗样本检测技术、防御蒸馏技术和随机化DL算法，其优缺点总结如表5所示。

表5 在无线通信中对抗样本的防御策略及其优缺点

Table 5 Defense strategy adversarial examples in wireless communication and their advantages and disadvantages

对抗攻击防御策略	优点	缺点	可防御对抗攻击方法
对抗样本训练	可用范围广	需要大量的对抗样本数据	对抗样本攻击、木马攻击和模型窃取攻击
统计方法	方法多样	需要做数据统计工作、操作复杂	木马攻击
蒸馏防御技术	通用、解决思路简易	该防御方法易被攻破，不稳固	对抗样本攻击、木马攻击和模型窃取攻击
随机化DL算法	可动态调整策略	需知道目标模型内部结构	模型窃取攻击、木马攻击和对抗样本攻击

(1) 对抗样本训练技术

利用对抗样本训练DNN是一种常用的防御对抗攻击并确保鲁棒性的方法[85]。对抗样本训练根据一种或多种攻击方法生成对抗样本,并使用标记的对抗性样本对DNN进行再训练。该防御机制已应用于保护无线信号分类器免受对抗样本攻击[45],并在训练过程中采用随机平滑提高了分类器在测试阶段后期对抗攻击的鲁棒性[91]。

然而,对抗训练中仍存在一些挑战,比如在一些通信场景中,对手使用不同于训练中使用的对抗样本攻击方式。除此之外,敌方还可以使用对抗样本训练的模型来制造新的扰动。同时,使用对抗样本进行训练通常会降低DL模型在无扰动信号上的性能。为了克服这些弊端,可以通过多个对抗样本进行大量训练[85]。此外,在训练阶段和测试阶段内可以建立认证防御机制[92],以保证在测试阶段存在对抗攻击时分类结果的统计意义。在存在对抗样本的情况下,文献[45]已经使用了经过认证的防御机制来为无线信号分类器提供性能保证。在训练过程中防御对抗攻击的另一种方法是通过自动编码器对基于DL的无线信号分类器进行预训练,这样训练后的模型对对抗样本的攻击同样具有鲁棒性[35]。

(2) 对抗样本检测技术

研究表明,对抗样本信号与干净样本信号有不同的统计特性,提取两者具有可辨别性的统计特征,依据这些可辨别的统计特征,判断输入样本的合法性。常见的检测方法有Softmax输出、峰值平均功率比、检测触发器等统计方法。

使用DNN分类器的Softmax输出是对抗样本检测的(Kolmogorov-Smirnov, KS)统计测试。DNN分类器的Softmax输出是神经网络最后一层,通过统计神经网络最后一层的数据从微观上来确定对抗样本是否带来了数据分布上的变化。通常输入数据的统计大小可以很好地决定统计测试的质量。该方法的有效性取决于波形的类型和传输信道的特性[93]。

射频样本的峰值平均功率比(Peak to average power ratio, PAPR)是一种检测对抗样本的方法,它利用射频数据的特性、接收信号的数字化射频样本的峰值平均功率比进行统计检验[93]。PAPR值在无线领域用来描述某个特定类信号的数据特征。因此,当输入信号的推断输出被归类到某一类时,若其PAPR统计量与高置信度相矛盾,则需要进一步分析评估,才能得出可靠的结论。要确定无线通信环境中是否存在对抗样本可使用KS两样本检验,通过收集样本输入的大小,计算PAPR值从而判断输入样本是否为干净样本。

检测敌对触发器的统计方法应插入敌对触发器(如木马、后门)利用聚类和绝对中位差(Median absolute deviation, MAD)等统计方法对训练数据进行异常值检测。MAD算法计算数据绝对值的中位数,即中值($|x_i-\hat{X}|$),$\hat{X}=\text{median}(X)$,从而发现离群值。如文献[94]所述,可以有效防御无线信号分类器上的木马攻击。

(3) 蒸馏防御技术

通过对复杂DL模型进行训练,保存训练得到的参数,并将参数用来训练较简单的DL模型,从而获得一个简化的网络模型的过程称为算法的蒸馏过程。即把庞大的复杂网络体系结构中学习到的知识特征转移到简易网络体系中去。此过程可以解决因大型复杂网络结构由于计算量偏大而无法得到有效应用的问题,据Hinton等[95]介绍,“蒸馏”可将复杂网络的知识迁移到简单网络上。Papernot等[96]使用网络蒸馏来防御对抗样本。防御性蒸馏也是梯度掩盖技术的一个特例。在蒸馏算法中,将原训练集合中的复杂网络知识转化为概率矢量,再用蒸馏方法将其抽取到更小的网络中,使其具有与复杂网络相似的精确度,同时还可以增强其他训练数据集合的泛化能力。当网络模型的输入端梯度幅度较大时,攻击者可以通过大的梯度来设计一个对抗样本,从而误导DL分类器的输出。防御性蒸馏算法可以让DL模型结构在蒸馏之后更加平滑,从而可以更好地防御来自训练样本以外的恶意样本。

（4）随机化DL算法

通过对分类器输出增加小的变化来随机化DL算法，即对DNN分类器输出增加轻微的变化，有意地执行一些轻微的错误操作，以欺骗对手并防御探索性攻击。例如，发射机可以故意利用忙碌信道发送信号或跳过空闲信道中的传输机会，从而使观察频谱状态的对手不能训练出可靠的代理替代模型。该防御机制在文献[86]中被应用于保护无线信号分类器免受生成探索性攻击，同时在文献[97]中用随机化DL算法防御规避攻击和诱导攻击。针对强化学习的对抗攻击的防御采用了类似的多样化方法来保护无线通信信号的正常传输[89]。防御者可以在不同的信道访问决策之间交替，通过比例、积分、微分（Proportional-integral-derivative, PID）控制器和模拟学习来增加对手的不确定性；另一种防御方法是在强化学习中采用正交策略，以防止对手在其模仿策略之间快速切换。除了这些防御策略外，通过监测奖励的变化并从对抗性干扰攻击中区分频谱环境变化来检测对抗样本也是可行的。

6 未来工作与展望

随着ML/DL成为当前和新兴通信系统5G和6G的核心技术手段，而ML/DL本身容易受到敌对对抗样本攻击的影响。为了在动态无线通信环境中应用DL实现智能运营和高效资源管理，需要考虑无线通信的独特特性，开发安全、有弹性和能防御对抗攻击的DL模型。考虑无线特性的对抗样本攻击模型有望成为DL驱动的现有新兴通信系统无线安全研究和开发的基础。

（1）研究和开发面向实际通信场景的标准化无线通信数据集

在无线通信领域，标准化的真实世界数据集能够充分代表真实通信场景，与CV和NLP等其他领域相比，无线领域只有少数几个公开可用的ML/DL数据集[89]。通常，这些数据集不包括对抗样本攻击数据。业界需研究更多公开可用的数据集，分别代表不同的场景，数据信息中不仅包括信道信息、干扰类型和波形时域图，而且还应包括对抗攻击，此举将有助于得出针对无线对抗攻击的DL模型的稳健性方案。从大量研究来看，无线通信中对抗攻击的工作集中在调制分类领域。这得益于调制分类的官方数据集，文献[98]是无线通信领域ML/DL研究的标准化数据集。但通过嵌入式无线电平台来实现对抗攻击和防御仍然是一大难题，并且评估真实的信道环境和无线电硬件对恶意样本影响对对抗样本攻击在现实环境进行部署至关重要。

（2）深入开展对抗样本产生机理研究进一步增强对抗样本特征的鲁棒性

在生成无线通信中的ML/DL模型时，对鲁棒特性的需求也是未来系统的一个重要考虑因素。因为现有的大多数ML/DL模型都没有使用鲁棒特征来进行构造。虽然鲁棒特征识别的研究正在进行，特别是在CV和NLP领域，但必须开发新的技术来识别最重要的特征，而且这些特征对无线通信系统中的对抗攻击具有鲁棒性。当前的研究多是从攻击者的视角来分析和评估对抗样本的有效性，而对于对抗样本生成机制的研究还不够全面。后续的研究可以从低机率区解释和线性解释两个角度来探讨对抗样本的数学描述模式，进而增强通信领域中对抗样本特征的鲁棒性。

（3）对抗样本防御技术从单一的数据层或模型层防御向可泛化性、可证明性的转变

设计和开发可认证的对抗样本防御机制是实现ML/DL在无线通信领域的全部潜力的一个重要组成部分。为了解决深度学习中的对抗样本攻击问题，基于模型和数据分析两个层面的研究可以有效防御对抗样本攻击。目前，数据层防御已有对抗训练、样本检测和输入预处理等技术，在模型层防御已有防御蒸馏、网络修剪和随机化操作等技术。两个方向互为补充，协同发展。现有的通信信号对抗样本防御技术大多针对特定算法和应用场景，防御方一直处于被动应对状态，缺少可泛化的对抗防御方案，无法为模型最低精度提供可靠保证，面对更强大的对手时变得无效。其中基于对偶方法和混合整数线性规划等数学理论是解决对抗样本防御机制可证明性的一条技术路径。

7　结束语

随着智能化技术在无线通信领域的蓬勃发展,以DNN为代表的人工智能算法存在缺乏鲁棒性、易受对抗样本攻击等缺点,使得用户对人工智能系统的决策结果无法完全信任,这是当前DL模型实际部署在无线通信应用领域的最大障碍。本文综述了国内外有关对抗样本攻击与防御的研究,并将现有研究进行归类整理。阐述了其生成机理,总结了无线通信场景下信号调制识别对抗样本攻击技术研究进展,同时,分析对比了调制识别对抗样本攻击典型的防御技术。在总结现有研究成果的基础上,指出了智能无线通信中对抗样本攻击亟待解决的技术难题和有价值的研究方向。

参考文献:

[1] ERPEK T, O'SHEA T J, SAGDUYU Y E, et al. Deep learning for wireless communications[C]//Proceedings of Development and Analysis of Deep Learning Architectures. Cham: Springer, 2020: 223-266.

[2] SIMEONE O.A very brief introduction to machine learning with applications to communication systems[J]. IEEE Transactions on Cognitive Communications and Networking, 2018,4(4): 648-664.

[3] WONG L J, CLARK IV W H, FLOWERS B, et al. The RFML ecosystem: A look at the unique challenges of applying deep learning to radio frequency applications[EB/OL]. (2020-10-01)[2023-10-15]. https://arxiv.org/abs/2010.00432.

[4] O'SHEA T J, CORGAN J, CLANCY T C. Convolutional radio modulation recognition networks[C]//Proceedings of Engineering Applications of Neural Networks. Cham: Springer International Publishing, 2016: 213-226.

[5] LIANG F, SHEN C, WU F.An iterative BP-CNN architecture for channel decoding[J]. IEEE Journal of Selected Topics in Signal Processing, 2018,12(1): 144-159.

[6] JAFARI H, OMOTERE O, ADESINA D, et al. IoT devices fingerprinting using deep learning[C]//Proceedings of IEEE Military Communications Conference (MILCOM). [S.l.]: IEEE, 2018.

[7] JIAN T, RENDON B C, OJUBA E, et al. Deep learning for RF fingerprinting: A massive experimental study[J]. IEEE Internet of Things Magazine, 2020,3(1): 50-57.

[8] MERCHANT K, REVAY S, STANTCHEV G, et al. Deep learning for RF device fingerprinting in cognitive communication networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2018,12(1): 160-167.

[9] DAVASLIOGLU K, SOLTANI S, ERPEK T, et al. DeepWiFi: Cognitive WiFi with deep learning[J]. IEEE Transactions on Mobile Computing, 2019, 20(2): 429-444.

[10] OMOTERE O, FULLER J, QIAN L, et al. Spectrum occupancy prediction in coexisting wireless systems using deep learning [C]//Proceedings of IEEE Vehicular Technology Conference (VTC-Fall). [S.l.]: IEEE, 2018.

[11] SUN H, CHEN X, SHI Q, et al. Learning to optimize: Training deep neural networks for wireless resource management[C]// Proceedings of IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) .[S.l.]: IEEE, 2017.

[12] COUSIK T S, SHAH V K, REED J H, et al. Fast initial access with deep learning for beam prediction in 5G mmWave networks[C]//Proceedings of MILCOM 2021—2021 IEEE Military Communications Conference (MILCOM). [S.l.]: IEEE, 2021: 664-669.

[13] ABUZAINAB N, ALRABEIAH M, ALKHATEEB A, et al. Deep learning for thz drones with flying intelligent surfaces: Beam and handoff prediction[C]//Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops). [S.l.]: IEEE, 2021.

[14] ZHOU Y, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: A convergence of communication and computing[J]. IEEE Communications Magazine, 2019, 57(5): 20-27.

[15] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of International Conference on Learning Representations (ICLR). [S.l.]: IEEE, 2014.

[16] VOROBEYCHIK Y, KANTARCIOGLU M. Adversarial machine learning[J]. Synthesis Lectures on Artificial Intelligence

and Machine Learning, 2017, 12(3): 1-169.

[17] SHI Y, SAGDUYU Y E, DAVASLIOGLU K, et al. Vulnerability detection and analysis in adversarial deep learning[M]. [S.l.]: Springer, 2018: 211-234.

[18] ZHOU Y, LIU L, WANG L, et al. Service-aware 6G: An intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.

[19] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2014-12-20)[2023-10-15]. https://arxiv.org/abs/1412.6572.

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

[21] XU H, MA Y, LIU H, et al. Adversarial attacks and defenses in images, graphs and text: A review[J]. International Journal of Automation and Computing, 2020,17: 151-178.

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

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

[24] SAGDUYU Y E, SHI Y, ERPEK T. Adversarial deep learning for over-the-air spectrum poisoning attacks[J]. IEEE Transactions on Mobile Computing, 2021,20(2): 306-319.

[25] SHI Y, DAVASLIOGLU K, SAGDUYU Y E. Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers[C]//Proceedings of ACM Workshop on Wireless Security and Machine Learning (WiseML). [S.l.]: ACM, 2020.

[26] 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.

[27] KOKALJ-FILIPOVIC S, MILLER R, MORMAN J. Targeted adversarial examples against RF deep classifiers[C]//Proceedings of ACM Workshop on Wireless Security and Machine Learning (WiseML). [S.l.]: ACM, 2019.

[28] KARUNARATNE S, KRIJESTORAC E, CABRIC D. Penetrating RF fingerprinting-based authentication with a generative adversarial attack[EB/OL]. (2020-11-03)[2023-10-15]. https://arxiv.org/abs/2011.01538v1.

[29] SAGDUYU Y E, SHI Y, ERPEK T. IoT network security from the perspective of adversarial deep learning[C]//Proceedings of IEEE International Conference on Sensing, Communication, and Networking (SECON). [S.l.]: IEEE, 2019.

[30] SHI Y, ERPEK T, SAGDUYU Y E, et al. Spectrum data poisoning with adversarial deep learning[C]//Proceedings of IEEE Military Communications Conference (MILCOM). [S.l.]: IEEE, 2018.

[31] 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 2020 54th Asilomar Conference on Signals, Systems, and Computers. New York: IEEE, 2020: 763-767.

[32] BAIR S, DELVECCHIO M, FLOWERS B, et al. On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition[C]//Proceedings of the ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2019: 25-30.

[33] LIU Q, GUO J, WEN C, et al. Adversarial attack on DL based massive MIMO CSI feedback[J].Journal of Communications and Networks, 2020,22(3): 230-235.

[34] WANG F, ZHONG C, GURSOY M C, et al. Defense strategies against adversarial jamming attacks via deep reinforcement learning[C]//Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS). New York: IEEE, 2020: 1-6.

[35] 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 (ICMCIS). New York: IEEE, 2019: 1-6.

[36] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction {APIs}[C]//Proceedings of the 25th USENIX Security Symposium (USENIX Security 16). New York: ACM, 2016: 601-618.

[37] SHI Y, SAGDUYU Y E, GRUSHIN A.How to steal a machine learning classifier with deep learning[C]//Proceedings of IEEE International Symposium on Technologies for Homeland Security (HST). [S.l.]: IEEE,2017.

[38] SHI Y, SAGDUYU Y E, DAVASLIOGLU K, et al. Active deep learning attacks under strict rate limitations for online API calls[C]//Proceedings of IEEE International Symposium on Technologies for Homeland Security (HST). [S.l.]: IEEE,2018.

[39] SHI Y, SAGDUYU Y E, DAVASLIOGLU K, et al. Generative adversarial networks for black-box API attacks with limited training data[C]//Proceedings of 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). [S.l.]: IEEE, 2018: 453-458.

[40] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of ACM Asia Conference on Computer and Communications Security (AsiaCCS). [S.l.]: ACM, 2017.

[41] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19)[2023-10-15]. https://arxiv.org/abs/1312.6199.

[42] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. Computer Science, 2015, 1412: 6572.

[43] SADEGHI M, LARSSON E G.Physical adversarial attacks against end-to-end autoencoder communication systems[J]. IEEE Communications Letters, 2019,23(5): 847-850.

[44] 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 (SSP). New York: IEEE, 2021: 590-594.

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

[46] RESTUCCIA F, D'ORO S, AL-SHAWABKA A, et al. Hacking the waveform: Generalized wireless adversarial deep learning[EB/OL]. (2020-05-05)[2023-10-15]. https://arxiv.org/abs/2005.02270.

[47] XU W, TRAPPE W, ZHANG Y, et al. The feasibility of launching and detecting jamming attacks in wireless networks[C]//Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing. [S.l.]: ACM, 2005: 46-57.

[48] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[EB/OL]. (2012-06-27)[2023-10-15]. https://arxiv.org/abs/1206.6389.

[49] SHI Y, SAGDUYU Y E. Evasion and causative attacks with adversarial deep learning[C]//Proceedings of MILCOM 2017 IEEE Military Communications Conference (MILCOM). New York: IEEE, 2017: 243-248.

[50] LUO Z, ZHAO S, LU Z, et al. When attackers meet AI: Learning-empowered attacks in cooperative spectrum sensing[J]. IEEE Transaction Mobeile Computing, 2020, 21(5): 1892-1908.

[51] LUO S, ZHAO Z, LU Y E, et al. Adversarial machine learning based partial-model attack in IoT[C]//Proceedings of ACM Workshop on Wireless Security and Machine Learning (WiseML). [S.l.]: ACM, 2020.

[52] WANG B, YAO Y, SHAN S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2019.

[53] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[EB/OL]. (2018-11-09)[2023-10-15]. https://arxiv.org/abs/1811.03728.

[54] DAVASLIOGLU K, SAGDUYU Y E. Trojan attacks on wireless signal classification with adversarial machine learning[C]//Proceedings of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). [S.l.]: IEEE, 2019.

[55] 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,2020,16: 1074-1087.

[56] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2016-07-08)[2023-10-15]. https://arxiv.org/abs/1607.02533v3.

[57] SAHAY R, BRINTON C G, LOVE D J. Frequency-based automated modulation classification in the presence of adversaries

[C]//Proceedings of ICC 2021—IEEE International Conference on Communications. New York: IEEE, 2021: 1-6.

[58] HAMEED M Z, GYORGY A, GUNDUZ D. Communication without interception: Defense against modulation detection[C]// Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP). [S.l.]: IEEE, 2019.

[59] BERIAN A, STAAB K, DITZLER G, et al. Adversarial filters for secure modulation classification[C]//Proceedings of the 2021 55th Asilomar Conference on Signals, Systems, and Computers. New York: IEEE, 2021: 361-367.

[60] LIN Y, ZHAO H, TU Y, et al. Threats of adversarial attacks in DNN-based modulation recognition[C]//Proceedings of IEEE Conference on Computer Communications (INFOCOM). [S.l.]: IEEE, 2020.

[61] 宣琦, 周晴, 崔慧, 等. 信号人工智能对抗攻击综合分析平台[J]. 信息安全学报, 2021, 6(4): 141-149.
XUAN Qi, ZHOU Qing, CUI Hui, et al. A comprehensive evaluation platform of adversarial attacks on artificial intelligence for signal[J]. Journal of Cyber Security, 2021, 6(4): 141-149.

[62] ZOU Y, ZHU J, YANG L, et al. Securing physical-layer communications for cognitive radio networks[J]. IEEE Communications Magazine, 2015, 53(9): 48-54.

[63] ZOU Y, ZHU J, WANG X, et al. A survey on wireless security: Technical challenges, recent advances, and future trends[J]. Proceedings of the IEEE, 2016, 104(9): 1727-1765.

[64] SAGDUYU Y E. Securing cognitive radio networks with dynamic trust against spectrum sensing data falsification[C]// Proceedings of IEEE Military Communications Conference. [S.l.]: IEEE, 2014.

[65] RAJENDRAN S, MEERT W, LENDERS V, et al. Saife: Unsupervised wireless spectrum anomaly detection with interpretable features[C]//Proceedings of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). [S.l.]: IEEE, 2018.

[66] VAN HUYNH N, NGUYEN D N, HOANG D T, et al. Jam me if you can: Defeating jammer with deep dueling neural network architecture and ambient backscattering augmented communications[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(11): 2603-2620.

[67] ABUZAINAB N, ERPEK T, DAVASLIOGLU K, et al. QoS and jamming-aware wireless networking using deep reinforcement learning[C]// Proceedings of MILCOM 2019 IEEE Military Communications Conference (MILCOM). New York: IEEE, 2019: 610-615.

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

[69] ZHOU Y, KANTARCIOGLU M, XI B. A survey of game theoretic approach for adversarial machine learning[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(3): e1259.

[70] WANG X, LI J, KUANG X, et al. The security of machine learning in an adversarial setting: A survey[J]. Journal of Parallel and Distributed Computing, 2019, 130: 12-23.

[71] OZDAG M. Adversarial attacks and defenses against deep neural networks: A survey[J]. Procedia Computer Science, 2018, 140: 152-161.

[72] FLOWERS B, BUEHRER R M, HEAEY W C. Evaluating adversarial evasion attacks in the context of wireless communications[J]. IEEE Transactions on Information Forensics and Security, 2020, 15(1): 1102-1113.

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

[74] HAMEED M Z, GYÖRGY A, GÜNDÜZ D. Communication without interception: Defense against modulation detection[C]// Proceedings of 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). [S.l.]: IEEE, 2019: 1-5.

[75] KE D, HUANG Z, WANG X, et al. Application of adversarial examples in communication modulation classification[C]// Proceedings of 2019 International Conference on Data Mining Workshops (ICDMW). New York: IEEE, 2019: 877-882.

[76] HAMEED M Z. New quality measures for adversarial attacks with applications to secure communication [EB/OL]. (2020-07-09)[2023-10-15]. https://spiral.imperial.ac.uk:8443/handle/10044/1/82214.

[77] KIM B, SAGDUYU Y E, ERPEK T, et al. Adversarial attacks with multiple antennas against deep learning-based modulation classifiers[C]//Proceedings of IEEE Globecom Workshops. [S.l.]: IEEE, 2020.

[78] DELVECCHIO M, FLOWERS B, HEADLEY W C. Effects of forward error correction on communications aware evasion attacks[C]//Proceedings of the 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. New York: IEEE, 2020: 1-7.

[79] RESTUCCIA F, D'ORO S, AL-SHAWABKA A, et al. Generalized wireless adversarial deep learning[C]//Proceedings of ACM Workshop on Wireless Security and Machine Learning (WiseML). [S.l.]: ACM, 2020.

[80] DELVECCHIO M, ARNDORFER V, HEADLEY W C. Investigating a spectral deception loss metric for training machine learning-based evasion attacks[C]//Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. New York: ACM, 2020: 43-48.

[81] HAMEED M Z, GYORGY A.Communication without interception: Defense against modulation detection[C]//Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP). [S.l.]: IEEE, 2019.

[82] YI J, GAMAL A E. Gradient-based adversarial deep modulation classification with data-driven subsampling[EB/OL]. (2021-04-03)[2023-10-15]. https://arxiv.org/abs/2104.06375.

[83] BAHRAMALI A, NASR M, HOUMANSADR A, et al. Robust adversarial attacks against DNN-based wireless communication systems[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2021: 126-140.

[84] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2017-06-19)[2023-10-15]. https://arxiv.org/abs/1706.06083.

[85] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018.

[86] ERPEK T T, SAGDUYU Y E, SHI Y. Deep learning for launching and mitigating wireless jamming attacks[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 5(1): 2-14.

[87] ZHONG C, WANG F, GURSOY M C, et al. Adversarial jamming attacks on deep reinforcement learning based dynamic multichannel access[C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). [S.l.]: IEEE, 2020.

[88] WANG F, ZHONG C, GURSOY M C, et al. Adversarial jamming attacks and defense strategies via adaptive deep reinforcement learning[EB/OL]. (2020-07-12)[2023-10-15]. https://arxiv.org/abs/2007.06055.

[89] SHI Y, SAGDUYU Y E, ERPEK T, et al. How to attack and defend 5G radio access network slicing with reinforcement learning[EB/OL]. (2021-01-14)[2023-10-15]. https://arxiv.org/abs/2101.05768v2.

[90] SHI Y, SAGDUYU Y E. Adversarial machine learning for flooding attacks on 5g radio access network slicing[C]//Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops). [S.l.]: IEEE, 2021.

[91] COHEN J, ROSENFELD E, KOLTER Z. Certified adversarial robustness via randomized smoothing[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2019: 1310-1320.

[92] RAGHUNATHAN A, STEINHARDT J, LIANG P. Certified defenses against adversarial examples[EB/OL]. (2018-01-29)[2023-10-15]. https://arxiv.org/abs/1801.09344.

[93] KOKALJ-FILIPOVIC S, MILLER R, VANHOY G. Adversarial examples in RF deep learning: Detection and physical robustness[C]//Proceedings of 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). [S.l.]: IEEE, 2019: 1-5.

[94] DAVASLIOGLU K, SAGDUYU Y E. Trojan attacks on wireless signal classification with adversarial machine learning[C]//Proceedings of 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN). [S.l.]: IEEE, 2019: 1-6.

[95] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2023-10-15]. https://arxiv.org/abs/1503.02531.

[96] 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 (SP). [S.l.]: IEEE, 2016: 582-597.

[97] SAGDUYU Y E, SHI Y, ERPEK T. Adversarial deep learning for over-the-air spectrum poisoning attacks[J]. IEEE Transactions on Mobile Computing, 2019, 20(2): 306-319.

[98] O'SHEA T J, WEST N. Radio machine learning dataset generation with gnu radio[J]. Proceedings of the GNU Radio Conference, 2016, 1(1): 15-20.

作者简介:

江汉(1977-),男,副教授,硕士生导师,研究方向:无线通信技术、通信对抗、认知电子战,E-mail: jh_forward@126.com。

胡林(1997-),**通信作者**,男,硕士研究生,研究方向:认知无线电、调制识别、对抗样本攻击与防御,E-mail: linshiyiedu@sina.com。

李文(1996-),男,博士研究生,研究方向:认知无线电、智能抗干扰、博弈对抗。

焦雨涛(1992-),男,博士,研究方向:移动区块链网络、算法机制设计和物联网联合学习。

徐煜华(1983-),男,教授,博士生导师,研究方向:认知无线电、智能频谱对抗、无人机集群通信和博弈论。

徐逸凡(1995-),男,讲师,研究方向:认知无线电、智能抗干扰、博弈对抗。

(编辑:陈珺)