高技术通讯 2023 年 第 33 卷 第 2 期:135-145

doi:10.3772/j.issn.1002-0470.2023.02.003

基于深度学习的无线电信号对抗样本检测研究①

徐东伟② 郝海洋 宣 琦③ 杨 浩 周 晴

(浙江工业大学信息工程学院 杭州 310023)

摘 要 针对无线电信号的攻击愈来愈频繁的情况,本文在数据流形理论基础上,使用深度神经网络(DNN)检测无线电信号对抗样本及其攻击方法。首先使用 5 种不同攻击方法对无线电信号进行攻击产生对抗样本,其次使用 3 种不同的神经网络检测对抗样本,最后用残差神经网络(ResNet)检测对抗样本的攻击方法。在信噪比(SNR)为 30 dB 和 20 dB的无线电信号数据上的实验结果表明,本文所使用的残差神经网络检测精度接近 100%,在信噪比为 10 dB 的无线电信号数据上的检测精度仍然在 90% 以上。结果表明本文所用的残差神经网络能有效检测无线电信号的对抗样本及其攻击方法。

关键词 对抗样本检测;数据流形;深度神经网络(DNN);残差神经网络(ResNet)

0 引 言

近年来,电子信息技术飞速发展,在很多领域都有广泛应用,如无人机、船舶、空中管制、卫星远程测量及物联网等。而且移动通信系统已经发展到 5G 商业部署阶段,6G 的研究工作也已经在部分国家展开。因此无线电信号种类及数量呈指数级增长,面对各种无线电信号的挑战,各种不同的分类方法也开始涌现[1]。与此同时,深度学习[2]逐渐应用于各个领域,如生物医学[3-4]、自然语言处理[5]、语音识别[6]和视觉场景[7]等。深度学习在通信领域的应用也取得了许多成果。文献[8]使用傅里叶变换将无线电信号具象为图片,并用图像领域的先进成果进行识别取得了 12 种信号 86.04% 的识别率。文献[9]通过电磁信号的星座图将电磁信号转为二维图像,接着用轻量级深度神经网络(deep neural network, DNN)进行识别,提升系统准确率的同时降低了计算成本。文献[10]将信号的 I/Q 两路分别提取出来构建二维图像,接着用卷积神经网络识别取得了对 8 种调制信号 85% 的识别结果。文献[11]在宽光信噪比(signal-to-noise-ratio, SNR)范围内对 4 种调制信号实现 100% 的识别率,这些方法是使用卷积神经网络对信号的眼图进行识别。文献[12]采集实测信号作为实验样本并且设计了 33 层残差神经网络(residual neural network, ResNet),用短时傅里叶变换将信号转换得到时频图再用神经网络处理,实验结果表明,该方法可对多进制相位调制信号实现 99.9% 的识别准确率。文献[13]利用开源深度学习软件库训练并运行了一个完整的通信系统,其中引入了一个基于深度学习网络的帧同步模块。文献[14]将电磁信号的波形域数据转化为眼图与矢量图形式,利用多端卷积神经网络对其进行识别,当信噪比为 5 dB 时,识别准确率可达 95%。文献[15]利用深度学习以端到端的方式处理无线频分复用信道,解决了信道失真的问题,证明了深度学习是用于无线通信中具有复杂信道失真和干扰的信道估计以及信号检测的有力工具。

然而在深度神经网络已成功应用于处理复杂问题的同时,持续的研究[16-18]表明它们在强对抗环境

① 国家自然科学基金面上项目(61973273)资助。
② 男,1985 年生,博士,副教授;研究方向:机器学习,深度学习;E-mail: dongweixu@zjut.edu.cn。
③ 通信作者,E-mail: xuanqi@zjut.edu.cn。
(收稿日期:2021-07-05)

下极易受到干扰。文献[19]开发了构造平滑的心电图追踪对抗样本,该样本无法被领域专家察觉,却能让应用于医学影像的深度学习模型检测到异常。此外,在信号领域,对抗攻击研究工作也取得了一定成果。文献[20]通过在信道上发送利用无线信道的开放性设计的扰动信号,将通信系统的误块率提高了几个数量级。文献[21]提出通过发射机对其频谱感测结果进行深度学习,以预测空闲时隙来进行数据传输的空中频谱中毒攻击。文献[22]针对无线通信中的深度学习应用提出了木马攻击,取得了较好的效果。文献[23,24]研究发现,攻击方天线数量的增加会使攻击成功率显著提高,可以更好地利用信道多样性来进行对抗攻击,并证实了调制分类器对空中对抗攻击的脆弱性。随着无线电信号攻击领域研究的日益增多,对无线电信号的对抗样本检测技术的需求也越来越急迫。如果能检测出无线电信号对抗样本由何种攻击方法生成,将对相关防御工作大有裨益。

无线电信号的对抗样本及其攻击方法检测,为深度学习领域带来了许多挑战和机遇。深度学习以从数据中学习特征的方式,在图像处理和语音识别方面也取得很大成功。但其在无线电信号的对抗样本及其攻击方法检测方面还存在很大的研究空白。本文使用深度神经网络检测无线电对抗样本及其攻击方法。使用的 ResNet 在检测对抗样本及其攻击方法方面有着出色的表现。作为参照,本文还使用了核密度(kernel density, KD)和局部内在维度数(local intrinsic dimensionality, LID)方法检测无线电对抗样本及其攻击方法。

1 分类模型和相关攻击方法

1.1 分类模型

本文用卷积神经网络进行调制识别,具体结构如图 1 所示,命名为 AlNet。它将作为信号调制分类器对无线电信号做调制识别。在用 AlNet 进行相关实验时会将无线电信号输入数据折叠为 32 × 32 的形状。

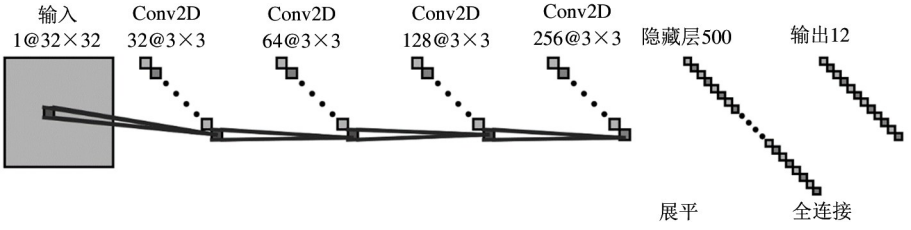

图 1 信号调制分类器结构图

1.2 攻击方法

由于已有的信号攻击工作成果较少,本文采用的攻击方法均从图像领域迁移而来,文献[25]建立了一个信号人工智能对抗攻击综合分析平台,详细评估了图像常用的攻击方法在信号领域的表现。以下是本文用到的 5 种攻击方法介绍。

快速梯度符号法(fast gradient sign method, FGSM)通过求出模型对输入的导数,然后用符号函数 sign()得到其具体的梯度方向,接着乘以一个步长,得到的“扰动”加在原来的输入上即得到在 FGSM 攻击下的信号。

$$M^* = M + \varepsilon \mathrm{sign}(\nabla_m J(\Theta, M, M_Z)) \quad (1)$$

其中,M^* 为对抗样本,M 为原始信号,ε 为扰动系数,$J(\Theta, M, M_Z)$ 为模型的损失函数。M_Z 为 M 的调制方法标签,Θ 为模型参数。

文献[26]提出 FGSM 变体基本迭代法(basic iterative method, BIM),采取多个小步骤进行多次迭代,并在每一步后调整方向。随后,文献[27]在此基础上又提出变体投影梯度下降法(projected gradient descent, PGD),在此基础上增加迭代轮数并添加一层随机化处理达到了更佳的分类效果。

文献[28]提出了基于雅可比矩阵的显著图攻击(Jacobian-based saliency map attack, JSMA),此方法首先计算给定信号的雅可比矩阵,通过寻找对输出结果影响最显著的信号的输入特征来添加扰动。文献[29]提出一种基于差分进化生成单像素的对

抗性扰动的单像素攻击法(one pixel attack, OPA),可以在最小攻击信息的条件下对网络进行欺骗。表1为本文中使用的攻击方法的总结。

表1 攻击方法

缩写	全称
FGSM	快速梯度符号法
BIM	基本迭代法
PGD	投影梯度下降法
JSMA	基于雅可比矩阵的显著图攻击
OPA	单像素攻击法

2 检测对抗样本原理及方法

2.1 背景原理

每个生成对抗样本的算法都可以改变预测的标签,而不改变潜在的真实标签,意味着人仍然可以正确地分类一个对抗样本,但模型不会。这可以从多种训练数据的角度来理解。许多高维数据集,如信号、图像,被认为位于一个低维流形上。文献[30]表明,通过仔细遍历数据流形,可以改变图像的底层真实标签。那么对输入不构成有明显的变化的对抗性扰动会将样本从它所属数据流形中推出来使其成为对抗样本。文献[31]在基于对抗样本位于接近数据子流形边缘的类边界的假设上对对抗样本进行研究。文献[30]证明 DNNs 只能在训练数据的小流形附近正确运行。本文在生成无线电信号对抗样本时所使用的方法由图像的攻击方法迁移而来,所以以上数据流形理论虽然是建立在图像数据上,但仍然可以迁移到无线电信号数据上。因此,本文认为无线电信号对抗样本也存在以上所述现象,即无线电信号对抗样本不存在于数据流形上,同时又靠近相应的数据流形。因为对抗样本的产生是对抗性扰动将原始样本从它所属数据流形中推出来的结果。

2.2 深度神经网络

根据2.1节数据流形理论,检测器可以通过检测在邻近类边界方向上稍微偏离数据流形中心的输入来识别敌对的例子。因此,检测器可以专注于检测输入摆脱数据流形的一个方向,即其中一个方向边界附近的一个类[32]。本文使用3种神经网络作为检测器检测无线电信号对抗样本。

深度神经网络本身拥有大量的自由参数,同时结合强大的正则化技术[33-34]、适应性矩估计(adaptive moment estimation, ADAM)优化算法、低成本高性能的显卡处理能力,并结合关键的神经网络架构创新,如卷积神经网络[35]和线性整流函数,使得DNN对输入的高维数据的特征学习能力大幅加强。

虽然 AlexNet 出现以来网络算法和体系结构得到改进,DNN 性能得到了显著提升,但是核心方法几乎没有改变。神经网络由一系列层组成,通过非线性的矩阵运算将每一层输入 h_0 映射到输出 h_1。公式如下所示。

$$h_1 = \sigma(\boldsymbol{W} \cdot h_0 + \boldsymbol{b}) \tag{2}$$

其中,权重系数矩阵 $\boldsymbol{W}$ 维度数为 $|h_0 \cdot h_1|$,偏置矩阵 $\boldsymbol{b}$ 维度数为 $|h_1|$,σ 为激活函数。通过权值共享减少神经网络中的参数。DNN 有前向传播和反向传播算法。前向传播算法就是根据输入从最开始的输入层,经过隐藏层一直到输出层得到输出的过程。反向传播算法计算输出值和实际值的误差,并将该误差从输出层经过隐藏层到输入层,在此过程中优化系数矩阵 $\boldsymbol{W}$ 和偏置矩阵 $\boldsymbol{b}$。

本文中作为检测器的第1种神经网络结构为AlNet。第2种神经网络借鉴于文献[36],由2个卷积层和2个全连接层组成。每层使用线性整流函数,输出层使用归一化指数函数。它已经被证明在信号的调制分类上有着极好的精度[36],这里被迁移作为检测信号对抗样本的检测器,命名为 O _ CNN。作为检测器的第3种神经网络是残差神经网络(ResNet),下面是详细介绍。

随着 AlexNet 出现以来网络算法和体系结构的改进,使用更深层次的网络成为可能,并可以使其性能提高。在计算机视觉空间中,ResNet 的思想已经越来越得到认可。如图2所示,在 ResNet 中,跳过或进行旁路连接的概念被大量使用,允许特征可以在网络的多层尺度和深度上传输。这使得计算机视觉性能显著改善,并且在时间序列音频数据中得到了有效使用。受此启发,在设计无线电信号对抗样本检测器时使用了 ResNet。Residual unit 和 Residual stack 如图2所示,检测对抗样本时的 ResNet 的网

络布局如表 2 所示。在网络的全连接区域使用自归一化神经网络,同时采用缩放指数线性单元激活函数。当检测对抗样本攻击方法时,会根据具体情况修改 Fc/Softmax 的输出维度。

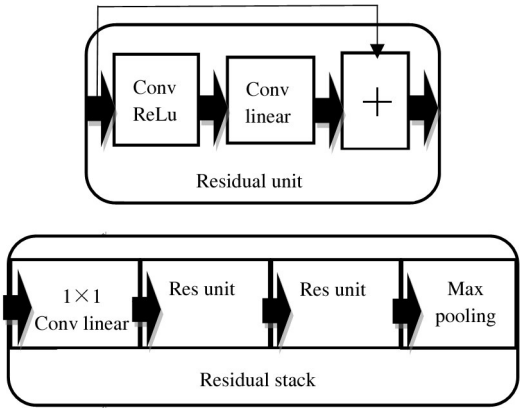

图 2　网络中使用的层次结构

表 2　ResNet 的网络布局

层级	输出维度
Input	512 · 2
Residual stack	32 · 256 · 1
Residual stack	32 · 128 · 1
Residual stack	32 · 64 · 1
Residual stack	32 · 32 · 1
Residual stack	32 · 16 · 1
Residual stack	32 · 8 · 1
Fc/SeLu	128
Fc/Softmax	2

在具体使用上述检测器对无线电信号数据进行相关实验时,也考虑了影响分类精度的因素,包括模型大小或深度、卷积层数量和卷积核大小以及数据集输入尺寸等。本文在模型的大小深度和卷积层数量及卷积核大小方面实验了多种情况,设置了不同的参数,最终选取了分类精度较好且计算消耗较少的模型结构。数据集的输入尺寸也尝试了多种类型,如在 ResNet 网络上经过实验证明数据集以 512 ×2 的尺存输入最好。

在进行检测信号对抗样本实验时也曾将循环神经网络中的长短期记忆(long short-term memory, LSTM)网络作为检测器,初步进行的一些实验结果表明其检测精度整体上弱于 ResNet、强于 AlNet,由于 LSTM 时间复杂度最高,所以没有在本文对其进一步探讨。

2.3　检测对抗样本的方法

作为参照本文使用了 KD 和 LID 技术检测对抗样本,根据 2.1 节中数据流形理论,无线电信号对抗样本是从其相对应的数据流形上推下来的,即无线电信号原始样本和其对抗样本与相应的数据流形存在着不同的“距离”,可以凭借“距离”不同来检测对抗样本。KD 和 LID 就是很好衡量此“距离”的方式,它们的具体计算方法可以参考文献[37]。

本文中检测对抗样本的实验过程如下。(1)使用 1.2 节所述的攻击方法分别对无线电信号原始样本进行攻击产生对抗样本。(2)将原始样本和其对应的对抗样本混合,制作出混合样本集,再将混合样本集按 8∶2 划分为训练集和验证集,混合前会给原始样本和对抗样本打上标签,而且会一直保留以便区分,标签不会对样本有任何影响。(3)当使用深度神经网络的方法检测对抗样本时,会将训练集和验证集送入 2.2 节中介绍的 3 种深度神经网络分别进行训练测试,验证集准确率作为此方法检测对抗样本的检测精度,并用于评估此方法效果。(4)当使用 KD 和 LID 方法检测对抗样本时,会分别提取训练集和验证集的数据 KD 和 LID 特征,得到 KD 和 LID 训练集特征集以及 KD 和 LID 验证集特征集,使用 KD 训练集特征集训练逻辑回归分类器,用 KD 验证集特征集得到逻辑回归分类器的验证集准确率,验证集准确率作为此方法检测对抗样本的检测精度,并用于评估此 KD 方法效果。LID 特征检测方法与 KD 相同。图 3 为检测对抗样本的过程。

检测对抗样本的攻击方法如下。(1)使用 1.2 节所述的攻击方法分别对无线电信号原始样本进行攻击产生对抗样本,这里将尽力确保不同攻击方法产生的对抗样本相对于原始样本的变化幅度相近。(2)将产生的对抗样本混合,制作出混合对抗样本数据集,再将混合对抗样本数据集按 8∶2 划分为训练集和验证集,混合前会给所有对抗样本打上标签,而且会一直保留以便区分,标签不会对样本有任何影响。(3)当使用深度神经网络的方法检测对抗样本的攻击方法时,会将训练集和验证集送入2.2节

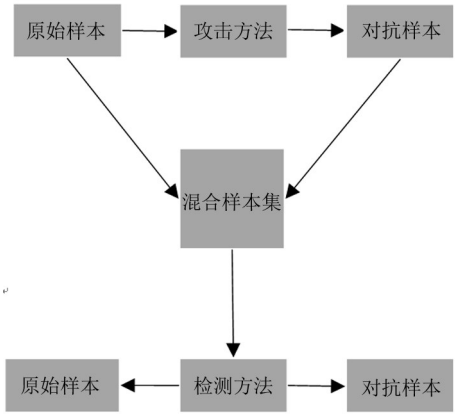

图 3　无线电对抗样本检测过程

中介绍的残差神经网络进行训练测试,验证集准确率作为此方法检测对抗样本的检测精度,并用于评估此 ResNet 方法效果。(4)当使用 LID 方法检测对抗样本的攻击方法时,会分别提取训练集和验证集的数据 LID 特征,得到 LID 训练集特征集以及 LID 验证集特征集,使用 LID 训练集特征集训练支持向量机(support vector machines,SVM)分类器,用 LID 验证集特征集得到 SVM 分类器的验证集准确率,验证集准确率作为此方法检测对抗样本的检测精度,并用于评估此 LID 方法效果。图 4 为检测无线电对抗样本的攻击方法过程。

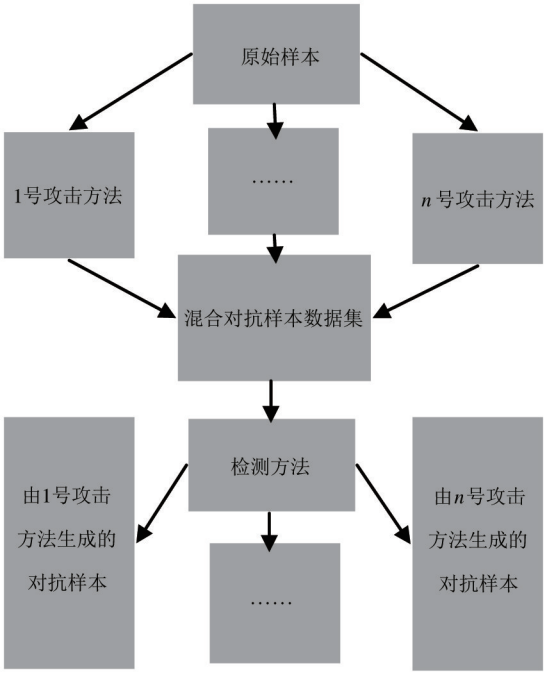

图 4　无线电对抗样本的攻击方法检测过程

3　实验结果与分析

3.1　实验设置

无线电信号数据集调制类别具体包括相移键控调制、频移键控调制、正交幅度调制和脉冲振幅调制,共 12 个小类别,即 BPSK、QPSK、8PSK、OQPSK、2FSK、4FSK、8FSK、16QAM、32QAM、64QAM、4PAM 和 8PAM。原始样本数据集是随机生成的,以保证传输比特的概率相等。脉冲整形滤波器采用升余弦滤波器和滚转系数,在[0.2, 0.7]范围内提取一个随机值。相位偏差在[$-\pi$,π]范围内随机选择,归一化载波频率偏移在[−0.1,0.1]范围内随机选择。每个调制类别的信噪比从 −20 dB均匀分布到 30 dB,间隔 2 dB。每个原始样本都是 IQ 信号,包含 64 个符号,每个符号的采样点数为 8,因此每个原始样本的采样点数为 512。信号训练集和信号验证集的大小分别为 312 000 × 512 × 2 和 156 000 × 512 × 2,每类调制信号样本量相同。

本文使用了上述数据集中信噪比为 10 dB、20 dB和 30 dB 的无线电数据集分别进行实验。每个单信噪比的信号训练集大小为 12 000 × 512 × 2。使用 1.1 节中介绍的神经网络进行调制识别。在 3 种信噪比上训练集精度都达到了 90% 以上。所有用于生成对抗样本的信号训练集数据都经过了筛选,也就是信号训练集数据中能被神经网络正确识别其类别的信号数据才用于本文中的实验,作为无线电信号原始样本集。

3.2　实验结果及分析

本文用 L2 范数来衡量攻击方法对原始样本的扰动幅度。L2 范数越大意味着对抗样本相对于原始样本变化越大。用检测精度衡量检测方法检测对抗样本的能力,精度越高表明检测能力越强。

检测对抗样本采用 2.3 节介绍的方法,包括使用深度神经网络以及提取无线电信号数据 KD 和 LID 特征的方法。选取了 2 个扰动级别进行实验。表 3 ~ 表 5 分别为在无线电数据集信噪比 30 dB、20 dB和 10 dB 上进行的检测对抗样本实验结果。

表 3 在信噪比为 30 dB 的数据集上的实验结果

攻击方法	检测方法	L2 范数(低)	检测精度	L2 范数(高)	检测精度
	ResNet		0.9989		0.9993
	AlNet		0.9719		0.9918
JSMA	O _ CNN	0.49	0.8540	1.16	0.9902
	LID		0.7644		0.7999
	KD		0.5493		0.6316
	ResNet		0.9994		0.9999
	AlNet		0.5000		0.5000
FGSM	O _ CNN	0.52	0.5000	1.00	0.9999
	LID		0.8552		0.8678
	KD		0.4994		0.7861
	ResNet		0.9999		0.9999
	AlNet		0.5000		0.9647
BIM	O _ CNN	0.52	0.5000	1.00	0.9715
	LID		0.7788		0.8353
	KD		0.5114		0.8029
	ResNet		0.9999		0.9999
	AlNet		0.5000		0.5000
PGD	O _ CNN	0.53	0.5000	0.98	0.5132
	LID		0.7788		0.8570
	KD		0.5114		0.8329
	ResNet		0.9858		0.9999
	AlNet		0.5000		0.5000
OPA	O _ CNN	0.58	0.5000	0.93	0.9189
	LID		0.6041		0.6466
	KD		0.5613		0.7127

表 4 在信噪比为 20 dB 的数据集上的实验结果

攻击方法	检测方法	L2 范数(低)	检测精度	L2 范数(高)	检测精度
	ResNet		0.9731		0.9982
	AlNet		0.8592		0.9841
JSMA	O _ CNN	0.57	0.7886	1.22	0.9783
	LID		0.8133		0.8813
	KD		0.5698		0.6952
	ResNet		0.9963		0.9999
	AlNet		0.5000		0.5000
FGSM	O _ CNN	0.61	0.5031	1.00	0.9565
	LID		0.8354		0.8568
	KD		0.5006		0.5129
	ResNet		0.9963		0.9994
	AlNet		0.5000		0.9572
BIM	O _ CNN	0.52	0.5037	1.00	0.9682
	LID		0.7913		0.8372
	KD		0.5006		0.8556

(续表4)

	ResNet		0.9908		0.9994
	AlNet		0.5000		0.9963
PGD	O _ CNN	0.52	0.5000	1.06	0.5000
	LID		0.8207		0.8488
	KD		0.6805		0.8256
	ResNet		0.9394		0.9810
	AlNet		0.9321		0.9743
OPA	O _ CNN	0.59	0.9308	0.96	0.9761
	LID		0.6010		0.6401
	KD		0.5520		0.7130

表5　在信噪比为10 dB的数据集上的实验结果

攻击方法	检测方法	L2范数(低)	检测精度	L2范数(高)	检测精度
	ResNet		0.9437		0.9454
	AlNet		0.5000		0.8476
JSMA	O _ CNN	0.58	0.6453	1.02	0.8944
	LID		0.8643		0.9091
	KD		0.6517		0.7202
	ResNet		0.6505		0.9123
	AlNet		0.5000		0.6216
FGSM	O _ CNN	0.53	0.5000	1.00	0.7170
	LID		0.8150		0.8188
	KD		0.5070		0.8342
	ResNet		0.7138		0.9520
	AlNet		0.5000		0.5749
BIM	O _ CNN	0.61	0.5000	1.00	0.5000
	LID		0.7593		0.7862
	KD		0.5007		0.8726
	ResNet		0.7305		0.9301
	AlNet		0.5000		0.5000
PGD	O _ CNN	0.61	0.5013	1.01	0.7164
	LID		0.8092		0.8169
	KD		0.7196		0.8307
	ResNet		0.9500		0.9930
	AlNet		0.9430		0.9872
OPA	O _ CNN	0.59	0.9432	0.94	0.9897
	LID		0.6575		0.6914
	KD		0.5115		0.6677

图5为在信噪比30 dB数据集上低扰动(L2范数低)情况下5种检测器检测5种对抗样本的能力分析。图6为在信噪比10 dB数据集上低扰动情况下5种检测器检测5种对抗样本的能力分析。图7为使用JSMA攻击方法分别在信噪比30 dB、20 dB和10 dB数据集上低扰动情况下进行的5种检测器检测对抗样本的能力分析。从实验结果中可以得到以下结论。

(1)在信噪比为20 dB和30 dB的数据集中使用ResNet方法检测对抗样本时,在低扰动和高扰动情况下都有着较好的效果,绝大多数情况下检测精度接近100%,意味着几乎可以完全检测出对抗样

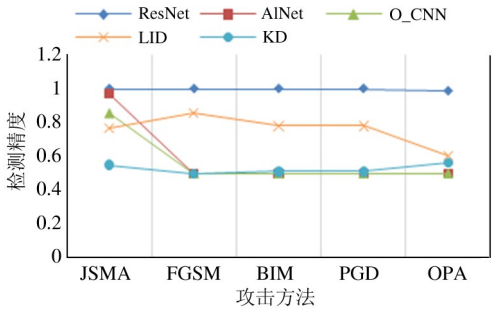

图 5　在信噪比 30 dB 数据集上低扰动情况下 5 种检测器检测 5 种对抗样本的能力分析

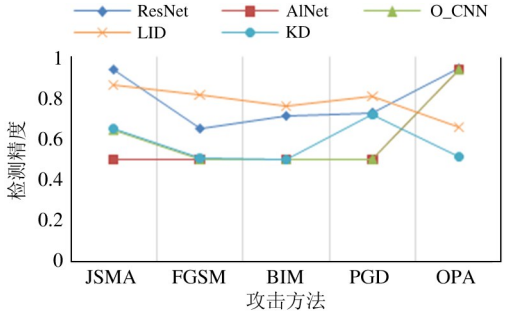

图 6　在信噪比 10 dB 数据集上低扰动情况下 5 种检测器检测 5 种对抗样本的能力分析

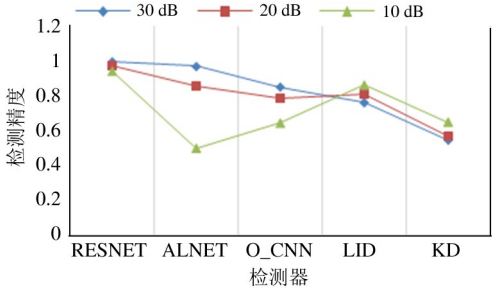

图 7　使用 JSMA 攻击方法分别在信噪比 30 dB、20 dB 和 10 dB 数据集上低扰动情况下进行的 5 种检测器检测对抗样本的能力分析

本，好于 KD 和 LID 方法检测效果。AlNet 和 O _ CNN 也有较好的表现，但检测效果表现出了明显的不稳定性，而且在检测效果较好的时候仍然落后于 ResNet。

（2）在信噪比为 10 dB 的数据集使用 ResNet 方法检测对抗样本时，由于噪声影响检测效果有所下降，在高扰动情况下的检测效果仍然很高，远好于 KD 和 LID 方法。低扰动情况下检测由 FGSM、BIM 和 PGD 攻击方法生成的对抗样本时，检测效果明显下降而且没有 LID 方法好。检测由 JSMA 和 OPA 攻击方法生成的对抗样本时，检测效果仍然较高并好于 KD 和 LID 方法。AlNet 和 O _ CNN 在检测由 JSMA 和 OPA 攻击方法生成的对抗样本时，检测效果较好，其他情况下较差。

（3）使用 LID 方法检测对抗样本时，除 OPA 攻击方法生成的对抗样本，其他检测效果较好。信噪比降低并没有对 LID 检测方法的检测能力造成明显影响，证明其抗噪能力较强。而 KD 方法的检测效果整体较差。

检测对抗样本的攻击方法采用 2.3 节介绍的使用深度神经网络 ResNet 和提取无线电数据 LID 特征 2 种方法，使用的对抗样本数据仍为检测对抗样本实验中所使用的对抗样本数据。这里将扰动级别低（L2 范数低）和扰动级别高（L2 范数高）的 5 种对抗样本数据分别混合进行实验以确保混合的对抗样本数据扰动幅度相近。

首先在 30 dB 上的数据集进行实验，在进行 L2 范数高的对抗样本攻击方法检测时，ResNet 的检测精度为 78.41%，LID 的检测精度为 44.31%。通过观察此时 ResNet 的混淆矩阵，如图 8 所示，可以看出 5 种攻击方法中主要是 FGSM 和 BIM 攻击产生的对抗样本之间无法分辨。

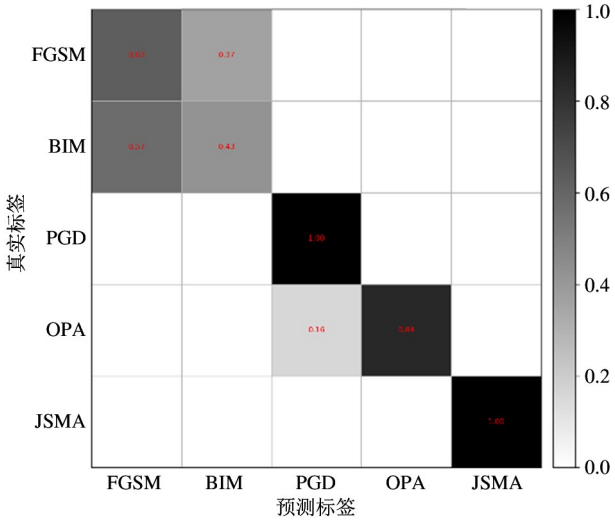

图 8　ResNet 的混淆矩阵图

由于文献[16]提出 FGSM 变体 BIM 是在 FGSM 基础上采取多个小步骤进行多次迭代，并在每一步

后调整方向,所以推测产生这种情况的原因主要是FGSM 和 BIM 的攻击方法原理上十分接近,导致在扰动不够高的情况下它们的攻击效果十分接近,ResNet 方法无法分辨它们攻击所产生的对抗样本。

由于文献[27]在 BIM 基础上提出变体 PGD,在 BIM 基础上增加迭代轮数并添加一层随机化处理,所以推测 PGD 也符合这种情况,即扰动较低时由于 FGSM、BIM 和 PGD 原理十分接近,导致 ResNet 方法无法辨出它们三者产生的对抗样本。

为了验证这一猜想,设计了以下实验。在信噪比为30 dB 的无线电数据集上生成 L2 在 2.4 左右的极高扰动的 3 种对抗样本数据集,它们分别由 FGSM、PGD 和 BIM 攻击方法生成。将它们混合后用 2.3 节中介绍的 ResNet 方法对它们进行分类检测,此时的检测精度为 99.80%。用检测对抗样本实验中由 FGSM、PGD 和 BIM 在信噪比为 30 dB 的数据集上攻击生成的对抗样本数据进行检测对抗样本攻击方法实验,这些对抗样本的 L2 在 0.5 左右。将它们混合后用 2.3 节中介绍的 ResNet 方法对它们进行分类检测,此时的检测精度为 33.33%。这个实验说明当扰动较低的时候,ResNet 方法将无法区分由 FGSM、PGD 和 BIM 攻击方法产生的对抗样本数据。所以在接下来的检测对抗样本的攻击方法实验中只保留 FGSM、JSMA 和 OPA 这 3 种攻击方法,这 3 种攻击方法的原理差别很大。表 6 为分别在信噪比为 30 dB、20 dB 和 10 dB 时进行的检测对抗样本的攻击方法的实验结果。

表 6　检测对抗样本的攻击方法实验结果

信噪比	检测方法	L2 范数(低)			检测精度	L2 范数(高)			检测精度
		FGSM	JSMA	OPA		FGSM	JSMA	OPA	
30 dB	ResNet	0.52	0.49	0.58	0.9956	1.00	1.16	0.93	0.9463
	LID				0.5613				0.5919
20 dB	ResNet	0.61	0.57	0.59	0.9877	1.00	1.22	0.96	0.9857
	LID				0.5731				0.5888
10 dB	ResNet	0.53	0.58	0.59	0.7294	1.00	1.02	0.94	0.9458
	LID				0.5483				0.4730

可以看出 ResNet 方法在数据集为 30 dB 和 20 dB上时,在高扰动和低扰动上的检测效果都非常好。在数据集为 10 dB 时,高扰动上仍然有很好的检测效果,低扰动时检测效果明显下降。LID 方法在检测对抗样本的攻击方法时明显落后于 ResNet 方法。

4　结 论

本文针对无线电信号对抗样本检测需求越来越大的情况,在数据流形理论基础上,通过使用深度神经网络作为检测器来区分无线电信号原始样本和对抗样本,并选择其中表现最好的残差神经网络进一步用于检测对抗样本的攻击方法。实验结果表明,残差神经网络在检测对抗样本上有较好的效果,并且在检测对抗样本的攻击方法上也有不错的效果。但本文方法在低信噪比和低扰动情况下的效果还有待提升,今后将继续探索这种情况下如何提升检测能力。

参考文献

[1] 黄震宇. 基于深度学习的无线电信号识别方法研究[D]. 西安:西安电子科技大学人工智能学院, 2018: 1-4.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015,521(7553): 436-444.

[3] SHEN D, WU G, SUK H I. Deep learning in medical image analysis[J]. Annual Review of Biomedical Engineering, 2017,19: 221-248.

[4] KERMANY D S, GOLDBAUM M, CAI W, et al. Identifying medical diagnoses and treatable diseases by image-

based deep learning[J]. Cell, 2018, 172(5): 1122-1131.

[5] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing [J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.

[6] LI J, DAI W, METZE F, et al. A comparison of deep learning methods for environmental sound detection[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017: 126-130.

[7] OWENS A, EFROS A A. Audio-visual scene analysis with self-supervised multisensory features[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich: IEEE,2018: 631-648.

[8] 周鑫, 何晓新, 郑昌文. 基于图像深度学习的无线电信号识别[J]. 通信学报, 2019, 40(7): 114-125.

[9] 张思成, 林云, 涂涯. 基于轻量级深度神经网络的电磁信号调制识别技术[J]. 通信学报, 2020, 41(11): 12-21.

[10] 陶冠宏, 廖开升, 周林. 一种基于深度学习的电磁信号建模与调制识别新方法[J]. 电子信息对抗技术, 2019,34(5):10-15.

[11] WANG D S, ZHANG M, LI Z, et al. Modulation format recognition and OSNR estimation using CNN-based deep learning[J]. IEEE Photonics Technology Letters, 2017, 29(19): 1667-1670.

[12] 吴佩军, 侯进, 吕志良, 等. 基于卷积神经网络的多进制相位调制信号识别算法[J]. 计算机应用与软件, 2019, 36(11): 202-209.

[13] DÖRNER S, CAMMERER S, HOYDIS J, et al. Deep learning based communication over the air[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 12(1):132-143.

[14] 查雄, 彭华, 秦鑫, 等. 基于多端卷积神经网络的调制识别方法[J]. 通信学报, 2019, 40(11): 30-37.

[15] YE H, LI G Y, JUANGB H. Power of deep learning for channel estimation and signal detection in OFDM systems [J]. IEEE Wireless Communications Letters, 2017, 7(1): 114-117.

[16] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2015-03-20)[2021-07-05]. https://arXiv.org/pdf/1412.6572V3.pdf.

[17] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015:427-436.

[18] SHARIF M, BHAGAVATULA S, BAUER L, et al. Adversarial generative nets: neural network attacks on state-of-the-art face recognition [EB/OL]. (2017-12-31)[2021-07-05]. https://arXiv.org/pdf/1801.00349V1.pdf.

[19] HAN X, HU Y, FOSCHINI L, et al. Deep learning models for electrocardiograms are susceptible to adversarial attack[J]. Nature Medicine, 2020, 26(3):360-363.

[20] SADEGHI M, LARSSON E G. Physical adversarial attacks against end-to-end autoencoder communication systems[J]. IEEE Communications Letters, 2019,23(5): 1-1.

[21] SAGDUYU Y, 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.

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

[23] KIM B, SAGDUYU Y E, ERPEK T, et al. Adversarial attacks with multiple antennas against deep learning-based modulation classifiers[C]//2020 IEEE Globecom Workshops. Taibei: IEEE, 2020: 1-6.

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

[25] 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): 141-148.

[26] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2017-02-11)[2021-07-05]. https://arXiv.org/pdf/1607.02533.pdf.

[27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards

deep learning models resistant to adversarial attacks[EB/OL]. (2017-08-09)[2021-07-05]. https://arXiv.org/pdf/1706.06083V2.pdf.

[28] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//2016 IEEE European Symposium on Security and Privacy (EuroS&P). Saarbruecken: IEEE, 2016: 372-387.

[29] SU J, VARGAS D V, KOUICHI S. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5):828-841.

[30] GARDNER J R, UPCHURCH P, KUSNERM J, et al. Deep manifold traversal: changing labels with convolutional features[EB/OL]. (2015-11-19)[2021-07-05]. https://arXiv.org/pdf/1511.06421V1.pdf.

[31] TANAY T, GRIFFIN L. A boundary tilting persepective on the phenomenon of adversarial examples[EB/OL]. (2016-08-27)[2021-07-05]. https://arXiv.org/pdf/1608.07690.pdf.

[32] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[EB/OL]. (2017-02-21)[2021-07-05]. https://arXiv.org/pdf/1702.04267.pdf.

[33] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[34] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning. Lille: JMLR, 2015:448-456.

[35] LeCUN Y, BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[36] O'SHEA T J, CORGAN J, CLANCY T C. Convolutional radio modulation recognition networks[C]// International Conference on Engineering Applications of Neural Networks. Aberdeen: EANN, 2016:213-226.

[37] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[EB/OL]. (2018-03-14)[2021-07-05]. https://arXiv.org/pdf/1801.02613.pdf.

Research on detection of radio signal adversarial samples based on deep learning

XU Dongwei, HAO Haiyang, XUAN Qi, YANG Hao, ZHOU Qing

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

Abstract

Aiming at the problem of the increasing frequent attacks on radio signals, based on data manifold theory, the deep neural network (DNN) is used to detect radio signal adversarial samples and their attack methods. First, five different attack methods are used to attack radio signals to generate adversarial samples. Second, three different neural networks are used to detect adversarial samples. Last, the residual neural network (ResNet) is used to detect adversarial samples' attack methods. The experimental results on radio signal data with signal-to-noise-ratio (SNR) of 30 dB and 20 dB show that the detection accuracy of the residual neural network used in this paper is close to 100%, while the experimental results on radio signal data with SNR of 10 dB show that the detection accuracy is still above 90%. The results show that the residual neural network used in this paper can effectively detect the adversarial samples of radio signals and their attack methods.

Key words: adversarial sample detection, data manifold, deep neural network (DNN), residual neural network (ResNet)