DOI:10.19651/j.cnki.emt.2212112

辐射源个体识别中的对抗攻击研究

刘丰汇[1,2] 张治中[1] 张 涛[2] 杨小蒙[1,2]
(1.南京信息工程大学电子与信息工程学院 南京 210044; 2.国防科技大学第六十三研究所 南京 210007)

摘 要: 基于深度学习的辐射源个体识别研究主要关注识别精度的提升,往往忽视了识别过程中对抗样本的威胁。针对上述问题,本文在增加辐射源个体类别并提升模型识别精度的同时分析研究了对抗样本对高识别率深度学习识别网络产生的影响。首先获取小样本 ADS-B 信号,通过数据随机切片进行数据增强;再对原有网络进行微调并加入卷积注意力模块提高模型对辐射源个体信号的识别率;最后使用 4 种攻击算法生成对抗样本并在辐射源个体识别网络上进行测试。除此之外,还将攻击前后的信号样本转化为图片进行可视化比较,以在攻击成功率和攻击隐蔽性之间权衡。实验结果表明,优化后的高识别率模型也容易受到对抗样本的攻击,基于动量的迭代攻击效果最好,相比于快速梯度下降的攻击方法的攻击效果高出 10%。

关键词: 辐射源个体;深度学习;对抗样本;注意力机制;数据随机切片
中图分类号: TN92;TP183 **文献标识码:** A **国家标准学科分类代码:** 510.5015

The research of adversarial attacks in specific emitter identification

Liu Fenghui[1,2] Zhang Zhizhong[1] Zhang Tao[2] Yang Xiaomeng[1,2]
(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. The 63rd Research Institute of National University of Defense Technology, Nanjing 210007, China)

Abstract: The research of specific emitter identification based on deep learning mainly focuses on the improvement of recognition accuracy, but often ignores the threat of adversarial samples in the recognition process. To solve the above problems, the experiment not only increases the category of emitter and improves the accuracy of model recognition, but also analyzes the impact of adversarial samples on deep learning recognition network with high recognition rate. In experiments, small samples of ADS-B signals were obtained, and the data were sliced randomly. Then fine tune the original network and add convolutional attention module to improve the recognition rate of the model. Finally, generate adversarial samples were created by using four adversarial attack algorithms and tested on the network which was trained in advance. Additionally, images of signal examples before and after the attack were compared to maintain a balance between the attack success rate and the attack concealment. The results show that the model with high recognition rate is also vulnerable to adversarial samples, the momentum iteration method has the best performance among four algorithms, and the attack performance of momentum iteration method is more than 10% higher than the fast gradient sign method.

Keywords: specific emitter;deep learning;adversarial example;attention mechanism;data random slicing

0 引 言

飞速发展的通信技术和加密技术使得无线设备广泛应用于多个领域之中。然而,开放的无线网络屏蔽不良信息或者非法设备的能力有限,这使无线通信很容易成为恶意软件攻击的对象。为了减少恶意用户潜在的威胁,文献[1]通过提取辐射源的射频指纹特征,用于识别不同辐射源设备。文献[2]中提出基于瞬态强度的射频指纹识别方法提高了识别率并缩减识别的时间。

但在实际通信时,密集复杂的电磁信号和较短的信号传输时间使得传统人为提取射频指纹特性的难度进一步加大。随着深度学习的兴起,由于其强大的拟合复杂数据和自主学习的能力,一些研究者将其应用于信号识别领域。2016 年 O'shea 等[3]在无线信号领域首次研究卷积神经网络(convolutional neural network,CNN)的应用,并与目前广泛使用的专家特征的方法进行比较,用实验证明了 CNN

收稿日期:2022-11-17

对信号识别的可行性。耿梦婕等[4]和陈小惠等[5]使用CNN进行辐射源个体识别,证明了CNN优秀的性能。

尽管神经网络在信号识别和个体识别领域都取得了傲人的效果,但由于神经网络的黑盒特性,研究者并不了解神经网络的学习过程,这种未知性无疑是增加了神经网络被攻击的风险。2014年Szegedy等[6]首次提出对抗样本的概念,他们在图片分类的神经网络中发现了神经网络的一个严重缺陷:通过对输入图片中加入人眼难以察觉的微小扰动,加入扰动后的图片被称作对抗样本,对抗样本可以轻易地使神经网络发生决策错误。Moosavi等[7]发现存在一种普适性的对抗样本,其迁移性使得不同网络被攻击的风险增加。Goodfellow等[8]提出了快速梯度下降法(fast gradient sign method,FGSM)来欺骗基于CNN的分类网络。Kurakin等[9]通过对FGSM算法加入迭代的思想,提出了基本迭代法(basic iterative method,BIM),通过不断地迭代来更新所生成的对抗样本从而优化攻击效果。

可以看出,目前对抗样本的研究在图像处理,语音识别,自然语言处理等领域都取得了丰富的理论和应用成果,但是在电磁空间领域的研究才刚刚起步。2019年Sadegh等[10]首次将对抗样本应用于调制识别的信号领域,并通过实验表明在无线通信系统的算法设计中,已经训练好的CNN模型非常容易受到对抗样本的攻击。在之后的研究中,Lin等[11]将基于标签计算梯度的传统对抗攻击方法应用到了信号调制识别当中,验证了基于信号调制识别的深度神经网络极易受到对抗样本的攻击。王超等[12]使用特征梯度攻击的算法攻击信号调制识别网络并取得良好的攻击效果。可见,对抗样本的存在使得基于深度学习的信号识别任务存在严重的安全隐患。但是上述研究中使用的数据集都是针对信号调制方式的识别且数据类别较少,没有考虑辐射源个体的识别问题和实际通信情况下多类别识别时的攻击情况。除此之外,上述研究中所攻击的目标模型识别率并不算很高。

为了更好地评估和模拟对抗样本对于性能良好的辐射源识别网络产生的影响,在数据集方面,实验采用更多种类的数据集,使用20类ADS-B辐射源信号作为数据集,并使用数据随机切片处理。在模型选择方面,为了解决模型识别率低的问题,分别对ResNet18、Vgg16的部分层参数调整,并加入空间注意力机制和通道注意力机制,使得网络识别精确度达到90%以上,Vgg16的识别率为91%左右;ResNet18识别率达到94.2%。然后使用4种不同的攻击算法来生成对抗样本来攻击目标模型并观察攻击效果。针对黑盒攻击,为了验证对抗攻击的泛化性,实验选用ResNet18作为黑盒攻击的替代模型。结果显示,随着扰动幅度增加,在白盒攻击和黑盒攻击的环境下,模型识别率分别下降60%和50%左右。表明高识别精度的模型在存在对抗样本的情况下也极易受到干扰。

1 数据处理与模型改进

1.1 信号数据随机切片

本文选择航空广播式自动相关监视系统ADS-B的20类辐射信号作为数据集训练模型并生成对抗样本,数据是在相同地点采集的来自20架某民航客机的信号组成,将20类信号划分为ADS-B-0到ADS-B-19。由于采集的均为小样本数据,使用数据随机切片进行数据增强。

由于原本数据集每类只有113~144条数据样本,但是每条数据样本却有4 096个采样点,为了方便后续目标模型的训练以及增加数据的特征平移不变性,提高模型的识别率,在数据输入分类器之前首先采取滑动窗口对原始数据进行切片处理,将每帧信号处理为$n(k)$个固定长度的子帧,从而将数据扩充。

数据集包括20类ADS-B辐射源的信号数据,一共为2 000×4 800×2,即20台设备,每一条样本采样点为4 800,信号分为I路和Q路。滑动窗口采取固定长度L,滑动窗口在长为4 800的信号序列随机滑动并截取窗口内的数据,考虑到不同L的取值会对识别结果造成影响,权衡计算资源后取$L=512$,即每个子帧数据结构为1×512×2。以ADS-B-0为例,具体滑动窗口切片示意如图1、2所示。

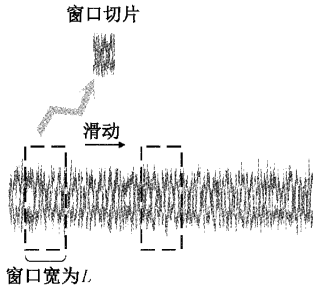

图1 数据样本切片

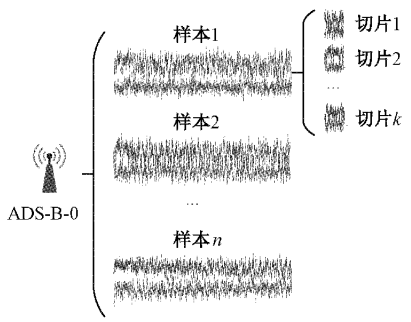

图2 ADS-B-0类数据切片

为了兼顾模型的训练速率并降低训练时数据特征冗余,训练时不需要将所有切片后的数据送入网络进行训练,综合考虑识别率和计算复杂度,每类每个样本随机切片45次,即数据总量扩充至45倍。

1.2 模型改进

测试对抗攻击的效果,选择识别率高的目标模型十分

重要，若模型识别率不理想，那么攻击的效果无法充分体现。本文模型选择 Vgg16 和 ResNet18，在网络中融合卷积注意力模块（convolutional block attention module，CBAM）[13]帮助网络提取信号特征，提高识别率。

1）CBAM 模块

CBAM 包括通道注意力模块和空间注意力模块。通道注意力模块如图 3 所示，通过平均池化层和最大池化层分别学习特征，并保持通道维度不变，压缩空间维度，学习到输入数据的有效信息，计算公式如下：

$$M_c(F)=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))=\sigma(W_1(W_0(F^c_{avg}))+W_1(W_0(F^c_{\max})))\tag{1}$$

其中，F^c_{avg} 和 $F^c_{\max}$ 分别为平均池化层和最大池化层学习到的特征，W_0 和 W_1 为全连接层降维和延展处理后的共享权重。$\sigma(\cdot)$ 为 sigmoid 激活函数。

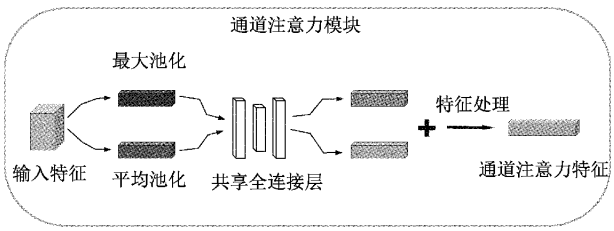

图 3　通道注意力模块

不同于通道注意力机制如图 4 所示，空间注意力机制学习过程中保持空间维度不变，压缩通道维度，从而学习到输入数据的位置信息，公式如下：

$$M_s(F)=\sigma(f^{7\times7}([AvgPool(F);MaxPool(F)]))=\sigma(f^{7\times7}(F^s_{avg};F^s_{\max}))\tag{2}$$

其中，F^s_{avg}，$F^s_{\max}$ 为平均池化层和最大池化层学习到的特征，$f^{7\times7}(\cdot)$ 为 7×7 卷积层，$\sigma(\cdot)$ 为 sigmoid 激活函数。

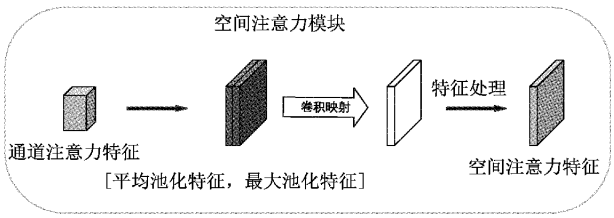

图 4　空间注意力模块

将两者进行串行结合，得到最终的 CBAM 模块。

2）改进后的目标模型和替代模型

针对白盒攻击，本文选用了 Vgg16 作为基础模型。为了使二维卷积层成功识别 512×2 的信号序列，修改了部分权重参数和网络层数，在网络每个卷积结构的 BN 层后，添加 CBAM 模块用于提高识别率。训练过程中通过不断调整输入数据的长度，宽度和高度来确保网络能成功提取信号的特征。图 5 为改进后 Vgg16 的基本卷积单元。

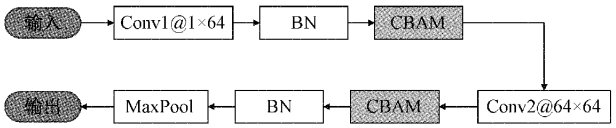

图 5　改进型 Vgg16 卷积结构

针对黑盒攻击，考虑到对抗样本的迁移特性，即生成的对抗样本不受各类 CNN 模型的限制，所以攻击方在无法获得目标模型的情况下，可以使用一个替代模型来模拟目标模型的决策边界。在黑盒的环境下，需要设计一个类似于目标模型的 CNN 替代模型来生成对抗样本从而将黑盒攻击转化为白盒攻击，文中为充分利用信号特征，直接将改进后的 ResNet18 作为黑盒攻击的替代模型。与改进型 Vgg16 类似，在残差模块 BN 层后添加 CBAM 模块，改进后残差块结构如图 6 所示。

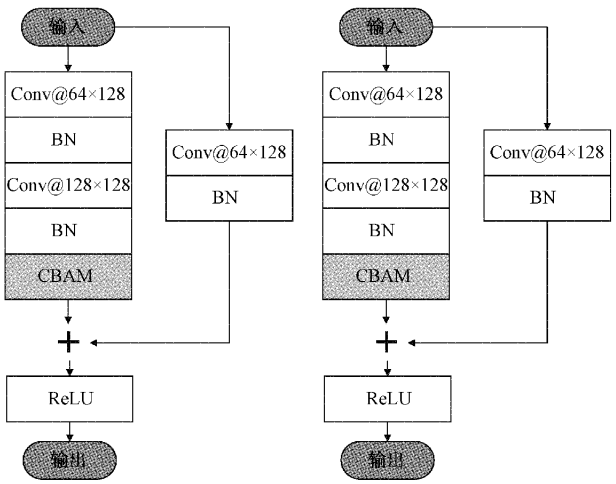

图 6　改进型 ResNet18 结构

2　对抗攻击算法

在给定一个训练好的目标模型和原始数据集的情况下，对抗攻击的思路可以看作是一个条件优化，X 和 X' 分别表示原始数据集和对抗样本，$l(\cdot)$ 表示目标模型，ε 表示加在原始数据集上的扰动大小，那么整体对抗样本的生成可以表示为：

$$\begin{aligned}&\varepsilon=X'-X\\&\min\|\varepsilon\|_p\\&\text{s.t. } l(X)\neq l(X')\end{aligned}\tag{3}$$

其核心是在成功欺骗模型的情况下使加入的扰动最小，以达到很难被察觉的目的，实验中，对抗攻击结构如图 7 所示。

2.1　FGSM 算法

Goodfellow 等[8]提出了一种快速生成对抗样本的方法，称为 FGSM 算法。该方法是典型的单步攻击的方法，它假设模型的损失函数与输入数据呈线性关系，所以通过计算模型损失函数 $J(\theta,x,l)$ 关于输入 x 的梯度方向，以求在梯度最陡的方向上增加扰动，其优点是能快速生成对

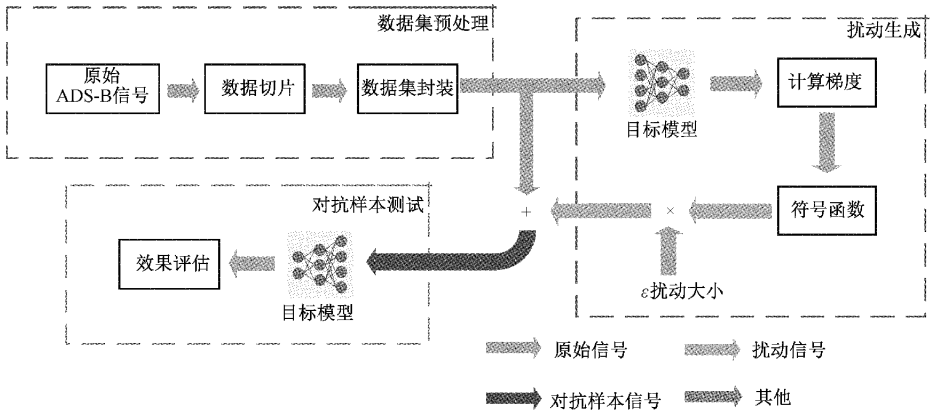

图 7 对抗攻击结构

抗样本,公式表示如下:

$$x' = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, l)) \tag{4}$$

其中,ε 为扰动大小,J 为模型损失函数,l 为分类模型,x' 为生成的对抗样本,x 为原始数据,$sign(\cdot)$ 为符号函数。

2.2 BIM 算法

FGSM 算法中假设目标模型损失函数与输入数据呈线性关系,但这种假设并不一定成立,导致加入的扰动较大但效果不明显。所以 Kurakin 等[9]在 FGSM 的基础上加入了迭代的思想,设计出 BIM 算法。该算法采用小步长,多次数的扰动不断更新对抗样本,公式表示如下:

$$\begin{cases} x_0 = x \\ x_{t+1} = Clip_{x,\varepsilon}\{x_t + \varepsilon \cdot sign(\nabla_x J(x_t, l))\} \end{cases} \tag{5}$$

其中,$Clip_{x,\varepsilon}\{Z\}$ 是将 Z 裁剪到 $[x-\varepsilon, x+\varepsilon]$ 的范围内。

2.3 投影梯度下降算法

投影梯度算法(projected gradient descent,PGD)[14]将投影的思想应用在梯度下降的方法之中,来确保公式的结果最终处于预先设定要的合理区间内。PGD 算法是一种能既能防御又能产生对抗样本的算法,其通过利用深度神经网络的鲁棒性和脆弱性,可以有效地结合攻击与防御,该算法的核心公式如下:

$$\min_{\theta} \rho(\theta)$$

$$\rho(\theta) = E_{(x,y)D}\left[\max_{\delta \in S} L(\theta, x+\delta, y)\right] \tag{6}$$

其中,S 是扰动允许的最大范围,$E_{(x,y)D}[L]$ 是定义的总体风险,D 是样本分布,该公式包含了内部最大化和外部最小化两个问题。内部最大化旨在找到使模型损失最大的数据扰动的攻击问题,外部最小化是找到使模型整体对抗损失最小的模型参数的防御问题。

2.4 动量迭代算法

BIM 算法的迭代攻击容易产生局部最优解从而出现过拟合的问题,所以,Dong 等[15]以神经网络优化器中的动量法为启发将动量的思想集成于 BIM 算法之上,提出动量迭代算法(momentum iterative method,MIM),该算法迭代的方向不仅与当前次有关,还与上一次迭代方向有关。加入动量因子之后,迭代更新的方向更加稳定,减少了陷入局部最优解的困境之中。算法描述如下:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J_\theta(x'_t, l)}{\| \nabla_x J_\theta(x'_t, l) \|_1} \tag{7}$$

将 x'_t 输入到目标函数 l,同时计算梯度 $\nabla_x J_\theta(x'_t, l)$,之后通过梯度是累计速度矢量来更新参数 g_{t+1},最后通过符号函数来更新 x'_{t+1}。

$$x'_{t+1} = x'_t + \varepsilon \cdot sign(\boldsymbol{g}_{t+1}) \tag{8}$$

其中,$\boldsymbol{g}_{t-1}$ 为每次梯度方向的速度向量,μ 为衰退值,其越大,前一次迭代的方向对当前次迭代方向的影响就越大。

3 实验结果与分析

3.1 实验基本设置

整体实验框架如图 7 所示,通过随机切片算法扩充信号样本之后封装成数据集,分别使用 4 种攻击算法生成对抗样本,在测试阶段攻击预先训练好的目标模型并比较攻击前后模型识别率的变化。

除了测试对抗样本的识别率,实验还引入了对抗样本的迁移率和黑盒泛化率来评估攻击效果。迁移率是指由模型 A 生成的对抗样本能成功欺骗模型 B 的样本个数与成功欺骗模型 A 的数量之比,算法迁移率越高,表示该算法生成的对抗样本更具有普适性。黑盒泛化率是指白盒模型 f_w 生成的对抗样本 x_{adv} 能成功欺骗黑盒模型 f_b 的样本数与总样本数之比,泛化率越高表示该算法生成的对抗样本在黑盒场景中攻击成功率越高。黑盒泛化率可表示为:

$$\frac{1}{|D_{ori}|} \sum_{(x_{adv}, y_{true})} (f_b(x_{adv}) \neq y_{true}) \tag{9}$$

其中,D_{ori} 为原始数据集,$D_{ori} = \{(x^1_{true}, y^1_{true}), \cdots, (x^N_{true}, y^N_{true})\}$。$x_{adv}$ 是由白盒攻击 f_w 生成的对抗样本。

3.2 目标模型识别率

实验采取改进后的 Vgg16 和 ResNet18 作为目标模型

和替代模型，并将两种网络模型最后一层的输出从 1 000 调至 20，以便与任务中的发射机类型数量相匹配。文中所有实验都是在 NVIDIA GeForce GTX 3060 上使用 GPU 进行，文中实验基于 Pytorch 框架，使用 Adam 优化算法来更新模型，分类交叉熵作为损失函数，ReLU 函数作为所有层的激活函数。经过 30 次迭代后模型逐渐收敛。实验保存识别率最高的一次训练权重并用来测试对抗样本。最终效果如图 8 和 9 所示。

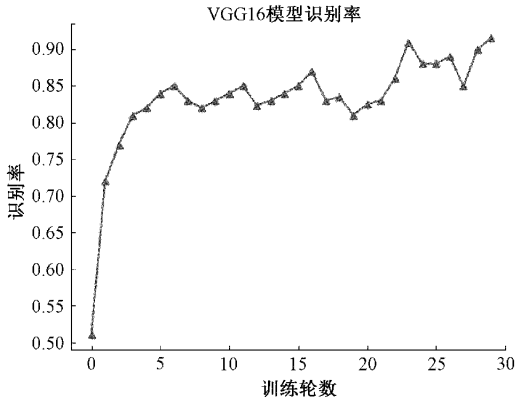

图 8　改进型 Vgg16 识别率

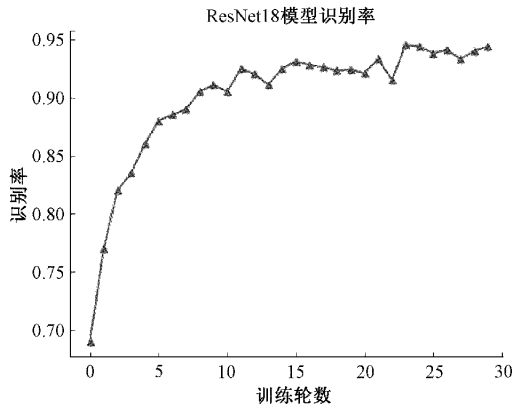

图 9　改进型 ResNet18 识别率

3.3　攻击效果分析

1)白盒攻击

为研究不同攻击对模型造成的影响，实验首先比较了 4 种攻击算法的白盒攻击，FGSM，BIM，PGD 和 MIM。图 10 显示了 4 种算法在以改进 VGG16 作为目标模型时的攻击效果，扰动大小设定的范围为 0～0.03。

如图 10 所示，当扰动 ε 设为 0 时，网络的识别率为 91%左右，随着扰动幅度不断增加，识别效果逐渐下降，当扰动大小为 0.03 时，模型整体识别率下降了 40%左右，可以看出模型对此类攻击比较敏感。随着扰动大小不断增加，迭代攻击的 PGD，MIM，BIM 算法的攻击效果都要优于单步攻击的 FGSM 算法。迭代算法中，在扰动大小小于 0.02 以前，PGD 算法和 BIM 算法效果基本一致，随着扰动增加，BIM 算法效果略微优于 PGD 算法，其中 MIM 算法

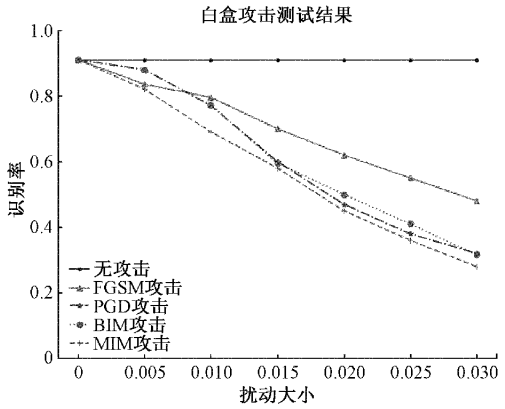

图 10　4 种算法白盒攻击效果

的攻击效果明显优于另外两种迭代算法，最终使模型识别率下降近 60%。所以在固定信噪比白盒攻击的情况下，选择迭代攻击的效果更好。

为了更近一步观察攻击的效果，实验中对 MIM 算法的攻击方式生成混淆矩阵，矩阵效果如图 11 和 12 所示。

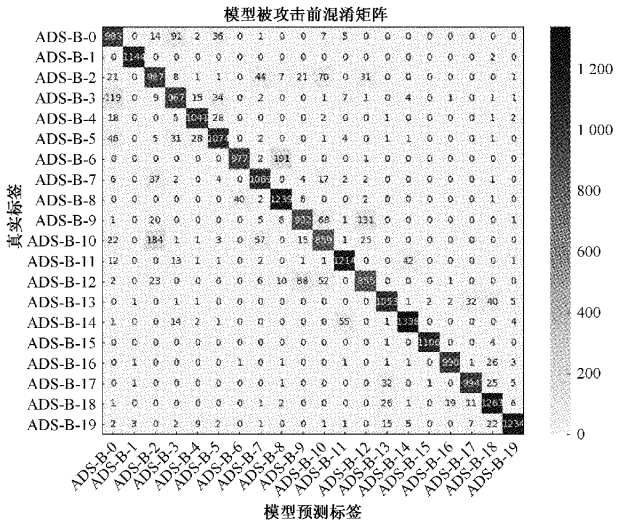

图 11　未添加扰动时混淆矩阵

可以看到，在没有扰动的情况下，ADS-B-6 和 ADS-B-8；ADS-B-10 和 ADS-B-2 相较于其他类有较为部分混淆。加入到扰动后的效果如图 12(b)所示，从矩阵中可以看出，当扰动大小为 0.05 的时候，ADS-B-2 的识别率依然很高，基本没有出现混淆，而 ADS-B-6 和 ADS-B-8 出现了十分严重的混淆，几乎所有的 ADS-B-6 信号被识别为 ADS-B-8。扰动幅度增加后，其他类也出现了较为严重的混淆。从实验中可以看出，不同设备的信号，由于辐射源信号个体特征不同，导致其抗干扰能力存在差别。

迁移率也是评判一个攻击算法好坏的重要指标，即针对目标模型 A 生成的对抗样本对目标模型 B 攻击的效果称之为迁移率，实验中，目标模型 A 使用的是改进后的 Vgg16，目标模型 B 使用的是改进后的 ResNet18。4 种方法的迁移率由图 13 所示。

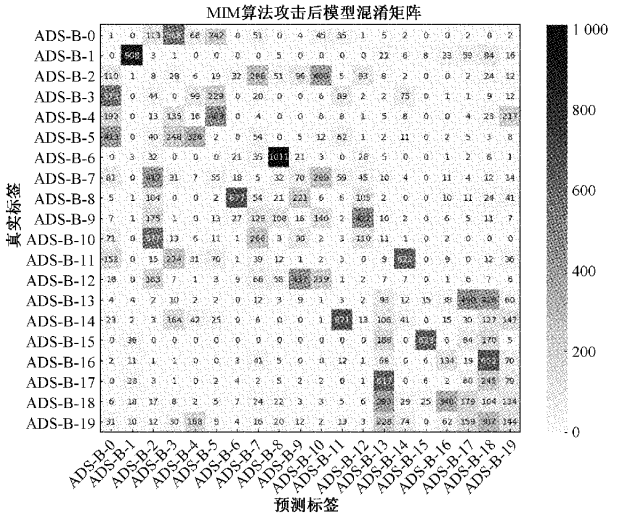

图 12 MIM 算法扰动大小为 0.05 时混淆矩阵

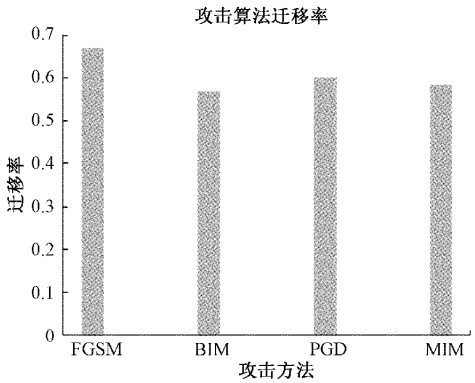

图 13 4 种算法的迁移率

根据图 13 可以看出,在 4 种算法之中,FGSM 算法的迁移率是最好的,其余 3 种方法的迁移率相差不大,即单步攻击算法虽然攻击效果不如迭代攻击算法,但是其迁移率却是优于迭代算法的。经过分析,其原因如下:由于迭代算法是需要不断访问目标模型的信息并更新对抗样本,所生成的对抗样本是更加针对当前模型的决策边界的,而单步攻击只访问一次,所以生成的对抗样本更具有普适性。所以,需要根据特定的情况选择合适的攻击算法,并且之后的新算法也需要在迁移率和攻击效果寻找最优解。

2)黑盒攻击

在实际环境中,由于目标模型的内部信息基本是不可获取的,所以黑盒环境的攻击更接近实际环境。目标模型采用改进后的 Vgg16。黑盒攻击效果如图 14 所示。

在黑盒攻击中,由于对模型信息掌握较少,所以并不能像白盒攻击那样具有针对性,只能采取替代模型进行攻击[16],所以使用较高的扰动才能达到相对比较明显的效果,文中替代模型使用的是改进型 ResNet18。实验中可以看到,当扰动较小的时候,4 种算法的攻击效果近乎一样,但是随着扰动幅度逐渐增加,模型的识别率也不断下降,

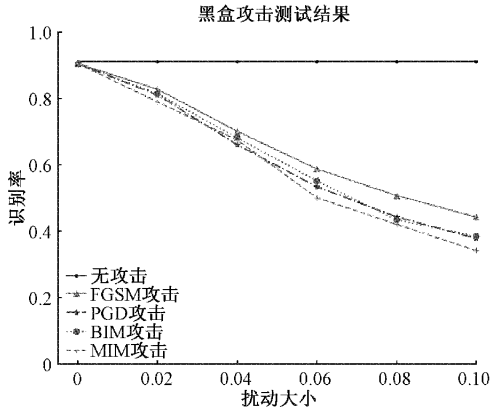

图 14 4 种算法的黑盒攻击效果

迭代攻击的效果也逐渐优于单步攻击的效果。迭代攻击中,MIM 算法的攻击效果比 BIM 算法和 PGD 算法更为突出,这主要是由于 BIM 算法和 PGD 算法的迭代性质需要不断访问目标模型信息,而黑盒攻击目标模型信息不可获取导致的,但是 MIM 算法引入动量因子,这不仅确保更新方向的稳定性,在保持攻击效果的同时兼顾迁移特性。图 15 为 4 种算法的黑盒泛化率。

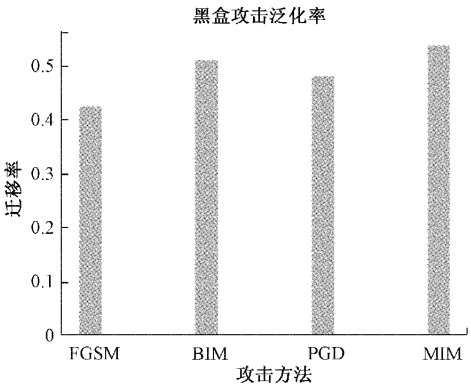

图 15 4 种算法的黑盒泛化率

可以看到 4 种算法中,FGSM 算法攻击较其他 3 种迭代算法泛化率较差,而在迭代算法中,MIM 算法的泛化率最优,说明 MIM 算法在黑盒环境中攻击的成功率最高,效果最好。

从不同算法攻击效果实验结果来看,迭代算法的攻击效果优于单步攻击算法,但是迭代算法的迁移率不如单步攻击算法;从黑盒白盒攻击效果来看,黑盒攻击虽然取得了一定的效果,但整体效果不如白盒攻击,并未达到预期效果,所以找到高效的黑盒攻击方法依然有着很大的难度和挑战。

3.4 攻击前后数据可视化

之前的实验说明了随着扰动幅度增加,对抗样本的攻击效果也越来越好,但同时增加了被检测到的风险;而扰动程度越低,对抗样本的伪装性能越高,被检测到的可能性越小,但可能不足以达到理想的攻击效果。很明显,在

实际情况中，需要在算法的攻击性和隐蔽性之间取得平衡。

文中将原始信号数据和攻击后的信号做对比，来验证添加的扰动是否足够小。图 16、17 显示了使用扰动大小为 0.03 和 0.05 时 FGSM 算法攻击前后信号数据的对比，“▲”形线为原始信号，“＋”形线为添加扰动后的对抗样本。

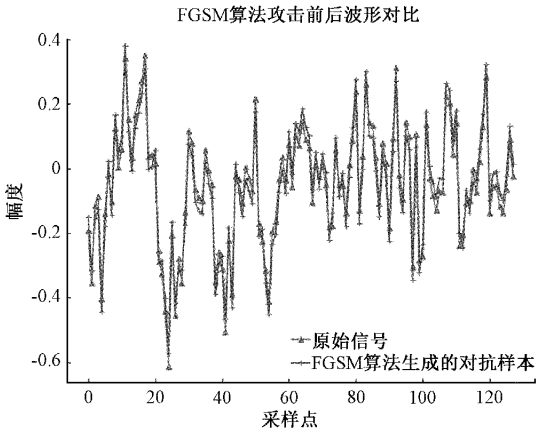

图 16　FGSM 算法扰动大小为 0.03 前后对比

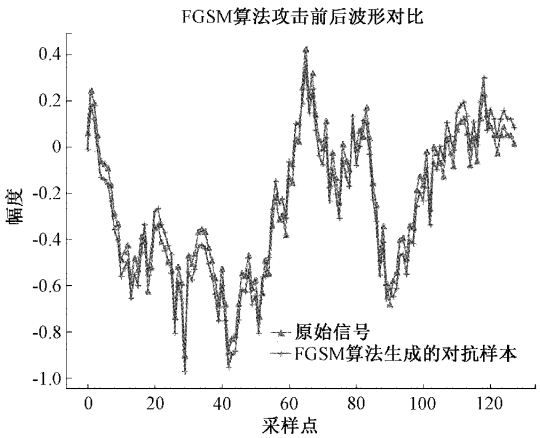

图 17　FGSM 算法扰动大小为 0.05 前后对比

从这两张图片中观察到，当扰动较小的时候，在保证误导模型的同时，扰动前后的信号波形依然近似，即幅度相位等特征未发生明显的变化，有较好的隐蔽性。当扰动增大的时候，虽然模型识别率明显下降，但是扰动前后的信号波形幅度上有着较为明显的差异。总的来说，在较低的扰动大小的情况下，以 MIM 算法为例所生成的对抗样本，既可以使模型发生判决错误，也能保证扰动后信号的波形具有较小的感知差异，表现出较为优秀的隐蔽性。

4　结　　论

在本实验中的 4 种攻击方法都对高识别精度的模型造成较为严重的干扰，其中 MIM 算法无论在白盒环境下还是黑盒环境下攻击效果都是最好的，较其他算法攻击效果高 4%～10%。FGSM 算法虽然攻击效果一般，但是作为单步攻击法普适性更好，具有更好的迁移性，迁移率较其他 3 种方法高出 10%左右。实验最后将攻击前后数据可视化来对比攻击前后效果。结论表明，对抗样本不仅对图片分类领域产生严重影响，还会对 ADS-B 系统的辐射信号识别造成严重的安全隐患，白盒攻击扰动大小为 0.03 时，模型识别率下降 60%左右；黑盒攻击扰动大小为 0.1 时，模型识别率下降 50%左右。除此之外，生成对抗样本时需要在攻击效果和隐蔽性之间权衡才能达到最佳攻击效果。本文实验在理想情况下对接收端输入信号进行干扰。考虑到实际通信环境敌方接收信号不可获取。下步工作将基于发送端进行对抗样本生成并考虑对抗样本在干扰敌方识别的前提下，降低对己方接收端的影响。

参考文献

[1] DANEV B, ZANETTI D, CAPKUN S. On physical-layer identification of wireless devices[J]. Acm Computing Surveys, 2012, 45(1):1-29.

[2] 田金鹏，刘燕平，刘小娟. 基于瞬态强度的射频指纹识别方法[J]. 电子测量技术，2016,39(4):58-61,65.

[3] O'SHEA T J, WEST N. Radio machine learning dataset generation with gnu radio[C]. Proceedings of the GNU Radio Conference,2016.

[4] 耿梦婕，张君毅. 基于神经网络的辐射源个体识别技术[J]. 电子测量技术，2019，42(21):137-142.

[5] 陈小惠，彭杰，薛毓楠. 基于复杂度的通信辐射源目标识别方法[J]. 国外电子测量技术，2021,40(5):22-26.

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

[7] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), IEEE, 2017.

[8] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. ArXiv Preprint,2014, ArXiv:1412.6572.

[9] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]. Boca Raton: Artificial Intelligence Safety and Security, 2018.

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

[11] LIN Y, ZHAO H, TU Y, et al. Threats of adversarial attacks in DNN-based modulation recognition [C]. IEEE INFOCOM 2020-IEEE

Conference on Computer Communications, IEEE, 2020: 2469-2478.

[12] 王超, 魏祥麟, 田青, 等. 基于特征梯度的调制识别深度网络对抗攻击方法[J]. 算机科学, 2021, 48(7): 25-32.

[13] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[14] MADRT A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. ArXiv Preprint, 2017, ArXiv:1706.06083.

[15] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9185-9193.

[16] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017: 506-519.

作者简介

刘丰汇,硕士研究生,主要研究方向为通信辐射源个体识别、对抗攻击、深度学习。

E-mail:fenghuiliu1998@163.com

张治中,博士,教授,主要研究方向为移动大数据、物联网、通信网测试及仪表技术等。

E-mail:zhangzz@nuist.edu.cn

张涛(通信作者),博士,副研究员,主要研究方向为物理层安全、无线传感器网络、机器学习。

E-mail: ztcool@126.com

杨小蒙,硕士研究生,主要研究方向为信号调制识别、深度学习等。

E-mail:1402420186@qq.com