计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300299

基于特征梯度的调制识别深度网络对抗攻击方法

王 超[1] 魏祥麟[2] 田 青[1] 焦 翔[1] 魏 楠[1] 段 强[2]
1 南京信息工程大学计算机与软件学院 南京 210044
2 国防科技大学第六十三研究所 南京 210007
(wangchao2020@nuist.edu.cn)

摘 要 基于深度神经网络(Deep Neural Network,DNN)的自动调制识别(Automatic Modulation Recognition,AMR)模型具有特征自提取、识别精度高、人工干预少的优势。但是,业界在设计面向AMR的DNN(AMR-oriented DNN,ADNN)模型时,往往仅关注识别精度,而忽视了对抗样本可能带来的安全威胁。为此,文中从人工智能安全的角度出发,探究了对抗样本对ADNN模型的安全威胁,并提出了一种新颖的基于特征梯度的对抗攻击方法。相比传统标签梯度的攻击方式,特征梯度攻击方法能够更有效地攻击ADNN提取的调制信号空时特征,且具有更好的迁移性。在公开数据集上的实验结果表明,无论白盒攻击还是黑盒攻击,所提出的基于特征梯度的对抗攻击方法的攻击效果和迁移性均优于当前的标签梯度攻击方法。

关键词:自动调制识别;调制信号;深度学习;神经网络;对抗样本
中图法分类号 TP183

Feature Gradient-based Adversarial Attack on Modulation Recognition-oriented Deep Neural Networks

WANG Chao[1],WEI Xiang-lin[2],TIAN Qing[1],JIAO Xiang[1],WEI Nan[1] and DUAN Qiang[2]
1 School of Computer and Software,Nanjing University of Information Science and Technology,Nanjing 210044,China
2 The 63rd Research Institute,National University of Defense Technology,Nanjing 210007,China

Abstract Deep neural network (DNN)-based automatic modulation recognition (AMR) outperforms traditional AMR methods in automatic feature extraction,recognition accuracy with less manual intervention. However,high recognition accuracy is the first priority of the practitioners when designing AMR-oriented DNN (ADNN) models while security is usually neglected. In this backdrop,from the perspective of the security of artificial intelligence,this paper presents a novel characteristic gradient-based adversarial attack method on ADNN models. Compared with traditional label gradient-based attack method,the proposed method can better attack the extracted temporal and spatial features by ADNN models. Experimental results on an open dataset show that the proposed method outperforms label gradient-based method in the attacking success ratio and transferability in both white-box and black-box attacks.

Keywords Automatic modulation recognition,Modulation signal,Deep learning,Neural Networks,Adversarial examples

1 引言

自动调制识别作为信号检测和信号解调之间的重要步骤,是解决日益拥挤和复杂的电磁空间的重要手段,也是缓解频谱资源短缺的重要前提。传统基于最大似然估计和统计模式识别的AMR方法依赖手工特征提取,其精确性较低,且严重依赖关于信号的先验知识。为此,学术界近年来尝试将深度神经网络(DNN)应用于自动调制识别。典型的方法包括卷积神经网络(Convolutional Neural Network,CNN)和图神经网络(Graph Convolutional Network,GCN)等,均取得了较好的识别精度[1-2]。

然而,深度学习在可解释性方面的缺陷使得面向AMR的DNN(ADNN)模型易受到对抗样本的威胁。研究发现,通过添加一个精心设计的微小扰动,就可以很容易地欺骗一个性能良好的深度网络模型,使模型网络以高置信度改变预测和输出。对于ADNN来说,对抗方可以通过在样本中添加微小的扰动,使得调制识别结果发生偏差,进而显著降低电磁空间感知的正确性。但是,当前ADNN的设计重点在于识别精度,对其安全性缺乏评估考量。

为此,本文针对ADNN模型的安全性问题,提出了一种基于特征梯度的ADNN对抗攻击方法,开展了如下工作:

(1)训练了3种不同架构的ADNN模型,均取得了较高

到稿日期:2021-03-30 返修日期:2021-05-06
基金项目:国家自然科学基金(61702273);江苏省自然科学基金(BK20170956)
This work was supported by the National Natural Science Foundation of China(61702273) and Natural Science Foundation of Jiangsu Province (BK20170956).
通信作者:魏祥麟(wei_xianglin@163.com)

的识别准确率。

(2)提出了一种新的基于特征梯度的 ADNN 对抗攻击方法,依赖 ADNN 提取的特征进行对抗样本生成,并设计了两种攻击算法。

(3)在公开数据集上开展了一系列对比实验。结果表明,相比基于标签梯度的攻击方法,本文所提方法攻击成功率更高,且迁移性更佳。

本文第 2 节介绍了深度学习在调制信号识别中的应用和对抗样本原理;第 3 节描述了基于特征梯度的对抗攻击方法;第 4 节分别从白盒攻击和黑盒攻击的角度,开展了一系列实验,并分析了实验结果;最后总结全文并展望未来。

2 相关工作

本节介绍了 ADNN 设计和对抗攻击的相关工作,主要包括 ADNN 设计、主流的对抗攻击方法以及对抗攻击在调制信号识别领域的应用等。

2.1 ADNN 模型

随着通信系统的日益复杂多样,无线信号数据比以往更复杂、更多样,具备了更强的随机性和异构性。基于以下 3 点考虑,业界开始将 DNN 应用到信号识别领域:1)DNN 模型需要大量的训练数据,而无线通信信号的海量特性很好地满足了 DNN 的数据需求;2)DNN 相比较于传统方法,可以自动提取调制信号特征,省去了手动选择特征可能引入的错误以及识别过程中对专家知识的依赖;3)DNN 可以比传统方法取得更加精准的识别效果。

Kato 等提出了一种针对异构网络流量的监督 DNN 模型[3]。O'shea 等证明了 DNN 模型能够被应用到模拟无线电的时间序列数据上进行分类,且准确率优于传统方法[4]。Wang 等提出了通过组合两种 CNN 模型来识别不同的调制模式[5]。Rajendran 等将长短期记忆网络(Long Short-Term Memory,LSTM)应用到调制识别中[6]。Tang 等提出了一种基于生成式对抗网络(GAN)的通信信号的调制分类算法,用于弥补数据不足对网络训练的影响[7]。Chen 等将自适应奇异值重构方法和残差网络运用于雷达信号的调制识别[8]。

2.2 对抗攻击

DNN 的不可解释性使其面临多种安全风险。Szegedy 等发现在输入样本中添加一些精心制作的人类难以察觉的微小扰动,能够显著降低 DNN 分类器的识别精度。这种添加过扰动的样本称为对抗样本(adversarial example)[9]。

根据对抗样本是否有目标可以将对抗攻击分为两类:目标攻击和无目标攻击。目标攻击就是指对抗样本除了能够欺骗模型,还需要将输入样本错误地分类到特定的类别。例如:在调制信号识别中,攻击者指定目标类别为 GFSK,则 8PSK,QPSK 或者是其他类别的信号在被攻击后都将被错误地分类为 GFSK。而无目标攻击则与有目标攻击相反,无目标攻击无须指定具体的攻击信号类别,即目标可以是除自身信号外的任意类型的信号。

根据攻击者对目标模型信息了解的多少也可以将对抗攻击分为白盒攻击和黑盒攻击。白盒攻击是假设攻击者对目标神经网络模型有充分的认识,包括输入样本、模型结构、模型参数和训练方法等,在这种情况下执行的攻击就是白盒攻击;黑盒攻击仅能在获取目标模型的输出置信度的情况下执行攻击,这增加了攻击的难度和复杂性。

当前,对抗样本研究大多集中于图像识别领域。Goodfellow 等提出了快速梯度符号法(Fast Gradient Sign Method,FGSM)来攻击深度模型,其思想是通过计算损失函数相对于输入样本本身的梯度来得到对抗样本[10]。Kurakin 等提出了迭代的 FGSM(Iterative Fast Gradient Sign Method,I-FGSM),采用迭代多次的方式产生对抗样本[11]。Dong 等将动量(momentum)引入迭代攻击中的梯度计算过程中,提出了 MI-FGSM,以提高模型每次迭代的稳定性和对抗样本的泛化性[12]。Moosavi-dezfooli 等提出了一种名为 DeepFool 的算法,该算法将深度分类模型替代为线性模型进行攻击[13]。Lin 等将 Nesterov 加速梯度引入到迭代攻击过程中,提出了 NI-FGSM,增加对抗样本的可迁移性[14]。Moosavi-dezfooli 等发现存在一种通用扰动,可以影响所有图像的模型分类结果[15]。Kurakin 等提出了在模型上执行对抗训练的方法来探索对抗样本对模型鲁棒性的影响[16]。Carlini 等提出了 3 种产生扰动的方式,利用 3 种不同的度量方式(L_1,L_2,L_∞)来避免模型的鲁棒性[17]。Athalye 等证明了即使是现实世界中的物体也可以用来欺骗分类模型[18]。

2.3 ADNN 模型对抗现状

目前,鲜有工作将对抗攻击应用到 ADNN。Lin 等将基于标签计算梯度的传统对抗方法应用到了调制信号识别当中,验证了 ADNN 容易受到对抗样本的攻击[19],Zhao 等将 Nesterov Adam 迭代法应用在调制识别中并降低了信号对抗样本的感知可见性[20]。但上述方法在执行黑盒攻击时仍采用替代模型的方法,并没有充分利用调制信号数据样本的特征。并且,其选择的调制信号识别的目标模型本身的识别准确率较低,仅为 70%左右。

3 基于特征梯度的 ADNN 对抗攻击方法

3.1 基本思路

传统标签梯度攻击方法使用样本的标签 y 计算梯度,其结构如图 1 所示。将原始信号样本 x_S 和对应的标签 y 输入目标模型的损失函数中,通过计算梯度和符号函数得到攻击方向,再乘以扰动大小 α 就可以得到对抗扰动,最后在原始样本基础上添加对抗扰动即可得到对抗样本 x^*。

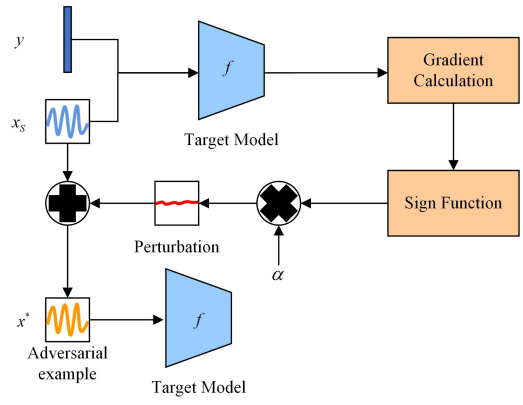

图 1 标签梯度攻击

Fig. 1 Label gradient-based adversarial attack

目前,基于标签梯度的攻击的方法主要有 FGSM,

I-FGSM，MI-FGSM 和 NI-FGSM（见 2.2 节），其中 I-FGSM，MI-FGSM 和 NI-FGSM 都是 FGSM 的变种。

针对 ADNN 在开展对抗攻击时，期望在信号中添加一个人类难以察觉的微小扰动，导致 ADNN 识别错误。假设原始信号样本为 x，所识别的类别结果为 y，而扰动为 η，其中 η 需要足够小，以满足 $\|\eta\|_\infty \leqslant \varepsilon$，则 FGSM 可以表达为：

$$\begin{cases}\eta=\varepsilon \cdot \operatorname{sign}(\nabla_x J(x,y)) \\ x^* = x+\eta\end{cases} \tag{1}$$

其中，J 为目标模型的损失函数；$\nabla_x J(x,y)$ 表示损失函数对样本 x 求导。

由于 FGSM 是单步式攻击，因此不能通过多次查询模型参数来更新对抗样本。I-FGSM 是 FGSM 的扩展，通过多次迭代和访问模型来更新对抗样本，但这要以增加时间和更复杂的计算为代价。I-FGSM 中的对抗样本每次迭代的扰动 α 较小，为 $\alpha=\varepsilon/N$，并在每次迭代后被截断，以确保它们在原始输入的 ε 附近，可以表述为：

$$\begin{cases}x_0^* = x \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \alpha \cdot \operatorname{sign}(\nabla_x J(x_n^*, y))\}\end{cases} \tag{2}$$

其中，$Clip_{x,\varepsilon}\{\ \}$ 表示将 x 限制在 $[x-\varepsilon, x+\varepsilon]$ 范围内。

MI-FGSM 在迭代中通过在损失函数的梯度方向上累积速度向量，加速梯度下降。MI-FGSM 引入动量并将其加入迭代攻击，来保证模型每个更新方向的稳定性，其可以表述为：

$$\begin{cases}x_0^* = x, g_0 = 0 \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^*, y)}{\|\nabla_{x_n^*} J(x_n^*, y)\|_1} \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \alpha \cdot \operatorname{sign}(g_{n+1})\}\end{cases} \tag{3}$$

其中，g_{n+1} 表示前 $n+1$ 次迭代产生的累加梯度，μ 为 g_n 的衰减因子。

NI-FGSM 与 MI-FGSM 类似，但将 Nesterov 加速梯度应用到了迭代攻击过程中，增强了对抗样本的迁移性能，可以表述为：

$$\begin{cases}x_0^* = x, g_0 = 0 \\ x_n^{\text{nes}} = x_n^* + \alpha \cdot \mu \cdot g_n \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^{\text{nes}}, y)}{\|\nabla_{x_n^*} J(x_n^{\text{nes}}, y)\|_1} \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \alpha \cdot \operatorname{sign}(g_{n+1})\}\end{cases} \tag{4}$$

其中，x_n^{nes} 为 Nesterov 项，参与梯度的计算。

相比图片等高维数据，频谱信号样本的信息量较小，如果采用 ADNN 模型提取的中间层特征作为攻击对象，可以提高攻击的精度，产生更加精细的调制信号对抗样本。针对同一类调制信号样本，不同的 ADNN 模型经过训练后，其中间层的输出特征往往具有相似性，而且样本的特征是可迁移的。因此，基于特征执行对抗攻击产生的扰动应该具有更好的可迁移性。

基于以上考虑，本文利用信号样本在模型中的特征计算梯度，用于攻击 ADNN 模型，以期取得更好的攻击成功率和可迁移性。当前，将特征层面的对抗攻击方法应用于 ADNN 模型的研究内容尚未见报道。

图 2 给出了本文设计的方法的整体流程。从图 2 可以看出，与标签梯度的攻击方式不同，所提方法在输入部分的输入分为了原始信号样本 x_S 和目标信号样本 x_T，此外，还需要将目标模型从 L 层进行截断，得到截断模型 f_L，将 x_S 和 x_T 分别输入 f_L 就得到了原始信号样本特征 $f_L(x_S)$ 和目标信号样本特征 $f_L(x_T)$，然后进行基于特征的梯度计算。

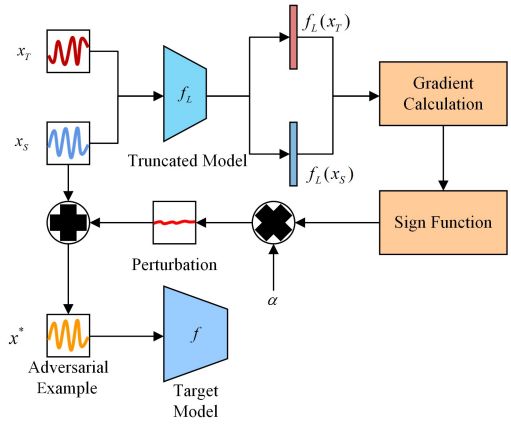

图 2　特征梯度攻击

Fig. 2　Feature gradient-base adversarial attack

3.2　攻击方法描述

为了在 ADNN 模型的特征空间上进行攻击，首先需要找到合适的特征空间。为了保证所选取的特征空间的信息量足够丰富，本文统一选择目标模型的截断层为模型最后的全连接层之前的激活层。因此，该攻击方法也可以称为激活攻击（Activation Attack，AA）。为了提高攻击的效率和增加对比实验，将 AA 算法分别与 MI-FGSM 和 NI-FGSM 算法相结合，组成了 AA-MI-FGSM 以及 AA-NI-FGSM 算法。对于 AA-MI-FGSM，其攻击过程为：

$$J_{AA}(x_T, x_S) = \|f_L(x_T) - f_L(x_S)\|_2 \tag{5}$$

$$\begin{cases}x_0^* = x_S, g_0 = 0 \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J_{AA}(x_T, x_n^*)}{\|\nabla_{x_n^*} J_{AA}(x_T, x_n^*)\|_1} \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \alpha \cdot \operatorname{sign}(g_{n+1})\}\end{cases} \tag{6}$$

其中，$\|\cdot\|_2$ 为 L2 范数，表示对抗样本特征和原始样本特征的相似度度量。

AA-MI-FGSM 方法的工作流程如算法 1 所示。

算法 1　AA-MI-FGSM

输入：原始信号样本 x_S；目标信号样本 x_T；截断模型 f_L；损失函数 J_{AA}；g_n 衰减因子 μ；扰动大小 ε 和迭代次数 N

输出：满足 $\|x^* - x_S\|_2 \leqslant \varepsilon$ 的对抗样本 x^*

1. $\alpha = \varepsilon/N$
2. $g_0 = 0$，$x_0^* = x_S$
3. x_T 输入 f_L，得到特征 $f_L(x_T)$
4. for n=0 to N−1 do
5.　　x_n^* 输入 f_L，得到特征 $f_L(x_n^*)$
6. $J_{AA}(x_T, x_n^*) = \|f_L(x_T) - f_L(x_n^*)\|_2$
7.　　计算 $\nabla_{x_n^*} J_{AA}(x_T, x_n^*)$
8.　　计算累加梯度，更新 g_{n+1}：

$$g_{n+1} = \mu \cdot g_n + \frac{\nabla_{x_n^*} J_{AA}(x_T, x_n^*)}{\|\nabla_{x_n^*} J_{AA}(x_T, x_n^*)\|_1}$$

9. 应用梯度方法，更新 x_{n+1}^{*}：

$$x_{n+1}^{*}=\text{Clip}_{x,\varepsilon}\{x_n^{*}+\alpha\cdot\text{sign}(g_{n+1})\}$$

10. end for

11. return $x^{*}=x_N^{*}$

AA-NI-FGSM 算法与 AA-MI-FGSM 算法类似，但在计算梯度前需要计算一个 Nesterov 项 x^{nes}，其工作流程如算法 2 所示。

算法 2 AA-NI-FGSM

输入：原始信号样本 x_S；目标信号样本 x_T；截断模型 f_L；损失函数 J_{AA}；g_n 衰减因子 μ；扰动大小 ε 和迭代次数 N

输出：满足 $\|x^{*}-x_S\|_2\leqslant\varepsilon$ 的对抗样本 x^{*}

1. $\alpha=\varepsilon/N$

2. $g_0=0$，$x_0^{*}=x_S$

3. x_T 输入 f_L，得到特征 $f_L(x_T)$

4. for n=0 to N−1 do

5. 计算 $x_n^{nes}=x_n^{*}+\alpha\cdot\mu\cdot g_n$

6. x_n^{nes} 输入 f_L，得到特征 $f_L(x_n^{nes})$

7. $J_{AA}(x_n^{nes},x_T)=\|f_L(x_n^{nes})-f_L(x_T)\|_2$

8. 计算 $\nabla_{x_n^{*}}J_{AA}(x_n^{nes},x_T)$

9. 计算累加梯度，更新 g_{n+1}：

$$g_{n+1}=\mu\cdot g_n+\frac{\nabla_{x_n^{*}}J(x_n^{nes},x_T)}{\|\nabla_{x_n^{*}}J(x_n^{nes},x_T)\|_1}$$

10. 应用梯度方法，更新 x_{n+1}^{*}：

$$x_{n+1}^{*}=\text{Clip}_{x,\varepsilon}\{x_n^{*}+\alpha\cdot\text{sign}(g_{n+1})\}$$

11. end for

12. return $x^{*}=x_N^{*}$

3.3 攻击方法分析

本文采取特征梯度攻击方式主要有以下 3 点考虑：

(1)深度特征空间表达的变化往往对网络模型的分类结果有很大的影响。本文的特征梯度扰动没有采用分类损失作为优化目标，而是利用分类网络截断模型中的特征空间，这是因为特征空间向量的维度一般较大，能够获取的信息也更丰富，对其执行攻击方法在理论上能够产生更细微的扰动，也能够满足调制信号对扰动大小的敏感性。

(2)深度模型的中间层特征已经被证明是可迁移的，因此可以推断在特征空间层面攻击产生的对抗样本也是可以迁移的。当前，深度模型的可解释性较弱，难以准确测量和描述网络模型在某一层所捕捉和学习到的特征。但是，可以认为不同的深度模型在输出分类结果前学习得到的特征空间是相似的，那么针对一个模型的抽象样本特征空间产生的扰动同样也会扰动另一个模型中获得的类似样本特征。

(3)可以假设两个不同的模型在相同分布的数据集上进行训练后具有相近的决策边界和类别方向。这与目标攻击方法息息相关，因为要使黑盒目标攻击成功，两种模型的决策边界与类别方向通常需要足够接近甚至相同。也就是说，如果在白盒模型中对抗样本的特征空间使分类器朝着某个目标样本方向移动，那么黑盒模型中的移动方向是相同或者相似的。

基于上述 3 点，我们认为本文所提出的基于特征梯度的对抗攻击算法相较于标签梯度的方法具有更好的攻击效果和迁移性。

4 实验和结果分析

4.1 实验设置

4.1.1 数据集

本文采用的数据集是 RADIOML 2016.10A，这也是调制识别领域的权威公开数据集[21]。该数据集是由不同 SNR 条件下的 11 种调制信号组成，每种信噪比下的数据都包含了 8 种数字信号（8PSK，QPSK，BPSK，GFSK，CPFSK，PAM4，QAM16，QAM64）和 3 种模拟调制信号（宽带调频（WBFM）、双边带调幅（AM-DSB）以及单边带调幅（AM-SSB））。

该数据集总共包含 220 000 个数据样本和 20 种信噪比（从 18 dB 到−20 dB，步长为 2 dB）。每个样本的维度为（128，2），包含 I 路和 Q 路两维数据，每一维度数据的长度都是 128。为了方便训练和验证模型，对该数据集随机划分，将 70%作为训练集，剩下的 30%作为测试集。在对抗攻击实验中，仅针对测试集中的数据进行对抗攻击处理。

4.1.2 ADNN 模型

对抗攻击的目标模型选择十分重要，如果一个 ADNN 模型的识别效果较差，那么攻击的效果就无法充分体现。考虑到频谱信号和图像具有不同的特征和参数，为了使调制数据集能适应和匹配卷积神经网络模型，针对白盒攻击，本文采用了 VGG 模型[22]，并将其网络结构针对上述数据集上进行了优化和调整。VGG 模型具体的网络结构如表 1 所列。

表 1 VGG 模型

Table 1 VGG model

Layers	Output Shape
Conv1D	(128,64)
Maxpooling1D	(64,64)
Conv1D	(64,64)
Maxpooling1D	(32,64)
Conv1D	(32,64)
Maxpooling1D	(16,64)
Conv1D	(16,64)
Maxpooling1D	(8,64)
Conv1D	(8,64)
Maxpooling1D	(4,64)
Conv1D	(4,64)
Maxpooling1D	(2,64)
Conv1D	(2,64)
Maxpooling1D	(1,64)
Flatten	(64)
Dense	(128)
Dropout	(128)
Dense	(128)
Dropout	(128)
Dense	(128)
Dropout	(128)
Dense	(11)

针对黑盒攻击，本文设计了两个 ADNN 模型。为了验证对抗攻击的可迁移效果，黑盒攻击的两个 ADNN 模型分别是 LSTM 和 ResNet 模型[23]。LSTM 相较于 VGG 模型，对调制信号序列数据具有更好的识别效果；ResNet 模型则是较为复杂的卷积神经网络模型，相对于 VGG 模型，具有更好的特征识别能力和分类效果。由于篇幅限制，具体的模型参数此处不再赘述。

4.1.3　评估指标

为了评估本文攻击方法的迁移性和泛化能力，针对生成的对抗样本定义如下评估指标：泛化率和迁移率。值得注意的是，在评估对抗样本时假设所有的信号样本都被白盒模型 f_w 和黑盒模型 f_b 正确分类。原始的数据集为 $D_{\text{orig}}=\{(x^{(1)},y_{\text{true}}^{(1)}),\cdots,(x^{(N)},y_{\text{true}}^{(N)})\}$；每种攻击方式，都会产生一个对抗数据集 $D_{\text{adv}}=\{(x_{\text{adv}}^{(1)},y_{\text{target}}^{(1)},y_{\text{true}}^{(1)}),\cdots,(x_{\text{adv}}^{(N)},y_{\text{target}}^{(N)},y_{\text{true}}^{(N)})\}$，其中数据 x_{adv} 和 y_{target} 由原始数据集 D_{orig} 在白盒模型 f_b 上执行目标攻击得到。

对抗样本的泛化率是指根据白盒模型 f_w 生成的对抗样本 x_{adv} 能使得黑盒模型 f_b 分类错误的数量与样本总数的比例，即：

$$\frac{1}{|D_{\text{orig}}|}\sum_{(x_{\text{adv}},y_{\text{true}})\in D_{\text{orig}}}1[(f_b(x_{\text{adv}}))\neq y_{\text{true}}] \tag{7}$$

泛化率越高，对抗样本在黑盒场景下的攻击成功率越高。

对抗样本的迁移率是指对抗数据集 D_{adv} 中既能欺骗白盒模型 f_w 的样本同时也能欺骗黑盒模型 f_b 的样本数量与成功欺骗白盒模型 f_w 的数量的比例。定义成功欺骗白盒模型的数据集为 $D_{f_b}\subseteq D_{\text{adv}}$。那么迁移率可以定义为：

$$\frac{1}{|D_{f_b}|}\sum_{(x_{\text{adv}},y_{\text{true}})\in D_{f_b}}1[(f_b(x_{\text{adv}}))\neq y_{\text{true}}] \tag{8}$$

这种评估方式直观地表示了在白盒攻击中产生的对抗样本同样在黑盒模型中起作用的可能性。

4.2　实验结果

4.2.1　目标模型的训练与结果

在模型训练阶段，为了控制训练的高效性和一致性，迭代次数、学习率等超参数均保持一致，迭代次数设置为 500，学习率为 0.001，并且设置了自动更新机制：如果测试集损失值连续 5 次没有下降，则学习率减半。此外，考虑到调制信号数据集由 20 种信噪比组成，每种信噪比的数据都具有不同的数据特征，训练模型时采取的方式是先将所有信噪比数据统一组合后作为一个数据集进行训练，然后在验证时分别在各个信噪比上单独验证其识别准确率。

VGG，LSTM 以及 ResNet 3 种 ADNN 模型的识别准确率如图 3 所示。

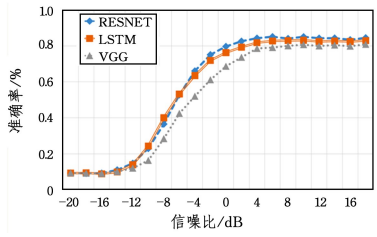

图 3　调制信号识别准确率

Fig. 3　Modulation Recognition accuracy

从图 3 可以看出，随着信噪比的增加，3 个模型的整体准确率呈上升趋势，这也与当前文献的结果一致。值得注意的是，调制信号信噪比为负值时，模型的识别准确率都很低，这是因为此时噪声的比例已经高于信号本身的功率，导致信号波形失真。如果只观察信噪比大于 4dB 的调制识别准确率，会发现实验所采用的 3 种目标模型的识别准确率大致近似，都可以达到 80%左右的精度，高于数据集作者设计的 VT-CNN2 模型[1]，可以满足对抗攻击所需目标模型的要求。

4.2.2　白盒攻击

由于对抗样本本身是为了使识别模型错误分类原始样本，根据这一特性，在进行白盒攻击时，本文只针对原始样本中分类正确的样本进行攻击。此外，由于调制信号具有不同信噪比这一特性，执行对抗攻击时也需要针对不同信噪比逐一执行攻击。

图 4 给出了白盒攻击的结果。对于白盒无目标攻击这种较为简单的攻击方式，大多数攻击方法都可以使 VGG 识别模型的准确率下降到接近 0%；但 FGSM 和 I-FGSM 两种攻击方法仅使模型识别准确率下降了 50%左右，明显弱于其他方法，其中的原因可能是 FGSM 作为唯一的非迭代单步攻击算法，尽管速度很快，但攻击效果不尽人意，此外，I-FGSM 虽然是迭代攻击方法，但可能由于其攻击速度较慢，在限制的迭代次数内未能成功攻击。

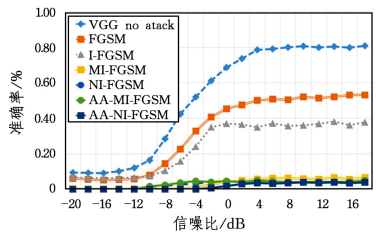

图 4　白盒无目标攻击

Fig. 4　White-box untargeted attack

图 5 给出了白盒的有目标攻击实验的结果。相较于无目标的对抗攻击，有目标的对抗样本的攻击成功率明显低于无目标的成功率。具体来说，FGSM 的攻击成功率基本上低于 10%。对于基于迭代攻击的非特征层面的 3 种攻击方法（I-FGSM，MI-FGSM，NI-FGSM）来说，在信噪比大于 0dB 的调制信号上的攻击效果明显好于 FGSM，整体攻击成功率可以达到 35%左右。本文提出的两种算法的攻击效果是最优的，其中 AA-MI-FGSM 可以达到 50%左右，而 AA-NI-FGSM 可以达到 62%。

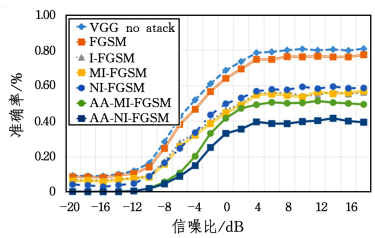

图 5　白盒有目标攻击

Fig. 5　White-box untargeted attack

从上述结果可以看出，在针对面向 AMR 的 DNN 模型的白盒攻击上，无论是有目标的还是无目标的攻击方法，基于迭代的攻击方法普遍优于 FGSM，而且基于特征梯度的对抗攻击方法表现最佳。

4.2.3　黑盒攻击

与理想实验环境不同，在实际的调制信号识别和对抗环境中，目标模型对于攻击者而言往往是不可见的，即通常都是

黑盒攻击的情形。这种情况下，对对抗样本的可迁移性和泛化能力有了较高的要求，更加考验对抗攻击方法的性能表现。

与传统的使用替代模型来替代目标黑盒模型的攻击方式不同，为了更好地验证对抗样本的可迁移性，本文所采用的黑盒攻击是直接将从 VGG 白盒攻击中生成的对抗样本迁移到黑盒模型中执行攻击的方式。此外，黑盒攻击分别在 LSTM 和 ResNet 两种不同的网络模型上进行测试，实验结果分别如图 6 和图 7 所示。

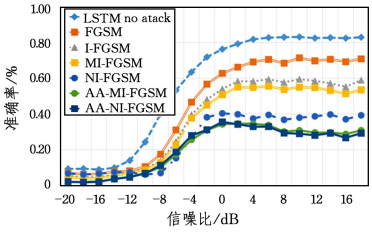

图 6 LSTM 黑盒无目标攻击

Fig. 6 Black-box untargeted attack on LSTM

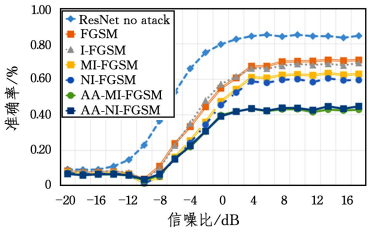

图 7 ResNet 黑盒无目标攻击

Fig. 7 Black-box untargeted attack on ResNet

从图 6 中可以看出，对于 LSTM 的黑盒模型，原本在白盒模型中能够使 VGG 网络的准确率下降到接近 0%的对抗样本，将其迁移到 LSTM 模型上时，攻击成功率有明显下降，尤其是基于标签梯度的攻击方法。相比之下，AA-MI-FGSM和AA-NI-FGSM仍然可以达到较好的对抗攻击效果，使 LSTM 模型的准确率下降到了 30%左右。从图 7 中也能得到类似的结论：基于特征梯度的攻击对抗样本在迁移到 ResNet 黑盒模型后还能保持较好的攻击效果。

图 8 和图 9 分别表示对抗样本在两种黑盒模型上的泛化率和迁移率的对比，其中由于 FGSM 方法的攻击效果较差，未将其列入对比方法中。从图 8 可以看出，从黑盒的泛化率，也就是对抗样本在黑盒模型上的攻击成功率上来看，基于特征的攻击方法 AA-MI-FGSM 和 AA-NI-FGSM 在攻击效果上要明显优于前面 3 种基于标签的迭代攻击方法。

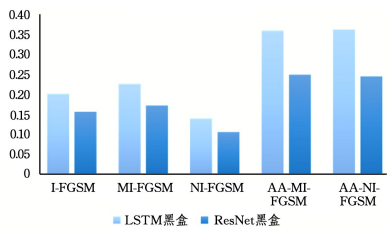

图 8 黑盒泛化率

Fig. 8 Black-box generalization rate

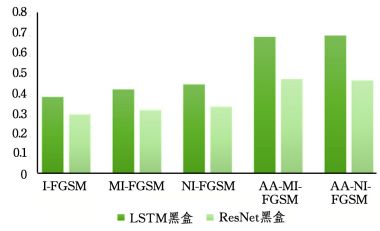

图 9 黑盒迁移率

Fig. 9 Black-box transfer rate

图 9 给出了在白盒模型上攻击成功的对抗样本同时也能成功攻击黑盒模型的比例，也就是迁移率的对比。从图中可以发现无论是 LSTM 模型还是 ResNet 模型，基于特征的两种攻击方法的黑盒迁移率同样高于传统方法，这说明所提出的基于特征的攻击方法具有优异的攻击迁移性能。

从上述实验中可以看出，所提出的方法在模型的泛化率和迁移率上都要优于标签梯度方法。

4.2.4 调制信号对抗样本分析

图 10 给出了在 VGG 模型上，当信噪比为 18 dB 时，原始样本中被成功攻击的不同种类的样本数量。

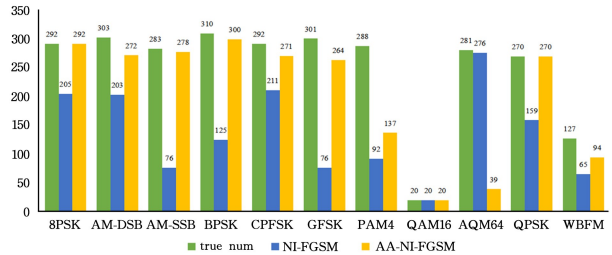

图 10 调制信号对抗样本数量

Fig. 10 Number of adversarial examples for each modulation mode

从整体上来看，除了 QAM16 和 QAM64，AA-NI-FGSM 攻击效果都优于 NI-FGSM。观察 NI-FGSM 方法会发现其对于 AM-SSB，BPSK，GFSK 以及 PAM4 的攻击效果都没有达到成功样本数量的 50%，这说明这几种信号类别对于 NI-FGSM 具有很高的稳定性和鲁棒性，不易被攻击；而对于 QAM64 样本而言，NI-FGSM 却可以成功攻击 98%的样本，说明 QAM64 信号样本对于 NI-FGSM 来说是敏感的和易攻击的。

观察 AA-NI-FGSM 的攻击结果会发现，其相比较于 NI-FGSM，在 AM-SSB，BPSK，GFSK 以及 PAM4 这几个调制类别上的攻击效果明显得到了较大提高，可以成功攻击样本数量的 80%。此外，除了 QAM64，在其他所有调制信号类别上 AA-NI-FGSM 相较于 NI-FGSM 在生成的对抗样本数量上都有不同程度的提高。可能的原因是基于特征梯度的攻击方法从特征空间中提取到了更丰富的信号样本信息，从而使其攻击效果相较于基于标签的方法更加有效。

结果表明，基于特征梯度的攻击方法在调制信号识别上相较于基于标签梯度的攻击方式，能够更好地弥补后者对于某些类别无法成功攻击的缺陷，这进一步说明了特征梯度攻击方法在 ADNN 模型对抗样本生成中的优越性。对于调制信号对抗样本评估而言，除了样本对模型的攻击成功率这个标准外，调制信号对抗样本本身相较于原始样本的扰动幅度

大小和扰动剧烈程度也是重要的评估标准。图 11(a)和图 11(b)给出了 QAM64 信号样本生成的对抗样本及其原始信号的曲线图，图 11(c)和图 11(d)则是 8PSK 信号样本生成的对抗样本及其原始信号的曲线图。图 11(a)和图 11(c)为基于标签梯度的信号扰动图像，图 11(b)和图 11(d)则为基于特征梯度的信号扰动图像，其中蓝色曲线代表未经过扰动的原始图像，橙色曲线表示经过对抗攻击添加扰动后的对抗信号图像。从图 11 可以看出，对于同样的一个信号样本，基于标签梯度的攻击方式所产生的扰动常常会有连续且剧烈的抖动，这往往不符合一个高信噪比调制信号的图像特征，很容易被检测出来；而对于特征梯度攻击的对抗样本信号图像，由于其添加的扰动在幅度和抖动程度上都更小，因此更不易被察觉和检测。

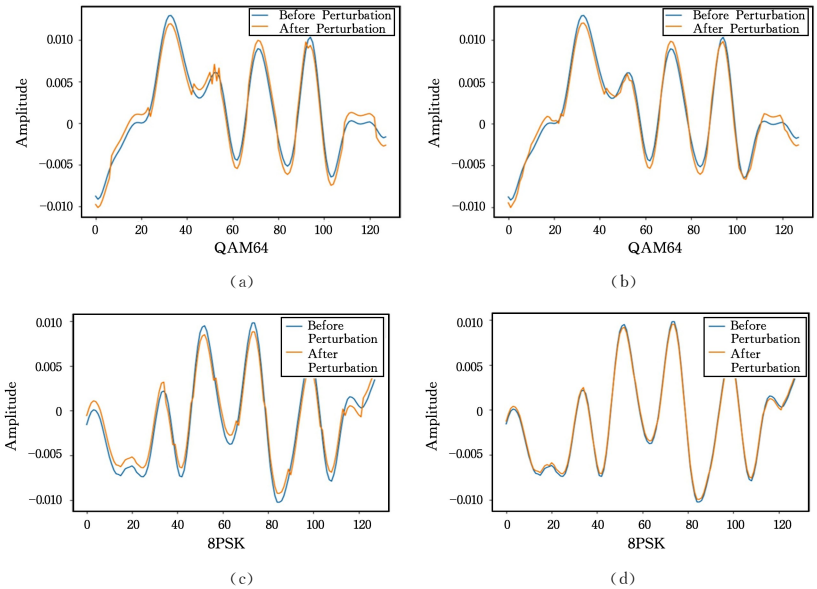

图 11　调制信号对抗样本扰动

Fig. 11　Modulation samples before and after introducing adversarial perturbation

结束语　本文探讨了面向自动调制信号识别的深度神经网络(ADNN)模型易于遭受梯度攻击的安全问题，提出了一种新的基于特征梯度的 ADNN 对抗攻击方法，并设计了两种攻击算法，即 AA-MI-FGSM 和 AA-NI-FGSM。相比传统基于标签的对抗攻击方法，所提方案可以更好地攻击 ADNN 提取到的信号中的稳定特征。在公开数据集上的大量实验表明，所提出的基于特征梯度的攻击方法在黑盒攻击和白盒攻击的场景下，在攻击成功率和可迁移性上均优于 4 种传统的基于标签梯度的攻击方法。此外，基于特征梯度的攻击方法所添加的扰动更加平滑且不易察觉。

虽然本文算法相较于传统的标签方法在信号对抗样本生成中具有许多优势，但仍存在一些有待改进之处：1)特征层的选择不唯一，不同模型的不同网络层次往往对信号数据特征的提取效果不一致，暂时不确定哪一层的特征对于对抗样本的生成效果最好，未来将设计最佳攻击层次选择方法。2)由于提取特征相比于直接运用标签需要更多的时间和空间，运行效率上略低于标签方法，因此如何降低攻击方法的时间复杂性是未来需要考虑的问题。

参 考 文 献

[1] O'SHEA T J，WEST N. Radio machine learning dataset generation with gnu radio[C]// Proceedings of the GNU Radio Conference. 2016：16.

[2] LIU Y，YANG C. Modulation recognition with graph convolutional network[J]. IEEE Wireless Communications Letters，2020，9(5)：624-627.

[3] KATO N，FADLULLAH Z M，MAO B，et al. The deep learning vision for heterogeneous network traffic control：Proposal，challenges，and future perspective [J]. IEEE Wireless Communications，2016，24(3)：146-153.

[4] 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.

[5] WANG Y，LIU M，YANG J，et al. Data-driven deep learning for automatic modulation recognition in cognitive radios[J]. IEEE Transactions on Vehicular Technology，2019，68(4)：4074-4077.

[6] 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.

[7] TANG B，TU Y，ZHANG Z，et al. Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks[J]. IEEE Access，2018，6：15713-15722.

[8] CHEN K，ZHANG S，ZHU L，et al. Modulation Recognition of Radar Signals Based on Adaptive Singular Value Reconstruction and Deep Residual Learning[J]. Sensors，2021，21(2)：449.

[9] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[J]. arXiv：1312. 6199，2013.

[10] GOODFELLOW I J，SHLENS J，SZEGEDYC. Explaining and harnessing adversarial examples [C]// ICML. 2015.

[11] KURAKIN A，GOODFELLOW I，BENGIO S. Adversarial examples in the physical world[J]. arXiv：1607. 02533，2016.

[12] 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.
[13] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2574-2582.
[14] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[J]. arXiv:1908.06281, 2019.
[15] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.
[16] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv:1611.01236, 2016.
[17] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017:39-57.
[18] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]// International Conference on Machine Learning. PMLR, 2018:284-293.
[19] LIN Y, ZHAO H, TU Y, et al. Threats of adversarial attacks in DNN-based modulation recognition[C]// IEEE Conference on Computer Communications (IEEE INFOCOM 2020). IEEE, 2020:2469-2478.
[20] ZHAO H, LIN Y, GAO S, et al. Evaluating and Improving Adversarial Attacks on DNN-Based Modulation Recognition[C]// 2020 IEEE Global Communications Conference (GLOBECOM 2020). IEEE, 2020:1-5.
[21] DeepSig. Deepsig dataset: Radioml 2016.10a [OL]. https://www.deepsig.io/datasets, 2016.
[22] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv:1409.1556, 2014.
[23] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

WANG Chao, born in 1997, postgraduate. His main research interests include deep learning and adversarial example.

WEI Xiang-lin, born in 1985, Ph.D, engineer. His main research interests include edge computing, deep learning and wireless network security.