doi:10.11887/j.cn.202305004 http://journal.nudt.edu.cn

利用主成分分析的通信调制识别通用对抗攻击方法*

柯 达[1],黄知涛[1,2],邓寿云[3],卢超奇[3]

(1. 国防科技大学 电子科学学院,湖南 长沙 410073;2. 国防科技大学 电子对抗学院,安徽 合肥 230037;
3. 中国人民解放军31433部队,辽宁 沈阳 110000)

摘 要:深度学习容易被对抗样本所攻击。以通信调制识别为例,在待传输的通信信号中加入对抗性扰动,可以有效防止非合作的用户利用深度学习方法识别信号的调制方式,进而提升通信安全。针对现有对抗样本生成技术难以满足自适应和实时性的问题,通过对数据集中抽取的小部分数据产生的对抗扰动进行主成分分析,得到适用于整个数据集的通用对抗扰动。通用对抗扰动的计算可以在离线条件下进行,然后实时添加到待发射的信号中,可以满足通信的实时性要求,实现降低非合作方调制识别准确率的目的。实验结果表明该方法相对基线方法具有更优的欺骗性能。

关键词:对抗样本;通用对抗扰动;通信调制识别

中图分类号:TN97 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:

文章编号:1001-2486(2023)05-030-08

听语音 聊科研 与作者互动

Universal adversarial attack method for communication modulation identification using principal component analysis

KE Da[1], *HUANG Zhitao*[1,2], *DENG Shouyun*[3], *LU Chaoqi*[3]

(1. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China;
2. College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China;
3. The PLA Unit 31433, Shengyang 110000, China)

Abstract: Deep learning is easily attacked by adversarial examples. Taking communication modulation recognition as an example, adding adversarial perturbations to the transmitted signal can effectively prevent non-cooperative users from utilizing the deep learning method to recognize the modulation of the signal. Thus, adversarial perturbations can help enhance communication security. To address the problem that the existing adversarial attack techniques are difficult to meet the adaptive and real-time requirements, the universal adversarial perturbation applicable to the whole dataset was obtained by the principal component analysis of the adversarial perturbation generated by a small part of the data extracted from the dataset. The computation of the universal adversarial perturbation can be carried out under offline conditions and then added to the signal to be transmitted in real time, which can satisfy the real-time requirements of communication and realize the purpose of reducing the accuracy of non-cooperative party modulation recognition. Experimental results show that the proposed method has better deception performance relative to the baseline method.

Keywords: adversarial examples; universal adversarial perturbation; communication modulation identification

通信信号自动调制识别(automatic modulation classification, AMC)技术作为信号检测和信号解调之间的重要步骤,可以自动识别接收信号的调制方式,提升认知电子战系统在非合作场景下实现高效频谱感知和频谱利用的能力[1],是适应复杂电磁环境的重要手段,也是近年来无线通信和电子侦察领域研究的重点课题。除传统的基于假设检验理论的AMC方法[2-4]和基于特征提取的方法[5-12],近年来学术界逐渐将目光转向了基于深度学习(deep learning, DL)的调制识别方法研究[13-18]。将DL方法应用到AMC问题中,DL可以自适应学习数据的有效表征,避免了人工设计特征的过程,加快了技术迭代的效率,同时大量的用频设备为数据驱动的DL方法提供了充足的数据支撑[19]。

但是,DL的脆弱性也一直为人诟病,现有DL

* **收稿日期**:2022-10-14
基金项目:国防科技大学青年科技创新奖资助项目(18/19-QNCXJ)
作者简介:柯达(1994—),男,贵州贵阳人,博士研究生,E-mail:1747884404@qq.com;
黄知涛(通信作者),男,湖北荆州人,教授,博士,博士生导师,E-mail:huangzhitao@nudt.edu.cn

模型虽然在许多领域取得了耀眼的成绩,但是存在一类针对 DL 模型的特殊攻击方式,可以实现对 DL 模型高效且隐蔽的攻击,这种攻击方式被称为“对抗样本”[20]。利用 DL 存在的固有缺陷,可以设计出一种特殊的极其微弱的扰动,将其添加到待处理的样本中,可以使原本表现良好的模型以很高的置信度对样本给出错误的处理结果[21]。由于添加的扰动极其微弱,人们难以察觉样本被做出了改动,因而难以预防。如果将对抗样本技术充分应用于通信信号波形的生成过程中,将会遏制未来智能化认知电子战系统的精确感知能力[22]。因此,有必要从通信信号对抗样本的生成和防御两方面展开研究。本文立足攻防结合,以攻促防,从对抗样本攻击的角度研究了一种通用对抗性扰动(universal adversarial perturbations, UAP)的生成方法,在基于 DL 的调制识别模型中初步验证了算法的可行性和有效性。

自 2014 年 Szegedy 等首次发现对抗样本现象后,次年,Goodfellow 等进一步解释了对抗样本存在的原因,并提出了快速计算对抗样本的快速梯度符号法[23](fast gradient sign method, FGSM)。Lü 等在文献[24]中提出了一种统一求解对抗样本的理论框架,为后续对抗样本的研究奠定了基础。前期关于对抗样本的研究主要集中在图像和语音领域中。2019 年,文献[25]首次研究了针对基于 DL 的通信信号调制识别模型的对抗样本攻击,验证了其可行性;文献[26]在通信信号对抗样本攻击的基础上,尝试了将对抗训练的思想引入对抗样本防御中。文献[18]中,Lin 等在开源调制识别数据集中验证了现有的多种对抗样本攻击方法的可行性。Kim 等在多种场景下考虑了通信信号对抗样本受信道影响的因素,提出了适应真实信道场景的对抗样本攻击方法[27-29]。尽管上述工作已经对调制识别任务中的对抗样本进行了充分的探索,但是上述工作采用的对抗样本攻击方法每次实施对抗样本攻击时均需要针对每一类调制每一个样本计算一次对抗性扰动,难以满足通信过程自适应和实时性的要求。UAP 可以针对目标模型,仅以部分样本为基础,生成一个特定的扰动,输入模型中的所有样本加上该扰动后,均能达到对抗攻击的效果。“通用”指的是生成的扰动可以破坏模型对尽可能多的调制方式与输入样本的识别过程。通用对抗性扰动可以在离线的条件下生成,实际通信过程中只需将该扰动添加到待发射的信号中,便可得到通信信号的对抗样本,满足实时性的要求。

文献[30]最早在图像处理问题中提出了 UAP 的生成方法,验证了 UAP 在不同数据之间存在很好的通用性,即使对于表现良好的神经网络,其输入的大部分数据加上微小的 UAP,都能被网络以很高的置信度识别成错误的结果;对 UAP 的存在性给出了经验性的解释,认为 UAP 揭示了基于 DL 的分类模型的高维决策边界之间存在几何相关性,即在输入空间中存在单一的方向可以破坏基于 DL 的分类器的识别过程。其核心思想是采用迭代的方法去逼近文中所提出的表征几何相关性的方向。本文在文献[30]的基础上,首先建立了通信调制识别的系统模型,然后针对调制识别模型计算了攻击所需的最小对抗扰动。由于最小对抗扰动的方向垂直于分类边界,所以该方向可以用于表征分类界面的特性。在此基础上,提出了基于主成分分析(principal component analysis, PCA)的 UAP 计算方法,通过对计算得到的若干最小对抗扰动进行主成分分析,得到最能表征目标模型分类界面几何相关性的方向,从而得到对整个数据集具有普适性的通用对抗扰动。

1　系统模型

对抗样本攻击即在正常的通信过程中添加一段精心设计的扰动,在尽可能不破坏合作通信过程的基础上,使得基于 DL 的非合作通信系统难以识别信号,从而达到保护合作通信的目的。对抗样本攻击的系统模型如图 1 所示,主要由通信发射机、接收机和通用对抗扰动 UAP 发射机构成。

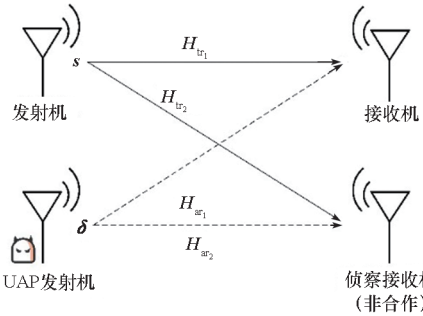

图 1　通信系统中对抗样本攻击的系统模型

Fig. 1　System model for combating a sample attack in a communication system

在正常的通信过程中,发射信号 $\boldsymbol{x}$ 经信道 $\boldsymbol{H}_{\mathrm{tr}}$ 传输至接收机,此时信号 $\boldsymbol{s}$ 容易被非合作系统截获。为了降低被智能侦察系统识别的概率,UAP 发射机会同时辐射通用对抗扰动 $\boldsymbol{\delta}$,此时接收方

和侦察系统接收到的信号分别为

$$\begin{cases} \boldsymbol{x}_{\mathrm{tr}} = \boldsymbol{H}_{\mathrm{tr}_1}(\boldsymbol{s}) + \boldsymbol{H}_{\mathrm{ar}_1}(\boldsymbol{\delta}) + \boldsymbol{n} \\ \boldsymbol{x}_{\mathrm{ar}} = \boldsymbol{H}_{\mathrm{tr}_2}(\boldsymbol{s}) + \boldsymbol{H}_{\mathrm{ar}_2}(\boldsymbol{\delta}) + \boldsymbol{n} \end{cases} \tag{1}$$

其中：$\boldsymbol{s}$ 为发射信号，$\boldsymbol{\delta}$ 为对抗性扰动 UAP；$\boldsymbol{H}_{\mathrm{tr}_i}$ 为 $\boldsymbol{s}$ 所经过的信道，$\boldsymbol{H}_{\mathrm{ar}_i}$ 为 UAP 传输的信道；$\boldsymbol{n}$ 为高斯白噪声。侦察系统会对接收信号 $\boldsymbol{x}_{\mathrm{ar}}$ 进行调制识别，并进一步进行解调解译等工作。在正常通信信号中假如对抗扰动 $\boldsymbol{\delta}$ 的目的就是在尽可能保证自身通信不受影响的前提下，使智能化侦察系统难以识别接收信号 $\boldsymbol{x}_{\mathrm{ar}}$ 的调制，所以对抗性扰动 UAP 的能量要尽可能小。

1.1 通信信号智能调制识别模型

通信系统中，往往需要经过调制将基带信息负载到载波上实现远距离传输。AMC 技术可以在未知先验信息的条件下，自动判断出通信信号的调制方式。

由于非合作系统无法知道接收信号是否含有对抗性扰动 $\boldsymbol{\delta}$，所以上述问题可以简化为

$$\boldsymbol{x}_{\mathrm{ar}} = \boldsymbol{H}_{\mathrm{tr}_2}(\boldsymbol{s}) + \boldsymbol{n} \tag{2}$$

基于深度学习的调制识别的目的就是设计一个基于深度学习的分类器 $f(\boldsymbol{x};\boldsymbol{\theta}):\mathcal{X}\rightarrow\mathbb{R}^C$。其中 $\mathcal{X}$ 为输入空间，$\boldsymbol{\theta}$ 为分类器 f 的参数，C 是待识别调制类型的数目。分类器 f 的识别结果为

$$k(\boldsymbol{x}) = \underset{k\in\{1,2,\cdots,C\}}{\arg\max} f_k(\boldsymbol{x};\boldsymbol{\theta}) \tag{3}$$

用于分类的深度神经网络主要有全连接网络（fully-connected network，FN）、卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN）。本文采用目前主流的卷积神经网络结构 ResNet18[31] 作为实验的对象，详细的网络结构和参数设置如图 2 所示。

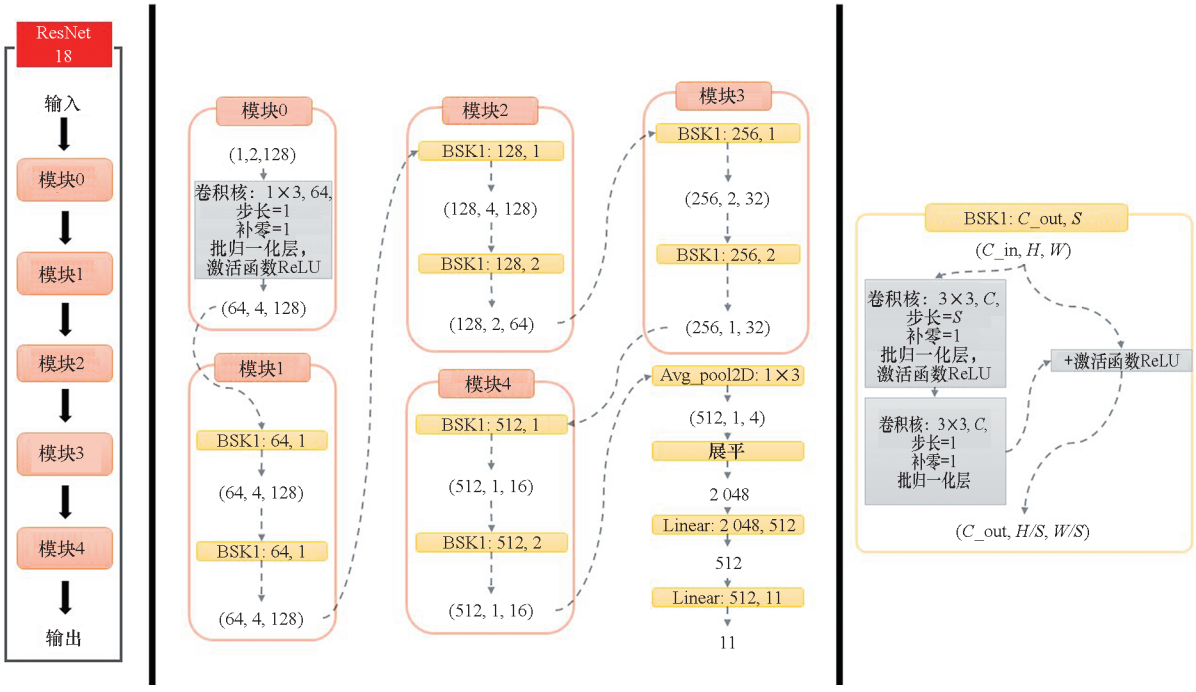

图 2 基于 ResNet18 的调制识别网络结构

Fig. 2 Structure diagram of the modulation recognition network based on ResNet18

1.2 基于 PCA 的通用对抗性扰动模型

深度学习自身也面临鲁棒性差的问题，特别是容易受到对抗样本攻击。如式（1）所示，若非合作通信系统接收信号中带有通用对抗扰动 $\boldsymbol{\delta}$，则智能侦察系统对接收信号 $\boldsymbol{x}_{\mathrm{ar}}$ 的调制识别准确率将会严重下降。通用对抗扰动的定义是

$$f(\boldsymbol{x}_{\mathrm{ar}} + \boldsymbol{\delta}) \neq f(\boldsymbol{x}_{\mathrm{ar}}) \tag{4}$$

并且要求式（4）对任意 $\boldsymbol{x}_{\mathrm{ar}} \in \mathcal{X}$ 均成立。设信号集 $\boldsymbol{X} = [\boldsymbol{x}_{\mathrm{ar}_1}, \boldsymbol{x}_{\mathrm{ar}_2}, \cdots, \boldsymbol{x}_{\mathrm{ar}_m}]^{\mathrm{T}}$ 抽样自接收信号的总体分布 $\mathcal{X}$。首先，分别计算 $\boldsymbol{X}$ 中的每个 $\boldsymbol{x}_{\mathrm{ar}_i}$ 对分类器 f 的最小对抗性扰动 $\boldsymbol{\delta}_i$[32]，得到最小对抗性扰动构成的集合 $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_m]^{\mathrm{T}}$。在 L_2 范数的约束下，$\boldsymbol{x}_{\mathrm{ar}_i}$ 的最小对抗性扰动的方向 $\boldsymbol{\delta}_i$ 可以理解为过 $\boldsymbol{x}_{\mathrm{ar}_i}$ 计算分类界面法向量。为简化问题，首先考虑最简单的情况，即二分类线性分类器，其原理如图 3 所示。

此时分类器 $f(\boldsymbol{x}) = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b}$，其中 $\boldsymbol{\omega}$ 为分类器权重，$\boldsymbol{b}$ 为分类器的偏置。则法向量 $\boldsymbol{\delta}_i$ 的表达式为

$$\boldsymbol{\delta}_i = -\frac{f(\boldsymbol{x}_i)}{\|\boldsymbol{\omega}\|_2^2}\boldsymbol{\omega} \tag{5}$$

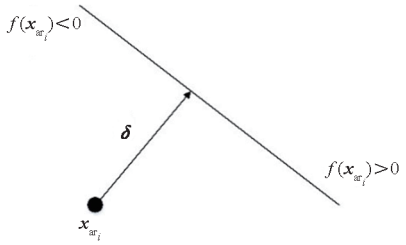

图 3　最小扰动原理图

Fig. 3　Schematic of the minimal perturbation

其中,$\|\cdot\|_2$ 表示 L_2 范数。推广到多分类器的情形,首先按照“1 对多”的思路将二分类器的情形推广到多分类器,即逐个计算样本点到各个分类界面的法向量距离,再找出其中最短的作为最小扰动。进一步地,将线性分类器推广到任意多分类器的情形如算法 1 所示[32]。

算法 1　任意多分类器的最小扰动计算方法

Alg. 1　Minimal perturbation calculation method for arbitrary multiple classifiers

输入:原始输入 $\boldsymbol{x}$,分类器 f

输出:最小功率扰动 $\boldsymbol{\delta}_{\min}$

初始化:$\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $j \leftarrow 0$

while $f(\boldsymbol{x}_j) \neq f(\boldsymbol{x}_0)$ **do**

　for $k = f(\boldsymbol{x}_0)$ **do**

　　$\boldsymbol{\omega}_k' \leftarrow \nabla f_k(\boldsymbol{x}_j) - \nabla f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_j)$

　　$f_k' \leftarrow f_k(\boldsymbol{x}_j) - f_{\hat{k}(\boldsymbol{x}_0)}(\boldsymbol{x}_j)$

　end for

　$\hat{l} \leftarrow \arg\min_{k \neq \hat{k}(\boldsymbol{x}_0)} \dfrac{|f_k'|}{\|\boldsymbol{\omega}_k'\|_2}$

　$\boldsymbol{r}_i \leftarrow \dfrac{|f_{\hat{l}}'|}{\|\boldsymbol{\omega}_{\hat{l}}'\|_2^2}\boldsymbol{\omega}_{\hat{l}}'$

　$\boldsymbol{x}_{j+1} \leftarrow \boldsymbol{x}_j + \boldsymbol{r}_j$

　$j \leftarrow j+1$

end while

返回 $\boldsymbol{\delta}_{\min} = \sum_j \boldsymbol{r}_j$

对每个样本计算得到的 $\boldsymbol{\delta}_{\min}$ 构成矩阵 $\boldsymbol{\Delta}$,再对矩阵 $\boldsymbol{\Delta}$ 进行奇异值分解,即

$$\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

计算矩阵 $\boldsymbol{\Delta}$ 的第一奇异值和对应的右奇异向量 $\boldsymbol{v}_1 = \boldsymbol{V}\boldsymbol{e}_1$,其中 $\boldsymbol{e}_1 = \underbrace{[1,0,\cdots,0]}_{m}^{\mathrm{T}}$。实际场景中,往往会对 $\boldsymbol{v}_1$ 的能量进行限制,则采用 $\boldsymbol{\delta} = p_{\max}\boldsymbol{v}_1$ 作为最终求解得到的通用扰动,其算法原理如图 4 所示,详细的算法过程如算法 2 所示。

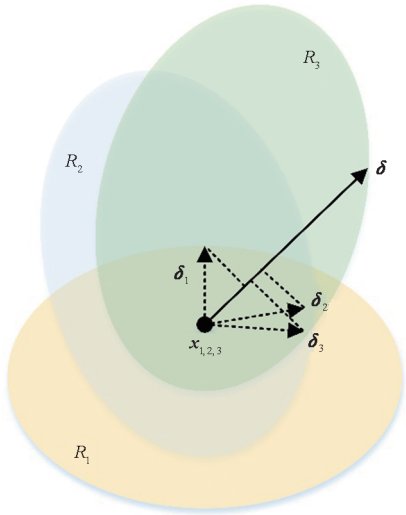

图 4　基于 PCA 的通用对抗扰动算法原理图

Fig. 4　Schematic of a general adversarial perturbation algorithm based on PCA

算法 2　基于 PCA 的 UAP 计算方法

Alg. 2　PCA-based calculation method for UAP

输入:$\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_m]^{\mathrm{T}}$,分类器 f, 最小功率对抗扰动的范数 $p_{\max}$

输出:通用对抗扰动 $\boldsymbol{\delta}$

计算 $\boldsymbol{\Delta}$ 的第一主成分向量 $\boldsymbol{v}_1$, i. e. $\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{v}_1 = \boldsymbol{V}\boldsymbol{e}_1$

$\boldsymbol{\delta} = p_{\max}\boldsymbol{v}_1$

2　实验验证

本节主要设计两个实验用于验证所提方法的性能。第一个实验是通过与算法 3[30] 对比,验证所提算法的先进性;第二个实验是分析不同抽样数量对本算法性能的影响。

2.1　基线方法介绍

采用文献[30]所提算法(见算法 3)作为对比的基线方法。设 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$ 为抽样自总体数据集的部分样本集。其算法原理为:对 $\boldsymbol{X}$ 中的每一个样本迭代地建立通用扰动 $\boldsymbol{\delta}$。如图 5 所示,每一轮迭代中都会把当前采样点对应的最小对抗扰动 $\boldsymbol{\delta}_i$ 送入当前的扰动采样点 $\boldsymbol{x}_i + \boldsymbol{\delta}$ 中,和之前计算出的通用扰动 $\boldsymbol{\delta}$ 构成新的通用扰动。其中 $\boldsymbol{x}_{1,2,3}$ 表示抽样的 3 个样本,R_1、R_2、R_3 分别表示 $\boldsymbol{x}_{1,2,3}$ 对应的 3 个分类界面。通过不断修正,促使最终的通用扰动 $\boldsymbol{\delta}$ 能够扰乱分类器对大部分样本的分类。

算法 3　通用对抗扰动计算方法

Alg. 3　Universal adversarial perturbation calculation methods

输入:原始输入 $\boldsymbol{x}$,分类器 f, 所需的最小扰动功率的范数 $p_{\max}$

输出:通用对抗扰动 $\boldsymbol{\delta}$

初始化:$\boldsymbol{x}_0 \leftarrow \boldsymbol{x}, \boldsymbol{\delta} \leftarrow 0$,

for $\boldsymbol{x}_i \in \boldsymbol{x}$ do

　if $f(\boldsymbol{x}_i + \boldsymbol{\delta}) = f(\boldsymbol{x}_i)$ do

　　$\Delta\boldsymbol{\delta}_i \leftarrow \arg\min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2$

　　s. t. $f(\boldsymbol{x}_i + \boldsymbol{\delta} + \boldsymbol{r}) \neq f(\boldsymbol{x}_i)$

　　$\boldsymbol{\delta} \leftarrow p_{\max}(\boldsymbol{\delta} + \Delta\boldsymbol{\delta}_i)$

　end if

end for

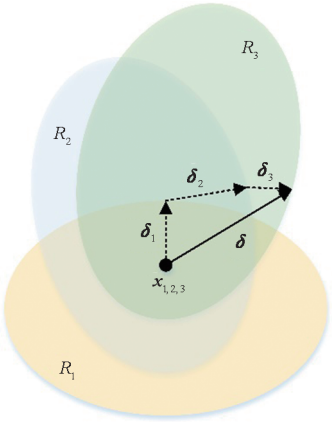

图 5　通用对抗扰动生成算法原理示意图

Fig. 5　Schematic diagram of the general adversarial disturbance generation algorithm

2.2　实验准备与数据集

所有实验均在 NVIDIA GeForce GTX 3090 GPU 上进行计算,通过 pytorch1.10 和 cuda11.3 实现。优化器为 Adam,使用交叉熵损失函数进行训练,总训练次数为 100 次,学习率为 10^{-5},设置 early stopping 策略。

实验所用数据来自开源通信信号调制识别数据集 RML2016.04C,该数据集包含 BPSK、QPSK、8PSK、16QAM、64QAM、BFSK、CPFSK、PAM4、WB-FM、AM-SSB、AM-DSB 共 11 种调制类型,覆盖信噪比 −20 ~ 18 dB,步进 2 dB。数据集共 162 060 个样本,每个样本包含 128 点 IQ 数据。训练集、验证集、测试集按照 7∶2∶1 的比例划分。

2.3　攻击性能对比

对于计算得到的通用对抗性扰动 $\boldsymbol{\delta}$,定义扰动 − 信号比(perturbation-signal ratio, PSR)衡量对抗性扰动相对于信号的强弱,用以评价对抗性扰动的“不可察觉性”,其计算公式为

$$PSR = 10\lg\left(\frac{P_\delta}{P_x}\right) \tag{7}$$

其中,P_δ 和 P_x 分别代表扰动和信号的功率。同样地,对于算法攻击性能的评价,定义欺骗率(fooling rate, FR)来评价算法的优劣,其计算公式为

$$FR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(f(\boldsymbol{x}_i + \boldsymbol{\delta}) \neq f(\boldsymbol{x}_i)\right) \tag{8}$$

其中,$f(\boldsymbol{x}_i)$ 为识别正确的结果,$\mathbb{1}(\cdot)$ 为指示函数,N 为总样本数。FR 的含义是,对于原本已经正确识别的样本 $\boldsymbol{x}_i$,加入通用扰动 $\boldsymbol{\delta}$ 后,被分类器误判,则认为欺骗成功。实验分别随机抽取 50、500、5 000 个样本生成了 PSR 为 −20 ~ 0 dB 的通用扰动,对已经训练完成的分类器进行攻击,测试样本数为 16 206,实验结果如图 6 所示。

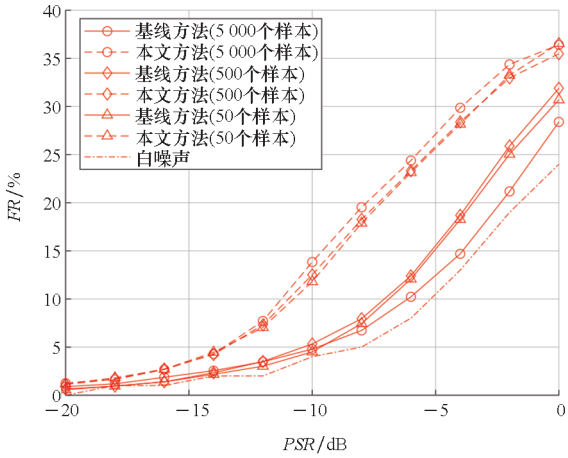

图 6　欺骗率随 PSR 变化曲线

Fig. 6　Fooling rate curve with PSR

本实验对比了基线方法和白噪声的攻击性能。实验结果表明,同等 PSR 下,所提出的算法攻击性能全面优于基线方法。当 PSR 大于 −15 dB后,所提方法的性能优势逐渐明显,同等欺骗率所需的扰动大小相对于基线方法可以降低 4 ~ 6 dB。当 $PSR = 0$ dB 时,所提方法的最高欺骗率为 36.55%,基线方法的最高欺骗率为 31.9%,所提方法性能提升了 4.65%。

实验分别抽样了 50、500、5 000 个样本用于生成通用扰动,从结果看来,随着抽样数的提升,算法性能有微弱的提升,但是并不明显。其中,当抽样数为 5 000 时,基线方法的性能反而比抽样数为 500 时差。根据图 5 所示的通用扰动算法原理,随着样本抽样数的增加,对通用扰动 $\boldsymbol{\delta}$ 的修正因素也会增加,但是每次修正因素只能保证对当

前样本是有效的,而无法保证对样本整体有效。基线方法并没有设计相应的机制来约束修正因素对样本整体的攻击性能,所以当抽样数由 500 增加到 5 000 时,造成了欺骗性能的下降。而本文提出的方法,核心思想是提取了多个扰动的主成分,有效地避免了由于样本数增加带来的干扰因素。

2.4　通用扰动攻击性能受抽样数的影响分析

为进一步探究抽样数与算法性能的影响,设计了本实验进行研究。首先,抽取 50 个样本计算得到算法 2 中的输入矩阵 $\boldsymbol{\Delta}=[\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\cdots,\boldsymbol{\delta}_m]^{\mathrm{T}}$,然后对矩阵 $\boldsymbol{\Delta}$ 进行奇异值分解。同步地,对 50 个高斯白噪声(Gaussian white noise, WGN)张成的矩阵计算奇异值。归一化的奇异值如图 7 所示由大到小进行排序。从图中可以看出,由 WGN 张成矩阵的奇异值按从大到小排列,奇异值曲线的下降速度缓慢,这符合随机噪声的特点,说明各个随机噪声向量之间不存在较强的相关性。相反,矩阵 $\boldsymbol{\Delta}$ 的奇异值经过了一个快速的下降过程,下降的拐点在第 5 个奇异值附近,说明少数的奇异值张成的向量便能近似地表征整个矩阵的主成分,这也解释了通用对抗扰动存在的合理性,即由对抗性扰动构成的矩阵 $\boldsymbol{\Delta}$ 中较大的奇异值张成的向量便能表征分类器脆弱性。

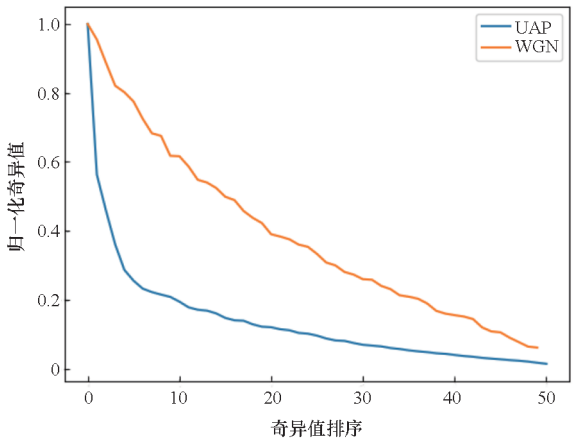

图 7　矩阵 $\boldsymbol{\Delta}$ 的奇异值

Fig. 7　Singular value of the $\boldsymbol{\Delta}$ matrix

定义每个奇异值占总奇异值之和的比重为奇异值贡献度,第 i 个奇异值的贡献度计算公式为

$$\eta_i=\frac{\sigma_i}{\sum_{k=1}^{m}\sigma_k} \tag{9}$$

通过计算,前 24 个奇异值的贡献度之和已经超过了 80%,这说明可以用少数几个奇异值和对应的左右奇异向量来近似描述原矩阵。

基于上述结论,分别抽样 1 ~ 50 个样本用于生成通用对抗扰动,固定其 *PSR* 为 −10 dB,攻击效果如图 8 所示。结果表明,随着抽样数的上升,欺骗率呈总体上升的趋势,当抽样数超过 25 后,欺骗率趋于稳定,实验结果与图 7 所示的推论相吻合,即仅需要 25 个样本所生成的通用对抗扰动便可实现较好的攻击性能,这也解释了 2.3 节中抽样数分别为 50、500、5 000 的性能差异并不明显的现象。

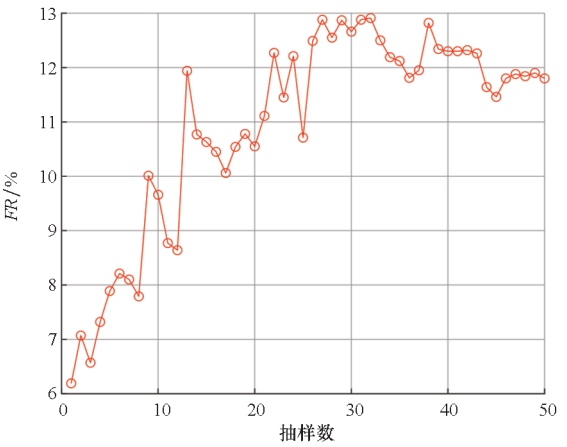

图 8　欺骗率随抽样数变化曲线

Fig. 8　Curve of fooling rate with the number of samples

3　结论

针对基于深度学习的通信信号调制识别的应用场景,对其容易受到对抗样本攻击的脆弱性进行了研究。

1)提出了一种基于 PCA 的改进的通用对抗扰动生成方法,该方法可基于少量的接收信号生成一个通用的对抗性扰动,该扰动可以降低识别器对所有输入信号的调制识别准确率。通用对抗性扰动可以在离线条件下产生,然后实时添加到通信过程中,能够满足通信过程的实时性要求。

2)采用卷积神经网络中的代表结构 ResNet18 在开源调制识别数据集中进行训练并识别。采用本文提出的通用对抗扰动生成方法,分别抽取 50、500、5 000 个样本生成 *PSR* 为 −20 ~ 0 dB 的通用扰动,并将其添加到 16 206 个测试样本中,测试其欺骗率并与基线方法对比。实验结果表明,本文方法相对于基线方法具有明显性能提升,同等欺骗率所需的扰动大小相对于基线方法可以降低 4 ~ 6 dB,最高欺骗率提升了 4.65%。

3)对通用对抗扰动受抽样数的影响进行了分析和实验验证。首先分析了通用对抗扰动的合理性,并分别抽样 1 ~ 50 个样本对分析结论进行验证。实验结果表明对所使用的数据集合分类器,所提算法仅需抽样 25 个样本便可生成稳定的通用扰动,与理论分析相吻合。证明本文算法具有更优的攻击性能。

参考文献(References)

[1] 焦翔, 魏祥麟, 薛羽, 等. 基于深度学习的自动调制识别研究[J]. 计算机科学, 2022, 49(5): 266 - 278.
JIAO X, WEI X L, XUE Y, et al. Automatic modulation recognition based on deep learning[J]. Computer Science, 2022, 49(5): 266 - 278. (in Chinese)

[2] PANAGIOTOU P, ANASTASOPOULOS A, POLYDOROS A. Likelihood ratio tests for modulation classification [C]// Proceedings of 21st Century Military Communications: Architectures and Technologies for Information Superiority, 2002.

[3] ZHAO Z J, LANG T. A MPSK modulation classification method based on the maximum likelihood criterion [C]// Proceedings of 7th International Conference on Signal Processing, 2004.

[4] WEI W, MENDEL J M. Maximum-likelihood classification for digital amplitude-phase modulations [J]. IEEE Transactions on Communications, 2000, 48(2): 189 - 193.

[5] DAHAP B I, LIAO H S, RAMADAN M. Simple and efficient algorithm for automatic modulation recognition for analogue and digital signals [M]//The Proceedings of the Second International Conference on Communications, Signal Processing, and Systems. Cham: Springer International Publishing, 2013: 345 - 357.

[6] AZZOUZ E E, NANDI A K. Automatic identification of digital modulation types [J]. Signal Processing, 1995, 47(1): 55 - 69.

[7] LIEDTKE F F. Computer simulation of an automatic classification procedure for digitally modulated communication signals with unknown parameters [J]. Signal Processing, 1984, 6(4): 311 - 323.

[8] UYS L Y, GOUWS M, STRYDOM J J, et al. The performance of feature-based classification of digital modulations under varying SNR and fading channel conditions[C]//Proceedings of 2017 IEEE AFRICON, 2017.

[9] NANDI A K, AZZOUZ E E. Algorithms for automatic modulation recognition of communication signals[J]. IEEE Transactions on Communications, 1998, 46(4): 431 - 436.

[10] SWAMI A, SADLER B M. Hierarchical digital modulation classification using cumulants [J]. IEEE Transactions on Communications, 2000, 48(3): 416 - 429.

[11] DOBRE O A, BAR-NESS Y, SU W. Higher-order cyclic cumulants for high order modulation classification [C]// Proceedings of IEEE Military Communications Conference, 2003.

[12] DAI W, WANG Y Z, WANG J. Joint power estimation and modulation classification using second-and higher statistics[C]//Proceedings of IEEE Wireless Communications and Networking Conference, 2002.

[13] POPOOLA J, OLST R. A novel modulation-sensing method[J]. IEEE Vehicular Technology Magazine, 2011, 6(3): 60 - 69.

[14] PARK C S, JANG W, NAH S P, et al. Automatic modulation recognition using support vector machine in software radio applications [C]//Proceedings of 9th International Conference on Advanced Communication Technology, 2007.

[15] WU H C, SAQUIB M, YUN Z F. Novel automatic modulation classification using cumulant features for communications via multipath channels [J]. IEEE Transactions on Wireless Communications, 2008, 7(8): 3098 - 3105.

[16] HUYNH-THE T, HUA C H, PHAM Q V, et al. MCNet: an efficient CNN architecture for robust automatic modulation classification [J]. IEEE Communications Letters, 2020, 24(4): 811 - 815.

[17] LIN Y, TU Y, DOU Z. An improved neural network pruning technology for automatic modulation classification in edge devices[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5703 - 5706.

[18] LIN Y, ZHAO H J, MA X F, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2021, 70(1): 389 - 401.

[19] LIANG Z, TAO M L, XIE J, et al. A radio signal recognition approach based on complex-valued CNN and self-attention mechanism [J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(3): 1358 - 1373.

[20] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. (2014 - 02 - 19) [2022 - 10 - 01]. https://arxiv.org/abs/1312.6199.

[21] 孙浩, 陈进, 雷琳, 等. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述[J]. 雷达学报, 2021, 10(4): 571 - 594.
SUN H, CHEN J, LEI L, et al. Adversarial robustness of deep convolutional neural network-based image recognition models: a review[J]. Journal of Radars, 2021, 10(4): 571 - 594. (in Chinese)

[22] HAIGH K, ANDRUSENKO J. Cognitive electronic warfare: an artificial intelligence approach [M]. Boston: Artech House, 2021.

[23] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2015 - 03 - 20) [2022 - 10 - 01]. https://arxiv.org/abs/1412.6572.

[24] LYU C C, HUANG K Z, LIANG H N. A unified gradient regularization family for adversarial examples [C]// Proceedings of IEEE International Conference on Data Mining, 2015.

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

[26] KE D, HUANG Z T, WANG X, et al. Application of adversarial examples in communication modulation classification[C]//Proceedings of International Conference on Data Mining Workshops, 2019.

[27] KIM B, SAGDUYU Y, ERPEK T, et al. Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond [C]//Proceedings of IEEE Statistical Signal Processing Workshop, 2021.

[28] 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, 2021.

[29] KIM B, SAGDUYU Y, DAVASLIOGLU K, et al. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers [J]. IEEE Transactions on Wireless Communications, 2022, 21(6): 3868 – 3880.

[30] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[32] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016.