一种针对无线电信号分类的对抗增强方法

宣 琦[1], 崔 慧[1], 徐鑫杰[1], 陈壮志[1], 郑仕链[2], 王 巍[2], 杨小牛[1,2]

[1]浙江工业大学网络空间安全研究院 杭州 中国 310012
[2]通信信息控制和安全技术重点实验室 嘉兴 中国 314033

摘要 深度学习模型依赖大量带类标的数据作为训练数据, 实际应用的各种无线电环境中收集并标记无线电信号需要消耗大量的人力物力, 极大地限制了深度学习模型在无线电信号识别中的应用。目前针对数据量不足带来的问题, 研究者们主要采用数据增强的方法, 即根据一些先验知识, 在保持已知信息的前提下, 对原始数据进行适当变换达到扩充数据集的效果。具体到分类任务, 在保持数据类别不变的前提下, 可以对训练集中的每个样本进行变换, 如在一定程度内的随机旋转、缩放、裁剪、左右翻转等, 这些变换对应着同一个目标在不同角度的观察结果, 并且增强效果有限。此外, 深度学习作为一个非常复杂的方法, 会面对各种安全问题。深度神经网络很容易受到对抗样本的攻击, 攻击者可以通过向良性数据中添加特定的扰动, 生成对抗样本, 使 DNN 模型出错。虽然这些伪造的样本对人类的判断没有影响, 但是对于深度学习模型来说是一个致命性的误导。聚焦到深度学习领域, 本论文提出一种针对无线电信号分类的对抗增强方法, 将对抗训练方法引入信号领域, 通过控制 *eps*、*iteration* 参数, 在数据集中添加算法精心设计的细微扰动生成靠近决策边界面的边界样本实现数据增强, 将边界样本与训练样本混合, 重新训练识别模型, 在提升模型识别精度的同时, 提升模型的防御能力。最终在多个分类模型、多个实际无线电信号数据集上的分类性能都有显著的提高, 同时防御性能也显著增强, 验证了本文提出的信号增强识别方法的有效性。

关键词 深度学习; 对抗训练; 调制识别; 数据增强
中图法分类号 TN92 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.05.10

An Adversarial Enhancement Method for Radio Signal Classification

XUAN Qi[1], CUI Hui[1], XU Xinjie[1], CHEN Zhuangzhi[1], ZHENG Shilian[2], WANG Wei[2], YANG Xiaoniu[1,2]

[1] Department of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310012, China
[2] Department of the Science and Technology on Communication Information Security Control Laboratory, Jiaxing 314033, China

Abstract Deep learning models have relied on a large number of labeled data as training data. Collecting and labeling radio signals in various environments require a lot of manpower and material resources, which will limit the application of deep learning models in radio signal modulation recognition. At present, in view of the problem caused by insufficient data quantity, researchers mainly use the method of data enhancement, that is, according to some prior knowledge, the original data is maintained with known information to achieve the effect of expanding the data set. Specifically for the classification task, on the premise of keeping the data category unchanged, each sample in the training set can be transformed, such as random rotation, zoom, cutting, left and right flipping to a certain extent. These transformations correspond to the observations of the same target at different angles, and the enhancement effect is limited. In addition, as a complex method, deep learning has faced various security problems. Deep neural networks are vulnerable to adversarial samples, and attackers can make the deep neural networks model wrong by adding specific perturbations to the benign data. Although these forged samples have no effect on human judgment, they are a fatal and misleading effect for deep learning models. Considering the importance of the problem, the adversarial enhancement method for radio signal classification was proposed. Adversarial training methods are introduced into the radio signal modulation recognition, by controlling the *eps*, *iteration* parameters, the data augmentation is achieved by adding subtle perturbations designed by the algorithm to generate the boundary sample near the decision boundary, the boundary sample was mixed with the training sample, retraining the radio signal classification model, which can improve the model's performance of defense while ensure the modulation recognition accuracy. Finally, the recognition accuracy are improved on multiple radio signal modulation recognition models and multiple radio signal data sets, and the defense performance is also significantly enhanced, which verifies the effectiveness of the signal data augmentation method proposed in this paper.

Key words deep learning; adversarial training; modulation recognition; data augmentation

通讯作者: 宣琦, 博士, 教授, Email:xuanqi@zjut.edu.cn。
本课题得到国家自然科学基金(No. U19B2016)资助。
收稿日期: 2021-03-16; 修改日期: 2021-05-26; 定稿日期: 2022-03-22

1 引言

目前, 无线电信号识别广泛应用于通信领域, 包括干扰识别[1]、频谱感知[2]和电子对抗[3]等。信号的调制识别问题本质上是一种典型的模式识别问题, 其目的是识别检测到的无线电信号的调制类型。早期的信号识别方法主要利用人工设计的特征进行信号判定, 精确度无法保证。深度学习技术具有从数据中自动学习“特征”的能力, 为解决信号识别这一问题提供了有效的方法, 越来越多的学者使用深度学习技术进行信号调制样式识别的研究[4-7]。然而, 在深度学习领域, 绝大部分的模型需要大量数据进行训练和学习, 随着更多应用场景的涌现, 我们越来越面临着样本数量不足的问题[8]。在信号识别中, 典型的挑战是在真实存在的各种环境中, 标记数据困难, 因此样本集有限, 不能很好地建立一个可靠的数据库, 样本集将直接影响训练效果[9]。

此外, 近几年来随着人工智能浪潮再次涌来, 研究者发现深度神经网络(deep neural networks, DNN)很容易受到对抗样本的攻击, 攻击者可以通过向良性数据中添加特定的扰动, 生成对抗样本, 使 DNN 模型出错[10-14]。虽然这些伪造的样本对人类的判断没有影响, 但是对于深度学习模型来说是一个致命性的误导。最近, 在图像领域成功实施的一系列对抗性攻击[15-17]证明了此问题是所有基于深度学习系统的安全隐患。图像数据一般为二维数据, 而信号数据一般为一维时序数据, 虽然它们的本质含义有着很大的差别, 但是表现形式却类似。因此, 可以将图像领域的攻击方法应用于无线电信号领域, 对抗防御技术的研究引起了通信领域学者们越来越多的关注。Sadeghi M 等人[18]解决了传统的快速梯度符号(fast gradient sign method, FGSM)算法存在的粗颗粒扰动且攻击成功率低的缺点, 并用二分法搜索确定最佳扰动系数, 在公开的无线电数据集中取得了良好的攻击效果; Y. Shi 等人[19]提出了一种对抗式的机器学习方法对无线电通信进行干扰攻击, 该方法可以支持攻击者根据检测结果对传输进行可靠的预测, 并有效地进行干扰传输。

针对当前无线电信号识别中部分信号样本数据量少导致基于深度学习的模型无法有效工作、存在安全隐患等问题, 本文提出了一种针对无线电信号分类的对抗增强方法, 通过在数据集中添加算法精心设计的细微扰动生成靠近决策边界面的边界样本实现数据增强, 同时, 在保证模型分类精度增强的情况下, 提升模型的防御能力。

本文主要的研究工作如下。

1) 提出了一种针对无线电信号分类的对抗增强方法。通过比较近些年基于深度学习的信号分类模型, 均衡分类精度与时间复杂度, 选取 1D-ResNet、2D_cnn 信号分类模型[4]。将对抗训练方法引入信号领域, 通常对抗训练在增强模型抗干扰能力的同时会降低模型的分类精度, 本文通过引入 *eps*、*iteration* 参数, 在样本上添加算法精心设计的细微扰动生成边界样本, 实现数据增强, 提升模型的分类准确率。

2) 提高模型的防御能力。将筛选后的逼近但未越过决策边界的边界样本与原始样本一并放入模型中训练, 增强模型的鲁棒性。使用 FGSM 对抗攻击算法攻击增强前后的模型, 生成对抗样本, 测试攻击成功率, 验证对抗增强方法的防御性能。

3) 利用多个模型、实际无线电信号数据集验证本文提出的无线电信号对抗增强方法的有效性, 同时探究方法的防御能力并分析结果。并与传统数据增强方法进行对比, 在多数情况下我们提出的方法均能实现增强且总体增强效果最佳, 同时模型对对抗样本的鲁棒性也显著增强, 验证了本文提出的针对无线电信号分类的对抗增强方法效果良好。

2 相关工作

2.1 信号识别

早期的信号识别方法主要利用人工设计的特征进行信号判定, 成本高且效率低下, 精确度无法保证。目前常采用的是特征提取方法, 通过提取与分类识别紧密相关的特征表示原始数据, 使特征在分类器上发挥更好的作用, 比如瞬时幅度、频率、星座图[20]、高阶累积量[21]等, 后续利用机器学习的分类器对信号进行识别分类。随着现代通信环境日渐复杂、新型通信方式层出不穷, 无线电信号呈现出海量、多维、动态等一些非结构化的特点, 使得基于专家先验知识的无线电信号特征提取和分类识别的应用面临着诸多的困难与局限。

近年来, 随着深度学习方法在生物医学[22]、语音识别[23]以及图像分类[24]等领域的成功应用, 深度学习技术因通过数据而学习“特征”的能力, 受到众多人士的热爱。深度学习的发展, 为解决信号识别这一问题提供了有效的方法, 研究者们已将多种深度学习算法应用于信号领域, 端到端的方式在信号识别领域的应用越来越多, 同时整体的方法也变得更加成熟。爱丁堡大学的 O’Shea 教授是较早使用深度学习技术进行端到端的调制识别研究的学者之一, 如文献 [4-5] 将不同的调制类型产生的同向正交

(in-phase and quadrature, IQ)数据作为单通道的图像数据进行处理, 并利用搭建的 DNN、改进的残差堆栈单元对信号调制类型进行识别; 文献[6]将 I、Q 数据的幅值与相位信息输入网络为两层, 节点数为 128 的长短期记忆网络(long short term memory, LSTM)网络提取信号特征, 并直接与全连接层相连, 实现对 11 类调制类型的识别; 文献[25]采用深度信念网络对认知无线电网络中的用户进行分类, 显著减少标记数据数量, 提高分类精度等级, 从而减少无线电中信道切换的时间; 文献[26]通过搭建 CNN_LSTM 网络捕获信号采样点之间的信息, 实现对无线电信号的分类。

2.2 数据增强

在机器学习中, 绝大部分的模型需要大量数据进行训练和学习, 训练样本的数量在实际应用中非常重要, 然而在现实世界中, 无线电信号的样本集有限并且标注样本困难, 无线电信号识别技术因此受到限制[9]。

目前针对数据量不足带来的问题, 研究者们主要采用数据增强的方法, 即根据一些先验知识, 在保持已知信息的前提下, 对原始数据进行适当变换达到扩充数据集的效果。具体到分类任务, 在保持数据类别不变的前提下, 可以对训练集中的每个样本进行变换, 如在一定程度内的随机旋转[27]、缩放、裁剪、左右翻转等, 这些变换对应着同一个目标在不同角度的观察结果; 对样本中的像素参加噪声扰动、比如椒盐噪声、高斯白噪声[27]; 还可以改变图像的亮度、清晰度、对比度、锐度等。除了上述启发式的变换方法, 近年来, 通过生成式对抗网络[28](generative adversarial networks, GAN)生成数据得到了越来越多的关注。GAN 可以生成像素级、复杂分布的逼真图像, 在计算机视觉领域取得巨大的成功[29]。Tang 等人[30]将调制信号数据集预处理, 转换为可以表征样本点密度信息的星座图, 通过 GAN 生成带标签的星座图样本从而扩充数据集、Harada 等人[31]将循环卷积网络与 GAN 结合生成时间序列数据、Zhang 等人[32]通过 GAN 生成脑电信号, 提高分类准确性。

然而, GAN 的生成器与判别器间的博弈思想使其训练很不稳定, 往往需要使用很多训练技巧才可以生成有效的样本, 在不同数据集上的泛化效果不佳, 并且通常情况下需要大量的数据才能训练好 GAN 模型, 对于实际应用中的小样本数据集并不能很好的适用。

2.3 对抗训练

深度网络在图像识别和特征表示方面取得了成功, 但它们往往对输入图像的微小扰动非常敏感。添加视觉上不可感知的噪声会导致图像分类失败, 这些添加噪声的图像, 通常被称为对抗样本[10]。由于输入数据的高维性和 DNN 的线性特性, 深度模型对对抗样本非常敏感, 使得决策边界在高维空间中容易受到攻击[11]。因此有了对抗训练的概念, 利用对抗样本修正深度模型, 提高模型的抗干扰能力[33]。虽然没有给出理论证明, 但研究表明对抗训练在现阶段是对抗攻击最有效的防御手段之一。

对抗训练即在每次迭代训练时, 通过向训练集中加入对抗性的样本对模型参数进行再训练。在实践中, 人们可以使用不同的方法来生成对抗样本, 如 FGSM[10]、投影梯度下降法(Projected Gradient Descent, PGD)[12]、基本迭代法(Basic Iterative Method, BIM)[11]等。Goodfellow 等人[10]首先提出对抗训练, 他们使用良性样本和通过 FGSM 算法生成的对抗样本一起训练神经网络, 增强神经网络的鲁棒性; 接着, 提出了使用由 PGD 算法生成的对抗样本进行对抗训练的方法[11]。根据实验结果, PGD 对抗训练可在 MNIST、CIFAR-10 和 ImageNet 等多个数据集上获得最高的准确度。但是由于生成 PGD 对抗样本需要大量的计算成本, 因此 PGD 对抗训练不是一种有效率的防御措施。此外, Szegedy 等人[33]表明将对抗样本和原样本混合放入模型训练能够使模型正则化。

3 方法

3.1 整体框架

针对基于深度学习模型的无线电信号分类系统依赖大量带类标的信号数据作为训练数据, 而实际应用的各种无线电环境中收集并标记信号需要消耗大量的人力物力, 深度学习系统的决策边界在高维空间中容易受到攻击等安全隐患问题, 本文提出一种针对无线电信号分类的对抗增强方法。

具体而言, 如图 1 的方法框架所示, (1)将信号数据集进行预处理, 并按 2:1 的比例划分训练集与测试集; (2)使用训练集预训练分类模型; (3)根据现有的模型决策边界, 通过引入 *eps*、*iteration* 参数, 使用 BIM 梯度攻击的方法对预训练好的模型进行攻击生成边界样本; (4)通过样本筛选机制, 保存梯度攻击成功前(逼近但未越过决策边界)的对抗样本, 将其归入边界样本集合中, 保存攻击成功(逼近且越过决策边界)的对抗样本, 将其归入强对抗样本集合中; (5)将边界样本与训练样本混合, 重新训练分类模型, 在保证原始模型精度增强的前提下提高模型的防御能力。实验中, 对数据增强前后的模型通过 FGSM 对抗攻击

算法测试攻击成功率。

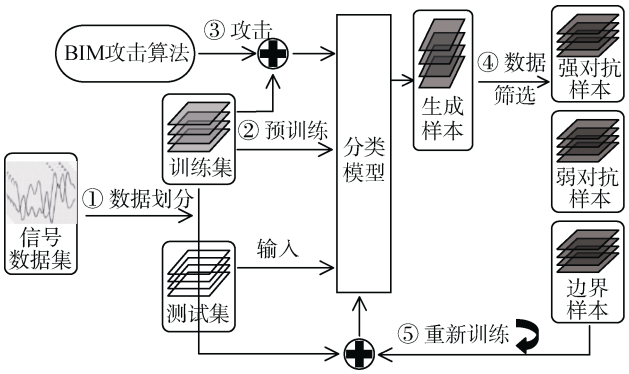


图 1 方法整体框架

Figure 1 The framework of the method

3.2 分类模型

本文使用 O'Shea 等人搭建的 2D_cnn[5]、1D_resnet[4]模型进行信号的调制识别。

2D_cnn 模型由两个卷积层和两个全连接层的 4 层网络构成, one-hot 输出层使用 softmax 激活, 其余每层均使用缩放指数线性单位(SELU)激活函数。1D_resnet 模型在残差块的基础上, 构建了残差堆栈单元, 每个残差堆栈单元包括一个用于在通道维度上做计算的卷积核大小为 1×1 的卷积层, 此外还包含两个残差块, 以及一个最大池化层。整个模型中包含 6 个残差堆栈单元, 数据每经过一个残差堆栈单元维度减半, 模型中所有卷积层的卷积核数量均为 32。在网络的全连接层中采用自归一化神经网络, SELU 激活函数, 平均响应缩放初始化(MRSA)和 Alpha Dropout。

3.3 对抗增强

3.3.1 样本生成机制

根据深度学习的流形假设, 样本空间是一个非常高维的空间, 但是我们所能掌握的有效样本其实在一个维度远远低于高维度样本空间的一个流形(manifold)空间内[34]。以经典的二分类问题为例, 深度学习模型通过在样本上训练, 学习出一个决策边界平面, 在决策边界平面的一侧的点都被识别为类别一, 在决策边界平面的另外一侧的点都被识别为类别二。

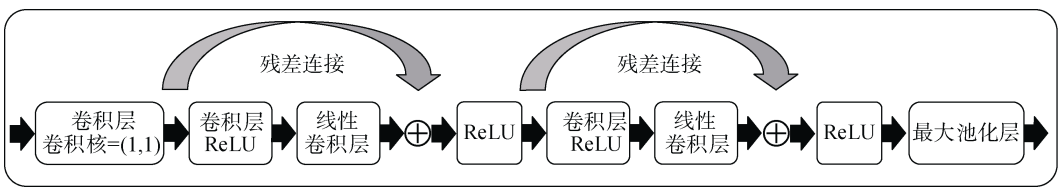


图 2 残差堆栈单元

Figure 2 The residual stack unit

本文结合攻击算法, 获取信号分类器的模型参数, 对训练样本进行梯度方向的调整。需要注意的是, 一般情况下攻击步长(能容忍的攻击范围)越大攻击越强, 这与数据增强目的相违背, 为了使生成的样本更加靠近原样本的决策边界面, 本文通过 *eps*、*iteration* 参数调节添加的扰动大小。其中, *eps* 可以控制每次添加的扰动, 使得边界样本距离决策边界面尽可能近, *iteration* 可以调节最终的扰动大小, 使得边界样本限制在一定范围内, 防止添加的扰动过大, 生成的样本抑制模型的性能。具体而言, 实验中确定对样本的目标攻击标签(其概率分布中排第二的标签), 在扰动前添加一个较小的系数 *eps* 减小扰动, 随后通过迭代形式在输入样本上有方向的添加扰动, 本文将攻击成功为目标标签的样本称为弱对抗样本, 弱对抗样本前一次迭代的样本为边界样本。边界样本非常接近于该任务的决策边界但未越过决策边界, 在有效增强模型的同时还可以增加模型的泛化能力。同时, 为了避免生成一个样本时间太长, 引入 *iteration* 参数控制迭代次数。强对抗样本即为通常意义的对抗样本, 与边界样本生成方式相同, 但不使用 *eps* 参数。

论文最终选用 BIM 攻击算法[11]生成信号样本。BIM 是 FGSM 方法[10]的变体, FGSM 为非迭代型白盒攻击方法, 通过保持扰动方向与梯度方向一致, 使损失函数值变化最大, 进而使分类器分类结果变化最大。FGSM 只需一步迭代就能生成对抗样本, 并且可以通过控制超参数ε生成任意无穷范数距离的对抗样本, 攻击表达式为:

$$X_{adv} = X + \varepsilon \cdot sign(\nabla_X J(X, Y)) \tag{1}$$

其中, X、X_{adv} 和 Y 分别表示原始样本、对抗样本和原始样本对应的标签, $Sign()$、∇_X、J 分别为符号函数、梯度函数以及损失函数。但是 FGSM 扰动自身抗干扰能力不强, 容易受到其他噪声的影响, 模型损失函数与模型输入并不是完全线性的, 这说明该算法生成的对抗样本扰动不是最优扰动。针对 FGSM 算法存在的问题, Goodfellow 等人[11]在 FGSM 算法基础

上提出了一种以多步迭代的方式生成对抗样本的方法 BIM。

3.3.2 样本筛选机制

边界样本需要与原样本标签一致, 在极少数情况下, 前一次样本可能被攻击到其他边界, 然后在最后一次攻击到目标类别边界, 因此需要删去边界样本预测标签与真实标签不同的样本。在经过初步筛选后, 对原样本、边界样本、弱对抗样本、对抗样本进行 L_2 距离计算, 验证生成的样本是否符合要求, 每类样本的距离要保证一定的区分度, 理论上来说, 边界样本与原样本的距离要远远小于其他类别样本与原样本的距离。

DNN 分类器生成的预测向量通常表示输入样本属于每个可能类别的概率分布。因此, 比较模型的原样本预测和生成样本的预测涉及到比较之间的概率分布向量。比较概率分布的方法有很多种, 如 L_1 范数、L_2 范数和 KL 散度等。本文选择 L_2 范数作为原样本预测向量与生成样本预测向量之间差异的自然度量:

$$L^{(V_{oi},V_{ai})}=\left\|g(V_{oi})-g(V_{ai})\right\|_2 \tag{2}$$

$g(v)$是 softmax 层生成的 DNN 模型的输出向量。L_2 分数越高, 意味着原样本预测与生成样本之间的差异越大。V_{oi} 为原始样本第 i 个数据的数值, V_{ai} 为生成样本第 i 个数据的数值。我们期望原样本预测向量与边界样本预测向量的差异最小。

4 实验结果与分析

4.1 实验设置

1) 实验平台: 本文中涉及实验的实验环境具体配置如下: i7-7700K 4.20GHzx8 (CPU), TITAN Xp 12GiBx2 (GPU), 16GBx4 DDR (内存), Ubuntu 16.04 (OS), Python 3.5, Tflearn-0.3.2, Tensorflow-gpu-1.3。

2) 数据集: 本文在三个无线电信号数据集上对所提的方法进行评估。**RML2016.10a**: 数据集为爱丁堡大学公开的调制信号数据集[5], 它使用 GNU Radio 合成信号样本。它包含 11 种调制类型, 信号信噪比(Signal-to-noise ratio, SNR)范围从–20dB 到 18dB, 间隔 2dB 均匀分布。每个无线电信号样本的采样点数为 128。训练集样本数为 176000, 测试集样本数为 44000。**2018.01.OSC**: 数据集为爱丁堡大学公开的调制信号数据集[4]。它包含 24 种调制类型, 信号信噪比(Signal-to-noise ratio, SNR)范围从–20dB 到 30dB, 间隔 2dB 均匀分布。每个无线电信号样本的采样点数为 1024。训练集样本数为 2040000, 测试集样本数为 510000。**Sig2019-12:** 数据集由团队仿真生成, 并在仿真时考虑了一个真实通信系统存在的一些影响。它包含 12 种调制类型: BPSK、QPSK、8PSK、OQPSK、2FSK、4FSK、8FSK、16QAM、32QAM、64QAM、4PAM 和 8PAM。原始信息数据以随机方式产生, 从而保证传输比特等概率取值。脉冲成形滤波器采用升余弦滤波器, 滚降系数在[0.2, 0.7]范围内随机取值。相位偏差在[–π, π]范围内随机取值, 归一化载波频偏(相对于采样频率)在[–0.1, 0.1]范围内随机选择。SNR 范围从–20dB 到 30dB, 间隔 2dB 均匀分布。每个数据样本包含 64 个符号, 每个符号的采样点数为 8, 因此每个样本的采样点数为 512。训练集和测试集的大小分别为 312000 和 156000, 每类调制信号样本量相同。实验中, 首先对数据进行最大最小归一化预处理, 并将信号数据划分为训练集与测试集, 相同性噪比的每类调制类型仅有 1000 个训练集样本。

4.2 对抗增强

本文选取 RML2016.10a 数据集的–10dB、0dB、10dB、18dB 数据, 2018.01.OSC、Sig2019-12 数据集的 0dB、10dB、18dB、30dB 数据进行实验。首先, 训练原数据集的分类模型, 保存最佳模型。通过目标攻击方法, 不断迭代添加扰动(扰动限制 5%以内), 设置 *eps* 为 0.004~0.006 减小扰动, 使样本更加靠近决策边界面, 迭代次数 *iteration* 小于 60。攻击完成后, 删去边界样本预测标签与真实标签不同的样本, 将筛选后的数据保存为边界样本。最后, 将边界样本与原数据集混合, 重新训练分类模型, 得到增强识别的分类精度。为了保证公平性, 增强前后分类模型的网络结构、测试集、学习率、训练次数等参数将保持一致。

由图 3~4 的时域波形图可知, 为了保证与原数据集的特征相似性, 原样本与生成的边界样本在时域的波形图肉眼几乎不可分辨, 强对抗样本与原样本的波形图有一定的区分度, 主要体现在波峰波谷的扰动。为了验证每类样本的距离有一定区分度, 经过初步筛选之后, 计算四类样本的概率分布, 对四类样本进行平均 L_2 距离计算。表 1 展示了 Sig2019-12 数据集的 10dB 样本使用 1D_resnet 模型, *eps* 为 0.001, *iteration* 为 50~100 时的不同样本间的距离。可以看到, 原样本与边界样本间的距离远小于原样本与强对抗样本间的距离。边界样本与原样本的距离最小, 验证了使用边界样本增强的合理性。

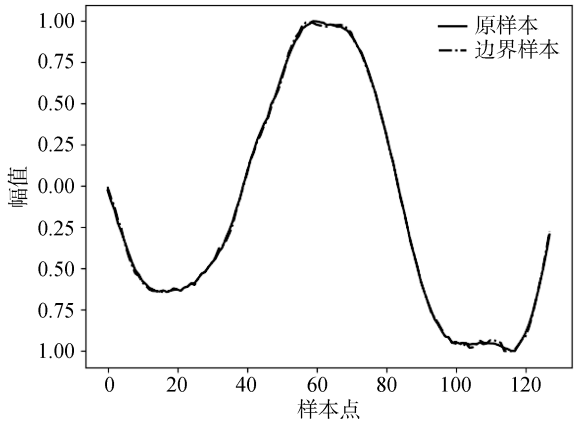


图 3　原样本与边界样本的波形图

Figure 3　The waveform diagram of original sample and boundary sample

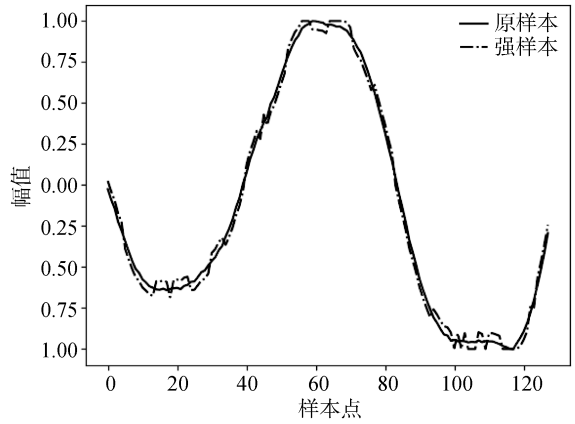


图 4　原样本与强对抗样本的波形图

Figure 4　The waveform diagram of original sample and strong adversarial example

表 1　样本间的距离

Table 1　The distance between samples

Iteration	边界-弱对抗	边界-强对抗	原始-边界	原始-强对抗
50	0.7820	1.0974	0.3167	1.3379
60	0.7819	1.0976	0.3169	1.3374
70	0.7819	1.0977	0.3172	1.3374
80	0.7820	1.0972	0.3165	1.3373
90	0.7816	1.0971	0.3169	1.3372
100	0.7819	1.0973	0.3167	1.3374

我们定义了精度上的相对改进率(RIMP)如下:

$$RIMP = (Accen - Accori)/Accori \tag{3}$$

其中, *Accen* 和 *Accori* 分别指增强模型和原始模型的分类精度。在实验中, 我们比较了一些传统的数据增强方法, 包括基于旋转的增强, 基于镜像的增强和基于高斯噪声的增强[27], 在表 2 中, 每种方法的最后一列给出了精度的相对改进率, 并将同一数据集、模型下的最高相对改进率加粗显示。

由表 2 结果可知, 对抗增强方法得到的总体增强效果最佳, 除了高信噪比数据的分类精度略有下降, 大多数情况下可以实现增强效果, 3 个数据集均在较低性噪比的样本下达到最高的增强识别效果。通常情况下, 性噪比越高, 加扰动的信号越容易被分类模型识别, 不利于增强, 高斯噪声增强与对抗增强的实验结果均验证了这一理论。具体而言, 针对不同的分类模型, 增强方法效果存在差异, 对抗增强方法在 1D_resnet 模型上表现良好, 在不同信噪比、3 个数据集的 12 种情况下, 7 种情况都达到了最高的相对改进率。高斯噪声方法增强效果较差, 大部分情况无法增强。对抗增强与高斯噪声增强都作为添加扰动的方法, 对抗增强比起添加高斯噪声的方法更加有效, 此实验结果也验证了高斯噪声产生的新样本只是随机分布在原数据集的某一个样本的附近, 而对抗攻击得到的边界样本处于模型的决策面附近, 使得模型更容易搜索到真实的决策边界。旋转增强方法与镜像增强方法的增强效果略低于对抗增强, 值得注意的是, 这两种方法在 2D_cnn 模型上表现良好, 作为一种合理的解释, 2D_cnn 模型的鲁棒性较好, 从而影响了对抗增强方法的适用性。不同于高斯噪声, 旋转与镜像对于原样本的改变较大, 从而让新生成的样本分布的更加均匀以此达到数据增强的效果。这些结果表明, 对抗增强方法在信号增强领域可以发挥积极的作用。

4.3　防御能力

为了验证所提方法的防御性能, 我们分别对增强前后的模型通过 FGSM 对抗攻击算法生成对抗样本, 测试攻击成功率。实验中设置攻击步长为 0.05, 同时以攻击成功率(预测标签与真实标签不同即为攻击成功)作为指标测试不同模型的防御性能。如表 3 所示, 我们在表中列出了模型增强前后的攻击成功率, 并将防御性能最好的结果加粗显示。可以发现, 使用对抗增强方法的模型有良好的防御性能, 在多数情况下攻击成功率最低, 且大部分对度增强效果相反, 多数实验结果呈现性噪比越高防御性能越好。对抗增强方法在 Sig2019-12 数据集、2D_cnn 分类模型的情况下增强效果不佳, 与之相反的是防御性能提升。

4.4　小样本

为了验证方法适用于更小的数据集, 均匀抽取 RML2016.10a、2018.01.OSC、Sig2019-12 数据集不同信噪比的 10%样本进行实验, 此时 RML2016.10a、2018.01.OSC、Sig2019-12 数据集每类样本数量分别

表 2 模型增强识别的实验结果

Table 2 The experimental results of enhancement recognition

数据集	模型	信噪比(dB)	原模型分类精度(%)		旋转增强分类精度(%)			镜像增强分类精度(%)			高斯噪声增强分类精度(%)			对抗增强分类精度(%)		
			训练集	测试集	训练集	测试集	RIMP	训练集	测试集	RIMP	训练集	测试集	RIMP	训练集	测试集	RIMP
RML2016.10a	**1D_resnet**	–10	38.82	29.73	53.86	27.45	–7.67	81.75	29.18	–1.85	58.49	33.00	**11.00**	43.84	32.86	10.53
		0	84.09	80.68	97.52	79.50	–1.46	99.03	80.27	–0.51	98.21	80.09	–0.73	98.85	83.09	**2.99**
		10	93.03	84.45	94.43	84.23	–0.26	94.64	84.27	–0.21	87.56	83.09	–1.61	95.27	87.41	**3.51**
		18	94.51	**84.95**	93.98	84.09	–1.01	94.46	84.09	–1.01	94.88	84.05	–1.06	95.68	84.09	–1.01
	2D_cnn	–10	91.84	26.00	61.56	25.50	–1.92	75.78	23.86	–8.23	47.35	26.32	1.23	38.66	27.18	**4.54**
		0	98.91	69.77	99.44	73.41	**5.22**	99.16	72.23	3.53	99.37	69.45	–0.46	93.26	68.45	–1.89
		10	96.70	77.27	95.62	80.36	**4.00**	96.33	80.36	**4.00**	97.67	78.59	1.71	95.15	78.68	1.82
		18	94.91	77.45	95.03	79.05	**2.07**	94.95	78.95	1.94	94.97	78.95	1.94	95.50	78.72	1.64
2018.01.OSC	**1D_resnet**	0	62.27	52.70	60.24	53.07	0.70	63.06	53.70	**1.90**	65.29	51.12	–3.00	63.51	52.82	0.23
		10	96.50	92.31	97.96	92.40	0.10	97.73	92.20	–0.12	96.11	90.75	–1.69	98.16	93.02	**0.77**
		18	96.70	95.32	98.42	96.33	**1.06**	98.61	96.29	1.02	97.42	95.72	0.42	98.24	96.06	0.78
		30	95.76	94.65	98.50	96.73	2.20	98.55	96.67	2.13	97.20	95.21	0.59	98.69	97.14	**2.63**
	2D_cnn	0	69.58	30.75	51.84	37.76	22.80	61.89	38.18	**24.16**	74.75	33.78	9.85	65.42	37.37	21.53
		10	75.97	64.38	77.00	66.70	3.60	82.60	65.82	2.24	73.84	65.28	1.40	88.11	67.45	**4.77**
		18	88.99	66.59	93.86	67.20	0.92	90.60	69.98	**5.09**	87.55	66.17	–0.63	88.88	69.39	4.20
		30	85.27	68.28	80.79	68.81	**0.78**	87.62	63.79	–6.58	81.57	66.86	–2.08	84.95	67.92	–0.53
Sig2019–12	**1D_resnet**	0	53.37	42.07	44.63	41.13	–2.23	52.16	41.07	–2.38	48.02	42.97	2.14	58.32	48.40	**15.05**
		10	97.51	84.83	99.30	85.33	0.59	99.94	85.37	0.64	99.72	85.67	0.99	98.18	89.67	**5.71**
		18	99.84	**93.37**	99.98	92.83	–0.58	99.80	91.03	–2.51	99.93	91.77	–1.71	99.29	92.87	–0.54
		30	99.66	93.57	99.64	93.20	–0.40	99.80	93.60	0.03	99.69	93.37	–0.21	99.79	97.07	**3.74**
	2D_cnn	0	100	21.73	94.68	24.70	13.67	99.98	27.43	**26.23**	100	20.13	–7.36	72.32	25.33	16.57
		10	66.03	51.03	99.98	51.90	**1.70**	99.97	49.63	–2.74	100	49.07	–3.84	100	47.60	–6.72
		18	100	63.47	99.99	67.60	6.51	100	69.67	**9.77**	100	63.87	0.63	99.95	60.17	–5.20
		30	99.99	64.77	99.99	71.77	**10.81**	99.98	71.07	9.73	99.97	58.40	–9.83	99.99	63.33	–2.22

表 3 防御性能测试

Table 3 The defense performance testing

数据集	模型	信噪比(dB)	原模型攻击成功率(%)	旋转增强攻击成功率(%)	镜像增强攻击成功率(%)	高斯噪声增强攻击成功率(%)	对抗增强攻击成功率(%)
RML2016.10a	1D_resnet	–10	69.72	76.55	**51.99**	77.61	74.19
		0	49.86	**33.64**	59.69	47.17	33.83
		10	27.05	**23.17**	23.45	31.69	29.98
		18	**17.56**	29.28	21.20	18.26	23.40
	2D_cnn	–10	81.80	83.97	85.53	77.98	**70.09**
		0	44.40	44.23	49.68	**42.08**	44.69
		10	11.53	13.35	11.72	13.28	**8.00**
		18	5.14	5.14	4.97	4.89	**2.95**
2018.01.OSC	1D_resnet	0	76.18	76.80	**74.68**	77.47	76.95
		10	42.35	36.80	38.72	55.47	**31.33**
		18	73.08	63.95	**56.25**	60.48	69.43
		30	79.69	71.58	68.60	72.36	**66.13**
	2D_cnn	0	82.00	83.03	82.96	83.78	**77.84**
		10	74.28	**69.82**	74.94	76.28	74.34
		18	70.93	73.21	73.57	**69.10**	71.44
		30	71.24	79.86	**70.27**	71.36	78.30

续表

数据集	模型	信噪比 (dB)	原模型 攻击成功率(%)	旋转增强 攻击成功率(%)	镜像增强 攻击成功率(%)	高斯噪声增强 攻击成功率(%)	对抗增强 攻击成功率(%)
Sig2019-12	1D_resnet	0	93.21	**84.44**	96.38	86.91	89.21
		10	66.07	64.14	**62.70**	65.64	62.88
		18	69.99	62.93	73.87	67.74	**53.47**
		30	74.51	70.43	73.57	73.19	**63.66**
	2D_cnn	0	99.54	99.78	99.63	99.18	**98.78**
		10	99.33	98.97	**97.54**	97.64	98.72
		18	97.02	96.28	98.73	95.90	**94.06**
		30	97.85	95.20	93.12	97.00	**91.97**

仅有 80、409、100 个。实验中使用相同的测试集测试增强前后模型的分类性能, 同时对增强前后的模型通过 FGSM 对抗攻击算法生成对抗样本, 测试攻击成功率。

由表 4 可知, 多个小样本数据集的增强识别效果良好, 大部分模型的攻击成功率下降, 验证了方法可适用于更小的数据集, 这在实际应用中有着很大的意义。具体而言, 对抗增强方法对于 RML2016.10a 小样本数据集适应性良好, 全部模型可以增强; 在2018.01.OSC小样本数据集上表现较差, 只增强了少数模型; Sig2019-12 小样本数据集的部分 2D_cnn 模型分类精度下降, 与之相反的是防御性能有较大的提升。我们猜想, 增强识别方法更适用于样本的采样点数较小、样本长度较短的数据集。

5 讨论与分析

本文将对抗训练方法引入信号领域, 通过控制 *eps*、*iteration* 参数, 在样本上添加算法精心设计的细微扰动, 生成边界样本, 在增强模型抗干扰能力的同时提高模型的分类精度。

表 4 小样本识别精度

Table 4 The recognition accuracy of small samples

数据集	模型	信噪比 (dB)	原模型 分类精度(%)	对抗增强 分类精度(%)	RIMP(%)	原模型 攻击成功率(%)	对抗增强 攻击成功率(%)
RML2016.10a	1D_resnet	–10	24.64	26.55	7.75	78.75	80.11
		0	58.09	62.05	6.82	67.95	49.89
		10	71.74	72.95	1.69	34.66	25.34
		18	70.18	70.86	0.97	33.86	19.20
	2D_cnn	–10	20.50	20.82	1.55	59.55	59.00
		0	42.73	42.82	0.21	53.86	59.32
		10	55.45	56.41	1.72	12.84	11.02
		18	57.41	57.55	0.24	7.38	5.80
2018.01.OSC	1D_resnet	0	22.60	29.20	29.20	96.25	92.92
		10	56.77	50.97	–10.22	87.50	86.50
		18	63.73	55.20	–13.39	77.83	70.08
		30	75.37	71.13	–5.62	73.25	70.75
	2D_cnn	0	12.53	12.20	–2.66	99.33	100
		10	18.27	17.63	–3.47	97.67	76.50
		18	18.17	18.73	3.12	85.75	99.92
		30	19.43	17.67	–9.09	67.17	73.17
Sig2019–12	1D_resnet	0	30.03	31.20	3.90	84.07	86.02
		10	56.63	56.27	–0.64	70.86	76.12
		18	67.63	74.63	10.35	75.76	69.41
		30	75.37	78.87	4.64	65.69	75.06
	2D_cnn	0	27.92	25.32	–9.33	81.19	77.61
		10	48.13	49.01	1.83	74.03	69.93
		18	54.96	46.15	–16.04	75.01	65.20
		30	44.52	47.74	7.24	57.87	56.35

传统的数据增强方式, 如对训练集中的每个样本进行随机旋转、左右翻转[27], 这些变换对应着同一个目标在不同角度的观察结果, 增强效果有限。目前使用 GAN 实现信号数据增强的方法引起了越来越多的学者注意[30-32], 然而GAN的训练很不稳定, 往往需要大量的训练技巧。本文所提方法训练稳定, 仅添加少量样本即可达到较好的增强效果, 但是也有不足之处, 如对 Sig2019-12 小样本数据集的部分 2D_cnn 模型没有起到增强作用, 仅大幅提升了防御性能。

6 结束语

本文提出了一种针对信号分类的对抗增强方法, 针对当前无线电信号识别中样本数据量少导致深度学习模型无法有效工作、深度学习系统的决策边界在高维空间中容易受到攻击等安全隐患问题, 利用对抗训练思想, 通过在数据集中添加算法精心设计的细微扰动实现数据增强, 在保证模型识别精度提升的情况下, 增加模型的防御能力。实验结果表明, 本文提出的方法具有较好的增强性能和防御性能, 提升多个无线电信号数据集模型分类精度的同时模型鲁棒性也显著增强。

未来工作可以考虑在实际应用中添加少量的弱对抗样本进一步的提高模型的鲁棒性, 根据实际需求, 权衡模型的分类精度与防御能力, 选取适当的边界样本与弱对抗样本。还可以使用多类攻击算法生成更多种类的对抗样本用于对抗训练, 从而提高模型鲁棒性。

参考文献

[1] Zhang X, Hu J H. Blind Interference Detection and Recognition for the Multi-Carrier Signal[J]. *The Journal of China Universities of Posts and Telecommunications*, 2017, 24(2): 48-56.

[2] Watson C M. Signal Detection and Digital Modulation Classification-Based Spectrum Sensing for Cognitive Radio[D]. Northeastern University Library, 2013. DOI:10.17760/d20003383.

[3] Iglesias V, Grajal J, Royer P, et al. Real-Time Low-Complexity Automatic Modulation Classifier for Pulsed Radar Signals[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2015, 51(1): 108-126.

[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] O'Shea T J, Corgan J, Clancy T C. Convolutional Radio Modulation Recognition Networks[C]. *Engineering Applications of Neural Networks*, 2016: 213-226.

[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] Yildirim Ö. A Novel Wavelet Sequence Based on Deep Bidirectional LSTM Network Model for ECG Signal Classification[J]. *Computers in Biology and Medicine*, 2018, 96: 189-202.

[8] Frid-Adar M, Klang E, Amitai M, et al. Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification[C]. *2018 IEEE 15th International Symposium on Biomedical Imaging*, 2018: 289-293.

[9] Dobre O A. Signal Identification for Emerging Intelligent Radios: Classical Problems and New Challenges[J]. *IEEE Instrumentation & Measurement Magazine*, 2015, 18(2): 11-18.

[10] Goodfellow I J , Shlens J , Szegedy C. Explaining and harnessing adversarial examples[EB/OL]. 2014: ArXiv Preprint ArXiv:1412.6572.

[11] Kurakin A , Goodfellow I , Bengio S. Adversarial examples in the physical world[EB/OL]. 2016: Computing Research Repository: 1607.02533.

[12] Madry A , Makelov A , Schmidt L , et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. 2017: Computing Research Repository:1706.06083.

[13] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[14] 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*, 2016: 372-387.

[15] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[16] Su J W, Vargas D V, Sakurai K. One Pixel Attack for Fooling Deep Neural Networks[J]. *IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828-841.

[17] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[18] Sadeghi M, Larsson E G. Adversarial Attacks on Deep-Learning Based Radio Signal Classification[J]. *IEEE Wireless Communications Letters*, 2019, 8(1): 213-216.

[19] Shi Y, Davaslioglu K, Sagduyu Y E. Generative Adversarial Network for Wireless Signal Spoofing[C]. *The ACM Workshop on Wireless Security and Machine Learning*, 2019: 55-60.

[20] Peng S L, Jiang H Y, Wang H X, et al. Modulation Classification Based on Signal Constellation Diagrams and Deep Learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(3): 718-727.

[21] Abdelmutalab A, Assaleh K, El-Tarhuni M. Automatic Modulation Classification Based on High Order Cumulants and Hierarchical Polynomial Classifiers[J]. *Physical Communication*, 2016, 21: 10-18.

[22] Shen D G, Wu G R, Suk H I. Deep Learning in Medical Image Analysis[J]. *Annual Review of Biomedical Engineering*, 2017, 19: 221-248.

[23] Qian Y M, Bi M X, Tan T, et al. Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, 24(12): 2263-2276.

[24] Xuan Q, Chen Z Z, Liu Y, et al. Multiview Generative Adversarial

Network and Its Application in Pearl Classification[J]. *IEEE Transactions on Industrial Electronics*, 2019, 66(10): 8244-8252.

[25] Mendis G J, Wei J, Madanayake A. Deep Learning-Based Automated Modulation Classification for Cognitive Radio[C]. *2016 IEEE International Conference on Communication Systems*, 2016: 1-6.

[26] 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.

[27] Huang L, Pan W J, Zhang Y, et al. Data Augmentation for Deep Learning-Based Radio Modulation Classification[J]. *IEEE Access*, 2019, 8: 1498-1506.

[28] I. Goodfellow, et al. Generative adversarial nets[J]. *in Proc. Adv. Neural Inf. Process. Syst*, 2014: 2672 – 2680.

[29] Yeh R A, Chen C, Lim T Y, et al. Semantic Image Inpainting with Deep Generative Models[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6882-6890.

[30] Tang B, Tu Y, Zhang Z Y, et al. Digital Signal Modulation Classification with Data Augmentation Using Generative Adversarial Nets in Cognitive Radio Networks[J]. *IEEE Access*, 2018, 6: 15713-15722.

[31] Haradal S, Hayashi H, Uchida S. Biosignal Data Augmentation Based on Generative Adversarial Networks[J]. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual International Conference*, 2018, 2018: 368-371.

[32] Zhang Q Q, Liu Y. Improving Brain Computer Interface Performance by Data Augmentation with Conditional Deep Convolutional Generative Adversarial Networks[EB/OL]. 2018: arXiv: 1806.07108[cs.HC]. https://arxiv.org/abs/1806.07108

[33] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB/OL]. 2014: ArXiv Preprint ArXiv:1312.6199.

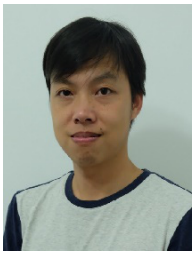
[34] Meng D Y, Chen H. MagNet: A Two-Pronged Defense Against Adversarial Examples[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 135-147.

宣琦 于 2008 年在浙江大学控制科学与工程专业获得博士学位。现任浙江工业大学网络空间安全研究院教授。研究领域为人工智能安全、网络数据挖掘、信号智能。研究兴趣包括: 人工智能、信号分析、网络科学。Email: xuanqi@zjut.edu.cn

崔慧 于 2019 年在浙江工业大学通信工程专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为深度学习、通信信号处理。研究兴趣包括: 人工智能、信号分析。Email: huicuizjut@qq.com

徐鑫杰 现在浙江工业大学自动化专业攻读学士学位。研究领域为深度学习、通信信号处理。研究兴趣包括: 人工智能、信号分析。Email: xxj1018@foxmail.com

陈壮志 于 2017 年在浙江工业大学电气工程及其自动化专业获得学士学位。现在浙江工业大学控制科学与工程专业攻读博士学位。研究领域为深度学习、机器视觉。研究兴趣包括: 人工智能、信号分析。Email: zzch@zjut.edu.cn



郑仕链 于 2014 年在西安电子科技大学通信与信息系统专业获得博士学位。现任通信信息控制和安全技术重点实验室副研究员。研究领域为认知无线电、压缩感知。研究兴趣包括: 深度学习、信号处理。Email: lianshizheng@126.com

王巍 于 2008 年在西安电子科技大学密码学专业获得博士学位。现任通信信息控制和安全技术重点实验室副主任, 研究员。研究领域为网络安全、网络通信。研究兴趣包括: 协议分析、人工智能。Email: wwzwh@163.com

杨小牛 于 1988 年在西安电子科技大学通信与电子系统专业获得硕士学位。现任通信信息控制和安全技术重点实验室主任。研究领域为通信信号处理与分析。研究兴趣包括: 软件无线电、智能信号处理、人工智能。Email: yxn2117@1126.com