引用格式：赵文硕, 张帅, 王轩瀚, 等. 面向对抗鲁棒的信号识别设计方法 [J]. 电子科技大学学报, 2025, 54(4): 618-625.
ZHAO W S, ZHANG S, WANG X H, et al. Design method for signal modulation recognition oriented to adversarial robustness[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(4): 618-625.

面向对抗鲁棒的信号识别设计方法

赵文硕，张　帅，王轩瀚*，宋井宽

（电子科技大学（深圳）高等研究院，深圳 518000）

摘要：深度学习在信号调制分类任务上取得了显著进展，然而，在实际应用中，深度神经网络已被证明存在内在脆弱性，容易受到对抗样本的攻击。对抗样本通过向输入添加细微扰动，致使模型产生错误的分类结果，给通信系统的安全性带来了严重的威胁与隐患。该文在对抗训练框架的基础上提出了一种防御方法：混合信号对抗训练（HSAT），以提高信号调制分类模型的鲁棒性。针对训练数据稀缺以及通过对抗样本训练所得网络表征能力不足的问题，提出一种基于线性插值的混合信号数据增强策略提升模型性能。同时，应用最大间隔损失函数替代交叉熵损失函数，增加模型决策边界距离，增强模型对于扰动输入的鲁棒性。通过对比当前先进的对抗攻击算法，验证该方法相较于传统对抗训练，在 3 种攻击算法上的对抗鲁棒性平均提升 7.07%，标准分类准确率仅下降 1.61%。

关键词：信号调制分类；对抗样本；鲁棒性；对抗训练
中图分类号：TP391.4　　**文献标志码：**A　　DOI: 10.12178/1001-0548.2024159

Design method for signal modulation recognition oriented to adversarial robustness

ZHAO Wenshuo, ZHANG Shuai, WANG Xuanhan*, and SONG Jingkuan

(Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518000, China)

Abstract: Deep learning has made significant progress in signal modulation classification tasks. However, in practical applications, deep neural networks have demonstrated inherent vulnerabilities, making them susceptible to adversarial attacks. Adversarial examples, created by adding subtle perturbations to inputs, can cause models to produce incorrect classification results, posing serious threats and risks to the security of communication systems. This paper proposes a novel defense method, Hybrid Signal Adversarial Training (HSAT), based on the adversarial training framework to enhance the robustness of signal modulation classification models. To address the issues of limited training data and insufficient network representation capabilities resulting from adversarial training, a mixed signal data augmentation strategy based on linear interpolation is proposed to improve model performance. Additionally, a maximum margin loss function is employed to replace the cross-entropy loss function, thereby increasing the distance of the model's decision boundaries and enhancing robustness against perturbed inputs. Through validation against current state-of-the-art adversarial attack algorithms, the proposed method demonstrates an average improvement of 7.07% in adversarial robustness across three attack algorithms, with only a 1.61% decrease in standard classification accuracy compared to traditional adversarial training.

Key words: signal modulation classification; adversarial examples; robustness; adversarial training

信号调制分类作为无线通信系统中的关键技术之一，对于通信安全、频谱监测管理以及现代电子战等领域具有至关重要的作用。随着技术的更新迭代，深度学习因其在处理复杂数据和提取高维特征方面的优势而被广泛应用于信号调制分类任务中。相较于传统方法，深度学习能够自动学习数据的内在规律和特征，无须依赖先验知识和人工特征，从而极大地提高调制分类的准确性和效率[1]。近年来，深度模型的引入满足了日益增长的通信系统对高效、可靠信号识别的需求，也推动了无线通信技

收稿日期：2024−07−01
作者简介：赵文硕，主要从事对抗攻击方面的研究。
*通信作者 E-mail: wxuanhan@hotmail.com

术的发展和创新。

然而，随着深度学习模型的广泛部署，其内在的脆弱性也开始暴露出来。深度神经网络（deep neural networks, DNN）已被证明极易受到对抗样本攻击的影响。这些对抗样本通过向正常的输入数据中添加细微的、人类不易察觉的扰动，可以轻易地导致信号分类错误。在调制分类领域，已有研究表明，如需构造出信号对抗样本，只需在干净信号样本上添加精心设计的微弱扰动。如对原始二进制相移键控（binary phase shift keying, BPSK）、正交相移键控（quadrature phase shift keying, QPSK）等调制方式的信号添加对抗扰动，形成对抗样本，对抗样本输入到分类器，就可能导致预先训练好的分类器将信号误分类为相正交振幅调制（quadrature amplitude modulation, 64QAM）或者其他调制方式的信号[2]。这一低鲁棒性、低可靠性的现象给通信系统的安全性带来了严重的威胁与隐患。

在图像领域，对抗样本和深度学习鲁棒性的研究已经取得了相对成熟的进展，提出了大量特定的攻击策略和防御方案。其中，对抗训练已成为增强网络对抗鲁棒性的主流方法之一，该方法通过将对抗样本融入常规训练过程从而提高网络的鲁棒性。然而，在信号调制分类领域，构建有效防御机制的研究尚处于起步阶段。由于信号数据的固有属性及其处理过程与图像数据不同，如何探索有效的防御策略增强信号数据处理的鲁棒性成为了一个亟需解决的问题。

本文提出了一种提升信号调制分类网络鲁棒性的方法：混合信号对抗训练，旨在提高对抗样本识别的成功率（鲁棒准确率）的同时，保持其在原始未扰动样本上的分类性能（标准准确率）。针对现有对抗性训练框架，本文做出了两项改进：首先，针对信号训练数据相对稀少以及通过对抗性样本训练所得网络表征的紧凑度较低的情况，本文提出一种基于线性插值的数据增强方法，同时对干净样本与对抗性样本进行数据扩充，增加训练集的样本量的同时提高网络表征的紧凑性。其次，采用最大间隔损失函数代替传统的交叉熵损失，通过扩大模型的决策边界增强模型对扰动输入的鲁棒性。

本文主要贡献如下。1）在标准对抗性训练框架的基础上提出了一种针对信号调制识别任务的鲁棒性增强方法。2）提出一种基于线性插值的混合信号数据增强方法，通过在训练过程中对干净信号样本与对抗样本进行扩充，有效解决训练数据稀缺与网络表征不足的问题。同时将交叉熵损失函数替换为最大间隔损失函数，提升模型对于扰动输入的鲁棒性。3）在公开信号数据集上进行实验验证，证明所提出的方法不仅能够保持原有的分类性能，同时显著地提高了网络对于对抗样本的鲁棒性。

1　相关工作

1.1　信号调制分类

应用深度网络的调制信号分类方法相较于传统基于特征的调制信号分类方法取得了更优越的性能。文献 [3-4] 将卷积神经网络（convolutional neural networks, CNN）应用于无线调制识别任务，其实验结果证实 CNN 展现出了更高的分类准确性，并与当前的基于专家的方法相比提供了更多灵活性。然而，无限地增加网络深度会导致梯度消失或爆炸等问题。文献 [5] 提出将残差神经网络应用于信号调制分类任务，通过在网络不同层之间创建联通路径来增强神经网络中的特征传播，极大地缓解了梯度消失或爆炸的问题。为进一步提高性能，文献 [6-7] 提出将循环神经网络（recurrent neural network, RNN）与长短期记忆网络（long short-term memory, LSTM）用于分类短信号调制模式，明确学习在各种噪声条件下不同时变信号的复杂关系。文献 [8] 利用深度学习中不同网络优势，对各种维度的信号特征进行提取，提出基于多通道特征融合网络的通信信号调制类型识别加强分类性能。本文着重研究卷积神经网络下信号调制分类鲁棒性的提升策略。

1.2　对抗攻击

对抗攻击的概念在文献 [9] 中提出，其中介绍了一种生成对抗样本的方法，即通过对输入图像施加精心设计的微小扰动。这些扰动虽然对人类几乎不可见，但却能够引起深度神经网络输出的显著变化。基于这个重要发现，研究者们陆续提出了多种不同的对抗攻击方法，如快速梯度符号方法[10]（fast gradient sign method, FGSM）、迭代式快速梯度符号方法[11]（I-FGSM）、基于动量的方法[12]、投影梯度下降[13]（projected gradient descent, PGD）等。这些方法在生成对抗性样本的过程中采用了相似的策略，即朝着梯度方向最大化损失函数。通过对损失函数关于输入数据的梯度进行计算，攻击者能够逐步对输入进行扰动，进而生成能使目标模型产生误分类的对抗样本。除了数字域对抗攻击，也有部分研究将重心转向其他域，文献 [14] 提出

BALA 方法，通过在背景区域添加对抗扰动块，在保持原始面部特征完整性的前提下达成对抗人脸识别攻击，实现了物理域的对抗攻击。

在信号领域，亦有部分工作尝试将对抗样本的思想迁移到具体应用中。文献 [15] 在信号数据上对比分析了 FGSM、PGD、BIM（basic iterative method）与 MIM（mometum iterative method）4 种对抗攻击算法的可行性和有效性；文献 [16] 基于 FGSM 算法评估了基于原始 I/Q 信号的调制分类漏洞；文献 [17] 使用 PGD 攻击方法与均匀随机噪声产生对抗样本，旨在将窃听者的识别准确性降到最低，同时确保预期接收者成功解码信号。

1.3 对抗鲁棒性

在图像领域，文献 [10] 提出使用 FGSM 方法通过单一梯度步骤生成对抗性样本，作为对抗性训练的早期形式。文献 [18] 针对 FGSM 方法添加随机化步骤提出 R+FGSM，极大地增强了攻击强度。随后，迭代式的方法[11] 通过采取多个更小的 FGSM 步骤对 FGSM 进行了改进，使得上述基于 FGSM 的对抗性训练失效。这种迭代式的对抗攻击通过随机重启提升了攻击强度，并被集成到对抗训练流程中。上述工作共同构建了当前广泛应用的针对 PGD 的对抗训练方法基础，该方法被证明是构建鲁棒网络的有效途径。此后，大量工作致力于提升 PGD 对抗训练框架的性能，如应用矩阵估计[19]、对数配对[20] 进行启发式防御，以及对多种类型的对抗攻击的泛化方法[21] 等。

对抗训练的难点在于如何在提高网络对对抗样本的鲁棒性的同时，保持其在干净样本上的原有分类性能。标准准确率与对抗性鲁棒性之间的权衡已经被多项研究广泛讨论。文献 [21] 指出，为了达到更好的对抗性鲁棒泛化效果，需要更高的样本复杂度。文献 [22] 从鲁棒目标的统计特性和在神经网络上优化鲁棒目标的动态过程的角度研究了这种权衡，并指出对抗训练需要更多数据才能实现较高的标准准确率。文献 [23] 在 SVHN、CIFAR-10 和 CIFAR-100 数据集上的实验结果进一步证实了这一点，即通过在 PGD 对抗性训练中增加训练样本量可以提高标准准确率。在信号调制分类领域，对抗鲁棒性的研究相对匮乏，文献 [24] 应用了标准对抗训练框架提升信号鲁棒性。本文在标准对抗训练的基础上进一步研究适用信号数据特点的对抗鲁棒性增强范式，提升信号系统的可靠性与安全性。

2 背景

2.1 任务及符号定义

对于一个由信号数据$x \in X$和对应标签$y \in Y$组成的数据分布D，调制信号分类任务的目的是学习一个映射函数$f: X \to Y$，使得给定x时，f输出相应的y。在深度学习背景下，一般使用深度学习模型作为f。训练的过程等价于最小化风险$E_{(x,y)\sim D}[L(x,y,\theta)]$，其中$L(\theta,x,y)$是损失函数，如交叉熵损失，$\theta \in R^p$是函数$f$的参数集合。由于期望值无法直接计算，因此常用的方法为最小化经验风险$\frac{1}{N}\sum_{i=1}^{N} L(x_i,y_i,\theta)$，只考虑从数据分布$D$中抽取的有限数量的信号样本，即$(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$。

2.2 对抗攻击与对抗训练

经验风险最小化框架可以在未受扰动的测试样本上展现出优越的识别性能，但无法保证模型在对抗样本上表现良好，即经验风险最小化框架缺乏对抗攻击的鲁棒性。因此，文献 [13] 提出了一种优化视角下的对抗鲁棒性，将鲁棒性定义为一个 min-max 优化问题。优化目标如下式：

$$\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},y)\sim D}\left[\max_{|\boldsymbol{\delta}|\leqslant\epsilon} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}+\boldsymbol{\delta}),y)\right] \tag{1}$$

式中，$\boldsymbol{\delta}$表示对抗扰动；$L(f_\theta(\boldsymbol{X}+\boldsymbol{\delta}),y)$代表模型损失函数，即表示模型在参数$\theta$下对于对抗样本$\boldsymbol{X}+\boldsymbol{\delta}$和真实标签$y$的分类性能。内层部分旨在最大化对抗样本的攻击强度，使原始信号样本添加扰动$\boldsymbol{\delta}$后最大限度地干扰分类网络的判断；而外层部分进行网络参数θ的学习。

对抗训练过程中需要选定一种特定的方法进行对抗样本生成，本文使用 PGD 攻击作为主要攻击算法。PGD 是一种迭代攻击，相比于普通的 FGSM 仅做一次迭代，PGD 在生成对抗样本的过程中进行多次迭代，每次走一小步且每次迭代都会将扰动投射到规定范围内，通过 PGD 算法生成的对抗样本在一阶对抗样本的类别中表现尤为突出。每一步迭代内，算法首先计算g_t，即t时刻的损失关于t时刻输入的梯度：

$$g_t = \nabla_{X_t} L(f_\theta(\boldsymbol{X}_t),y) \tag{2}$$

式中，∇表示梯度计算。据此，$t+1$时刻的对抗样本可根据t时刻的输入及t时刻的梯度求出：

$$\boldsymbol{X}_{t+1}=\Pi_{X+S}\left(\boldsymbol{X}_t+\epsilon\left(\frac{g_t}{|g_t|}\right)\right) \tag{3}$$

式中，ϵ代表扰动步长；S代表扰动最大范围；Π_{X+S}代表投射操作，即将扰动后的输入投射回到规定的范围S内。

2.3　信号调制识别网络模型

本文采用深度卷积网络作为分类模型并针对一维信号处理任务进行结构设计。模型输入为双通道信号$\boldsymbol{x}\in\boldsymbol{X}$，尺寸为 2×2 048。输入信号首先通过一个包含 128 个卷积核的一维卷积层捕捉信号初级特征。接下来，模型依次通过 12 个卷积块进行深层次的特征提取，每个卷积块都包含一维卷积层、批量归一化层和 ReLU 激活函数。每通过 4 次卷积后进行一次最大池化，有效降低特征维度并减少计算复杂度。此外，模型在最后几个卷积块中采用渐进式通道扩展策略，逐步增强网络的表征能力。最终，使用两个全连接层将特征映射到分类标签上。全连接层间使用 Dropout 与批归一化技术减少过拟合。网络结构示意图如图 1 所示。

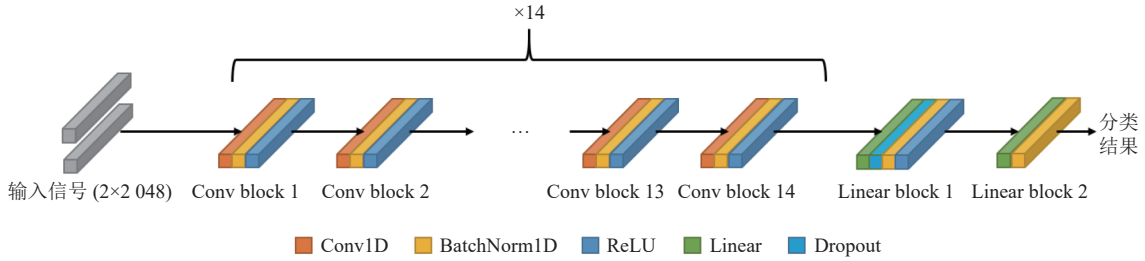

图 1　信号调制分类卷积神经网络结构

3　提升鲁棒性的设计方法

本文遵循经典对抗训练框架，步骤组成如下。

1）在数据分布D上采样一个批次的数据(x_i,y_i)。

2）将未受扰动的信号数据（干净样本）x_i送入分类器f，计算其标准损失$\mathcal{L}_c$：

$$\mathcal{L}_c=L(f_\theta,x_i,y_i) \tag{4}$$

3）应用攻击算法（如 PGD），依据标签y_i对信号样本x_i生成其对抗样本$\widehat{x_i}=\mathrm{PGD}(x_i,y_i)$。

4）将对抗样本$\widehat{x_i}$送入分类器f，计算其对抗损失$\mathcal{L}_a$：

$$\mathcal{L}_a=L(f_\theta,\widehat{x_i},y_i) \tag{5}$$

5）计算融合损失$\mathcal{L}=(\mathcal{L}_c+\mathcal{L}_a)/2$。依据融合损失进行反向梯度传播，更新网络参数$\theta$。

由于训练过程中同时应用了干净样本与对抗样本，模型期望在提升对抗鲁棒性的同时保持其原有的分类能力。然而，由于信号训练数据相对稀缺、网络表征不足，简单地应用标准的对抗训练范式会导致模型在干净样本与对抗样本上表现不佳。本文在经典对抗训练框架下做出两点改进。

3.1　混合信号数据增强

本文提出一种基于线性插值的混合信号数据增强方法，并将对抗训练与其结合，混合数据增强方法流程如下。

首先，从数据集中抽取两对样本$(x_i,y_i)\sim D$和$\left(x_j,y_j\right)\sim D$，然后在输入空间中进行随机线性插值：

$$\widetilde{x}=\lambda x_i+(1-\lambda)x_j \tag{6}$$

式中，λ在每次更新时从 Beta 分布中随机采样。将$\widetilde{x}$送入网络f_θ进行训练：

$$L=\lambda L(f_\theta(\widetilde{x}),y_i)+(1-\lambda)L\left(f_\theta(\widetilde{x}),y_j\right) \tag{7}$$

相较于图像数据，信号数据相对稀少，而依据文献 [22]，对抗训练框架需要更大量的训练样本来达到更高的准确率。混合信号数据增强可以在已有训练集上创造大量新数据样本，显著增加训练集样本数量，因此有助于提高模型准确率。

应用混合信号数据增强的另一个动机来自于信息压缩角度。文献 [25] 研究了深度网络学习中特征信息压缩强度与网络泛化性之间的关系。当深度网络对特征进行更强的压缩时，泛化性能更强。受此启发，本文设置了一组实验来评估调制分类对抗训练对特征信息压缩的影响。首先，预训练一个分类模型（使用标准训练流程或 PGD 对抗训练），然后冻结模型参数，研究这些冻结的表征能否成功地预测固定随机标签。如果模型较强地压缩了表征，则将更难拟合随机标签。具体地，在冻结模型的表征之后添加一个双层 MLP 拟合随机二进制标

签，并使用相同的超参数对模型进行了 200 轮次训练。对于拟合 10 000 个随机标签的样本，最终准确率为 81.08%（标准模型）和 94.00%（PGD 对抗性训练模型），这表示对抗性训练使得模型的压缩程度大大降低。而基线性插值生成的混合信号能学习到更压缩的特征，因此，将混合信号数据增强与对抗训练结合可能会缓解对抗训练的不利影响。使用与上述相同的实验设置，将混合信号融入对抗训练中得到的准确率为 78.13%。结果表明对抗训练导致网络学习到的表征被压缩程度更低，而与混合信号结合的对抗训练模型显著缓解了这一现象，从而有助于提高标准准确率。

3.2 最大间隔损失

在标准对抗训练框架中，标准损失$\mathcal{L}_c$与对抗损失$\mathcal{L}_a$均使用交叉熵损失函数进行优化。对抗训练的任务目标是增强模型对输入扰动的鲁棒性，因此，对模型的决策边界和分类确信度更敏感。本文提出使用最大间隔损失（hinge loss）代替交叉熵损失，进一步提高模型在对抗攻击环境下的性能。最大间隔损失函数定义如下：

$$H=\max\left(0,1+\max_{j\neq y}\left(f(x)_j-f(x)_y\right)\right) \tag{8}$$

式中，$f(x)_y$表示正确类别的分数；$f(x)_j$表示所有错误类别的最高分数；y表示真实类标签。

最大间隔损失函数旨在确保正确类别的分数至少比任何错误类别的分数高出一个固定边界。与交叉熵损失相比，其增大了正确类别和错误类别之间的决策边界。在对抗训练的任务背景下，应用最大间隔损失可以促使模型在分类时保持更大的安全边界，从而提高模型在面对扰动数据时的性能表现，提升模型对抗鲁棒性的同时也最大化了模型在未受攻击数据上的性能。综上所述，本文提出的改进对抗训练框架流程如下。

1）在数据分布D上采样一个批次的数据(x_i,y_i)。

2）将未受扰动的信号数据进行混合信号数据增强，得到扩充样本集合$\left(x_i',y_i'\right)$。将干净样本送入分类器f，计算其标准损失$\mathcal{L}_c$，其中损失函数使用最大间隔损失H：

$$\mathcal{L}_c=H\left(f_\theta,\mathrm{Mix},x_i',y_i'\right) \tag{9}$$

3）应用攻击算法（如 PGD）依据标签对扩充后信号样本生成其对抗样本$\widehat{x_i}=\mathrm{PGD}\left(x_i',y_i'\right)$。

4）将对抗样本$\widehat{x_i}$送入分类器f，计算其对抗损失$\mathcal{L}_a$，其中损失函数使用最大间隔损失H：

$$\mathcal{L}_a=H\left(f_\theta,\mathrm{Mix},\widehat{x_i'},y_i'\right) \tag{10}$$

5）计算融合损失$\mathcal{L}=(\mathcal{L}_c+\mathcal{L}_a)/2$。依据融合损失进行反向梯度传播，更新网络参数$\theta$。

4 实验结果与分析

4.1 数据集

本文使用公开信号数据集 Panoradio HF 作为实验数据。数据集中涵盖 18 种不同波形的无线电信号：morse、psk31、psk63、qpsk31、rtty45_170、rtty50_170、rtty100_850、olivia8_250、olivia16_500、olivia16_1000、olivia32_1000、dominoex11、mt63_1000、navtex、usb、lsb、am、fax；每个信号均为 2 048 维 IQ 双通道数据。数据的生成过程如下：首先利用标准软件对语音、音乐及文本信息进行调制处理；随后将调制后的信号分割成若干短时序列；每个时序列均添加高斯噪声的干扰以及随机频率和相位偏移的影响。通过此过程合成的数据与实际接收到的信号具有较高相似度，数据集相关参数如表 1 所示。

表 1 数据集相关参数

参数	取值
采样频率/kHz	6
随机频率偏移范围/Hz	±250
信噪比（SNR）/dB	−10∶25

4.2 实验设置

实验环境为 64 位 Ubuntu 18.04 LTS 系统，模型通过 PyTorch 框架构建，利用 NVIDIA RTX 3090 显卡加速训练过程。网络优化采用随机梯度下降（stochastic gradient descent, SGD）优化器，动量参数设置为 0.9，权重衰减参数设置为 0.000 2。损失函数使用交叉熵。混合信号数据增强中，λ在 beta 分布$(-1,1)$中随机采样。PGD 攻击中，扰动界限ϵ设置为 0.5，截断界限S设置为$(-10,10)$，迭代轮次k设为 20。批处理大小为 512，进行 200 轮次训练。

4.3 评价指标

实验采用标准准确率（standard accuracy, SA）和鲁棒准确率（robust accuracy, RA）两个评价指标衡量网络性能。其中， SA 反映分类模型在未受扰动的干净样本上的表现，理想状态下，经过对抗

训练的模型应能维持较高的 SA，表明其基本的分类能力未受损害。RA 则衡量了模型在面对由 FGSM、PGD 等方法生成的对抗样本时的表现，RA 的提升意味模型具备更强的抵御对抗性攻击的能力，展现出更优的鲁棒性。

4.4　实验结果及分析

本文依据 2.3 节所述网络结构训练了 4 个模型进行实验对比：标准分类模型（原始模型）、使用 FGSM 攻击方法进行对抗训练（adversarial training, AT）的模型、使用 PGD-20 攻击方法进行对抗训练的模型，以及本文提出的混合信号对抗训练模型。训练后的模型依据 4.3 节中的评价指标进行实验，其中鲁棒准确率方面选取了 3 种攻击方法：FGSM 攻击、7 轮迭代 PGD 攻击及 20 轮迭代 PGD 攻击。实验结果如表 2 所示。另外，本文对比了对抗训练前后干净样本上各个调制类别的准确率（P），召回率（R）变化情况，如表 3 所示。

表 2　不同鲁棒模型实验结果　　%

训练方式	SA	RA-FGSM	RA-PGD(7)	RA-PGD(20)
原始模型	92.28	19.84	4.87	4.04
FGSM AT	84.22	51.63	10.24	8.91
PGD AT	86.48	56.29	42.27	41.99
本文	**90.67**	**61.51**	**51.57**	**48.70**

表 3　调制类别分类统计指标

信号类型	P-clean	P-AT	R-clean	R-AT
morse	0.81	0.95	0.98	0.98
psk31	0.80	0.92	0.84	0.90
psk63	0.76	0.91	0.98	0.95
qpsk31	0.94	0.83	0.59	0.87
rtty45_170	0.91	0.75	0.84	0.72
rtty50_170	0.86	0.76	0.91	0.76
rtty100_850	1.00	0.99	1.00	1.00
olivia8_250	0.89	0.82	0.94	0.83
olivia16_500	0.89	0.75	0.89	0.77
olivia16_1000	1.00	0.98	0.99	0.96
olivia32_1000	0.98	0.94	0.93	0.83
dominoex11	0.96	0.87	0.93	0.86
mt63_1000	0.97	0.99	0.99	0.99
navtex	0.98	0.93	1.00	0.96
usb	1.00	0.99	0.97	0.94
lsb	0.98	0.96	0.99	0.98
am	0.87	0.86	0.81	0.89
fax	0.98	0.94	0.91	0.92

标准准确率方面，原始模型达到了最高的 92.28%。相较之下，通过对抗训练方法训练的模型在标准准确率上都有所下降。其中使用 FGSM 对抗训练的模型的准确率相比原始模型下降了 8.06%；而使用 PGD-20 对抗训练的模型的准确率相比原始模型下降了 5.80%。本文提出的混合信号对抗训练模型的标准准确率从 92.28% 下降到 90.67%，降幅仅为 1.61%，在所有对抗训练模型中维持了最高的标准准确率。具体地，在未经对抗训练的原始模型上，某些调制类型如 rtty100_850 和 usb 表现出较高的分类准确性和召回率（1.00）。在经过对抗训练后仍保持在非常高的水平，如 rtty100_850 的准确率和召回率在对抗训练后分别为 0.99 和 1.00。通过对比 P-clean 和 P-AT，可以观察到在部分情况下，准确率在经过对抗训练后甚至有所提高。如 morse 从 0.81 提升至 0.95，psk31 从 0.80 提升至 0.92，这来自于混合信号数据增强带来的收益，即大量对训练集样本量的扩充与网络表征紧凑度的提升。另一方面，部分调制类别经对抗训练后的准确率、召回率有所下降。如 rtty45_170 的准确率从 0.91 下降到 0.75，这可能是由于对抗训练导致了模型对于某些特定类型的对抗信号样本产生了过拟合，导致对未扰动信号的分类性能下降。

鲁棒准确率方面，由于多步迭代攻击相较于单步攻击更有效，因此所有模型中 FGSM 的攻击强度普遍小于 PGD-7 和 PGD-20。而随着 PGD 迭代次数 k 增加，攻击效果也逐渐增强，故 PGD-7 的攻击效果通常小于 PGD-20。值得注意的是，在对抗训练中使用 FGSM 攻击方法训练的模型在 PGD 攻击下防御能力较差。相反，使用 PGD 攻击方法进行训练的模型则在防御 FGSM 攻击时表现出较强的鲁棒性。因此，复杂的多步攻击训练方法（如 PGD）能够提供更泛化的鲁棒性能。

在未经过对抗训练的原始模型上，鲁棒准确率最低，其中 PGD-20 的鲁棒准确率仅为 4.04%，表明基础模型在面对对抗攻击时极具脆弱性。经过对抗训练的模型鲁棒准确率都得到显著提升，本文提出的混合信号对抗训练方法在所有攻击类型下均实现了最高的鲁棒准确率。相较于标准对抗训练模型，本文方法在 3 种攻击下的鲁棒准确率平均提升约 7.07%。

为探究混合信号数据增强与最大间隔损失函数对于对抗训练方案的影响，本文了设计两组消融实验：1）固定应用混合信号数据增强方法，对比以下 4 个模型的标准准确率（SA）与对抗鲁棒性（RA）：使用交叉熵损失函数训练的标准分类模

型（CE-Clean）、使用最大间隔损失函数训练的标准分类模型（Hinge-Clean）、使用交叉熵损失函数进行对抗训练的鲁棒模型（CE-AT），以及本文提出的混合信号对抗训练模型。2）固定应用最大间隔损失函数，对比以下 4 个模型的标准准确率（SA）与对抗鲁棒性（RA）：使用未经数据增强的原始数据集训练的标准分类模型（Clean）、使用混合信号数据增强方法训练的标准分类模型（Mix-Clean）、使用未经数据增强的原始数据集进行对抗训练的鲁棒模型（AT），以及本文提出的混合信号对抗训练模型。实验结果如表 4 和表 5 所示。

表 4　损失函数对比实验结果

%

训练方式	SA	RA-FGSM	RA-PGD(7)	RA-PGD(20)
CE-Clean	92.45	17.43	3.14	2.97
Hinge-Clean	92.28	19.84	4.87	4.04
CE-AT	88.23	56.93	43.82	43.63
本文	**90.67**	**61.51**	**51.57**	**48.70**

表 5　数据增强方法对比实验结果

%

训练方式	SA	RA-FGSM	RA-PGD(7)	RA-PGD(20)
Clean	91.83	18.23	3.01	2.36
Mix-Clean	91.28	18.84	3.87	3.04
AT	84.84	51.64	41.78	40.65
本文	**90.67**	**61.51**	**51.57**	**48.70**

依据表 4，本文采用的最大间隔损失模型在对抗鲁棒性方面优于交叉熵损失模型。在未经对抗训练的模型上，最大间隔损失模型在标准准确率上与交叉熵损失模型相近；且最大间隔损失模型在所有攻击方法下的鲁棒性均优于交叉熵模型。同样，在对抗训练模型上最大间隔损失也表现出了较强防御能力，如在 PGD 攻击（7 轮和 20 轮）下鲁棒准确率分别比交叉熵损失模型高出 7.75% 和 5.07%。最大间隔损失通过在训练过程中扩大决策边界，增强了模型对小幅输入扰动的鲁棒性。

依据表 5，本文提出的混合信号数据增强方法显著提升了对抗训练后模型对干净样本数据的识别能力。与未应用数据增强的对抗训练方案（AT）相比，本文方法在标准准确率上提高了 5.83%。混合信号数据增强与对抗训练的结合有效缓解了标准对抗训练导致网络学习表征压缩程度更低的不利影响，从而提升了标准准确率。综上所述，本文提出的对抗训练方案在面对对抗攻击时提供了更有效的防御效果，增强了实际应用中的可靠性和安全性。

为探究混合信号对抗训练方案对于训练效率的影响，本文依据 2.3 节所述网络结构训练了 4 个模型进行训练耗时实验对比：标准分类模型（原始模型）、使用 FGSM 攻击方法进行对抗训练的模型、使用 PGD-20 攻击方法进行对抗训练的模型以及本文提出的混合信号对抗训练模型。

依据表 6，施加对抗训练会显著增加模型训练的时间开销。其中，PGD 对抗训练比 FGSM 耗时更长，因为 PGD 攻击方法需在每个输入样本上执行多轮梯度迭代更新，显著增加了计算复杂度。本文提出的混合信号对抗训练方案在包含数据增强的情况下，仅比 PGD 对抗训练多耗时 8.13 s，展现了良好的训练效率，且保证了更高的对抗鲁棒性和标准准确率，具有较强的实际应用价值。

表 6　不同鲁棒模型训练单轮耗时

s

鲁棒模型	原始模型	FGSM AT	PGD AT	本文
单轮训练耗时	113.45	245.23	297.52	305.65

5　总结语

本文在对抗训练框架的基础上提出了一种结合数据增强的防御方法：混合信号对抗训练，提高信号调制分类模型的鲁棒性。针对训练数据稀缺以及通过对抗样本训练所得网络表征能力不足的问题，采用混合信号数据增强策略增强模型性能。此外，本文采用最大化间隔损失代替传统的交叉熵损失，通过扩大分类决策边界提升模型对于扰动输入的鲁棒性，提升对抗样本的识别精度。实验结果证明，本文提出的方法不仅显著提升了模型面对多种对抗攻击算法时的防御能力，且能够保持良好的标准分类性能。未来的研究将会进一步针对信号数据提出更有效的防御策略，如自适应对抗训练、预训练等。此外，如何应对不断提出的先进对抗攻击算法也是下一步研究的重点。

参 考 文 献

[1] 李辉, 龚晓峰, 雒瑞森. 基于时频融合的深度学习调制识别算法[J]. 电讯技术, 2024, 64(1): 22-28.
LI H, GONG X F, LUO R S. A deep learning modulation recognition algorithm based on time-frequency fusion[J]. Telecommunication Engineering, 2024, 64(1): 22-28.

[2] 江汉, 胡林, 李文, 等. 信号调制识别的对抗样本攻防技术研究进展[J]. 数据采集与处理, 2023, 38(6): 1235-1256.
JIANG H, HU L, LI W, et al. Research progress of adversarial attack and defense for signal modulation recognition[J]. Journal of Data Acquisition and Processing,

2023, 38(6): 1235-1256.

[3] MA K, ZHOU Y B, CHEN J Y. CNN-based automatic modulation recognition of wireless signal[C]//2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE). Dalian: IEEE, 2020: 654-659.

[4] 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.

[5] WEI Z X, JU Y, SONG M. A method of underwater acoustic signal classification based on deep neural network[C]//5th International Conference on Information Science and Control Engineering (ICISCE). Zhengzhou: IEEE, 2018: 46-50.

[6] LIU X Y, YANG D Y, ALY E G. Deep neural network architectures for modulation classification[C]//51st Asilomar Conference on Signals, Systems, and Computers. Pacific Grove: IEEE, 2017: 915-919.

[7] ZHANG B, CHEN G, JIANG C. Research on modulation recognition method in low SNR based on LSTM[C]// Journal of Physics: Conference Series. Harbin: IOP Publishing, 2022: 012003.

[8] 黄杰. 基于深度学习的通信信号调制类型识别[D]. 成都: 电子科技大学, 2023.
HUANG J. Recognition of communication signal modulation types based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2023.

[9] CHRISTIAN S, WOJCIECH Z, ILYA S, et al. Intriguing properties of neural networks[EB/OL]. [2024-06-08]. https://arxiv.org/abs/1312.6199.

[10] IAN J G, JONATHON S, CHRISTIAN S. Explaining and harnessing adversarial examples[EB/OL]. [2024-06-08]. https://arxiv.org/abs/1412.6572.

[11] ALEXEY K, IAN J. G, SAMY B. Adversarial machine learning at scale[EB/OL]. [2024-06-08]. https://arxiv.org/abs/1611.01236.

[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. Salt Lake City: IEEE, 2018: 9185-9193.

[13] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. （2017-06-20） [2024-06-08]. http://arxiv.org/abs/1706.06083.

[14] 张晨晨, 王帅, 王文一, 等. 针对人脸识别卷积神经网络的局部背景区域对抗攻击[J]. 光电工程, 2023, 50(1): 113-125.
ZHANG C C, WANG S, WANG W Y, et al. Adversarial background attacks in a limited area for CNN based facerecognition[J]. Opto-Electron Eng, 2023, 50(1): 113-125.

[15] 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. Toronto: IEEE, 2020: 2469-2478.

[16] FLOWERS B, BUEHRER R M, HEADLEY W C. Evaluating adversarial evasion attacks in the context of wireless communications[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1102-1113.

[17] MUHAMMAD Z H, ANDRAS G, DENIZ G. Communication without interception: Defense against modulation detection[C]//2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Ottawa: IEEE, 2019: 1-5.

[18] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses [EB/OL]. [2024-06-08]. https://arxiv.org/abs/1705.07204.

[19] YANG Y Z, ZHANG G, DINA K, et al. Me-net: Towards effective adversarial robustness with matrix estimation [EB/OL]. [2024-06-08]. https://arxiv.org/abs/1905.11971.

[20] MOSBACH M, ANDRIUSHCHENKO M, TROST T, et al. Logit pairing methods can fool gradient-based attacks [EB/OL]. [2024-06-08]. https://arxiv.org/abs/1810.12042.

[21] TRAMER F, BONEH D. Adversarial training and robustness for multiple perturbations[J]. Advances in Neural Information Processing Systems, 2019, 32: 5866-5876.

[22] CARMON Y, RAGHUNATHAN A, SCHMIDT L, et al. Unlabeled data improves adversarial robustness[J]. Advances in neural information processing systems, 2019, 32: 11192-11203.

[23] LAMB A, VERMA V, KANNALA J, et al. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2019: 95-103.

[24] MAROTO J, BOVET G, FROSSARD P. SafeAMC: Adversarial training for robust modulation recognition models[EB/OL]. [2024-06-08]. https://arxiv.org/abs/2105.13746.

[25] SHWARTZ-ZIV R, TISHBY N. Opening the black box of deep neural networks via information[EB/OL]. [2024-06-08]. https://arxiv.org/abs/1703.00810.

编　辑　刘飞阳