第 40 卷 第 4 期
2021 年 7 月

应用声学
Journal of Applied Acoustics

Vol. 40, No. 4
July, 2021

◇ 研究报告 ◇

时频谱图和数据增强的水声信号深度学习目标识别方法*

刘 峰† 罗再磊 沈同圣 赵德鑫

(军事科学院 国防科技创新研究院 北京 100071)

摘要: 水声目标识别一直是水声领域研究的重点问题之一,深度学习方法可以有效地解决目标识别问题,然而,水声样本的稀少限制了该方法的应用。该文提出一种基于数据增强的水声信号深度学习目标识别方法,该方法以Mel功率谱作为网络的输入特征,通过对原始信号在时域和时频域的拉伸和掩蔽等变换,实现数据扩展和增加泛化性能的目的,最后,利用改进的VGG网络模型实现目标分类。实验结果表明,该文方法得到的水下目标识别准确率(95.2%)要优于其他4种对比方法,证明了该文提出的网络模型和数据增强方法均有助于提高目标分类性能。

关键词: 水声目标识别;卷积神经网络;数据增强;Mel功率谱

中图法分类号: TP391.4 **文献标识码:** A **文章编号:** 1000-310X(2021)04-0518-07

DOI: 10.11684/j.issn.1000-310X.2021.04.004

Deep learning target recognition method of underwater acoustic signal based on data augmentation and time-frequency spectrum

LIU Feng LUO Zailei SHEN Tongsheng ZHAO Dexin

(*National Lnnovation Lnstitute of Defense Technology, Academy of Military Sciences, Beijing* 100071, *China*)

Abstract: Underwater acoustic target recognition has always been one of the key issues in the field of underwater acoustic research. Deep learning methods can effectively solve the problem of target recognition. However, the scarcity of underwater acoustic samples limits the application of this method. This paper proposes a deep learning target recognition method for underwater acoustic signals based on data enhancement. This method uses Mel power spectrum as the input feature of the network, and in order to increase the generalization performance of the method, data augmentation is achieved by stretching and masking the original signal in the time domain and time-frequency domain. Finally, using an improved VGG network model to achieve target classification. The experimental results show that the underwater target recognition accuracy (95.2%) obtained by this method is better than the other four comparison methods, which demonstrates that the network model and data enhancement method proposed in this paper can help to improve the target classification performance.

Keywords: Underwater acoustic target recognition; Convolutional neural network; Data augmentation; Mel spectrogram

2020-09-27 收稿; 2021-01-10 定稿

*国家自然科学基金项目 (41906169)

作者简介: 刘峰 (1988–), 男, 黑龙江哈尔滨人, 博士, 助理研究员, 研究方向: 声信号处理, 目标识别。

†通信作者 E-mail: liufeng_cv@126.com

0 引言

随着海洋的战略地位日益突显，各国都在积极开发利用海洋资源和空间。声波是目前在海洋中唯一能够进行远距离传播的能量形式，水声目标识别对于海洋开发、国防安全有着重大意义，现已成为水声领域的研究热点之一。水下目标自动识别主要包括特征提取与构建分类器两大部分。当前主流的特征提取方法包括时域波形结构分析、频域谱估计以及时频域分析3个方面。时间域的分布可由峰-峰值、过零点分布、波列面积和波长差分布等特征进行描述[1]。频域谱估计可提取信号的频率、功率、包络等特征，以及利用高阶谱分析非高斯信号的特征[2]。这类方法原理简单、易于实现，仅通过采集到的原始水声信号即可获得，但是提取的特征需要一定的先验知识进行信号预处理，在时变的海洋环境下泛化性较弱。时频分析方法提供了时间域与频率域的联合分布信息，可以清楚地描述信号频率随时间变化的关系，是目前应用效果最好、应用最广的特征提取方法，常用的方法包括短时傅里叶变换(Short time Fourier transform, STFT)[3]、梅尔频率倒谱系数(Mel-Frequency cepstral coefficients, MFCC)[4−5]、希尔伯特-黄变换(Hilbert-Huang transform, HHT)等。在特征提取之后再训练隐马尔可夫模型(Hidden Markov model, HMM)、支持向量机、K近邻、神经网络等分类器以实现水下目标的识别。

近年来，随着计算机硬件技术、信号处理技术的进一步发展，以机器学习(Machine learning, ML)、深度学习(Deep learning, DL)、大数据(Big date)等为代表的人工智能(Artificial intelligence, AI)技术，已经在语声识别、图像理解、机器翻译等多个方面取得了长足的进展[6]，这为水声目标识别提供了新的解决思路。借助现代计算机技术、信号处理技术、人工智能技术等，开展基于深度学习和大数据分析的水声信号智能化目标识别技术研究，可有效提高自主识别系统的泛化能力和环境适应性。水声信号在传播过程中受环境影响较大，存在着数据获取困难、样本数据少、噪声干扰强等特点，在实际的应用场景中，很难针对每一种水下目标收集到足量的数据，因此当收集到的数据量不足以支撑深度神经网络的训练需求时，如何利用少量数据实现目标识别是当前研究所面临的难题。

Kamal等[7]首先将深度置信网络模型应用于水声信号被动目标识别任务中，在40个类别的目标共1000个测试样本的测试集上取得了90.23%的分类正确率，验证了深度学习模型的有效性。王强等[8]利用卷积神经网络(Convolutional neural network, CNN)对3类水下目标噪声数据进行分类识别，并与支持向量机方法进行对比。随着深度学习的发展，目标识别的网络构架逐渐成熟，基于ResNet[9]和DenseNet[10]等方法的网络模型性能显著优于早期的基于VGG[11]、AlexNet[12]等架构，这主要是因为ResNet很好地解决了训练过程中的梯度消失问题。然而，在水声目标识别任务中，可用数据规模通常较小，训练这样的深层架构会导致训练样本的过度拟合，目前最先进的分类方法仍然主要由VGG架构产生。McDonnell等[13]采用了取自计算机视觉领域的VGG架构，以声谱图作为网络输入，在声场景分类方面取得了良好的效果。Koutini等[14]通过调整不同网络层中CNN的感受野增强模型的泛化能力，实现对不同场景中的声目标信号进行分类，通过对比多种网络模型的性能，基于VGG网络的改进结构取得了最好的分类效果。

本文以Mel功率谱(Mel spectrum)作为水声信号的特征提取方法，提出了一套适用于小样本水声信号的目标识别方法，利用多种数据增广技术并结合深度学习网络进行仿真验证。结果表明，在数据样本匮乏和样本分布不平衡情况下的水声目标识别方面，本文方法具有明显优势。

1 本文方法

本文方法如图1所示，处理流程主要分为3个步骤：(1) 将原始信号提取Mel功率谱作为特征；(2) 采用数据增强方法，分别从时域信号和时频谱图两个方面进行扩展；(3) 利用改进的VGG网络对时扩展后的频谱图进行特征学习和训练，实现目标分类。

1.1 Mel功率谱特征提取

在声频信号处理中，构建特征向量和设计分类器通常被视为两个独立的问题。MFCC特征受到人类听觉系统和语声感知生理学的启发，被用作声频分析任务的主要声学特征之一，由于其滤除的信息较多，Mel频谱作为一种特征提取方式在使用神经网络作为分类器时被广泛使用。

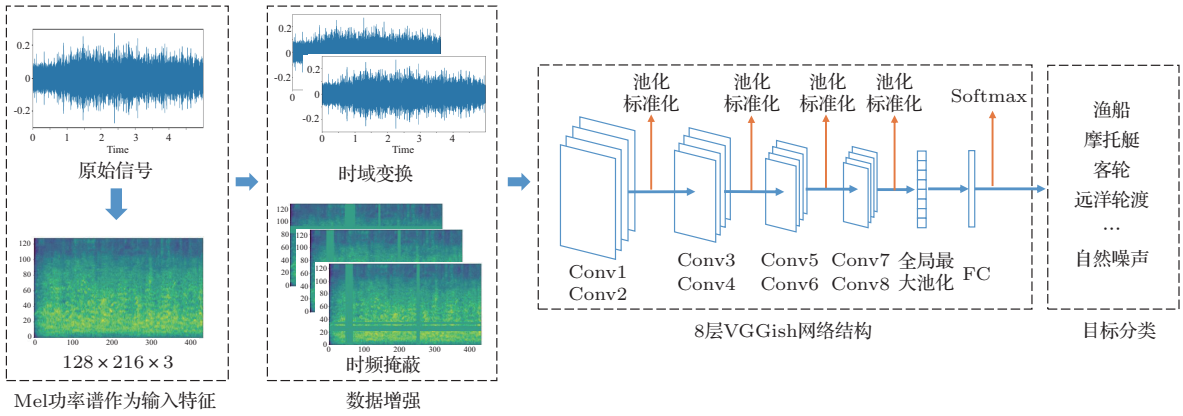


图 1　本文方法的网络结构示意图

Fig. 1　Schematic diagram of the method in this paper

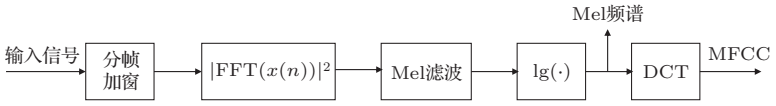


图 2　Mel 频谱特征提取过程

Fig. 2　The process of Mel-spectrogram feature extraction

Mel 频谱的提取过程如图 2 所示，首先经过分帧加窗等预处理确定待处理的信号范围，再利用快速傅里叶变换 (Fast Fourier Transform, FFT) 把水声信号由时域变换到时频域，用一组 Mel 尺度的三角形滤波器组进行带通滤波，再利用对数变换将这些幅值谱投影到一组缩小的频带上，然后用离散余弦变换 (Discrete Cosine transform, DCT) 进行近似的白化和压缩，生成 MFCC 特征。由于 DCT 变换删除了较多的信息，破坏了谱图的空间关系，而省略 DCT 变换得到的 Mel 频谱，在深度学习模型中通常能得到更好的分类效果。

在依赖频谱图作为深度学习的输入特征时，谱图的分辨率严重影响最终的目标分类结果，然而，时间分辨率和频率分辨率是相互矛盾的。在处理过程中，宽窗具有良好的频率分辨率，但时间分辨率会受到影响，窄窗则相反。到目前为止，这个问题并没有得到统一的共识，根据不同的实际问题，研究者们选择了不同的频带范围，较为普遍的趋势是将 Mel 频带的数目限制在 [60, 200] 范围内。

1.2　数据增强

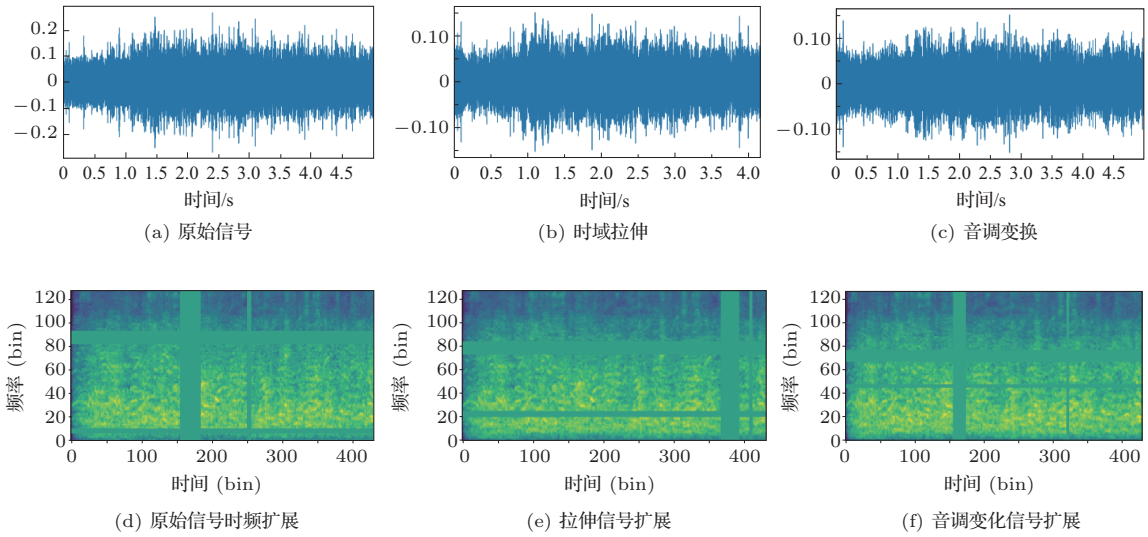
数据增强是深度学习中常用的技术之一，主要用于增加训练数据集，使数据集尽可能多样化，使训练模型具有较强的泛化能力。由于水声信号本身就面临着严重的数据样匮乏、样本完备性不足等问题，数据增强方法有助于改善目标分类结果。

本文分别从原始信号和频谱图两方面进行数据的增强。在时域信号上，采用时域拉伸和音调变换进行数据增强，然后将变换后的信号转换成 Mel 频谱作为深度神经网络的输入。在频谱图上，借鉴 SpecAugment[15] 方法进行数据增强，该方法将声信号的增强问题转化为视觉问题进行处理，通过时空干扰和随机掩蔽技术对频谱图实现增强，该方法能较好地应对时间方向上的变形和频率信息的部分损失，具有较强的鲁棒性和泛化性，这里忽略了时间扭曲变换，只采用时间掩蔽和频率掩蔽进行变换。具体变换方法如下：

时域拉伸：放慢或加快声频样本 (同时保持音调不变)。每个样本用两个参数进行时间拉伸：{0.8, 1.2}；

音调变换：提高或降低声频样本的音高 (同时保持持续时间不变)，每段声频样本的音调变换比例为 $\{-2, 2\}$；

时间掩蔽：在频谱图中使 t 个连续时间步长 $[t_0, t_0+t)$ 被图像均值掩蔽，其中 t 为掩蔽时长，其取值从 0 到时间掩码参数 T 的均匀分布中随机选择，t_0 为起始时间，从 $[0, \tau - t)$ 中选择，τ 为信号帧长，T 取值范围与帧长呈正相关；



(a) 原始信号　(b) 时域拉伸　(c) 音调变换

(d) 原始信号时频扩展　(e) 拉伸信号扩展　(f) 音调变化信号扩展

图 3　数据增强示意图

Fig. 3　The schematic of data augmentation

频率掩蔽：使 f 个连续的 Mel 频谱通道 $[f_0, f_0+f)$ 被掩蔽，其中 f 为掩蔽频段，其取值从 0 到频率掩码参数 F 的均匀分布中随机选择，f_0 为起始频率，从 $[0, \upsilon - f]$ 中选择，υ 是 Mel 频谱的总频带数，F 取值与 Mel 频带数呈正相关。

根据上述参数设置，时域拉伸和音调变换可扩展数量为 4，时间掩蔽和频率掩蔽产生的数据增强可以表示在同一频谱图中，这里选择扩展的倍数为 5。因此，通过上述处理方法，结合原始信号共产生了 10 倍的数据增强。部分增强结果如图 3 所示。

1.3　网络构建

本文参考 VGG 网络作为基础模型，通过修改其网络层中的部分参数以适应水声信号的分类任务。网络由 8 个卷积层组成，每个卷积层通过一组滤波器对前一个卷积层的输出进行卷积，用以捕获特征图中的局部信息，激活函数选择 ReLU，滤波器大小为 3×3[12]。在每个卷积层之后应用批处理归一化 (Batch normalization) 和 2×2 的最大值池化，达到降低特征图维度并避免过拟合的目的。对最后一个卷积层 [16] 的特征图应用全局最大池化操作，将特征图生成为一维向量，最后通过全连接层和 softmax 分类器输出不同目标类别的概率，实现分类。深度神经网络通过多层结构自动提取声谱图中的特征信息，通过有监督的线性与非线性的组合获取数据的高层统计特征，达到减少人工参与、实现数据驱动的目的。由于声信号的采样率通常较高，且时域信号包含的信息有限，通常以人工提取的频谱图作为网络的输入数据，本文将人工特征提取和深度网络模型相结合实现水声信号的识别，网络结构和参数如表 1 所示。

表 1　本文采用的基于 VGG 模型改进的网络构架

Table 1　The network architecture based on improved VGG

特征图尺寸	本文网络结构	参数数量
T×128	Mel 频谱	4.7 M
$T/2$ × 64	$\begin{bmatrix} 3\times3, BN \\ 3\times3, BN \end{bmatrix}$, 64 2×2 最大值池化	640, 256 36928, 256
$T/4$×32	$\begin{bmatrix} 3\times3, BN \\ 3\times3, BN \end{bmatrix}$, 128 2×2 最大值池化	73856, 512 147584, 512
$T/8$×16	$\begin{bmatrix} 3\times3, BN \\ 3\times3, BN \end{bmatrix}$, 256 2×2 最大值池化	295168, 1024 590080, 1024
$T/16$×8	$\begin{bmatrix} 3\times3, BN \\ 3\times3, BN \end{bmatrix}$, 512 2×2 最大值池化	1180160, 2048 2359808, 2048
4×4	全局最大值池化	
	全连接层, softmax	2565

2　仿真实验

2.1　实验数据

为了评估本文的方法，使用 ShipsEar 数据集 [17] 进行仿真验证，该数据集中共包含 90 段声

频记录，4 类不同的船只目标和 1 类环境噪声，数据是利用自容式水听器对码头上往来的船只噪声信号进行记录，以采集不同船速下的噪声以及与进坞或离坞时的空化噪声。由于数据是在真实开放水域中采集的，部分信号中混杂了人说话声、自然背景噪声，偶尔也会记录到海洋哺乳动物的声音。最后，该数据集由 5 类 wav 格式的 90 条记录组成。每个类别包含一个或多个目标，每个声频片段的持续时间从 15 s 到 10 min 不等。经初步处理后，消除了背景噪声干扰强烈和模糊不清的信号。对数据进行预处理，去除空白信号，并将原始信号按照 5 s 时长进行分帧和标注，共生成 1956 个标注样本。将所有样本按照 7 : 1.5 : 1.5 的比例进行随机分割，得到训练集、验证集和测试集数据样本分别为 1370 个、293 个和 293 个。

详细信息如表 2 所示，第一列是声信号目标类别，第二列是每类对应的细分船只，第三列是每类目标的帧数。

表 2　待测试数据集中的目标类别及帧数

Table 2　Target category and number of frames in the data set to be tested

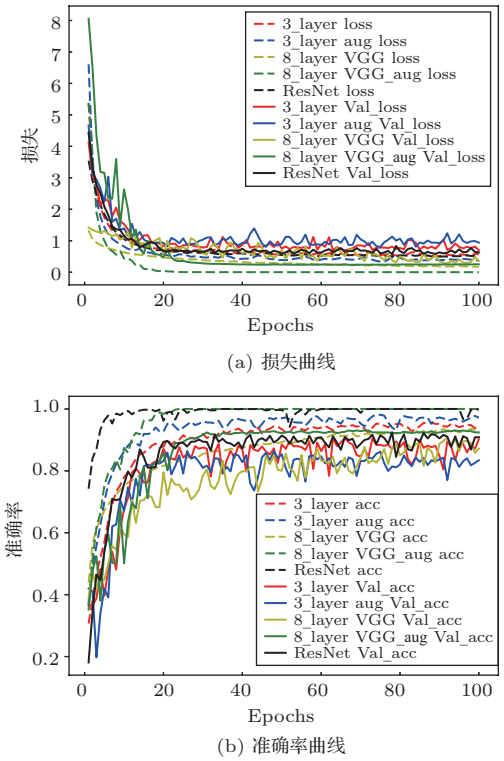
类别	细分船只	帧数
Little boat	fishing boats, trawlers, mussel boats, tugboats, dredgers	98/28/ /95/23/52
Moto boat	motorboats, pilot boats, sailboats	195/26/76
Passenger	passenger ferries	703
Ocean boat	ocean liners, ro-ro vessels	174/261
Nature Noise	background noise	225

结合本文所述方法和网络模型，采用 Tensorflow2.1 和 Librosa[16] 模块对声频信号进行处理，具体参数如下：

特征提取：将输入数据下采样到 22050 Hz，STFT 采样窗口为 2048，步长移动率为 25%，然后在加权谱图上利用 Mel 滤波器组，生成 128 bin 和 43 帧/s 的 Mel 频谱；

数据增强：时域拉伸的尺度选择为 rate = {0.8, 1.2}，音调变换中音高 pitch = { − 2, 2}，在时频谱图掩蔽中，时间掩蔽的最大值为 $T = 30$，时间掩蔽数 $m\_T = 2$，频率掩蔽最大值为 $F = 13$，频率掩蔽数 $m\_F = 2$，输入信号通过数据增强变换后，共产生 10 倍的数据增强；

网络训练：将输入数据归一化为零均值和单位方差，采用 Adam 进行优化，损失函数为交叉熵 (Cross entropy)，在卷积层中，采用基于高斯分布的方法进行随机初始化，以 ReLU 作为激活函数，使用 softmax 函数来获得每个目标类别的概率。训练中，共采用 100 个 Epoch，batch_size = 256，初始学习速率为 1×10^{-4}，50 ～ 100 迭代周期时线性衰减，最小达到 5×10^{-6}。



(a) 损失曲线

(b) 准确率曲线

图 4　不同方法的目标分类性能对比

Fig. 4　Comparison of target classification performance of different methods

2.2　实验结果

本文分别验证数据增强和改进的 VGG 网络模型对于水声信号的分类识别结果，如图 4 所示，图 4(a) 为训练周期中的训练损失曲线，图 4(b) 为识别率曲线，其中，虚线表示训练集的测试结果，实线表示测试集结果。本文共对比了 5 种方法，分别为

(1) 红色曲线表示3层的CNN网络，记为3_CNN；

(2) 蓝色曲线为数据增强条件下的3层CNN网络，记为3_CNN_Aug；

(3) 黑色曲线为利用ResNet网络进行迁移学习的测试结果，该网络以ImageNet训练权重进行初始化，通过添加全连接层进行目标分类，利用本文的数据集对后30层网络进行微调实现迁移学习，记为ResNet；

(4) 黄色曲线为采用改进的VGG网络进行的测试，该网络共包括8个卷基层，记为8_VGGish；

(5) 绿色曲线为本文方法，采用数据增强和改进VGG网络的测试结果，记为8_VGGish_Aug。

从图4中分析可知，分别对比3_CNN、8_VGGish方法和3_CNN_Aug、8_VGGish_Aug方法，在相同的网络参数下，在一定范围内更深层的网络结构可以取得更好的分类性能，8_VGGish、8_VGGish_Aug方法以VGG模型为基础，选取8层网络进行测试，取得了较好的效果。对比ResNet和8_VGGish方法，ResNet方法为目前流行的ResNet-50网络，但是由于水声信号谱图中的纹理、梯度等特征不明显，细节信息较少，过深的网络容易造成梯度消失，ResNet网络的性能相比于VGGish较差。最后分别对比3_CNN、ResNet和3_CNN_Aug、8_VGGish_Aug方法，在相同网络结构下，数据增强后的分类性能均有了一定的提高。综上所述，网络结构的改进和数据增强均有助于分类性能的提高，本文所提的方法取得了最好的分类性能。表3列出了不同分类方法的对比，通过数据增强和网络模型的构建后，最终取得了95.2%的分类准确率。

表3　不同方法的分类准确率

Table 3　The classification accuracy of different methods

方法	分类准确率/%
3_CNN	80.6
3_CNN_Aug	83.5
ResNet	82.1
8_VGGish	90.9
8_VGGish_Aug	95.2

图5为本文方法识别结果的混淆矩阵(Confusion matrix)，可用来呈现算法性能的可视化效果，每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；每一行代表了数据的真实归属类别，每一行的数据总数表示该类别的数据实例的数目。如第二行的Moto boat，共有测试样本68个，正确分类结果为54个，误分类为Passanger、Ocean boat和Nature Noise的个数分别为2个、11个和1个，通过混淆矩阵能够很快地分析每个类别的误分类情况。

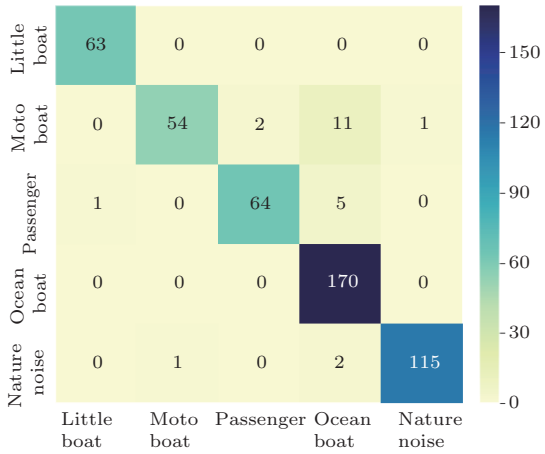


图5　5类目标分类的混淆矩阵

Fig. 5　Confusion matrices for five types of targets

3　结论

本文以典型的船舶类水下辐射噪声信号为研究对象，以水声信号的分类识别为目的，研究了深度学习方法在水声信号分类识别领域的应用能力。提取Mel功率谱图作为特征，构建基于VGG网络的分类模型框架，验证了在数据增强条件下的分类性能，最终的分类性能达到95%。但是本文采用的试验数据较少且训练集、测试集中的数据属于同一次试验采集，在实际情况下的水声信号种类更多，环境噪声也更复杂。因此，深度学习方法在更加复杂环境下的识别应用还有待进一步进行研究。

参　考　文　献

[1] Wang S, Zeng X. Robust underwater noise targets classification using auditory inspired time–frequency analysis[J]. Applied Acoustics, 2014, 78: 68–76.

[2] 程玉胜, 张宝华, 高鑫, 等. 船舶辐射噪声解调谱相位耦合特性与应用 [J]. 声学学报, 2012, 37(1): 25–29.

Cheng Yusheng, Zhang Baohua, Gao Xin, et al. Phase-coupling characteristics of ship radiated-noise demodulation spectrum and application[J]. Acta Acustica, 2013, 32(1): 36–36.

[3] Leal N, Leal E, Sanchez G. Marine vessel recognition by acoustic signature[J]. ARPN Journal of Engineering and Applied Sciences, 2015, 10(20): 9633–9639.

[4] Wang W, Li S, Yang J, et al. Feature extraction of underwater target in auditory sensation area based on MFCC[C]//2016 IEEE/OES China Ocean Acoustics (COA). IEEE, 2016: 1–6.

[5] 张少康, 田德艳. 水下声目标的梅尔倒谱系数智能分类方法 [J]. 应用声学, 2019, 38(2): 267–272.
Zhang Shaokang, Tian Deyan. Intelligent classification method of Mel frequency cepstrum coefficient for underwater acoustic targets[J]. Journal of Applied Acoustics, 2019, 38(2): 267–272.

[6] Zhao Z Q, Zheng P, Xu S, et al. Object detection with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212–3232.

[7] Kamal S, Mujeeb A, Supriya M H. Novel class detection of underwater targets using self-organizing neural networks[C]//2015 IEEE Underwater Technology (UT). IEEE, 2015: 1–5.

[8] 王强, 曾向阳. 深度学习方法及其在水下目标识别中的应用 [C]// 中国声学学会水声学分会 2015 年学术会议论文集, 2015.

[9] Targ S, Almeida D, Lyman K. Resnet in resnet: generalizing residual architectures[J]. arXiv preprint, arXiv: 1603.08029, 2016.

[10] Iandola F, Moskewicz M, Karayev S, et al. Densenet: implementing efficient convnet descriptor pyramids[J]. arXiv preprint, arXiv: 1404.1869, 2014.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 2014.

[12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, 2012: 1097–1105.

[13] McDonnell M D, Gao W. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths[C]//ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 141–145.

[14] Koutini K, Eghbal-Zadeh H, Dorfer M, et al. The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification[C]//2019 27th European signal processing conference (EUSIPCO). IEEE, 2019: 1–5.

[15] Park D S, Chan W, Zhang Y, et al. Specaugment: a simple data augmentation method for automatic speech recognition[J]. arXiv preprint, arXiv: 1904.08779, 2019.

[16] McFee B, Raffel C, Liang D, et al. Librosa: audio and music signal analysis in python[C]//Proceedings of the 14th Python in Science Conference, 2015, 8: 18–25.

[17] Santos-Domínguez D, Torres-Guijarro S, Cardenal-López A, et al. ShipsEar: an underwater vessel noise database[J]. Applied Acoustics, 2016, 113: 64–69.