DeepESC 网络的环境声分类方法研究

阴法明[1]，王诗佳[2]，赵 力[2]
(1. 南京信息职业技术学院通信学院，江苏南京 210023；2. 东南大学信息科学与工程学院，江苏南京 210096)

摘要：为进一步提升环境声分类的识别率，提出了一种仿深度隐藏身份特征 (Deep Hidden Identity Feature, DeepID)网络连接方式的卷积神经网络——深度环境声分类网络(Deep Environment Sound Classification, DeepESC)。DeepESC 网络共有六层——三层卷积层、两层全连层以及一层聚合层，为使网络在自动抽取高层次特征的同时能有效地兼顾低层次特征，网络将三层卷积层的输出聚合为一层，该层充分包含不同层次的特征，提升了卷积神经网络的特征表达能力。ESC-10 和 ESC-50 数据集上的仿真结果表明：在相同的识别框架下，与随机森林分类器相比，本文网络识别率分别平均提升了 7.6%和 22.4%，与传统的卷积神经网络相比，识别率分别平均提升 4%和 2%，仿真实验验证了本文分类器的有效性。

关键词：卷积神经网络；环境声分类；DeepID 网络
中图分类号：TB52+9 **文献标识码**：A **文章编号**：1000-3630(2019)-05-0590-04
DOI 编码：10.16300/j.cnki.1000-3630.2019.05.018

Environmental sound classification using DeepESC convolutional neural networks

YIN Fa-ming[1], WANG Shi-jia[2], ZHAO Li[2]
(1. *Nanjing College of Information Technology*, *Nanjing* 210023, *Jiangsu*, *China*;
2. *School of Information Science and Engineering, Southeast University*, *Nanjing* 210096, *Jiangsu*, *China*)

Abstract: To improve the accuracy of environmental sound classification, a new convolutional neural network named DeepESC, which imitates the connection of DeepID network, is proposed. DeepESC is composed of three convolution layers, two fully connected layers and one concatenate layer. To extract both high-level features and low-level features effectively, a concatenate layer is designed to join all convolution layers' output together, which comprises all features of different levels in the DeepESC network. Experimental results on ESC-10 and ESC-50 data sets show that, compared with random forest classification in same conditions, the accuracy of DeepESC is improved by 7.6% and 22.4% respectively, and by 4% and 2% respectively compared with the traditional convolutional neural network.

Key words: convolution networks; environmental sound classification; DeepID network

0 引 言

由于镜头角度固定、光线偏弱等原因，传统的人工视觉系统领域的监控系统的性能受到较多限制，而基于环境声的系统往往能够稳定工作，弥补视觉监控系统的不足。在环境声的系统中，环境声识别是研究的重点，开展针对环境声识别的研究具有较强的实际意义。

在环境声分类中，分类器的选择在一定程度上决定了系统的性能，因此，国内外学者针对该问题进行了大量的研究。在以往的研究中，通常以随机森林(Random Forest)[1]、支持向量机(Support Vector Machine, SVM)[2]和高斯混合模型(Gaussian Mixed Model, GMM)[2]作为主流分类器进行识别。尽管这些传统的分类器已经取得了一定的效果，但离人们的期望仍有一定的差距。

在环境声识别领域，一些研究者尝试卷积神经网络算法(Convolutional Neural Networks, CNN)[3-5]，并取得一定的成果。比如，具有独特的网络结构和特征提取算法的 DeepID 网络[4]，在人脸识别领域达到了 99%的成功率。但环境声分类问题不同于人脸识别，环境声片段是一维的时间序列数据，而人脸图像则是具有特殊拓扑结构的二维数据，因此，DeepID 网络并不能直接应用于环境声分类问题。基于此，本文首先将一维的环境声数据转换为二维的梅尔倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)图像，并使用卷积神经网络作为分类器，采用 DeepID 特有的网络连接方式组织网络，从而构建了可以直接用于环境声分类的 DeepESC 网络。

收稿日期：2018-05-13；**修回日期**：2018-07-06
基金项目：国家自然科学基金(61571106)
作者简介：阴法明(1980－)，男，山东肥城人，硕士，副教授，研究方向为信号处理。
通讯作者：阴法明，E-mail: yinfm@njcit.cn

此外，传统的声音事件特征以 MFCC 为主[6-7]，为进一步挖掘 MFCC 内在特征，发挥图像的多通道优势，本文在 MFCC 图像的基础上，提取出 MFCC 的 1 阶至 5 阶差分特征，再加上原 MFCC 图像，总共形成 6 通道图像特征，构成最终的输入特征。数据集 ESC-10 和 ESC-50 上的仿真实验验证了本文模型的有效性。

1　相关理论

1.1　卷积神经网络

一个典型的卷积神经网络由输入层、若干卷积层和池化层、少量的全连层和最后一层输出层(分类器)组成。卷积层和池化层一般交替出现。卷积层的作用是提取图像的特征；池化层的作用是对特征图进行压缩，降低计算复杂度，提高特征提取的鲁棒性。卷积层和池化层一般交替出现在网络中，全连接层负责把提取的特征图连接起来，最后通过分类器得到最终的分类结果。一张特征图中的所有元素都是通过一个卷积核计算得出的，也即一张特征图共享了相同的权重和偏置项。这一结构使得卷积神经网络能够利用输入数据的二维结构。与其他深度学习结构相比，卷积神经网络在图像和语音识别方面能够给出更好的结果。

卷积神经网络的低层卷积层所抽取的特征，往往是局部的，高层卷积层抽取的特征源于低层卷积层的输出，层数越高学到的特征就越全局化。在实际应用中，往往使用多层卷积，然后再使用全连接层进行训练[7]。

1.2　DeepID 网络

DeepID 网络包括 8 层网络结构：4 个卷积层，3 个池化层，1 个全连接层。全连接得到的是 160 特征向量，最后根据 160 维向量进行 SVM 或者 Softmax 分类。为了克服多层卷积导致的局部特征丢失的问题，DeepID 网络 3 个池化层的输出与第 4 个卷积层的输出连接后传播至全连接层，使得网络既能利用局部特征，又能利用全局特征。

2　环境声分类网络 DeepESC

环境声片段的 MFCC 图像与传统图像相比，仅有单通道，像素级的信息相对较少，并且局部相关性强。传统 CNN 的各卷积层在逐层细化提取图像特征的同时，也在丢失粗粒度、低层次的特征，这使得原本像素信息相对较少的 MFCC 图像在 CNN 网络中最顶层的信息维度偏低。

本文通过参考 DeepID 卷积神经网络的结构，针对环境声分类问题，构造出应用于环境声分类问题的 DeepESC 卷积神经网络。用 I^l 表示第 l 层的输入，K^l 表示 DeepESC 第 l 层卷积层的卷积核，第 l 层的特征图 F^l 可以表示为

$$F^l=(K^l\times I^l)=\sum_m\sum_n I^l(i-m,j-n)K^l(m,n) \tag{1}$$

通过把前三层卷积层所提取出的特征图互相连接在一起，可以得到新的特征图。但由于三层卷积层的特征图具有不同的维度，因此按式(2)将特征图展开为一维特征：

$$F^l_{\text{flat}}(i+j\times m+c\times m\times n)=F^l(i,j,c) \tag{2}$$

其中，m, n 表示第 l 层卷积核的尺寸，i 和 j 分别表示像素索引，c 表示特征图的通道数。

再将展平的各层特征图连接，得到最终的融合特征图：

$$F_{\text{final}}=(F^0_{\text{flat}},F^1_{\text{flat}},\cdots,F^l_{\text{flat}}) \tag{3}$$

从式(3)可知，所有卷积层提取所得的特征图融合在一起获得了 F_{final}，最终 F_{final} 作为融合特征输入 DeepESC 的全连层进行分类识别。DeepESC 的网络结构见图 1，各层参数如表 1 所示。

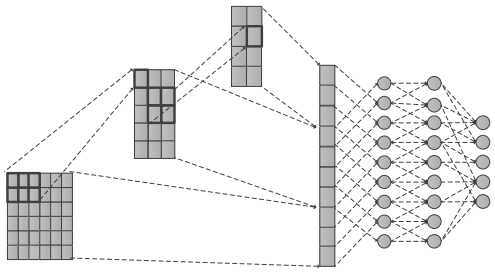

图 1　DeepESC 网络结构
Fig.1　DeepESC network structure

表 1　DeepESC 网络结构参数
Table 1　DeepESC network parameters

No.	层	维度		通道数	卷积核	步长
		行	列			
0	输入	60	41	6	-	-
1	卷积	58	40	64	3,2	3, 2
2	池化	29	20	64	2,2	2, 2
3	卷积	28	18	96	2,3	2, 3
4	池化	14	9	96	2,2	2, 2
5	卷积	12	8	128	3,2	3, 2
6	池化	6	4	128	2,2	2, 2
7	平坦	1	148 480	1		-
8	平坦	1	48 384	1		-
9	平坦	1	12 288	1		-
10	级联	1	209 152	1		-
11	全连接	1	2 048	1		-
12	全连接	1	2 048	1		-
13	Softmax	1	x	1		-

由于本文所用数据量较小，且 DeepESC 网络层数较多，在训练过程中产生了较强的过拟合现象。为对抗过拟合，本文采用 Dropout 算法[8]，根据卷积层以及全连层的过拟合程度不同，分别对全连接层、DeepESC 卷积层进行比例为 0.5 和 0.2 的 Dropout 算法处理。

3 实验仿真

3.1 数据集

本文采用公开数据集 ESC-10 以及 ESC-50[9]。ESC-50 数据集是 2 000 个环境音频样本集合，每个样本长度是 5 s，共 50 类声音，采样率为 44.1 kHz，适用于环境声音分类算法测试。ESC-10 数据集是 ESC-50 的子数据集，包含 10 个类别，每个类别 40 个样本，共 400 个环境声样本，总时长为 33 min。

神经网络容易出现过拟合现象，因此需要更多的训练数据。本文采用了文献[10]和文献[11]中的方法，根据环境声数据的类别，对样本进行不同程度的移调和时间伸缩，以此扩充数据集。由此，ESC-10 数据集被扩大了 10 倍，ESC-50 数据集被扩大了 4 倍。进行数据扩充后的 ESC-10 和 ESC-50 数据集被用于提取梅尔频谱特征，并进行分段形成最终的样本集合。ESC-10 数据集最终共包含 1500 个样本，ESC-50 则含有 7 200 个样本。

3.2 实验相关参数设计

预处理及特征提取：为提高算法的有效性，首先通过端点检测去除样本语音的静默片段。然后以 22.050 kHz 的频率对样本进行重采样，对样本分帧并计算快速傅里叶变换(Fast Fourier Transform, FFT)，其中，FFT 点数为 512，帧重叠率为 50%。之后，使用 60 个子带滤波器组成梅尔滤波器组，计算得到梅尔频谱，并将其分为等长的若干段，段重叠率为 50%，以段作为单元进行识别。每段共 41 帧，时长约 930 ms。在梅尔频谱图像基础上，利用 Librosa 软件包[12]提取其 1 阶至 5 阶的差分特征，最终构成 6 通道的图像输入特征。

训练网络：本文采用目前流行的深度学习框架 Caffe 搭建训练网络[13]。在深层神经网络(Deep Neural Networks, DNN)中超参数的选择对网络的训练乃至最后网络的收敛状态有着极大的影响[14]。目前，只能通过启发式搜索来寻找一个较优解[15]的方式选择网络的超参数。通过多次实验与比较，最终确定的网络超参数见表 2。

对比分类器及其参数：(1) 随机森林分类器，最大深度为 6，最大估计量为 100[9]；(2) CNN，两层卷积层，卷积核尺寸分别为(57, 6)和(1, 3)，后接池化层的池化核尺寸均为(2, 2)，最后为两层具有 5 000 个神经元的全连层[16]；(3) DNN，共 5 层神经元数目为 384 的全连层，Dropout 比率为 0.5[17]

评估标准：环境声识别中，以国际上通用的准确率作为评估指标。

表 2 训练超参数表

Table 2 Hyper parameters for training

超参	数据集	
	ESC-10	ESC-50
Dropout 比率	0.2	0.3
动量	0.90	0.95
学习速率	0.002 5	0.000 5
学习策略	Fixed	Fixed
优化方法	SGD	ADAM[18]
正则化	L2 Norm	L2 Norm
权重衰减	0.000 5	0.000 1
批尺寸	64	512
迭代次数	50	80

3.3 对比实验

本文模型最终的分类准确率通过五折交叉验证得到，其中，每份验证集中均不包含扩充数据集中的音频片段，只包含原始的音频片段，扩充的环境声片段只用于训练网络。

为使模型评估更具对比度，在相同特征的基础上(MFCC)，将 DeepESC 网络与随机森林(Random Forests)分类器以及传统 CNN 分类器[16]，在相同数据集 ESC-10 和 ESC-50 上进行了比较。此外，为了对比卷积层提取特征的作用，本文构建了一个 5 层深层神经网络，并在 ESC 数据集上训练测试。

表 3 给出了 4 种分类器在 ESC-10 数据集和 ESC-50 数据集上的实验结果。与随机森林分类器相比，在 2 个数据集上，DeepESC 分别提升了 7.6%，22.4%，卷积神经网络在环境声分类问题上所表现出的性能优于传统分类器；与 DNN 相比，DeepESC 网络的识别率分别提升了 17.5%，23.6%。由于具有卷积层，因此 DeepESC 网络识别率属于深层神经网络 CNN，卷积神经网络由于具有局部区域连接、权值共享、降采样的结构特点，使其在图像处理和语音识别领域表现出色。与传统 CNN 相比，DeepID 网络通过连接各个卷积层的输出，融合了多个层次的特征，从而能更大程度上地保留特征信息[19]。而本文在 DeepID 网络的基础上增加两层全连层构成 DeepESC 网络，该结构能保留不同维度的信息，并

增加 Softmax 层，使得 DeepESC 能直接对环境声进行分类，改变了 DeepID 仅提取特征而不进行分类的模式。因此，DeepESC 较传统 CNN 识别率分别提高了 4%和 2%。

表 3　不同分类器的识别率对比
Table 3　Accuracy comparison of different classifiers

数据集	随机森林	DNN	卷积网络	DeepESC
ESC-10	74.7%	64.8%	79.0%	82.3%
ESC-50	43.3%	42.1%	64.2%	65.7%

从整体的计算复杂度和空间复杂度来看，DNN 的空间复杂度约为 10^6 的量级，卷积网络则为 10^7，DeepESC 也同样为 10^7。在同样使用 GPU 计算的情况下，三种神经网络的前向推理所耗费的时间基本相同，都为 10 ms 左右。可见，在牺牲了一定的存储空间下，DeepESC 通过增加网络容量，提高了识别的精度。

4　结 语

本文尝试利用卷积神经网络解决环境声分类问题，并取得了优于传统模型的识别率，从而证明了卷积神经网络对环境声分类的可行性。此外，在传统卷积神经网络的基础上，通过参考 DeepID 的特殊网络连接方式，构建适用于环境声分类的 DeepESC 网络。实验结果表明，DeepESC 网络以特殊的网络连接方式获取了更多层次的特征，并且由此达到比传统卷积神经网络更高的分类识别率，在环境声分类问题上有较好的应用前景。

参 考 文 献

[1] PHAN H. Random regression forests for acoustic event detection and classification[J]. IEEE ACM Transactions on Audio Speech & Language Processing, 2015, **23**(1): 20-31.

[2] ZIEGER C, OMOLOGO M. Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm[C]//INTERSPEECH 2008, Conference of the International Speech Communication Association, Brisbane, Australia, September. DBLP, 2008: 115-118.

[3] HAN Y, LEE K. Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation[J]. ArXiv Preprint ArXiv, 2016: 1607.02383.

[4] ELIZALDE B, KUMAR A, SHAH A, et al. Experiments on the DCASE Challenge 2016: acoustic scene classification and sound event detection in real life recording[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop(DCASE2016). Budapest, Hungary, 2016: 20-24.

[5] ZÖHRER M, PERNKOPF F. Gated recurrent networks applied to acoustic scene classification and acoustic event detection[C]// Presented at the Detection and Classification of Acoustic Scenes and Events 2016 (DCASE 2016), 2016: 115-119.

[6] VU, TOAN H., AND JIA-CHING WANG. Acoustic scene and event recognition using recurrent neural networks[C]//Detection and Classification of Acoustic Scenes and Events 2016, Budapest, Hungary, 2016.

[7] 陶锐，孙彦景，刘卫东. 多重水印快速加密技术在图像深度传感器中的应用[J]. 传感技术学报, 2018, **31**(12): 159-164.
TAO Rui，SUN Yanjing，LIU Weidong. Application of multi watermark fast encryption technology in image depth transduce[J]. Chinese Journal of Sensors And Actuators, 2018, **31**(12): 159-164.

[8] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, **15**(1): 1929-1958.

[9] PICZAK K J. ESC: Dataset for environmental sound classification [C]//ACM International Conference on Multimedia, ACM, 2015:1015-1018.

[10] SUN Y, WANG X, TANG X. Deeply learned face representations are sparse, selective, and robust[C]//Computer Vision & Pattern Recognition. 2015: 2892–2900.

[11] Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models[J]. Journal of Time Series Analysis, 1994, **15**(2): 183-202.

[12] MCFEE B, RAFFEL C, LIANG D, et al. Librosa: Audio and music signal analysis in Python[C]//Proc. of the 14th Python in Science Conf. (SCIPY 2015), 2015: 18-24.

[13] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]//Acm International Conference on Multimedia, 2014: 675–678.

[14] POVEY D, ZHANG X , KHUDANPUR S . Parallel training of deep neural networks with natural gradient and parameter averaging[C]// Computing Research Repository(CoRR 2014), 2014: 1410-7455.

[15] BERGSTRA J, BENGIO Y. Random search for Hyper-Parameter optimization[J]. Journal of Machine Learning Research, 2012, **13**(1): 281-305.

[16] PICZAK K J. Environmental sound classification with convolutional neural networks[C]//2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015: 1-6.

[17] HERTEL L, PHAN H, MERTINS A. Comparing time and frequency domain for audio event recognition using deep learning[C]//2016 International Joint Conference on Neural Networks (IJCNN). Vancouver, BC, 2016: 3407-3411.

[18] Diederik P. Kingma, Jimmy Ba. Adam: A method for stochastic optimization[J]. ArXiv Preprint ArXiv, 2014: 1412. 6980.

[19] 陶锐. 面向电子票据认证的数字水印加密算法研究[D]. 中国矿业大学, 2018.
TAO Rui. Research on digital watermarking encryption algorithm for electronic bill authentication[D]. China University of Mining and Technology, 2018.