ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 960－968
DOI:10. 16337/j. 1004-9037. 2021. 05. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多级残差网络的环境声音分类方法

曾金芳，李友明，杨恢先，张　钰，胡雅欣

（湘潭大学物理与光电工程学院，湘潭 411105）

摘　要：为了对环境声音进行更好的识别和分类，提出了基于多级残差网络（Multilevel residual network，Mul-EnvResNet）的环境声音分类方法。对声音事件进行时标和基频压扩之后，提取其梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCCs），以及它们的差分作为特征参数送入Mul-EnvResNet对声音事件进行分类。实验数据集采用ESC-50，将Mul-EnvResNet模型与端到端的卷积神经网络（EnvNet）、基于注意力机制的循环神经网络（Attention based convolutional recurrent neural network，ACRNN），以及受限卷积玻尔兹曼机的无监督滤波器组模型（Convolutional restricted Boltzmann machine，ConvRBM）进行对比实验。实验结果表明，Mul-EnvResNet取得了89.32％的最佳分类准确率，相较上述3种模型在分类准确率上分别有18.32％、3.22％、2.82％的提升，相较于其他的声音分类方法也均有明显的优势。

关键词：环境声音分类；多级残差网络；时标压扩；基频压扩

中图分类号：TN912　　**文献标志码：**A

Environmental Sound Classification Method Based on Multilevel Residual Network

ZENG Jinfang，LI Youming，YANG Huixian，ZHANG Yu，HU Yaxin

（School of Physics and Optoelectronics，Xiang Tan University，Xiangtan 411105，China）

Abstract：To better identify and classify environmental sound，a multilevel residual network（Mul-EnvResNet）is proposed for environmental sound classification. After time stretch and pitch shift for sound events，the Mel-frequency cepstral coefficients（MFCCs）and their deltas are extracted as feature parameters and sent into the Mul-EnvResNet to classify sound events. The experimental data set uses ESC-50，Mul-EnvResNet is compared with the end-to-end convolutional neural network（EnvNet），the attention based convolutional recurrent neural network（ACRNN）and the unsupervised filterbank learning using convolutional restricted Boltzmann machine（ConvRBM）. The experimental results show that，Mul-EnvResNet achieves the best accuracy rate of 89.32％ in terms of classification accuracy，compared with the above three models，the classification accuracy has been improved by 18.32％，3.22％ and 2.82％，respectively，which also has obvious advantages compared with other sound classification methods.

Key words: environmental sound classification; multilevel residual network; time stretch; baseband stretch

---

基金项目：国家自然科学基金(62071411)资助项目；湖南省自然科学基金(2018JJ3486)资助项目。
收稿日期：2020-11-25；**修订日期：**2021-02-28

引　言

声音在人类与环境的互动中起着至关重要的作用,因此,对于人工智能来说,机器或计算机能够像人类一样理解声音是极其关键的。近年来,声音事件的检测和分类研究引起了学者的极大兴趣。目前,关于环境声音分类(Environmental sound classification,ESC)的研究越来越多,并广泛应用于智能家居、助听器、情景感知[1]和安防监控[2]等。ESC任务早期尝试使用信号处理技术,包括矩阵分解[3]、梅尔频率倒谱系数(Mel-frequency cepstral coefficients,MFCCs)特征集、高斯混合模型(Gaussian mixed model,GMM)[4]和受限卷积玻尔兹曼机的无监督滤波器组模型(Convolutional restricted Boltzmann machine,ConvRBM)[5]等。之后许多学者提出将机器学习算法应用于声音事件识别,如Phan等[6]提出使用随机森林算法,Zieger等[7]提出使用支持向量机(Support vector machine, SVM)的方法等。近年来,随着深度学习的兴起与发展,基于深度学习的方法也被广泛用于ESC任务,如Tokozume等[8]构建了基于卷积神经网络(Convolutional neural network, CNN)的端到端分类系统(EnvNet),Zhang等[9]提出基于注意力机制的卷积循环神经网络(Attention based convolutional recurrent neural network, ACRNN)系统,以及刘亚荣等[10]提出结合美尔谱系数(Mel-frequency spectral coefficient, MFSC)与CNN对环境声音进行分类等。

尽管上述声音分类方法已经很好地应用于不同领域,但是它们也具有一定的局限性。随机森林在某些噪音较大的分类或回归问题上容易出现过拟合现象,且参数较复杂,模型训练和预测都比较慢[11]。对于SVM算法,如果数据的特征数多于样本数,则SVM的表现很差,且经典的SVM算法只能解决小样本下的二分类问题。用CNN作为环境声音的分类器,相较于以往的方法在分类效果上确实能取得更佳的准确率,然而之前提出的CNN模型层数均比较浅,例如,Piczak[11]首次提出用于声音分类的CNN模型时,网络结构只有2个卷积层和3个全连接层。如今越来越深层次的CNN模型[12-15]被提出,并在ImageNet或其他基准数据集上取得了更好的性能,这些模型的结果揭示了网络深度的重要性,证实在一定程度上更深的网络能够产生更好的效果。然而实验表明,深层次的CNN在ESC任务上较难训练,因此,为了避免通过单纯增加网络层数所带来的问题,本文提出使用多级残差网络(Multilevel residual network, Mul-EnvResNet)进行训练,其相较以往在ESC任务上使用的网络深度更深,且是首次提出将残差网络应用于ESC任务。另外,ESC任务的标记数据相对稀缺是CNN难以在较简单模型上改进的重要原因,虽然近年来已经发布了一些新的数据集[16],但它们仍然比可供研究的数据集要小得多。一个解决音频数据稀缺较好的方案是对数据进行时标和基频压扩,也就是说,对样本集合进行一种或多种变换,从而产生新的、额外的样本数据。

针对上述问题,本文提出了基于Mul-EnvResNet的ESC方法,主要贡献总结如下:

(1) 将残差网络用于ESC,提出了残差网络模型EnvResNet,相较于其他用于ESC任务的CNN模型大大加深了网络结构的层数,从而能够提取更深层次的重要特征,提高分类准确率。

(2) 在设计的EnvResNet基础上,继续构建短连接,提出了一种Mul-EnvResNet模型,进一步提高了模型的学习能力。

1　ESC技术

本文提出的基于Mul-EnvResNet的ESC流程如图1所示。首先,对声音信号进行时标和基频压扩,随后提取其MFCCs以及它的差分,并将提取的特征集按一定比例分为训练集和验证集,以便后续对模

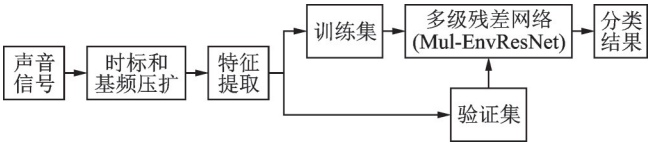

图1　基于Mul-EnvResNet的ESC流程图

Fig.1　ESC process based on Mul-EnvResNet

型的训练和测试，接着将训练集送入Mul-EnvResNet进行训练，最后输出验证数据集在模型上的泛化性能。

2 残差块结构与原理

到目前为止，已经有各种信号处理和机器学习技术应用于ESC，包括矩阵分解、字典学习、小波滤波器组和深度神经网络[17]等。之前提出的CNN模型均层数较少，而越来越多的研究证明深层次的网络提取的特征往往比从浅层网络中提取得更好。但随着网络的不断深化，两个问题不可避免，梯度消失和梯度爆炸[18]。对数据进行合理的初始化和正则化在一定程度上可以解决这两个问题，但却会产生新的难点，就是网络性能的退化，即随着深度加深时错误率也将跟着上升。为了解决这个问题，He等[15]提出了深度残差网络(Deep residual network, ResNet)，通过在原始卷积层外部加入短连接构成残差块，如图2所示。

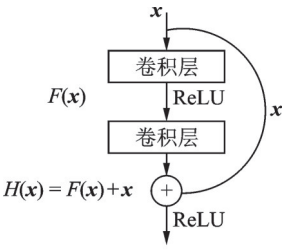

图2 残差块的结构

Fig.2 Structure of residual block

残差块函数可被描述为

$$\boldsymbol{y}_l = \boldsymbol{x}_l + F(\boldsymbol{x}_l) \tag{1}$$

$$\boldsymbol{x}_{l+1} = f(\boldsymbol{y}_l) \tag{2}$$

式中：$\boldsymbol{x}_l$为第l个残差块的输入，$\boldsymbol{x}_{l+1}$为第l个残差块的输出，$F(\boldsymbol{x}_l)$为残差映射，f表示修正线性单元(Rectified linear unit, ReLU)。

$$\mathrm{ReLU}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} & \boldsymbol{x} > 0 \\ 0 & \boldsymbol{x} \leqslant 0 \end{cases} \tag{3}$$

相比于传统的神经网络激活函数，如tanh双曲函数和逻辑函数等，ReLU不需要进行指数运算使得网络模型整体计算成本下降，且收敛较快，能够更高效地进行梯度下降和反向传播，避免梯度爆炸和梯度消失问题。此外，在与原输入值$\boldsymbol{x}$相加并执行激活函数前，可以看到$H(\boldsymbol{x}) = F(\boldsymbol{x}) + \boldsymbol{x}$，即$F(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$，当$F(\boldsymbol{x}) = 0$时，得到$H(\boldsymbol{x}) = \boldsymbol{x}$，这就实现了恒等映射，从而使得网络随深度加深时至少不会退化。ResNet在ILSVRC—2015年度比赛中获得第1名，证明了其相对于其他网络结构的优势。

3 网络结构设计

3.1 残差网络结构设计

本文在设计Mul-EnvResNet之前，为了选取较优的模型设计策略，首先提出了一个残差网络并将其命名为EnvResNet，模型结构如图3所示。

EnvResNet的输入维度为60×41×2，首先，使用32个维度为3×3的卷积核来提取局部特征，如图3(a)所示。然后利用5个残差块提取更深层次的信息，5个残差块输出端使用的滤波器个数分别为64、128、128、256、256，每个残差块有一个短连接，这里短连接不再是$\boldsymbol{x}$，而是设计成一个卷积层$h(\boldsymbol{x})$，这样可以保证即使$F(\boldsymbol{x}) = 0$时，残差块也至少会有一个卷积层$h(\boldsymbol{x})$在运作，从而可以避免输入数据和输出数据的通道个数必须相等的问题，如图3(b)所示，$\boldsymbol{x}_{l+1}$的计算式为

$$\boldsymbol{x}_{l+1} = F(\boldsymbol{x}_l) + h(\boldsymbol{x}_l) \tag{4}$$

式中：$\boldsymbol{x}_l$为残差块的输入，$\boldsymbol{x}_{l+1}$为残差块的输出，$F(\boldsymbol{x}_l)$为残差映射，$h(\boldsymbol{x}_l)$表示短连接。在每一层的输出端都使用ReLU作为激活函数，经过5个残差块后进入池化层，池化的目的是保留主要的特征，同时减少参数和计算量，防止过拟合，提高模型泛化能力。接着，使用长短期记忆网络(Long short-term memo-

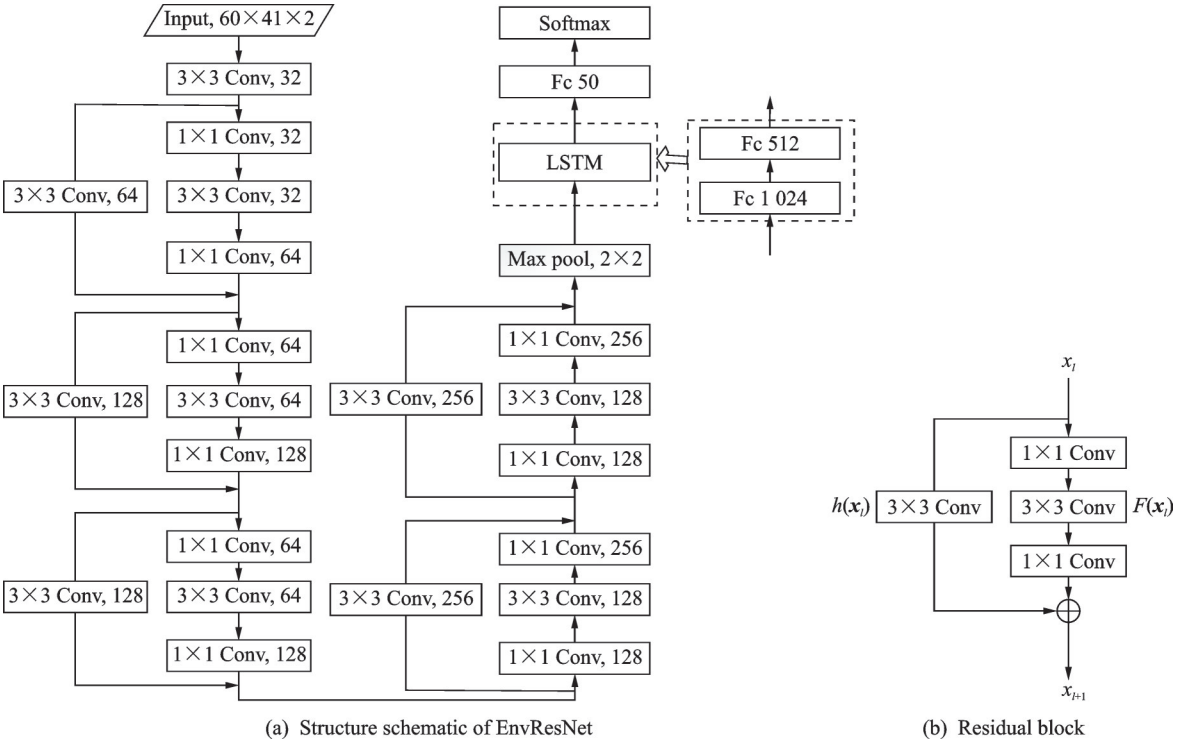

图3 EnvResNet结构与残差块

Fig.3 Structure of EnvResNet and residual block

ry, LSTM)对池化输出做进一步的处理,LSTM适合于处理与时间序列高度相关的问题。最后通过全连接层进入Softmax层对结果进行分类,Softmax的计算为

$$S_i = \frac{e^{V_i}}{\sum_i^C e^{V_i}} \tag{5}$$

式中:V_i为输入Softmax层的向量,也是前一级的输出;C为总的类别数量;i为当前类别的索引;S_i为当前元素对应的Softmax输出,Softmax将多分类的输出数值归一化为相对概率,便于清晰地理解和比较。

3.2 Mul-EnvResNet结构设计

本文最终提出的Mul-EnvResNet模型是通过对EnvResNet模型的后4个残差块进一步构建短连接而形成,如图4所示。

图4(b)中,2个子残差块和一个短连接组成了一个多级残差块。假设多级残差块的输入为x,经过2个子残差块得到的输出分别为$\boldsymbol{y}_1$和$\boldsymbol{y}_2$,经过多级残差块的最终输出为$\boldsymbol{y}_3$,则有

$$\boldsymbol{y}_1 = F_1(\boldsymbol{x}) + h_1(\boldsymbol{x}) \tag{6}$$

$$\boldsymbol{y}_2 = F_2(\boldsymbol{y}_1) + h_2(\boldsymbol{y}_1) = F_2(F_1(\boldsymbol{x}) + h_1(\boldsymbol{x})) + h_2(F_1(\boldsymbol{x}) + h_1(\boldsymbol{x})) \tag{7}$$

$$\boldsymbol{y}_3 = h_3(\boldsymbol{x}) + \boldsymbol{y}_2 = h_3(\boldsymbol{x}) + F_2(F_1(\boldsymbol{x}) + h_1(\boldsymbol{x})) + h_2(F_1(\boldsymbol{x}) + h_1(\boldsymbol{x})) \tag{8}$$

从输出$\boldsymbol{y}_2$中可以分析出,后一个子残差块会对前一个子残差块的输出进行残差映射,若残差块过

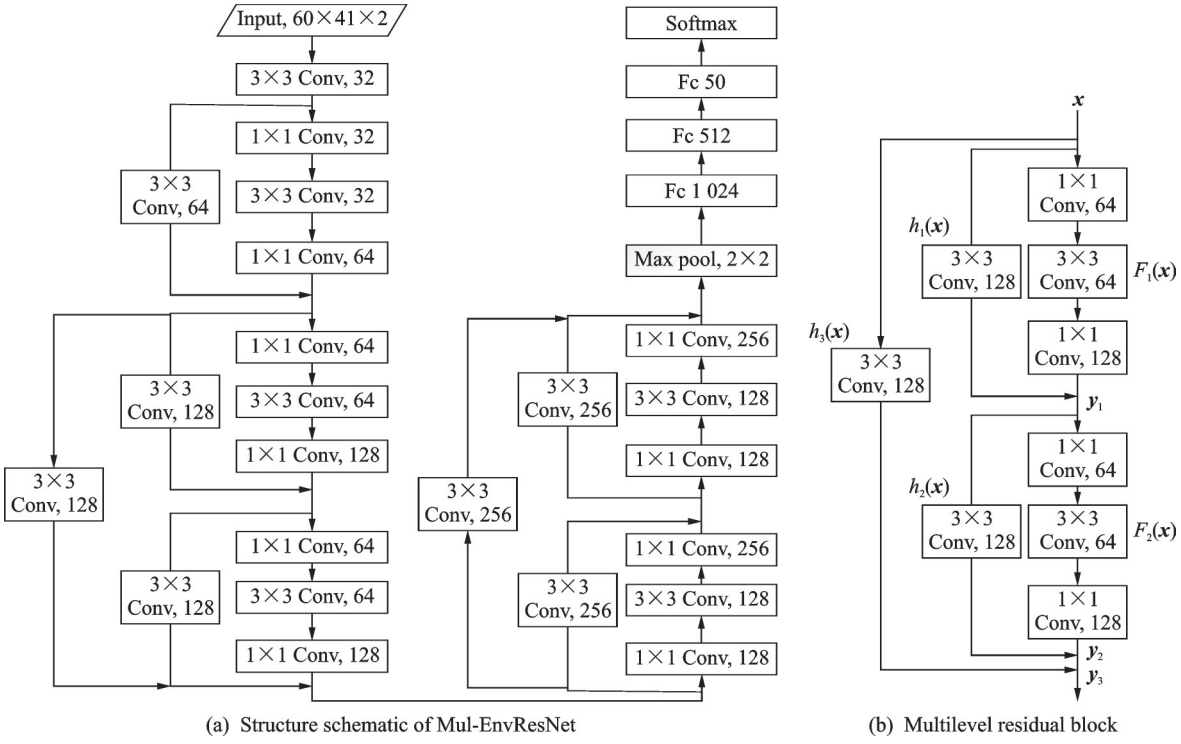

图4 Mul-EnvResNet结构与多级残差块

Fig.4 Structure of Mul-EnvResNet and multilevel residual block

多时,将逐层增加参数量并增大利用梯度下降法进行反向传播时的计算难度,从而影响模型最终的训练效果。设计成多级残差之后,从输出$\boldsymbol{y}_3$中可以分析出,当模型去拟合2个连续的子残差块比较困难时,反向传播会逐渐让$\boldsymbol{y}_2$趋近于0,那么训练的网络可以转换为去学习$\boldsymbol{y}_3=h_3(\boldsymbol{x})$,在相同的计算条件下,拟合一个卷积层$h_3(\boldsymbol{x})$显然比拟合2个子残差块容易[15]。所以Mul-EnvResNet在2个子残差块拟合较好时,能够充分利用子残差块内部卷积层特征向量的相关信息,而当2个子残差块训练较困难时,则可以转为学习一个较简单的卷积层,从而保证模型能产生较好的训练效果。

4 实验与结果分析

4.1 实验准备

本文使用公共数据库ESC-50[19]来完成ESC任务。ESC-50数据集包含有2 000个带标记的环境音频记录,分为50个语义类(每类40个样本),也可分为5大类,即动物、自然声和水声、人类非言语声音、室内声音和城市噪音;每个音频记录时长为5 s,适合于ESC任务的基准测试方法。

在进行特征提取之前,本文对音频记录进行了数据扩容,每个音频长度都是5 s,通过复制将其转换为10 s的录音,即将每个音频的持续时间加倍以增加训练样本的时长。在音频信号处理领域存在两个基本的数据扩容方法:时标压扩和基频压扩,前者是在时间维度上的一个尺度变换,后者是对音调的一个调整,而音调的高低取决于频率,频率越高音调越高,因此基频压扩可以看作是对频率的一个尺度变换。通过数据扩容,增加了训练样本的变化情况,能够改进训练后模型的泛化能力,从而对环境声音进行更好分类。

在对数据进行时标和基频压扩之后,首先对音频信号进行预处理,预处理包括滤波、A/D转换、预加重、分帧和加窗等,其中预加重用于优化和改善整个声音信噪比和数据的高频部分,从而使声音信号数据的频谱更平缓[20-21]。然后重新采样到22 050 Hz并归一化形成60个波段,从中提取MFCCs,并使用Librosa(一个用于音频、音乐分析与处理的Python第三方库,包含时频处理、特征提取和绘制声音图形等功能)将这些谱图分割成41帧。最后,对提取的特征进行差分并与提取的MFCCs进行垂直叠加作为网络的2通道输入(60×41×2)。

4.2 实验设置

本文实验的硬件环境选用CPU为Core i5-8400,显卡为GeForce GTX 1080Ti的服务器(11 GB显存)。在Ubuntu18.04操作系统上进行网络模型的训练和测试,深度学习框架为Keras2.2.4(一个由Python编写的开源人工神经网络库,可以作为Tensorflow和Theano的高阶应用程序接口,进行深度学习模型的设计、调试、评估、应用和可视化等)。为了验证模型中各模块的有效性,本文对各种设计策略进行了多组对照实验。为保证实验的公平性,对各组策略采用相同的训练参数设置。输入维度为60×41×2,批处理大小为64,迭代次数为100。优化器选用动量梯度下降法,动量为0.9,权重衰减为10^{-4},初始学习率为10^{-2},并逐渐降低。

4.3 实验结果与分析

在对残差网络模型EnvResNet进行设计时,对池化后是使用LSTM还是2个全连接层(分别为1 024和512个神经元)做对比实验,如图3(a)所示,同时对残差块的短连接使用不同的卷积核大小做对比实验,如图3(b)所示,因此有多种不同的设计组合。实验结果如表1所示。

表1 不同模型和不同卷积核大小的短连接下的准确率

Table 1 Accuracy of different models and shortcut with different convolution kernel sizes

短连接的卷积核大小	EnvResNet-LSTM/%	EnvResNet-Fc/%
1×1	81.31	84.61
3×3	83.28	88.35
5×5	83.12	87.54

由表1可知,在池化后使用两个全连接层比使用LSTM时性能更好,识别率有3%~5%的提升,且在实验过程中使用LSTM还出现了过拟合的现象,原因可能是经过残差后的特征已经失去时间序列的优势,从而使用全连接逐步缩小神经元的个数效果更佳。此外,可以看出,无论使用何种模型策略,短连接使用3×3的卷积核时,模型的识别率相较使用1×1与5×5的卷积核作为短连接时有1%~4%的提升。这里选用3×3的卷积核,一方面尽管在CNN中,卷积核越大,感知域越大,获取到上层的信息便会越多,但是大的卷积核会导致计算量激增,从而计算性能降低,反而不利于提升网络的深度,另一方面选用3×3的卷积核也不会如1×1的卷积核那样,使得提取的特征之间联系较小。

通过分析以上实验得出的结果,本文提出的EnvResNet在末尾的池化层之后选用全连接层、残差块的短连接选择3×3的卷积核时能达到最好的分类效果。而本文最终提出的Mul-EnvResNet则是通过对EnvResNet模型的后4个残差块进一步构建短连接而形成,能够实现更好的分类效果。Mul-EnvResNet的训练和测试曲线如图5所示。由于整个数据集并不大,因此划分到验证集的数据不是很多,从而导致了测试数据的准确率略微高于训练数据的准确率,但从图5中可以看出,总体的拟合效果良好,且并未出现过拟合的现象。

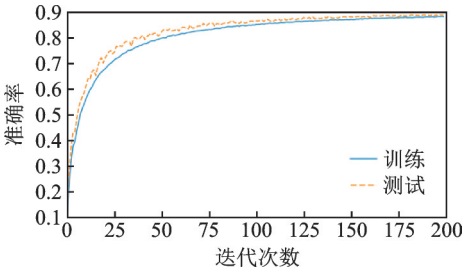

图5 Mul-EnvResNet训练和测试曲线图

Fig.5 Multilevel residual network training and test curves

Mul-EnvResNet与EnvResNet的实验结果对比如表2所示。由表2中数据可知，Mul-EnvResNet的分类准确率更高，作为代价，Mul-EnvResNet模型的参数量增加，且反向传播的计算变得更为复杂，训练难度加大，导致训练时间更长，且训练时的迭代次数更多，迭代200次才趋于收敛，较EnvResNet多迭代了100次。但模型训练结束后，两个模型对于验证集的分类时间是基本一致的，故从分类准确率上考虑，最终选用模型Mul-EnvResNet总体性能更好。

表2 不同模型下分类准确率和训练时间

Table 2 Classification accuracy and training time under different models

模型	分类准确率/%	训练时间/h
EnvResNet	88.35	11.6
Mul-EnvResNet	89.32	27.5

将本文提出的Mul-EnvResNet模型与CNN（Piczak FBEs-CNN[11]、logmel-CNN[8]、logmel-CNN ⊗EnvNet[8]）、音频数据相移（PEFBEs[22]、FBEs⊕PEFBEs[22]）、基于注意力机制的循环神经网络（ACRNN[9]）及ConvRBM[5]等其他声音分类方法进行比较，结果如表3所示。

表3 ESC-50上各模型对比实验结果

Table 3 Camparison of experimental results of various models on ESC-50

模型	分类准确率/%
Piczak FBEs-CNN[11]	64.50
logmel-CNN[8]	66.50
logmel-CNN ⊗EnvNet[8]	71.00
PEFBEs[22]	73.25
FBEs⊕PEFBEs[22]	84.15
ACRNN[9]	86.10
ConvRBM[5]	86.50
Mul-EnvResNet	89.32

表3中Piczak FBEs-CNN是首次将CNN用于ESC的模型，由2个卷积层和3个全连接层组成，而logmel-CNN和logmel-CNN ⊗EnvNet则是在Piczak FBEs-CNN的基础上进行了数据扩容。PEFBEs与FBEs⊕PEFBEs使用的网络模型仍是Piczak FBEs-CNN，但在特征提取时充分利用了相位信息，在一定程度上得到了更好的训练效果。基于注意力机制的循环神经网络ACRNN及ConvRBM分别取得了86.10%和86.50%的分类准确率，相较之前的模型具有一定优势，但本文提出的Mul-EnvResNet性能明显优于其他的声音分类方法，具有最高的分类准确率。

5 结束语

通过对复杂环境下声音分类技术进行研究，本文将残差网络用于ESC，提出了残差网络模型EnvResNet，在相同数据集下，与以往的CNN模型及其他的声音分类方法进行比较，取得了更高的分类准确率。此外，还提出了一个Mul-EnvResNet模型，进一步提高了模型在环境声音上的学习和泛化能力，取得了最高的分类准确率。但该模型在训练效率上并没有达到理想的效果，训练时间相对较长，这是今后工作需要重点解决的问题。另外，今后的工作也会把提出的Mul-EnvResNet模型应用到不同数据集上，如UrbanSound8K和RWCP数据集，并结合生成对抗网络和知识图谱等相关技术，对ESC技术做进一步研究。

参考文献：

[1] TANG Baolong, LI Yuanqing, LI Xuesheng, et al. Deep CNN framework for environmental sound classification using weighting filters[C]//Proceedings of IEEE International Conference on Mechatronics and Automation. [S.l.]: IEEE, 2019: 2297-2302.

[2] 胡涛，张超，程炳，等. 卷积神经网络在异常声音识别中的研究[J]. 信号处理，2018，34(3): 357-367.

HU Tao, ZHANG Chao, CHENG Bing, et al. Research on convolutional neural networks in abnormal sound recognition[J]. Signal Processing, 2018, 34 (3): 357-367.

[3] BISOT V, SERIZEL R, ESSID S, et al. Feature learning with matrix factorization applied to acoustic scene classification[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2017, 25(6): 1216-1229.

[4] LIU Z , WU Z , LI T , et al. GMM and CNN hybrid method for short utterance speaker recognition[J]. IEEE Transactions on Industrial Informatics, 2018, 14(7): 3244-3252.

[5] SAILOR H B, AGRAWAL D M. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification[C]//Proceedings of Interspeech. Stockholm, Sweden:[s.n.], 2017.

[6] PHAN H, MAAB M, MAZUR R, et al. Random regression forests for acoustic event detection and classification[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(1): 20-31.

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

[8] TOKOZUME Y, HARADA T. Learning environmental sounds with end-to-end convolutional neural network[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).[S.l.]: IEEE, 2017.

[9] ZHANG Z, XU S, QIAO T, et al. Attention based convolutional recurrent neural network for environmental sound classification[C]//Proceedings of Chinese Conference on Pattern Recognition and Computer Vision.[S.l.]: Springer, 2019.

[10] 刘亚荣,黄昕哲,谢晓兰,等. 美尔谱系数与卷积神经网络相组合的环境声音识别方法[J]. 信号处理, 2020, 36(6): 1020-1028.
LIU Yarong, HUANG Xinzhe, XIE Xiaolan, et al. Environmental sound recognition method combining meir spectral coefficients and convolutional neural network[J]. Journal of Signal Processing, 2020, 36(6): 1020-1028.

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

[12] LEE C Y, XIE S, GALLAGHER P, et al. Deeply-supervised nets[C]//Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics.[S.l.]: PMLR, 2015.

[13] HUANG G, LIU Z, MAATEN L V D, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).[S.l.]:IEEE, 2017.

[14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .[S.l.]:IEEE, 2015.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2016: 770-778.

[16] ADAVANNE S, VIRTANEN T. Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network[C]//Proceedings of Workshop on Detection & Classification of Acoustic Scenes & Events. Munich, Germany: [s.n.], 2017.

[17] 王诗佳.基于深度学习的声音事件识别研究[D].南京:东南大学, 2018.
WANG Shijia. Research on sound event recognition based on deep learning[D]. Nanjing: Southeast University, 2018.

[18] SALAMON J, BELLO J P. Feature learning with deep scattering for urban sound analysis[C]//Proceedings of 2015 23rd European Signal Processing Conference (EUSIPCO).[S.l.]: IEEE, 2015.

[19] PICZAK K J . ESC: Dataset for environmental sound classification[C]//Proceedings of ACM International Conference on Multimedia.[S.l.]: ACM, 2015.

[20] 曾剑飞.低信噪比条件下的语音端点检测算法研究[D].广州:华南理工大学,2019.
ZENG Jianfei. Research on speech endpoint detection algorithm under low SNR[D]. Guangzhou: South China University of Technology, 2019.

[21] 陈旺.语音端点检测的鲁棒性研究[D].广州:广州大学,2019.
CHEN Wang. Research on robustness of speech endpoint detection[D]. Guangzhou: Guangzhou University, 2019.

[22] TAK R N, AGRAWAL D M, PATIL H A. Novel phase encoded mel filterbank energies for environmental sound classification[C]//Proceedings of International Conference on Pattern Recognition and Machine Intelligence. [S.l.]: Springer, 2017.

作者简介:

曾金芳(1978-),**通信作者**,女,博士,讲师,研究方向:智能信号处理、声音信号处理,E-mail:zengjinfang@xtu.edu.cn。

李友明(1995-),男,硕士研究生,研究方向:深度学习、声音信号处理,E-mail:15074943298@163.com。

杨恢先(1963-),男,教授,研究方向:图形图像处理、嵌入式系统应用,E-mail:yanghx@xtu.edu.cn。

张钰(1996-),女,硕士研究生,研究方向:深度学习、声音信号处理,E-mail:156323134@qq.com。

胡雅欣(2000-),女,本科,研究方向:深度学习,E-mail:501572217@qq.com。

(编辑:张彤)