基于多通道特征和混合注意力的环境声音分类

周　帅[1],李　理[1,2],彭章君[1],黄鹏程[1]
(1. 西南科技大学 计算机科学与技术学院,四川 绵阳 621000;
2. 四川省自主可控人工智能工程技术中心,四川 绵阳 621000)

摘　要:环境声音分类(ESC)已成为非常重要的研究方向,但由于环境声音种类繁多,无法进行统一表征,加之易受噪声的干扰,使得 ESC 任务变得复杂。为了提高 ESC 任务的识别精度,提出了基于多通道特征和混合注意力模型的分类方法。首先,将 ESC 信号进行时频转换并使用多种滤波器提取频谱特征,将其重构为三通道特征图。多通道特征可以利用特征之间的互补性,弥补单一特征信息表征不足的缺点;其次,引入了一种由通道和时频注意力模块组成的混合分类模型,通道注意力模块计算特征图并对不同通道分配权重,含有更多有效信息且对该类声音分辨较好的通道特征则会被分配更多的权重,时频注意力模块会重点关注时域和频域中更有效的信息。该方法可较好地抑制背景噪声,消除冗余,提高收敛速度和分类精度。对比实验表明,在 ESC-10,ESC-50 数据集上的识别精度分别达到了 96.25% 和 89.56%,在 Urbansound8k 的数据集上达到 98.40%。

关键词:环境声音分类;多通道特征;通道注意力;时频注意力;混合注意力模型;深度模型
中图分类号:TP391.42　　**文献标识码**:A　　**文章编号**:1673-629X(2023)08-0043-08
doi:10.3969/j.issn.1673-629X.2023.08.007

Environmental Sound Classification Based on Multi-channel Features and Mixed Attention

ZHOU Shuai[1], LI Li[1,2], PENG Zhang-jun[1], HUANG Peng-cheng[1]
(1. School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621000, China;
2. Sichuan Autonomous Controllable Artificial Intelligence Engineering Technology Center, Mianyang 621000, China)

Abstract: Environmental sound classification (ESC) has become a very important research direction. However, the task of ESC becomes complicated due to the variety of environmental sounds, which cannot be characterized uniformly, and the susceptibility to noise. In order to improve the recognition accuracy of ESC task, a classification method based on multi-channel feature and mixed attention model is proposed. Firstly, the ESC signal is converted into time-frequency, and the spectral features are extracted by a variety of filters, which are reconstructed into a three-channel feature map. Multi-channel features can make use of the complementarity between features to make up for the lack of single feature information representation. Secondly, a hybrid classification model consisting of channels and time-frequency attention modules is introduced. The channel attention module calculates the feature map and assigns weights to different channels. The channel features with more valid information and better resolution for this type of sound will be assigned more weights. The time-frequency attention module will focus on more valid information in the time domain and frequency domain. The proposed method can suppress the background noise, eliminate the redundancy, and improve the convergence speed and classification accuracy. The comparison experiment shows that the recognition accuracy reaches 96.25% and 89.56% on ESC-10 and ESC-50 datasets respectively, and 98.40% on Urbansound8k datasets.

Key words: environmental sound classification; multi-channel feature; channel attention; time-frequency attention; mixed attention model; deep model

收稿日期:2022-09-02　　**修回日期**:2023-01-05
基金项目:国家自然科学基金(U21A20157);国家重点研发计划(2019YFB1310501)
作者简介:周　帅(1997-),男,硕士研究生,研究方向为语音识别、机器视觉;通讯作者:李　理(1981-),男,CCF 专业会员(16160M),副教授,博士,研究方向为语音识别、机器视觉、先进控制与建模。

0 引 言

音频分类主要集中在三大应用领域,首先是音乐信息检索(MIR),其次是自动语音识别(ASR),然后是环境声音分类(ESC)。环境声音分类作为音频识别的一个重要应用场景,被广泛应用于噪声检测、智能声控、场景分析、多媒体检索等分支领域。ESC 可以使得机器高效准确地解析背景环境,将算力集中到高价值的有效的信息。但是由于环境声音包含了室内、室外、人声、其他动物声、风声、雨声等多种声音,不能仅仅被描述为单一语义的语音或音乐,故将语音识别(Automatic Speech Recognition,ASR)和音乐信息检索(Music Information Retrieval,MIR)等其他模型直接移植到 ESC 分类中会导致分类精度差,效率低。因此,开发一种专用的高效 ESC 辨识与分类算法有着非常重要的意义。

ESC 环境声音分类算法主要包括环境声音特征提取和分类模型搭建。对于特征提取,往往是通过对连续语音先进行预加重、分帧、加窗等操作,然后从每一帧中提取特征作为模型的输入。对于传统的环境声音识别,人工提取特征向量,如梅尔频率倒谱系数(MFCC)、梅尔谱图特征、小波变换等,然后通过机器学习算法完成特征分类,例如,Pillos A 等人[1]使用基于 MFCC 和多层感知器进行实时环境识别,准确率为 74.5%。后面也出现了较多仅以 MFCC 作为单一特征进行分类的环境声音方法[2-5],但是由于环境声音大多是非平稳信号,没有有意义的模式或子结构,因此使用单一特征可能会导致无法捕获有关环境音频事件的重要信息。而聚合特征相比单一特征可以达到不同特征之间的优势互补效果,增强特征的表达能力,从而提供更好的性能。

传统的分类器主要被设计用于对缺乏时间和频率不变性的音频数据进行分类。近年来,深度学习(DL)模型在解决复杂数据的分类问题方面已被证明比传统分类器更准确和高效。卷积神经网络(CNN)是 DL 使用最广泛的模型,作为端到端分类器在图像分类[6-7]和声学场景分类[8-9]等应用中显示出巨大的优势。Abdoli S[10]提出了一种用于环境声音分类的端到端一维 CNN,其网络架构取决于音频信号的长度,模型由三到五个卷积层组成,平均准确率达到 89%。Zhu B 等人[11]以原始波形为输入,使用一组具有不同卷积滤波器大小和步长的分离并行神经网络,以学习具有多时相分辨率的特征表示。一维卷积神经网络模型不需要对音频做预处理,直接输入网络,由网络自行学习其特征,这使得 ESC 任务以端到端的方式运行成为可能。而端到端的模型降低了网络的可解释性,对于音频来说,频域特征和时域特征同样重要,所以目前大量的研究工作集中在二维卷积网络。Piczak K[12]首先提出通过计算声音的时频表示,将声音信号转为频谱图输入到二维卷积神经网络中;之后又出现了将多特征融合作为特 2D-CNN 的特征输入。Su Y 等人[13]评估了六种基本的频域声学特征和两种感知激励声学特征在 2D-CNN 模型上最佳的聚合特征组合,同时也提出了 TSCNN-DS[5]模型,一种四层卷积神经网络,分别由两个 LMC 和 MC 特征集训练的 LMCNet 和 MCNet 网络,最后通过 DS 证据理论融合两个网络的 Softmax 层的输出。

可以发现,当前 ESC 主流方法要么是采用单一特征作为模型的输入[14-16],然后对网络结构进行优化,要么是采用融合特征与 CNN 网络相结合进行分类[17],但是单一特征存在对信息表征不足从而会影响分类的准确性和泛化性的问题,而仅仅采用融合特征也会带来冗余信息的干扰,导致分类精度不高。

针对上述问题,该文提出将多通道特征作为模型的输入,与单一特征相比,多通道特征可以有效扩大可用数据规模,并利用特征之间的互补性进行特征融合描述与表征,可使其更具判别能力。引入了一种具有混合注意力的深度网络,但网络层数的加深需要更多的数据来训练,因此通过音高增加、时间拉伸、静音修剪、添加噪声等方式进行了有效数据扩充,以防止网络过拟合。由于使用了多通道特征与数据增强操作,会导致信息的冗余,从而降低分类精度,而引入的通道注意力和时频注意力组成的混合注意力机制则能较好地消除冗余信息和背景噪声的影响,提高模型整体的分类性能。

1 方 法

1.1 多特征数据提取

环境声音包括人声、动物声、机械声等各种各样的声音类别,且夹杂着大量的噪声,直接将声音输入网络或仅提取单一特征作为分类特征会导致分类精度不高,且模型的泛化性很差。为了弥补单一特征信息表征不足的缺点,该文提出利用多特征数据组成多通道作为分类特征。多特征数据能更全面且更有效地表征声音的特征。

(1)MFCC 特征提取。

在语音识别、说话人辨识等领域,很常用的一种音频特征叫做梅尔倒谱系数(Mel-scale Frequency Cepstral Coefficients,MFCC)。MFCC 的特点是使用一组用来创建梅尔倒谱的关键系数,使得它的倒频谱和人类非线性的听觉系统更为接近。MFCC 提取的过程如下:

(a)对声音进行预处理。

(b)通过傅里叶变换,将声音信号从时域转换到频域。

(c)构造一个梅尔滤波器组,并与能量谱进行点积运算。梅尔滤波器的作用是将能量谱转换为更接近人耳机理的梅尔频率。Mel 频率可以用公式表达如下:

$$m = 2\ 595 \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

$$f = 700(10^{\frac{m}{2\ 595}} - 1)$$

其中,f 表示频率。

(d)进行 DCT 变换。

$$\mathrm{mfcc}(i,n) = \sum_{m=1}^{M} \log[H(i,m)] \cdot \cos[\frac{\pi \cdot n \cdot (2m-1)}{2M}] \qquad (2)$$

其中,H 为二维矩阵能量谱和梅尔滤波器二维数组的乘积,M 代表梅尔滤波器个数,i 代表第几帧数据(取值为 1 ~ 301),n 代表第 i 帧的第 n 列。

(2)GFCC 特征提取。

Gammatone 滤波器是一种基于标准耳蜗结构的滤波器,用来模拟人耳听觉频率响应,Gammaton 滤波器可以模仿基底膜的分频特性,谱峰比三角滤波器平缓,能够改善三角滤波器能量不足的问题,并且它有完整的幅度和相位信息,可以弥补 MFCC 丢失的相位信息,并且其倒谱系数具有很好的抗噪性。其时域表达形式为:

$$g(f,t) = kt^{a-1}\mathrm{e}^{-2\pi bt}\cos(2\pi ft + \varphi)t \geqslant 0 \qquad (3)$$

其中,f 为中心频率;k 为滤波器增益;a 为滤波器阶数;φ 为相位;b 为衰减因子。

GFCC 和 MFCC 之间有两个主要区别。MFCC 基于 mel 比例,而 GFCC 基于 ERB 比例:

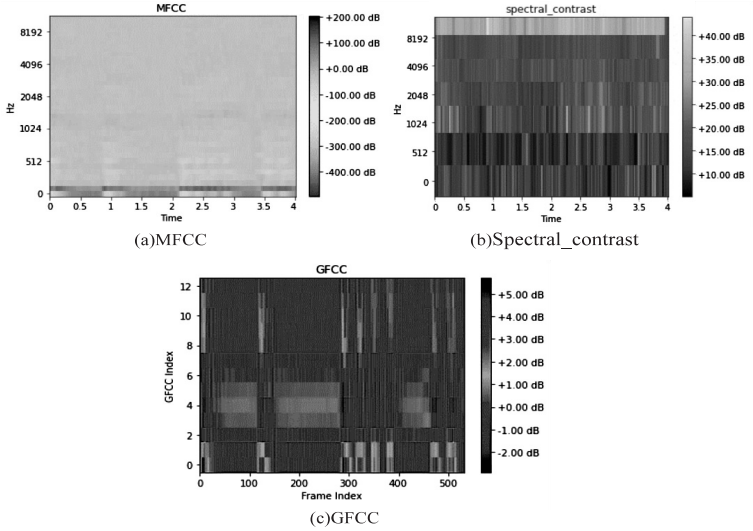
$$\mathrm{ERB}(f) = 24.7 \times (4.37\frac{f}{1\ 000} + 1) \qquad (4)$$

(3)Spectral_contrast 特征提取。

光谱对比度特征表示光谱峰/谷及其差异的强度。提取的方法是首先将声波分割为 200 ms 的帧,重叠 100 ms,执行 FFT 以获取频谱。然后,应用倍频程滤波器将频率划分为子带,估计谱峰、谷及其差的强度。最后,将估计结果转换到对数域后,使用 Karhunen-loeve 变换将原始光谱对比度特征映射到正交空间,并消除不同维度之间的相关性。Spectral _ contrast 相较 MFCC,保留了更多子带信息。

图 1 是提取的玻璃碎声的 MFCC、Spectral _ contrast、GFCC 特征图,其大小为 64×429,每一段声音重采样为 44 100 Hz,然后分为 200 ms,重叠 50%,每一帧的大小约为 100 ms。最终组成 3×64×429 的三通道特征图输入网络。

采用 MFCC、Spectral_contrast、GFCC 组成的三通道特征能很好地弥补单一特征信息表征不足的缺点。同时,又能利用不同特征之间的特点进行优势互补,例如,MFCC 作为说话人识别最常用的特征,但它却对声音高频部分不敏感,而环境声中却涵盖各种频率范围的声音。除此之外,MFCC 采用的是三角滤波器,三角滤波器的抗干扰能力较差,因此在具有大量噪声的环境中分类效果较差。该文同时采用 Spectral_contrast、GFCC 与 MFCC 组合的方式组成三通道特征,采用不同的滤波器组,提取不同的特征,不仅提高了模型的抗干扰能力,同时对在不同频段的声音都有较好的分类效果。对于 ESC 任务,能对不同的声音进行很好地表征,从而提高分类精度。



(a)MFCC　(b)Spectral_contrast

(c)GFCC

图 1　三种特征图

图 2 展示了环境声音处理的总体流程。首先对声音进行分帧、加窗等预处理,提取三种不同的声学特征 Spectral_contrast、GFCC 与 MFCC,然后,将提取的特征输入分类模型,最后做出预测。

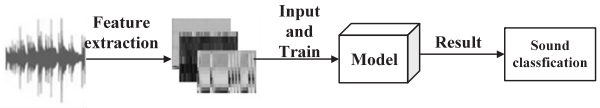


图 2 整体流程

1.2 建立分类模型

提出的 ESC 分类模型是基于残差网络(Resnet)的,Resnet 的主要优点是容易优化,并且通过增加相当的深度来提高准确度。由于其内部的残差块使用了跳跃链接,缓解了深度神经网络中增加深度带来的梯度消失的问题。ESC 分类模型如图 3 所示,模型主要由卷积层、批量归一化层、残差层、混合注意力块以及全连接层组成。首先,对音频提取多特征数据组成三通道特征图输入分类模型,然后经过一个卷积层,卷积核大小为 7×7,卷积核个数为 64,步长为 2,padding 为 3,再通过一个 BN 层,接下来是进入四个残差层,残差层的结构如虚线框里面所示,每一个残差层由三个瓶颈层组成,每一个瓶颈层又由卷积层和批归一化层组成。每一个残差层并行接入一个混合注意力模块,注意力模块的结构如图 4 所示,最后经过全局平均池化后进入全连接层。

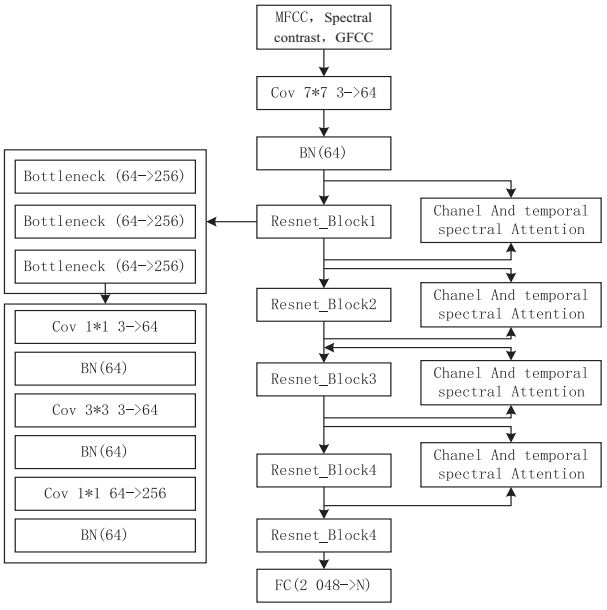


图 3 基于 Resnet 的 ESC 分类模型

1.3 混合注意力模块(MIA_block)设置

混合注意力模块由通道注意力模块和时频注意力模块组成(见图 4),在 ESC 任务中,音频信号首先被转换成了频谱图,音频的两个维度特征分别对应时间和频率,不同时间维度里面频段的重要程度不同,因此,时频注意力模块可以让模型重点关注时域和频域中重要的信息,同时提出的是基于多通道的特征,添加的通道注意力模块能分配给通道不同权重,包含关键信息的通道就能得到更多的关注。混合注意力采用的是一个串行的方式,特征图首先通过通道注意力模块再进入时频注意力模块。使得模型首先进行通道的选择,其次关注通道上时域和频域上的重要信息。

混合注意力的公式如下所示:

$$M_c(\boldsymbol{F}) = \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\max}))) \tag{5}$$

$$M_s(\boldsymbol{F}) = \sigma(f^{3\times3}([\text{AvgPool}(\boldsymbol{F}); \text{maxPool}(\boldsymbol{F})])) = \sigma(f^{3\times3}(F^s_{\text{avg}}; F^s_{\max})) \tag{6}$$

$$F'' = M_s(\boldsymbol{F}) \otimes M_c(\boldsymbol{F}) \otimes \boldsymbol{F} \tag{7}$$

其中,式(5)是通道注意力计算公式,$\boldsymbol{F}$ 表示输入矩阵,大小为(W,H,C),对应由 MFCC、GFCC、Spectral_contrast 组成的多通道输入。首先对输入 $\boldsymbol{F}$ 分别进行全局平均池化和全局最大池化,输出矩阵大小为 $(1\times1\times C)$,其中 $W_0 \in R^{\frac{C}{r}\times C}$,$W_1 \in R^{C\times\frac{C}{r}}$ 表示在上一步池化操作后接入两个全连接层,第一个 FC 层起到降维的作用,参超数 r 设置为 8,第二个 FC 层恢复到原始的大小即 $(1\times1\times C)$,然后将得到的两个向量进行相加,再通过 sigmoid 函数激活,最后得到通道域注意力。

式(6)表示时频注意力公式,和通道注意力一样,首先对输入 $\boldsymbol{F}$ 分别做全局平均池化和全局最大池化,然后将得到的向量($W\times H\times1$)进行拼接,此时得到的矩阵大小为($W\times H\times2$),然后再做一个卷积操作,最后通过 sigmoid 激活。式(7)表示混合注意力的计算公式,$M_c(\boldsymbol{F}) \otimes \boldsymbol{F}$ 表示得到的通道注意力权重图,与时频注意力相乘即得到混合注意力权重图。

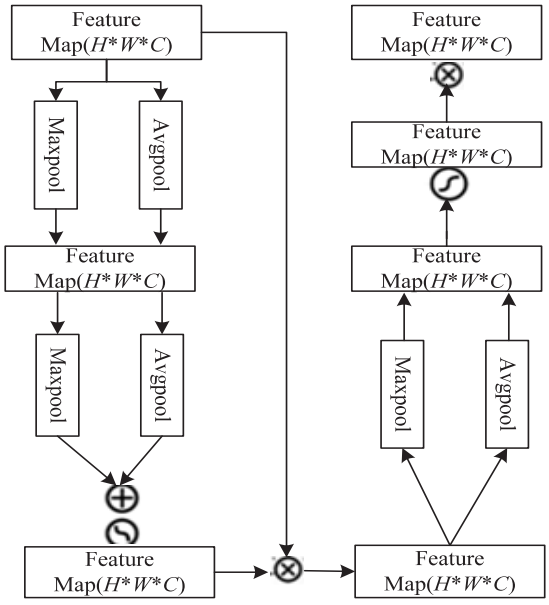


图 4 混合注意力模块

2 实验与结果分析

2.1 数据集选择

实验所用的数据集为环境声音分类任务中最为广泛使用的三个公开数据集,包括 ESC-50、ESC-10、Ur-

bansound8k。这些数据集涵盖了室内室外多种场景的多种环境声。例如,ESC-10 包括 400 个录音,其中共计 10 个类别。ESC-50 由五个大类的声音组成,每个大类又包含 10 个小类,总计 50 个环境声音类别。Urbansound8k 由 8 732 个音频文件组成,共 10 个类别,包括冷气机声、汽车喇叭声、儿童玩耍声、狗吠声、钻孔声、发动机空转声、枪射击声、手持式凿岩机声、警笛声、街头音乐声。数据集的详细信息如表 1 所示。以上三个数据集均采用五折交叉验证训练和测试模型的性能。

表 1　数据集详细信息

数据集	类别	总样本数量	总时长/min
ESC-10	10	400	33
ESC-50	50	2 000	168
Urbansound8k	10	8 732	582

2.2　数据增强

ESC-50 总共涵盖 50 类不同的声音,样本数量为 2 000,较少的数据量容易造成网络过拟合,而 Urbansound8k 虽然样本数相比 ESC-50 更多,但此数据集中包含多个时长仅为 1 s 的数据,并且音频中静音片段也较多,导致这些数据没有包含足够的数据量,也会导致模型训练时学习率不足以及出现过拟合的现象。因此,该文采用数据增强的方法,如下所述:

(a)音高增加:数据集中每个音频信号的音调增加 2。

(b)负音高偏移:在提高可用声音片段的音调后,在这种方法中,变形数据集的音调减 2。

(c)静音修剪:这是一种独特的数据增强技术,其中音频剪辑的静音部分被修剪,并且仅保留包含声音的部分。

(d)快速时间拉伸:将数据集的每个声音片段的时间瞬间拉伸 1.20 倍。

(e)慢速时间拉伸:时间拉伸系数是将音频录制速度降低 0.70 倍。

(f)添加白噪声:对于这种增强方法,声音片段受到值为 0.05 的白噪声的影响。混合显示为 $y + 0.05 * Wn$。

2.3　实验环境

软件方面,所有的实验都在 Python3.8.13 的环境下进行,采用 Pytorch 框架实现分类模型的搭建,音频的预处理和特征提取使用 Librosa 和 Spafe 库,操作系统为 Ubantu20.04,硬件条件:CPU 采用 Intel Xeon Gold 6130,显卡为丽台 P5000,16 GB 显存,运行内存为 128 GB。

2.4　评价指标

采用的评价指标都为准确率(Accuracy),即分类正确的样本占总样本个数的比例,公式如下:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{PN}} \tag{8}$$

其中,TP 为被模型预测为正的正样本的数量,FP 为被模型预测为正的负样本的数量,FN 表示被模型预测为负的正样本的数量,TN 表示被模型预测为负的负样本的数量。

2.5　实验结果与分析

2.5.1　不同特征对分类精度的影响

环境声音种类繁多且很难找到统一特征来区分所有声音,为了证明多特征组合是否优于单一特征,以及最优的特征组合,通过实验比较了不同手动特征对于模型分类精度的影响。表 2 列出了不同特征组合得出的结果。

表 2　采用不同特征的实验结果　　%

特征	ESC-10	ESC-50	Urbansound8k
MFCC	94.25	83.72	91.78
Spectral_contrast	80.82	68.56	78.86
GFCC	83.65	72.76	82.00
MFCC+Spectral_contrast	95.60	85.26	92.48
GFCC+ Spectral_contrast	80.98	70.24	80.26
MFCC+GFCC	95.46	83.42	92.20
MFCC+Spectral_contrast+GFCC	96.25	89.56	98.40

由表 2 可知,当把 MFCC、Spectral_contrast、GFCC 用作组合特征作为模型的输入时,在 3 个数据集上的分类精度均达到了最佳。当仅使用单一特征时,MFCC 的表现明显优于其他两类特征,Spectral_contrast 作为在音乐分类任务中使用最为广泛的特征,在环境声音中却表现最差,这可能与音乐的音阶有关。得益于 MFCC 的分类优势,在两通道组合特征时 MFCC+Spectral_contrast、MFCC+GFCC 的表现优于 Spectral_contrast+GFCC。实验证明,仅仅使用一种特征不足以有效地支持环境声音分类,当把三个单一特征组合成三通道特征作为模型输入时,在每一个数据集上的分类效果都比其它方式的特征输入的效果要好。

2.5.2　注意力模块的使用

注意力机制的作用就是找到真正感兴趣的区域,加以处理,使其更好地完成任务。注意力机制近几年在图像、自然语言处理(Natural Language Processing,NLP)和计算机视觉(Computer Vision,CV)等领域中都取得了重要的突破,被证明有益于提高模型的性能。该文通过提出改进的时频注意力和通道注意力构成混合注意力模块,通道注意力能关注对于特定的某种声

音,可判断哪种通道特征相比起来更有区分度,比如人声,MFCC 通道的特征可能被分配更大的权重。时频注意力则能让 CNN 重点关注时域和频域上更重要的信息,对于噪声以及静音则能有更好的抑制作用。注意力的使用如表 3 所示。

表 3 不同注意力的使用对结果的影响 %

Attention	ESC-10	ESC-50	Urbansound8k
无注意力	94.15	85.69	94.80
通道注意力	94.15	86.23	94.86
时频注意力	96.13	87.52	95.16
混合注意力	96.25	89.56	98.40

该文比较了不使用注意力机制,使用单一通道或者时频注意力,以及使用混合注意力对模型分类精度的影响,如表 3 所示,实验证明使用注意力模块会提高模型的分类精度。当仅使用通道注意力时,提升微乎其微,仅使用时频注意力模块时,有小部分的提升,但当使用混合注意力时效果达到最佳,相比不使用注意力机制,在三个数据集上分别提升 2.1 百分点,3.87 百分点,3.6 百分点。

表 4 混合注意力位置对结果的影响 %

插入位置	ESC-10	ESC-50	Urbansound8k
模型输入时	95.25	86.13	90.21
Resnet_Blocks	96.25	89.56	98.40
bottleneck	96.85	89.20	93.00
模型输出时	94.15	85.70	90.25

表 4 展示了混合注意力插入不同位置的实验结果。由表 4 可以看出,当 Resnet_Block 时效果最为显著,虽然插入瓶颈块也会有较好的效果,但是这会带来较大的参数量提升,相比提升的精度,这是不值得的,插入模型输入时效果提升不大,而插入模型输出端时,几乎没有效果提升。

2.5.3 消融实验

为了证明提出的多通道特征和混合注意力对于模型的分类精度有所提高,在 ESC-50 数据集上进行了消融实验,结果如表 5 所示。

表 5 消融实验

基本模型	多通道特征	混合注意力	精度/%
√			82.50
√	√		85.69
√	√	√	89.56

由表 5 可知,当采用多通道特征时,分类精度提升 3.19 百分点,这是因为采用 MFCC、Spectral_contrast、GFCC 组成的三通道特征能很好地弥补单一特征信息表征不足的缺点。MFCC 提取特征时采用的是三角滤波器,对噪声的抗干扰性较差,而 GFCC 特征采用的是高通滤波器提取特征,能够改善三角滤波器能量不足的问题,同时 Spectral_contrast 特征相较 MFCC,保留了更多子带信息,因此融合的三通道特征不同特征之间的特点进行优势互补,能够提高网络的识别能力。此外,加入了混合注意力模块后,精度提高了 3.87 百分点,通道注意力模块计算特征图并对不同通道分配权重,含有更多有效信息且对该类声音分辨较好的通道特征则会被分配更多的权重,时频注意力模块会重点关注时域和频域中更有效的信息。同时对两个维度进行注意力分配增强了注意力机制对模型性能的提升效果,且对于噪声以及静音能有更好的抑制作用。通过表 5 可知,提出的多通道特征和混合注意力都对模型的分类性能有所提升。

2.5.4 文中方法总体精度展示

采用五折交叉验证的方法在 ESC-10、ESC-50、Urbansound8k 三个数据集上进行了性能评估,图 5 ~ 图 7 分别展示了文中方法在 ESC-10、ESC-50、Urbansound8k 三个数据集上得出的混淆矩阵,图 5 中,仅有三个样本被错误分类。

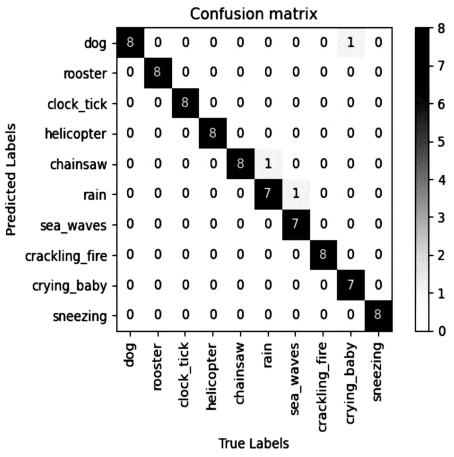


图 5 ESC-10 数据集混淆矩阵

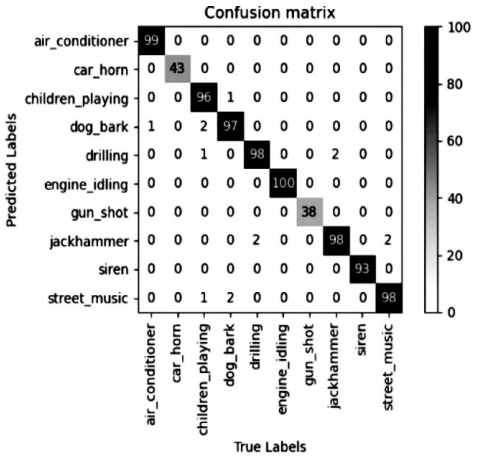


图 6 Urbansound8k 数据集混淆矩阵

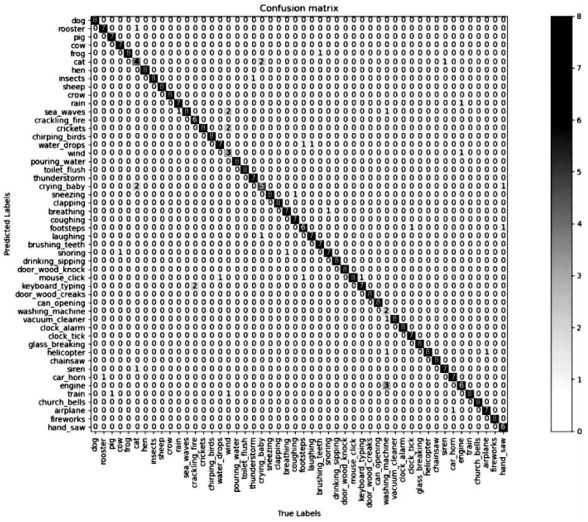


图 7　ESC-50 数据集混淆矩阵

图 6 中,可发现标签为手提钻声有两次被误判成为了钻孔声,同样钻孔声也有两次被误判成为了手提钻声,究其原因,可能是现实生活中两种声音相似度高,导致它们特征共同点多。对于 ESC-50 数据集,在其 50 个类别中,只有洗衣机声、风声、猫叫声的分类精度相对偏低,其中洗衣机声为 25%(2/8)、风声为 37.5%(3/8)、猫叫声为 50%(4/8),其中洗衣机声容易被误判为引擎声,风声容易被误判为海浪声和蟋蟀声,猫叫声容易被误判为小孩哭声,除此之外,其余的大多的分类精度基本都在 80% 以上。

2.5.5　文中方法与其他方法对比

表 6 展示了提出的环境声音分类方法与目前先进的方法的比较。由表 6 可知,在 ESC-10、ESC-50、Urbansound8k 三个数据集上,目前已有方法能达到的较好的精度分别为 95.80%、87.70% 和 96.83%,而文中方法在三个数据集上的分类精度分别能达到 96.25%、89.56% 和 98.40%,相比目前较优的结果分别提高了 0.45 百分点、1.86 百分点和 1.57 百分点。

表 6　文中方法与现有方法精度比较　　%

年份	模型	ESC-10	ESC-50	Urbansound8k
2015	PiczakCNN(Baseline model)[13]	80.50	64.90	73.70
2017	EnvNet v2[18]	91.30	84.70	78.30
2018	CNN+Audio Augmentation+Mix-up[19]	91.70	83.90	83.70
2019	DCNN+Temporal structure+Augmentation[20]	94.20	86.50	–
2020	ESResNet-Attention[21]	94.25	83.15	96.83
2021	MCTA-CNN[22]	95.80	87.70	–
2022	Ours	96.25	89.56	98.40

3　结束语

提出了一种新的环境声音分类算法,该算法由多通道特征输入与基于混合注意力机制的深度卷积神经网络组成。为了解决单一特征信息对于复杂环境声音表征不足的问题,添加了 MFCC、Spectral_contrast 和 GFCC 组成三通道特征,较好解决了音频数据中因噪声干扰导致识别率低和高频区分辨率低的问题。设计的通道注意力加时频注意力的混合注意力模型,兼顾了对于特定声音更好的识别效果对时域和频域有效信息的更好关注,并通过混合注意力机制有效提升了特征表示能力,较好地抑制了环境中的背景噪声,消除了冗余信息的干扰。为了得到更佳的识别效果,采用了多种数据增强技术防止模型过拟合。实验表明,提出的 ESC 分类模型在 ESC-10、ESC-50、Urbansound8k 三个数据集上的分类精度分别达到了 96.25%、89.56%、98.40%,有效提高了环境声音识别的准确性。

参考文献:

[1] PILLOS A, ALGHAMIDI K, ALZAMEL N, et al. A real-time environmental sound recognition system for the android OS[C]//2022 challenge on detection and classification of acoustic scenes and events (DCASE). Budapest: Tampere University of Technology, 2016: 75-79.

[2] MASSOUDI M, VERMA S, JAIN R. Urban sound classification using CNN[C]//2021 6th international conference on inventive computation technologies (ICICT). Wuhan: IEEE, 2021: 583-589.

[3] LU J, MA R, LIU G, et al. Deep convolutional neural network with transfer learning for environmental sound classification[C]//2021 international conference on computer, control and robotics (ICCCR). Shanghai: IEEE, 2021: 242-245.

[4] PENG N, CHEN A, ZHOU G, et al. Environment sound classification based on visual multi-feature fusion and GRU-AWS[J]. IEEE Access, 2020, 8: 191100-191114.

[5] SU Y, ZHANG K, WANG J, et al. Environment sound classification using a two-stream CNN based on decision-level fusion[J]. Sensors, 2019, 19(7): 1733.

[6] 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. Las Vegas:

IEEE,2016:770-778.
[7] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii:IEEE,2017:4700-4708.
[8] DORFER M,LEHNER B,EGHBAL-ZADEH H,et al. Acoustic scene classification with fully convolutional neural networks and I-vectors[J]. Proceedings of the Detection and Classification of Acoustic Scenes and Events,2018,80(6):90.
[9] EGHBAL-ZADEH H,KOUTINI K,WIDMER G. Acoustic scene classification and audio tagging with receptive-field-regularized CNNs[J]. Tech. Rep,DCASE 2019 Challenge,2019,38(7):50.
[10] ABDOLI S,CARDINAL P,KOERICH A L. End-to-end environmental sound classification using a 1D convolutional neural network[J]. Expert Systems with Applications,2019,136:252-263.
[11] ZHU B,XU K,WANG D,et al. Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features[C]// Pacific rim conference on multimedia. Hefei:Springer,2018:528-537.
[12] SU Y,ZHANG K,WANG J,et al. Performance analysis of multiple aggregated acoustic features for environment sound classification[J]. Applied Acoustics,2020,158:107050.
[13] PICZAK K J. Environmental sound classification with convolutional neural networks[C]//2015 IEEE 25th international workshop on machine learning for signal processing (MLSP). Boston:IEEE,2015:1-6.
[14] 蒋翠清,邵宏波. 基于 MFCC 与改进 ACF 的汽车声音识别算法研究[J]. 计算机技术与发展,2015,25(2):140-143.
[15] 史秋莹,郑铁然. 基于深度学习的环境声音识别[J]. 智能计算机与应用,2018,8(5):34-37.
[16] 刘 慧,李小霞,何宏森. 基于多分辨率特征和时频注意力的环境声音分类[J]. 计算机应用研究,2021,38(12):3569-3573.
[17] 吴佳赛,高振斌. 基于双二流卷积和多特征融合的 D-S 声音分类[J]. 计算机应用研究,2022,39(3):693-698.
[18] TOKOZUME Y,HARADA T. Learning environmental sounds with end-to-end convolutional neural network[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans:IEEE,2017:2721-2725.
[19] ZHANG Z,XU S,CAO S,et al. Deep convolutional neural network with mixup for environmental sound classification[C]//Chinese conference on pattern recognition and computer vision (PRCV). Guangzhou:Springer,2018:356-367.
[20] ZHANG Z,XU S,ZHANG S,et al. Learning attentive representations for environmental sound classification[J]. IEEE Access,2019,7:130327-130339.
[21] GUZHOV A,RAUE F,HEES J,et al. Esresnet:environmental sound classification based on visual domain models[C]//2020 25th international conference on pattern recognition (ICPR). Milan:IEEE,2021:4933-4940.
[22] WANG Y,FENG C,ANDERSON D V. A multi-channel temporal attention convolutional neural network model for environmental sound classification[C]//ICASSP 2021-2021 IEEE international conference on acoustics,speech and signal processing (ICASSP). Toronto:IEEE,2021:930-934.