ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 730－740
DOI:10. 16337/j. 1004-9037. 2025. 03. 013

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

基于迁移学习卷积记忆网络的多声音事件检测

陈鹏飞，夏秀渝

(四川大学电子信息学院，成都 610065)

摘　要：针对多声音事件检测任务中强标注数据集有限、真实场景下检测性能急剧恶化的问题，提出了基于迁移学习卷积记忆网络的多声音事件检测方法。首先，该方法使用带有预训练权重的卷积块提取音频数据的局部特征，再将局部特征和方位特征一并送入残差特征增强模块进行特征融合和通道降维处理。接着将提取到的融合特征送入采用正则化方法的记忆网络，以进一步学习音频数据中的时序信息。实验结果显示，与DCASE挑战赛冠军系统模型相比，该方法在DCASE 2016 Task3数据集的开发集和评估集上，错误率分别降低了0.277和0.106，F_1分数分别提高了22.6%和6.6%；在DCASE 2017 Task3数据集的开发集和评估集上，错误率分别降低了0.22和0.123，F_1分数分别提高了17.2%和14.4%。

关键词：多声音事件检测；迁移学习；特征增强；记忆网络；正则化

中图分类号：TP391.4　　**文献标志码：**A

Polyphonic Sound Event Detection Based on Transfer Learning Convolutional Retentive Network

CHEN Pengfei, XIA Xiuyu

(School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Aiming at the problems of limited strong annotation datasets and the sharp degradation of detection performance in real-world scenarios for polyphonic sound event detection tasks, a method for polyphonic sound event detection based on Transfer learning convolutional retentive network is proposed. Firstly, the method utilizes convolutional blocks with pre-trained weights to extract local features of audio data. Subsequently, the local features, along with orientation features, are input into the residual feature enhancement module for feature fusion and channel dimension reduction. The fused features are then fed into the retentive network with regularization methods to further learn the temporal information in the audio data. Experimental results demonstrate that, compared to the champion system model of the DCASE challenge, the method achieves a reduction in error rates by 0.277 and 0.106, and an increase in F_1 scores by 22.6% and 6.6% on the development and evaluation sets of the DCASE 2016 Task3 dataset, respectively. On the development and evaluation sets of the DCASE 2017 Task3 dataset, the error rates are reduced by 0.22 and 0.123, and the F_1 scores increase by 17.2% and 14.4%, respectively.

Key words: polyphonic sound event detection; Transfer learning; feature enhancement; retentive network; regularization

收稿日期：2024-01-30；**修订日期：**2024-03-19

引 言

声音事件检测(Sound event detection,SED)指对现实环境中的声音事件进行种类判断并标注出起始时间和终止时间。相比于图像或视频,利用声音进行事件检测具有许多优势,比如不受光线和遮挡的影响、需要的计算资源更少等。声音事件检测也是模拟人类听觉感知的重要研究课题。根据同一时刻发生事件的种类数量,声音事件检测被分为单声音事件检测和多声音事件检测;按衡量单位的不同,可以划分为基于事件和基于片段两种方式。本文研究的是基于片段的多声音事件检测。

深度学习的发展使得基于神经网络的声音检测模型逐渐超越传统分类器,包括深度神经网络(Deep neural network, DNN)[1]、卷积神经网络(Convolutional neural network, CNN)[2]和循环神经网络(Recurrent neural network, RNN)[3]。Cakir等[4]提出的卷积循环神经网络(Convolutional RNN, CRNN)不仅使用CNN用于捕捉时频局部特征,还结合RNN进行序列识别,在多音事件检测任务中表现出色。为了在真实音频场景下取得更好的声音事件检测性能,一些新的模型方法被提出,包括空洞卷积[5]、Transformer[6]以及胶囊网络[7]。文献[8]中提出了一种简单无参数网络模型,通过引入简单无参模块和注意力机制,帮助模型聚焦深层特征的能力,从而增强了网络对不同声音特征的辨别能力;Wang等[9]利用扩张卷积来捕获长期依赖关系,并将标准卷积得到的细粒度特征与扩张卷积得到的时序特征进行融合,在充分利用相邻信息的同时增加了感受野的大小。由于声音事件蕴含着空间信息,所以引入与空间信息相关的声源方位特征有助于提高音频事件检测的性能。Koyama等[10]将提取出的音频源信号与模拟的空间脉冲响应进行卷积,获得增强的方位特征信息用于训练,提高了音频检测与定位的整体性能。为了解决SED任务训练数据不足的问题,杨利平等[11]提出了一种利用弱标签数据的空间-通道特征表征与自注意池化声音事件检测方法;刘臣等[12]提出了一种复合数据扩增技术使模型获得了更好的泛化能力。研究者们还采用半监督学习的方法来同时利用弱标注和无标注的数据,Zheng等[13]提出使用协同训练的方法来解决SED训练数据不足问题。这类半监督学习方法的网络往往比较复杂和庞大,需要更多的计算资源。解决带标签数据量不足的另一个热门方法便是迁移学习。相关研究表明,迁移学习方法不仅适用于图像处理相关的任务[14],在音频识别领域[15]也取得了良好的效果。

目前多声音事件检测任务仍存在以下两个难点:(1)由于人工注释的强标注数据制作成本高昂,因此针对这类任务的强标注数据集非常有限;(2)由于真实环境的复杂性,声音事件检测系统在真实场景下的性能会急剧下降。尽管更复杂的系统模型可以用来提高性能,但这往往需要较大的参数量和较高的计算复杂度作为代价。针对上述问题,提出了一种基于迁移学习卷积记忆网络(Transfer learning convolutional retentive network,TCRETNET)的多声音事件检测方法。实验表明,该方法在小型公开数据集和真实音频场景的限制条件下表现出有竞争力的检测性能。

1 TCRETNET网络模型

1.1 模型概述

图1展示了TCRETNET方法的总体框架。它主要由预训练卷积块(ConvBlocks)、残差特征增强模块(Residual feature enhancement

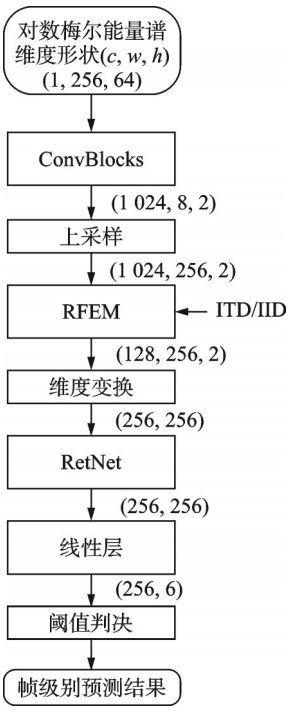

图1 TCRETNET结构框图

Fig.1 Structure diagram of TCRETNET

module,RFEM)和采用注意力正则化策略(DropKey)改进的记忆网络(Retentive network,RetNet)3部分组成。

图1中输入的对数梅尔能量谱的维度形状为(c,w,h),其中c为通道数,w为帧数,h为特征维数。首先提取样本音频双声道对数梅尔能量谱,将其平均作为网络的输入特征,采用带有预训练初始化权重的ConvBlocks进行微调训练提取局部特征。随后,将ConvBlocks输出的特征送入RFEM模块中与方位特征(双耳时间差(Interaural time difference,ITD)和耳间强度差(Interaural intensity difference,IID))进行融合训练,得到更加适用于下游任务(音频事件检测)的高维特征。之后,将该高维特征送入能对长期依赖关系进行建模的记忆网络,进一步捕获音频的时序信息。最后,通过线性激活层输出原始的类别概率分布,使用阈值判决得到最终的帧级别预测结果。

1.2 PANNs和预训练卷积块

在音频模式识别领域,通过在大规模数据集上进行预训练的系统已经在一些任务上取得了良好的泛化效果。Kong等[16]提出了一种在大规模音频数据集AudioSet[17]上进行训练的预训练音频神经网络(Pretrained audio neural networks,PANNs),这些PANNs在被迁移至多个不同音频模式识别任务时表现出了优异的性能。TCRETNET迁移了PANNs用于局部特征的提取,使用的PANN为CNN14。构建预训练卷积块ConvBlocks的方法和迁移学习策略为:取原始预训练模型CNN14的前10个卷积层作为ConvBlocks,训练时对所有卷积层都进行预训练权重初始化,通过学习新的目标数据集来微调更新卷积层的参数。图2展示了原始CNN14和预训练卷积块ConvBlocks的模型框图。

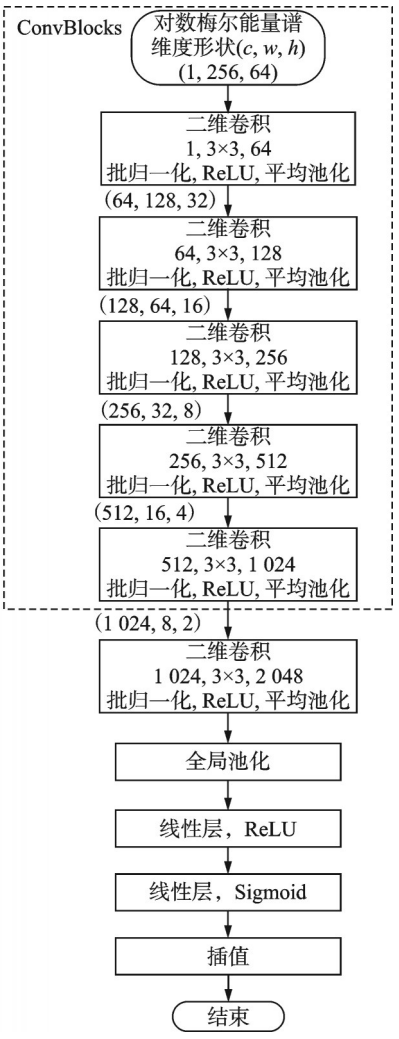

图2 CNN14和ConvBlocks结构框图

Fig.2 Structure diagrams of CNN14 and ConvBlocks

图2中的1、3×3和64表示二维卷积层的参数,分别为:输入通道数、卷积核大小和输出通道数。虚线框部分为预训练卷积块ConvBlocks。该卷积块能够保留大部分底层卷积层的通用特征,并通过微调策略来学习适应特定任务的特征。

1.3 残差特征增强模块

为了使网络学习到更适合下游任务的特征,设计了一种残差特征增强模块RFEM。图3展示了RFEM的结构框图。不同声音事件的声源方位特点具有显著差异。利用声源方位特征也有助于对不同的声音事件进行检测和分类。2017年,文献[18]提出利用声源方位特征双耳时间差ITD和耳间强度差IID来实现对复杂声音环境中主导音的提取。利用数据集音频具有双声道的特点,TCRETNET从样本中提取出ITD和IID作为网络的辅助输入特征,特征具体计算方式为:设左、右声道的信号分别为$x_{\mathrm{L}}(t)$,$x_{\mathrm{R}}(t)$,通过短时傅里叶变换得到短时谱$X_{\mathrm{L}}(i,k)$,$X_{\mathrm{R}}(i,k)$,分别表示信号第i帧,第k个频点的左右声道信号频谱(信号采样率为44 100 kHz,窗口长度取2 048个样点,频点总数为1 025)。按照式(1)和(2)分别提取每个时频单元的方位特征IID和ITD,有

$$\mathrm{IID}(i,k)=20\times\log_{10}\left\|\frac{X_{\mathrm{L}}(i,k)}{X_{\mathrm{R}}(i,k)}\right\| \tag{1}$$

式中:IID(i,k)表示第i个时间帧、第k个频点的耳间强度差;$\|\cdot\|$表示对复数比值$\frac{X_{\mathrm{L}}(i,k)}{X_{\mathrm{R}}(i,k)}$取模(即复数的绝对值)。

$$\mathrm{ITD}(i,k)=-\frac{\Delta\psi}{2k\pi}=-\frac{\psi_{\mathrm{L}}-\psi_{\mathrm{R}}}{2k\pi} \tag{2}$$

式中:ITD(i,k)表示第i个时间帧、第k个频点的双耳时间差;$\Delta\psi$表示耳间相位差;ψ_{L}、ψ_{R}分别表示左、右声道信号在第i个时间帧、第k个频率点的相位;$\psi_{\mathrm{L}}=\arctan\left\{\frac{\mathrm{Im}[X_{\mathrm{L}}(i,k)]}{\mathrm{Re}[X_{\mathrm{L}}(i,k)]}\right\}$,$\psi_{\mathrm{R}}=\arctan\left\{\frac{\mathrm{Im}[X_{\mathrm{R}}(i,k)]}{\mathrm{Re}[X_{\mathrm{R}}(i,k)]}\right\}$。

在图3中,首先对ITD和IID在特征维度上进行平均池化和卷积池化,然后将新的方位特征与预训练卷积块输出的特征在通道维度上进行拼接,最后将其送入带有残差结构的卷积层进行通道降维和进一步的特征提取操作。该模块有如下优点:(1)通道降维。通过减少输出特征通道数,减少了模型的参数量和计算复杂度。这种降维不仅能加快训练,还能丢弃来自迁移学习源域的冗余信息,有助于模型更好地泛化到新的数据集。(2)特征融合。将双耳时间差ITD和耳间强度差IID特征融合到模块中,使网络可以捕捉到信号中来自左右声道的方位特征信息,提高模型对不同音频事件的区分能力。(3)残差结构。残差结构的使用有助于减轻更深的网络中梯度消失的问题,从而提高模型的性能。

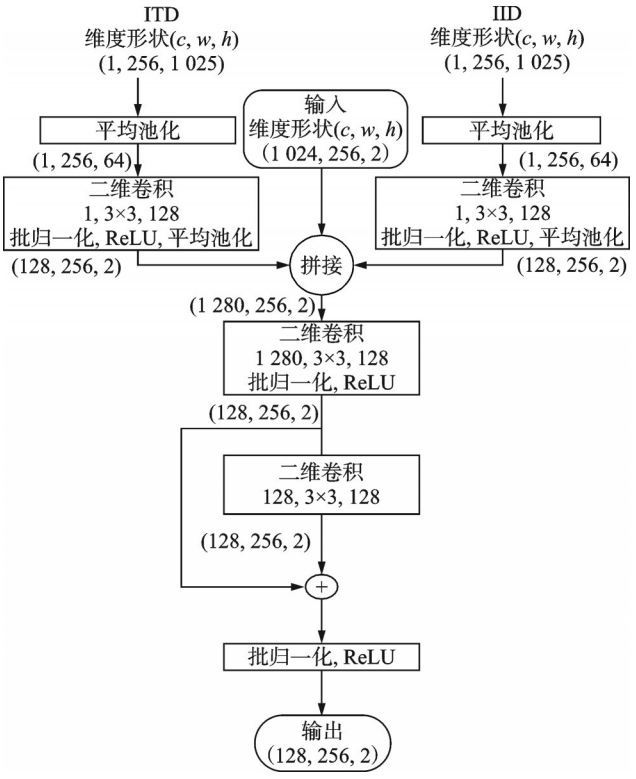

图3 RFEM结构框图

Fig.3 Structure diagram of RFEM

1.4 使用正则化策略的记忆网络

声音是随时间变化的信号。为了更好地捕捉和理解时间相关的信息,使用时序网络来帮助网络捕获和学习长期依赖关系十分有必要。Sun等[19]提出的RetNet,其同时具备训练可并行、推理成本低和良好性能的优点,在处理序列任务被认为具有超越Transformer的潜力。DropKey是由Li等[20]在2022年提出的一种自注意力正则化技术,有助于缓解小训练样本导致的过拟合问题。TCRETNET将RetNet网络应用于声音事件检测任务,并引入DropKey技术进行改进。图4展示了改进的RetNet,其由门限多尺度记忆层(Multi-scale retention,MSR)和使用了高斯误差线性单元(Gaussian error linear unit,GELU)激活函数的前馈神经网络(Feedforward neural network,FFN)组成。原始记忆模块在单个时间步n的输出为

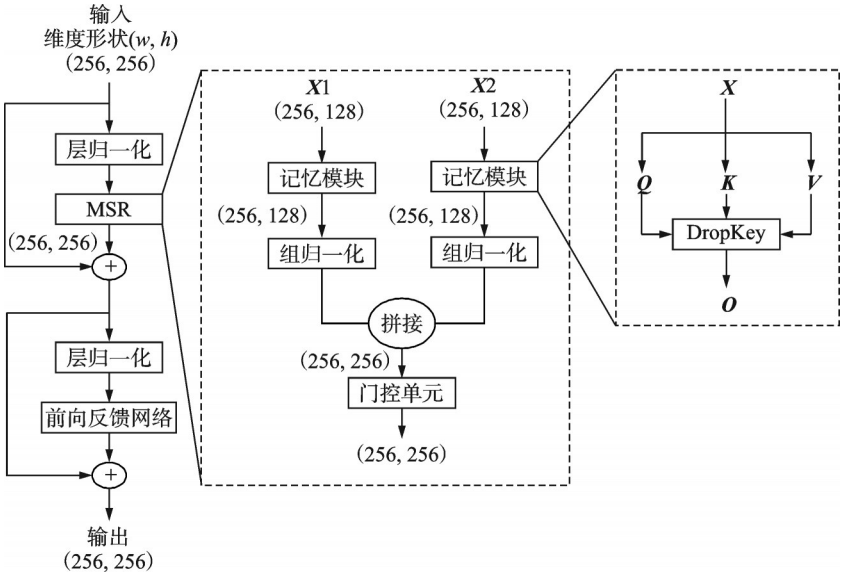

图4 引入DropKey的RetNet结构框图

Fig.4 Structure diagram of RetNet structure with DropKey

$$\boldsymbol{o}_n=\sum_{m=1}^{n}\gamma^{n-m}(\boldsymbol{Q}_n\mathrm{e}^{\mathrm{i}n\theta})(\boldsymbol{K}_m\mathrm{e}^{\mathrm{i}m\theta})^{\dagger}\boldsymbol{v}_m \tag{3}$$

式中：$\boldsymbol{o}_n$表示第n个步数时的输出向量；m表示从第一个位置到当前第n个位置中间任一步数；γ为一个衰减因子，是介于0和1之间的标量；$\boldsymbol{Q}_n$为第n个位置的查询向量；$\boldsymbol{K}_m$为第m个位置的键向量；$\boldsymbol{v}_m$为第m个位置的值向量；$\mathrm{e}^{\mathrm{i}n\theta}$和$\mathrm{e}^{\mathrm{i}m\theta}$为旋转因子；$\theta$为旋转角度参数；$\boldsymbol{Q}_n\mathrm{e}^{\mathrm{i}n\theta}$和$\boldsymbol{K}_m\mathrm{e}^{\mathrm{i}m\theta}$表示对$\boldsymbol{Q}_n$和$\boldsymbol{K}_m$应用旋转式位置编码[21]，“†”表示共轭转置操作。在训练时所有时间步是可以并行计算的，矩阵表达如下

$$\begin{aligned}&\boldsymbol{Q}=(\boldsymbol{X}\boldsymbol{W}_Q)\odot\boldsymbol{\Theta},\boldsymbol{K}=(\boldsymbol{X}\boldsymbol{W}_K)\odot\bar{\boldsymbol{\Theta}}\\&\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V\end{aligned} \tag{4}$$

式中：$\boldsymbol{X}$表示输入序列；$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$是可学习的权重矩阵，分别将输入$\boldsymbol{X}$投影到查询$\boldsymbol{Q}$、键$\boldsymbol{K}$、$\boldsymbol{V}$矩阵空间；$\boldsymbol{\Theta}$为复数位置编码矩阵；$\bar{\boldsymbol{\Theta}}$为$\boldsymbol{\Theta}$矩阵的共轭形式。

$$\boldsymbol{\Theta}_n=\mathrm{e}^{\mathrm{i}n\theta}\qquad D_{nm}=\begin{cases}\gamma^{n-m} & n\geqslant m\\ 0 & n<m\end{cases} \tag{5}$$

式中：$\boldsymbol{\Theta}_n$表示第n时间步的编码；D_{nm}为一个下三角矩阵中某一位置的值，仅当行索引n大于等于列索引m时元素非零。

$$\mathrm{Retention}(\boldsymbol{X})=(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\odot\boldsymbol{D})\boldsymbol{V} \tag{6}$$

式中：Retention($\boldsymbol{X}$)表示原始记忆模块的输出矩阵；$\boldsymbol{D}\in\mathbf{R}^{|w|\times|w|}$为将因果掩蔽和沿相对距离的指数衰减相结合的矩阵，$|w|$表示序列长度。引入DropKey策略，设注意力权重矩阵$\boldsymbol{A}$和掩码矩阵$\boldsymbol{M}_r$为

$$\begin{aligned}&\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\\&\boldsymbol{M}_r\sim\mathrm{Bernoulli}(r)\end{aligned} \tag{7}$$

式中：$\boldsymbol{A},\boldsymbol{M}_r\in\mathbf{R}^{|w|\times|w|}$，$\boldsymbol{M}_r$为一个服从伯努利分布的二进制掩码矩阵，矩阵中每个元素为0或1的概率为r，设置为0.5。生成注意力掩蔽矩阵$\boldsymbol{M}_{\mathrm{mask}}$，并计算得到新的权重矩阵$\boldsymbol{A}_{\mathrm{DropKey}}$，有

$$\begin{aligned}&\boldsymbol{M}_{\mathrm{mask}}=-10^{12}\boldsymbol{M}_r\\&\boldsymbol{A}_{\mathrm{DropKey}}=\boldsymbol{A}+\boldsymbol{M}_{\mathrm{mask}}\end{aligned} \tag{8}$$

最终引入正则化策略的记忆模块的输出矩阵表达为

$$\text{Retention}(\boldsymbol{X})=(\boldsymbol{A}_{\text{DropKey}}\odot\boldsymbol{D})\boldsymbol{V} \tag{9}$$

2 实验设置

2.1 数据集

为了满足小数据集、真实音频环境的前提条件,实验采用DCASE2017任务3的TUT-sound-events-2017[22]和DCASE2016任务3的TUT-sound-events-2016[23]数据集进行实验。两个数据集都包含了开发集和评估集两个部分,其中开发集使用四折交叉验证法来进行训练和验证。两个数据集均为双声道音频,每条录音3~5 min,采样率为44.1 kHz,分辨率为24位,标签由人工标注,包含了事件开始时间、事件终止时间和事件类别等必要标签。两个数据集均采集自现实场景,TUT-sound-events-2017的声学场景为街道,TUT-sound-events-2016的声学场景包含室内和居住区(室外)两个部分。

2.2 参数设置

模型训练采用Adam优化器,初始学习率Lr设置为0.000 1,每批次训练样本数量BatchSize为16,训练周期Epoch为80。训练时使用二元交叉熵损失函数(Binary crossentropy loss,BCE Loss)计算损失值,输出原始概率分布后经过的判决阈值设置为0.5。分帧时窗口大小WindowsSize设置为2 048个采样点,跳跃大小HopSize为1 024个采样点,每帧的对应时间长度约为0.023 s。对每条分帧后的音频作预处理:以连续的256帧为单位进行切分得到送入网络的样本数据,末尾不足256帧的部分直接舍弃。

2.3 评价指标

本文采用基于片段的F_1得分和错误率ER对模型性能进行评价。F_1得分计算方式为

$$F_1=\frac{2P\cdot R}{P+R}\times 100\% \tag{10}$$

式中:$P=\dfrac{\sum\text{TP}}{\sum\text{TP}+\sum\text{FP}}$,表示精确率;$R=\dfrac{\sum\text{TP}}{\sum\text{TP}+\sum\text{FN}}$,表示召回率。其中,$\sum$TP表示所有被模型正确预测为正类的片段总数;$\sum$FP表示所有被模型错误预测为正类的片段总数;$\sum$FN表示所有真实正类被模型漏检(未预测到)的片段总数。F_1得分越接近1表明模型性能越好。ER的计算方式为

$$\text{ER}=\frac{\sum_{t=1}^{T}S(t)+\sum_{t=1}^{T}I(t)+\sum_{t=1}^{T}D(t)}{\sum_{t=1}^{T}N(t)} \tag{11}$$

式中:t表示第t帧;替换错误$S(t)=\text{Min}(\text{FN}(t),\text{FP}(t))$,表示模型的预测事件类型与实际事件类型不一致;插入错误$I(t)=\text{Max}(0,\text{FP}(t)-\text{FN}(t))$,表示实际处于非活动状态的事件类型被错误地预测为活动状态;删除错误$D(t)=\text{Max}(0,\text{FN}(t)-\text{FP}(t))$,表示实际处于活动状态的事件类型被错误地预测为非活动状态;$N(t)$表示第t帧中真实标签为正的音频事件总量。ER越接近0,表明模型性能越好。

3 实验结果与分析

为验证提出的TCRETNET模型性能,本文进行了一系列实验来证明所提方法对多声音事件检测的有效性。

3.1 预测结果可视化分析

对预测结果进行可视化分析有利于更准确地分析模型。本文在TUT-sound-events-2017数据集上对TCRETNET模型的预测结果进行了可视化分析。图5展示了对于评估集中音频“a123.wav”0～58.88 s(对应前2 560帧)的帧级别预测可视化图。图6展示了在开发集和评估集上的混淆矩阵。

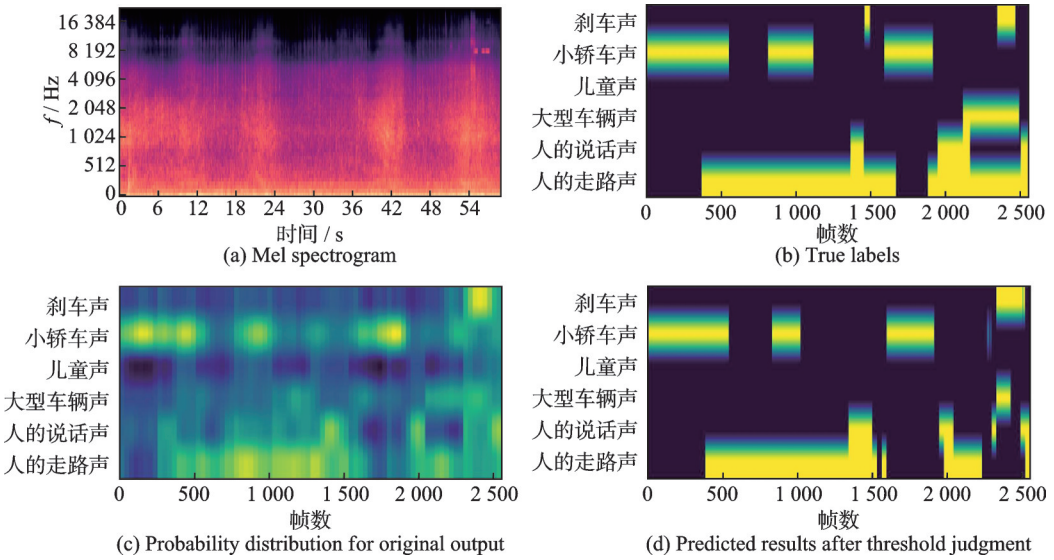

图5 帧级别预测可视化图

Fig.5 Frame level prediction visualization

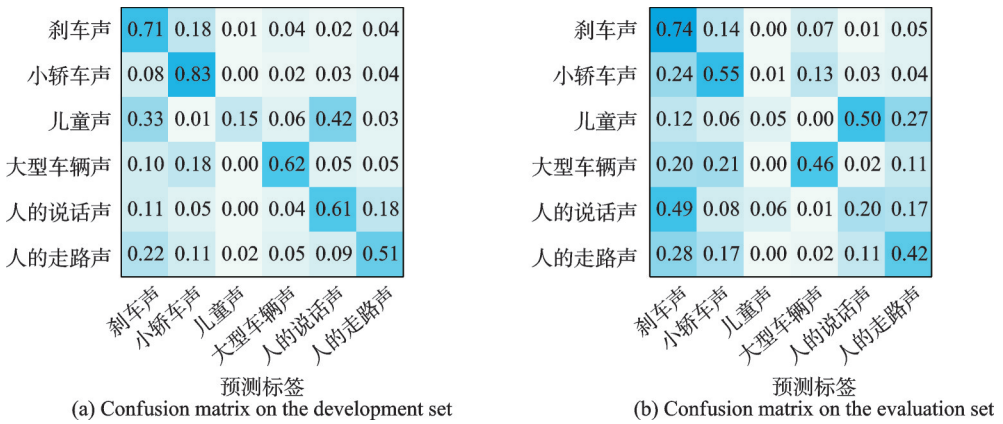

图6 在开发集和评估集上的混淆矩阵

Fig.6 Confusion matrix on development and evaluation sets

从图5可以看出,TCRETNET可以正确检测到大部分的小轿车声、人的走路声和刹车声,但针对人的说话声检测效果不佳。从图6可以看出,预测精度最高的事件为刹车声、小轿车声和大型车辆声,精度最低的事件为儿童声、人的说话声和人的走路声。出现该结果的可能原因:一方面是由于用于微调训练的公开数据集规模小、类别不平衡,对应种类的样本越多,模型检测精度越高,反之检测精度越低;另一方面由于掩蔽效应的存在,较弱的声音通常会被较强的声音所掩蔽,从而影响到模型的检测结果。

3.2 消融实验与分析

为了验证所提方法的效果,本文在TUT-sound-events-2017数据集上进行消融实验,总共设置8组

实验,表1展示了这8组消融实验在开发集和评估集上的最终结果。第1组“CNN14”表示使用图2所示的原始预训练模型CNN14,卷积层预训练初始化权重,线性层随机初始化权重,进行微调训练;第2组“RetNet”表示在TCRETNET模型的基础上剥离ConvBlocks和RFEM模块,即仅使用改进的RetNet时序网络;第3组“TCRETNET(no pretrained)”表示在TCRETNET模型的基础上,ConvBlocks进行随机初始化权重,不使用预训练初始化权重;第4组“TCRETNET(no RFEM)”表示在TCRETNET模型的基础上剥离RFEM模块;第5组“TCRETNET(no ITD,IID)”表示在TCRETNET模型的基础上,RFEM模块不进行ITD和IID特征融合;第6组“TCRETNET(BiGRU)”表示在TCRETNET模型的基础上将RetNet网络替换为双向门控循环单元(Bidirectional gated recurrent unit ,BiGRU);第7组“TCRETNET(no DropKey)”表示在TCRETNET模型的基础上,RetNet不使用DropKey正则化方法;第8组“TCRETNET”表示完整的TCRETNET模型,即使用了ConvBlocks(预训练权重初始化)、RFEM模块以及改进的RetNet网络。表2展示了第1、4、8组模型的参数量和计算复杂度。图7展示了各组模型在开发集上F_1得分和ER的学习曲线。

表1 消融实验结果

Table 1 Results of ablation experiment

实验组	模型	Development dataset		Evaluation dataset	
		ER	F_1/%	ER	F_1/%
1	CNN14	0.698	47.2	0.754	42.3
2	RetNet	0.686	51.6	0.810	37.9
3	TCRETNET(no pretrained)	0.612	57.2	0.720	48.6
4	TCRETNET(no RFEM)	0.451	72.5	0.728	54.1
5	TCRETNET(no ITD,IID)	0.423	73.8	0.691	54.8
6	TCRETNET(BiGRU)	0.528	65.2	0.705	51.1
7	TCRETNET(no DropKey)	0.404	74.5	0.673	54.2
8	TCRETNET	0.380	76.2	0.668	56.1

表2 第1、4、8组实验的模型参数量和计算复杂度对比

Table 2 Comparison of model parameter quantity and computational complexity for the first, the fourth, and the eighth experiments

模型	模型参数量/10^6	模型计算复杂度/Gmac
CNN14	151.29	5.19
TCRETNET(no RFEM)	91.78	5.34
TCRETNET	22.05	4.47

观察图7可以发现,第4、5、7、8组模型由于迁移学习的影响,第1轮的ER和F_1得分就能分别达到低于0.6和高于60%的水平,并且较少的轮次内便能达到模型的最佳性能;第6组模型虽然也应用了迁移学习,但是由于BiGRU无法并行训练、缺少自注意力机制的结构缺陷所以收敛速度缓慢。分析表1,都以第8组的完整模型为基准,可以看出:第1组模型虽然使用了预训练初始化权重,但是由于源域和目标域的数据集差异、过拟合以及忽略时序信息的原因而表现不佳;第2组模型虽然捕获了时序信息,但由于缺失细粒度的时频局部信息而表现出较差的性能;第3组模型由于ConvBlocks使用了随机初始化权重,没有使用预训练权重和微调方法,因此检测性能欠佳,验证了迁移学习的有效性;第4组模型验证了RFEM模块的有效性;第5组模型验证了融合ITD和IID特征的有效性;第6组模型用BiGRU网络来对时序信息进行建模,其检测性能出现了明显的下降,证明了改进的RetNet的有效性;第7组验证了

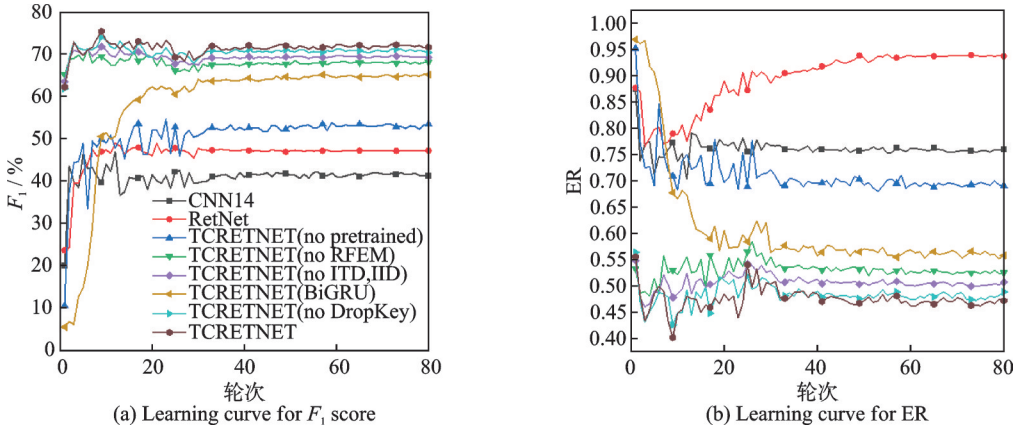

图7 消融实验中各组模型在开发集上F_1得分和ER的学习曲线

Fig.7 Learning curves of F_1 scores and ER for each model in ablation experiment on development set

正则化方法DropKey的有效性。通过分析表2可以进一步验证RFEM模块中的通道降维操作能够有效减少参数量和计算复杂度,从而达到缓解过拟合、提升模型泛化能力的效果。

3.3 TCRETNET与其他方法的性能对比

为了更加客观准确地评价TCRETNET模型的性能,本文在TUT-sound-events-2017和TUT-sound-events-2016两个公开数据集上进行实验,并将实验结果与基线系统、冠军系统模型以及其他现有的多声音事件检测方法进行对比。表3和表4展示了在两个数据集上的对比实验结果。表3和表4中"*"表示当年挑战赛冠军所使用的方法,"—"表示方法对应的论文中没有给出该项指标数据。从表3和表4可以看出,与DCASE挑战赛当年获胜系统模型相比,TCRETNET模型在DCASE 2016

表3 不同模型在TUT-sound-events-2016数据集上的对比实验结果

Table 3 Comparative experimental results of different models on the TUT-sound-events-2016 dataset

模型名称	Development dataset		Evaluation dataset	
	ER	F_1/%	ER	F_1/%
DCASE2016 baseline	0.910	23.7	0.877	34.3
RNN[24]*	0.880	34.7	0.805	47.8
MS-RNN[25]	0.820	31.5	—	—
AMS[26]	—	—	0.782	48.7
MS-FCN[27]	0.778	42.0	0.933	25.4
TCRETNET	0.603	57.3	0.699	54.4

表4 不同模型在TUT-sound-events-2017数据集上的对比实验结果

Table 4 Comparative experimental results of different models on the TUT-sound-events-2017 dataset

模型名称	Development dataset		Evaluation dataset	
	ER	F_1/%	ER	F_1/%
DCASE2017 baseline	0.690	56.7	0.936	42.8
CRNN[28]*	0.600	59.0	0.791	41.7
MS-FCN	0.571	61.2	0.784	48.6
DenseNet[29]	—	—	0.752	49.1
AMCSA[30]	—	—	0.680	49.6
TCRETNET	0.380	76.2	0.668	56.1

Task3数据集的开发集和评估集上,错误率分别降低了0.277和0.106,F_1分数分别提高了22.6%和6.6%;在DCASE 2017 Task3数据集的开发集和评估集上,错误率分别降低了0.22和0.123,F_1分数分别提高了17.2%和14.4%;和现存的一些其他模型方法相比,TCRETNET模型也表现出了更好的检测性能。

4 结束语

针对多声音事件检测任务中强标注数据集不足、真实场景下性能恶化的问题,提出了基于迁移学习卷积记忆网络TCRETNET。该方法利用预训练卷积层ConvBlocks和REFM模块增强了对局部特征的提取能力,减小了模型冗余,并通过结合能够捕获时序信息改进的RetNet网络,进一步提高了检测精度。在公开数据集上的实验结果证明,TCRETNET能够显著提升在真实环境和小数据集条件下对多声音事件的检测效果。

参考文献:

[1] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[2] 王泽雨, 王国靖. 基于改进卷积神经网络的声音事件分类模型[J]. 信息技术与信息化, 2023, (5): 181-184.
WANG Zeyu, WANG Guojing. Sound event classification model based on improved convolutional neural network[J]. Information Technology & Informatization, 2023, (5): 181-184.

[3] ZHOU J. Sound event detection in multichannel audio LSTM network[EB/OL]. (2017-06-07). https://arxiv. org/pdf/1706.02293.pdf.

[4] CAKIR E, PARASCANDOLO G, HEITTOLA T, et al. Convolutional recurrent neural networks for polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(6): 1291-1303.

[5] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 472-480.

[6] 濮子俊, 张寿明. 基于特征融合与Transformer模型的声音事件定位与检测算法研究[J]. 计算机工程与科学, 2023, 45(6): 1097-1105.
PU Zijun, ZHANG Shouming. A sound event localization and detection algorithm based on feature fusion and Transformer model[J]. Computer Engineering & Science, 2023, 45(6): 1097-1105.

[7] 李海涛, 杨树国. 基于自注意力路由胶囊网络的多音事件检测[J]. 青岛科技大学学报(自然科学版), 2022, 43(5):121-126.
LI Haitao, YANG Shuguo. Polyphonic sound event detection based on self-attention routing capsule network[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2022, 43(5): 121-126.

[8] 许春冬, 汪雄, 闵源. 融合注意力机制的SimNet声音事件定位与检测算法[J]. 国外电子测量技术, 2023, 42(8): 33-39.
XU Chundong, WANG Xiong, MIN Yuan. SimNet sound event localization and detection algorithm incorporating attention mechanism[J]. Foreign Electronic Measurement Technology, 2023, 42(8): 33-39.

[9] WANG Y, ZHAO G, XIONG K, et al. MSFF-Net: Multi-scale feature fusing networks with dilated mixed convolution and cascaded parallel framework for sound event detection[J]. Digital Signal Processing, 2022, 122: 103319.

[10] KOYAMA Y, SHIGEMI K, TAKAHASHI M, et al. Spatial data augmentation with simulated room impulse responses for sound event localization and detection[C]//Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 8872-8876.

[11] 杨利平, 侯振威, 辜小花, 等. 弱标签声音事件检测的空间-通道特征表征与自注意池化[J]. 电子学报, 2023, 51(2): 297-306.
YANG Liping, HOU Zhenwei, GU Xiaohua, et al. Spatial-channel feature representation and self-attention pooling for weakly-labeled sound event detection[J]. Acta Electronica Sinica, 2023, 51(2): 297-306.

[12] 刘臣, 倪仁健, 周立欣. 多任务实时声音事件检测卷积模型与复合数据扩增[J]. 计算机应用研究, 2023, 40(4): 1080-1087.
LIU Chen, NI Renjie, ZHOU Lixin. Multi-task real time CNN model and combined data augmentation for sound event detection[J]. Application Research of Computers, 2023, 40(4): 1080-1087.

[13] ZHENG X, SONG Y, DAI L R, et al. An effective mutual mean teaching based domain adaptation method for sound event detection[C]//Proceedings of Interspeech. Brno: [s.n.], 2021: 556-560.

[14] 林佳伟, 王士同. 用于迁移学习的多尺度领域对抗网络[J]. 数据采集与处理, 2022, 37(3): 555-565.
LIN Jiawei, WANG Shitong. Multi-scale domain adversarial network for transfer learning[J]. Journal of Data Acquisition and Processing, 2022, 37(3): 555-565.

[15] CHEN S, WU Y, WANG C, et al. Beats: Audio pre-training with acoustic tokenizers[EB/OL]. (2022-12-28). https://arxiv.org/pdf/2212.09058.pdf.

[16] KONG Q, CAO Y, IQBAL T, et al. PANNS: Large-scale pretrained audio neural networks for audio pattern recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.

[17] GEMMEKE J F, ELLIS D P W, FREEDMAN D, et al. Audio set: An ontology and human-labeled dataset for audio events [C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017: 776-780.

[18] 吕菲, 夏秀渝. 基于方位特征的听觉选择性注意计算模型研究[J]. 自动化学报, 2017, 43(4): 634-644.
LYU Fei, XIA Xiuyu. Study on computational model of auditory selective attention with orientation feature[J]. Acta Automatica Sinica, 2017, 43(4): 634-644.

[19] SUN Y, DONG L, HUANG S, et al. Retentive network: A successor to transformer for large language models[EB/OL]. (2023-07-17). https://arxiv.org/pdf/2307.08621.pdf.

[20] LI B, HU Y, NIE X, et al. DropKey for vision transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 22700-22709.

[21] SUN Y, DONG L, PATRA B, et al. A length-extrapolatable transformer[EB/OL]. (2022-12-20). https://arxiv.org/pdf/2212.10554.pdf.

[22] MESAROS A, HEITTOLA T, DIMENT A, et al. DCASE 2017 challenge setup: Tasks, datasets and baseline system[C]//Proceedings of DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events. [S. l.]: IEEE, 2017.

[23] MESAROS A, HEITTOLA T, VIRTANEN T. TUT database for acoustic scene classification and sound event detection [C]//Proceedings of 2016 24th European Signal Processing Conference (EUSIPCO). Budapest: IEEE, 2016: 1128-1132.

[24] ADAVANNE S, PARASCANDOLO G, PERTILÄ P, et al. Sound event detection in multichannel audio using spatial and harmonic features[EB/OL]. (2017-06-07). https://arxiv.org/pdf/1706.02293.pdf.

[25] LU R, DUAN Z, ZHANG C. Multi-scale recurrent neural network for sound event detection[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Cambridge: IEEE, 2018: 131-135.

[26] DING W, HE L. Adaptive multi-scale detection of acoustic events[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 294-306.

[27] WANG Y, ZHAO G, XIONG K, et al. Multi-scale and single-scale fully convolutional networks for sound event detection[J]. Neurocomputing, 2021, 421: 51-65.

[28] ADAVANNE S, VIRTANEN T. A report on sound event detection with different binaural features[EB/OL]. (2017-10-09). https://arxiv.org/pdf/1710.02997.pdf.

[29] ZHE H, YING L. Fully convolutional densenet based polyphonic sound event detection[C]//Proceedings of 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB). Fuzhou, China: IEEE, 2018: 1-6.

[30] WANG M, YAO Y, QIU H, et al. Adaptive memory-controlled self-attention for polyphonic sound event detection[J]. Symmetry, 2022, 14(2): 366.

作者简介:

陈鹏飞(1999-),男,硕士研究生,研究方向:多声音事件检测, E-mail: 358660877@qq.com。

夏秀渝(1970-),**通信作者**,女,副教授,研究方向:声音事件检测、语音分离、计算听觉场景分析等, E-mail:xiaxy@163.com。

(编辑:刘彦东)