ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 6,Nov. 2025,pp. 1556－1567
DOI:10. 16337/j. 1004-9037. 2025. 06. 014

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

一种基于特征融合的声音事件检测方法

赵　明，陈　睿

（上海海事大学信息工程学院，上海 201306）

摘　要：现有的基于深度学习的声音事件检测方法多使用传统的二维卷积，然而其平移不变性的特点并不适用于声音信号，这使得模型难以检测复杂的声音事件。针对上述问题，本文提出一种基于特征融合的混合卷积神经网络模型，通过计算频谱图的分布来自适应地生成卷积核，动态地提取与声音信号保持物理一致性的局部特征；同时并行地使用自注意力算法提取全局特征，捕获频谱图的长距离特征关联；为消除局部特征与全局特征的语义差异，将两种不同的特征表示有效结合，提出一种特征融合模块。为进一步提升模型对声音事件的检测性能，提出一种基于多尺度注意力机制的双向门控单元，对融合后的特征信息进行充分整合，突出事件帧并抑制背景帧。在DCASE2020数据集上的实验结果表明，本文方法的F_1分数达到52.57%，优于现有的其他方法。

关键词：声音信号；声音事件检测；深度学习；卷积神经网络；特征融合

中图分类号：TN912　　**文献标志码：**A

Sound Event Detection Method Based on Feature Fusion

ZHAO Ming, CHEN Rui

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Most existing deep learning-based sound event detection methods adopt the conventional 2D convolution. However, its inherent translation invariance property is incompatible with audio signals, rendering the model incompetent in detecting complex sound events. To address the issue, a hybrid convolutional neural network based on feature fusion is proposed. Specifically, by calculating the distribution of the audio spectrogram and adaptively generating convolutional kernels, the proposed model dynamically extracts local features that maintain physical consistency with the audio signal. Meanwhile, the self-attention mechanism is employed in parallel to capture long-distance feature dependencies of the spectrogram. To eliminate the semantic gap between local and global features, a feature fusion module is designed to effectively integrate these two distinct feature representations. Furthermore, to further enhance the detection performance of the proposed model, an improved bidirectional gated recurrent unit based on a multi-scale attention mechanism is proposed to fully refine the fused feature information, which emphasizes event-related frames and suppresses background frames. Experiment results on the DCASE2020 dataset indicate that the proposed model has achieved an F_1-score of 52.57%, which outperforms other existing methods.

Key words: audio signal; sound event detection; deep learning; convolutional neural network; feature fusion

---

基金项目：国家自然科学基金(62101316)。
收稿日期：2024-12-25；**修订日期：**2025-03-19

引　　言

人类的听觉系统可以在复杂的声学场景中高效分离不同的声音并关注感兴趣的声音,这种能够在嘈杂环境下专注于目标声音的现象被称为“鸡尾酒会效应”,其原因是听觉系统根据时频特征的差异性将声音划分为了不同的听觉对象[1],因此人们即使多次听到单类声音也能将其感知为一个连续的整体,同时听到多类声音时也能专注于目标声音。声音事件检测(Sound event detection, SED)作为音频处理领域一项重要的智能化技术,通过模拟人类的听觉系统对不同声音事件的时频模式进行建模,识别出输入音频中每个时间步发生的声音事件的类别[2],从而赋予机器感知声学场景的“听觉”。目前,SED技术已被应用于环境监测[3]、语音识别[4]、智能安防[5]、医疗检测[6]以及交通安全[7]等多个领域,展现出广阔的应用前景。然而,SED仍面临着诸多挑战:(1)受环境噪声的干扰,声音数据的信噪比往往偏低,影响了对目标事件的检测;(2)不同类别的声音事件可以并发存在,使得声音数据的时频模式极为复杂,增加了检测难度;(3)人工标注声音事件的时间边界相当困难且价格高昂,因此只有少量强标签数据可用,导致模型泛化性受限。

基于上述问题,国内外研究人员开展了大量研究。相较于传统的机器学习模型,基于深度学习的深度神经网络具备强大的非线性拟合能力,并且可以从声音信号里自动学习高维的特征表示,因此成为SED领域的主流方法[8-9]。通过结合卷积神经网络(Convolutional neural network, CNN)的帧级特征提取能力和循环神经网络(Recurrent neural network, RNN)的帧间联系建模能力,卷积循环神经网络(Convolutional RNN, CRNN)具备了可观的性能并成为深度学习方法里的常用模型[10-11]。但受限于模型本身的特征表示能力,常规的CRNN模型难以识别复杂的声音事件,且无法并行训练。Guo等[12]提出构建多分辨率的卷积神经网络来获取较大的感受野,提高对不同持续时长的声音信号的感知性。Vesperini等[13]尝试用胶囊网络[14]来取代传统的二维卷积,实现对不同类别声音信号的自适应建模。Kong等[15]尝试用Transformer[16]代替CRNN模型中的RNN,但由于Transformer需要较多的训练数据,其性能提升未达预期。Miyazaki等[17]尝试将CNN与Transformer的一种被用于语音识别的变体Conformer[18]进行结合,该方法在由坦佩雷大学主办的一项国际权威声学研究平台DCASE(Detection and classification of acoustic scenes and events)2020挑战任务中取得第一名。Kim等[19]提出构建基于频率分布的频率动态卷积从而实现对不同类别声音信号的动态特征提取,与Vesperini等[13]提出的胶囊网络具有更好的可解释性与性能,但卷积核的多维性与全局特征被忽略,限制了模型的性能。近来,研究人员[20-22]提出构建CNN-Conformer混合网络模型来提取同时包含局部细节与全局联系的高维特征,相关实验证明了全局特征的重要性[23],它使得模型能更好地分辨不同持续时长的声音信号。然而,现有的CNN-Conformer模型大多使用串行结构,将两种特征进行简单叠加往往会面临特征恶化的问题[24];另一方面,卷积神经网络由于具备平移不变性而被广泛用于图像识别[25-26],但对于SED任务所用的输入即频谱图而言,其频率轴并不满足平移不变性,因为声音信号的类别取决于频率分布,继续使用传统的二维卷积将限制模型的性能。

针对上述问题,本文提出一种基于特征融合的混合神经网络模型,主要贡献如下:(1)提出一种与声音信号保持物理一致性的多维动态卷积算法,有效提升对时频特征的学习能力;(2)将Conformer并行使用,提出特征融合模块将局部与全局特征有效结合,尽可能地保留优质特征;(3)提出基于多尺度注意力机制的双向门控单元,对融合后的高维特征信息进行充分整合,突出事件帧并抑制背景帧,进一步提高模型的检测性能。

1 网络设计

1.1 网络整体结构设计

本文所提出的深度神经网络的整体结构如图1所示，使用由声音信号转换得到的对数梅尔谱图作为输入。用于提取局部特征的卷积部分由一层下采样二维卷积模块和D^1层多维动态卷积模块组成，实现对不同类别声音信号的自适应特征提取；用于提取全局特征的部分由D^2层Conformer组成，从而建立频谱图各帧之间的关联；特征融合模块用于消除两种特征之间的语义差异并将其有效结合，并尽可能地保留优质特征。基于多尺度注意力机制的双向门控单元对融合后的高维特征信息进行充分整合，从帧级（局部）与音频级（全局）两种不同的时间尺度上综合考虑声音事件的分布情况，通过调整特征权重的方式突出更倾向于发生事件的帧并压制与检测无关的背景帧。

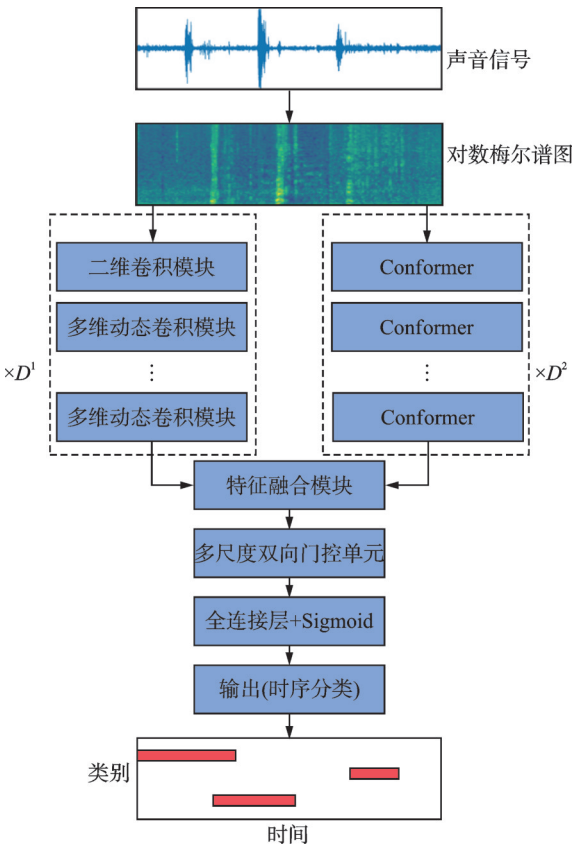

图1 所提出深度神经网络的整体结构

Fig.1 Overall structure of the proposed deep neural network

1.2 多维动态卷积

多维动态卷积的结构如图2所示，由生成动态卷积核和选择动态卷积核两条通路组成，分别表示为

$$y_i(t,f)=(\alpha_{ci}\odot\alpha_{si}\odot W_i)*\boldsymbol{x}(t,f) \tag{1}$$

$$y(t,f)=\sum_{i=1}^{N}\alpha_{wi}\odot y_i \tag{2}$$

式中：$\boldsymbol{x}$和y分别表示多维动态卷积的输入与输出；t和f分别表示输入特征图的时间与频率；$\odot(\cdot)$表示元素相乘；“$*$”表示卷积操作；W_i表示第i组基础卷积核的权重参数；y_i表示第i组动态卷积核的输出；N表示基础卷积核的组数（等于动态卷积核的组数）；α_{ci}、α_{si}和α_{wi}分别表示由输入$\boldsymbol{x}$得到的3种关于卷积核不同维度的注意力权值，其中α_{ci}和α_{si}分别用于为基础卷积核的通道与空间维度调整权重，从而生成N组动态卷积核，α_{wi}用于为每组动态卷积核的输出调整权重，从而为特征图$\boldsymbol{x}$的每一段频域选择与其相适应的卷积核组。

从图2可见，多维动态卷积先为输入特征图生成N组动态卷积核。具体来说，设输入$\boldsymbol{x}\in\mathbf{R}^{T\times F\times C}$，其中，$C$为特征图的通道数，$T$和$F$分别表示时间与频率。首先使用全局平均池化（Global average pool，GAP）对特征图的整体时频信息进行压缩，得到全局池化向量$\boldsymbol{X}^1\in\mathbf{R}^{1\times1\times C}$，即

$$\boldsymbol{X}^1=\frac{1}{T\times F}\sum_{t=1}^{T}\sum_{f=1}^{F}\boldsymbol{x}(t,f) \tag{3}$$

再使用两组全连接层（Fully connected layer，FC）进行特征降维与特征映射，分别得到关于基础卷积核通道和空间维度的注意力权值，即

$$\alpha_{ci}=\sigma(\boldsymbol{W}_c\delta(\boldsymbol{W}_f\boldsymbol{X}^1)) \tag{4}$$

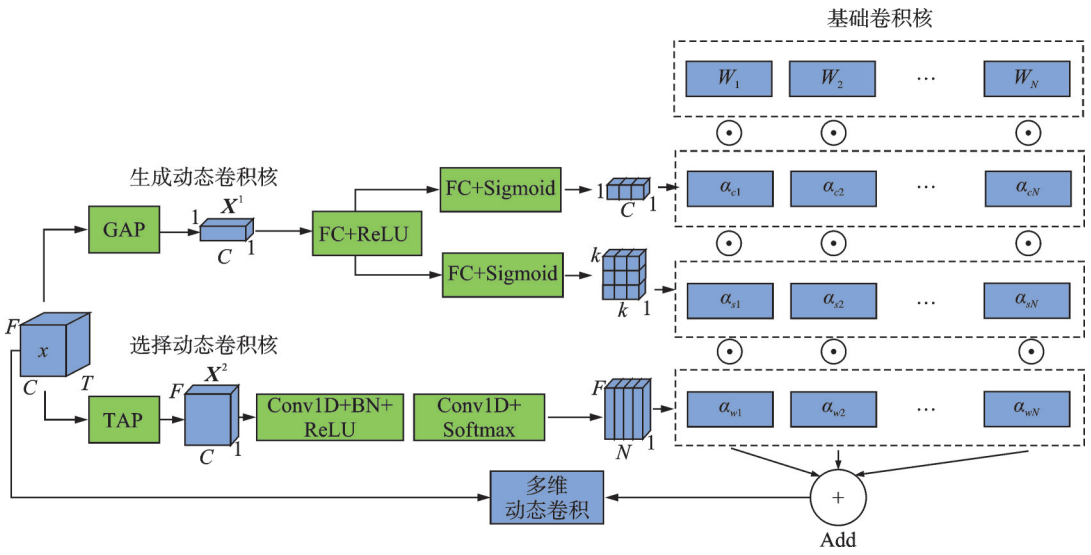

图2　多维动态卷积的结构图

Fig.2　Structure diagram of channel-spatial dynamic convolution

$$\alpha_{si}=\sigma(\boldsymbol{W}_{\mathrm{s}}\delta(\boldsymbol{W}_{\mathrm{f}}\boldsymbol{X}^{1})) \tag{5}$$

式中:$\delta(\cdot)$和$\sigma(\cdot)$分别表示ReLU和Sigmoid激活函数;$\boldsymbol{W}_{\mathrm{f}}\in\mathrm{R}^{C\times C/r}$表示用于特征降维的全连接层权重参数,$\boldsymbol{W}_{\mathrm{c}}\in\mathrm{R}^{C/r\times C}$和$\boldsymbol{W}_{\mathrm{s}}\in\mathrm{R}^{C/r\times k\times k}$则分别表示用于映射到卷积核通道与空间维度的全连接层权重参数;r表示通道缩放率;k表示卷积核大小。最后,如式(1)所示,每组基础卷积核的通道与空间维度的权重参数分别与α_{ci}和α_{si}相乘,从而得到适用于该特征图的N组动态卷积核,并将其应用于$\boldsymbol{x}$得到N组候选输出。

然后,多维动态卷积为输入特征图的每一段频域选择与其相适应的动态卷积核组。首先,对同一输入$\boldsymbol{x}\in\mathrm{R}^{T\times F\times C}$使用时域平均池化(Temporal average pool, TAP)得到与频率相关的池化向量$\boldsymbol{X}^{2}\in\mathrm{R}^{1\times F\times C}$,即

$$\boldsymbol{X}^{2}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}(t,f) \tag{6}$$

再使用两组一维卷积(1D convolution, Conv1D)进行压缩与映射,得到基于频率分布的注意力权值,即

$$\alpha_{wi}=s(\boldsymbol{\mathcal{F}}^{2}(\delta(\mathrm{BN}(\boldsymbol{\mathcal{F}}^{1}(\boldsymbol{X}^{2}))))) \tag{7}$$

式中:$s(\cdot)$表示Softmax激活函数,BN$(\cdot)$表示批归一化(Batch norm, BN);$\boldsymbol{\mathcal{F}}^{1}\in\mathrm{R}^{C\times C/r}$和$\boldsymbol{\mathcal{F}}^{2}\in\mathrm{R}^{C/r\times N}$分别表示所用的两组一维卷积。最后,如式(2)所示,使用该注意力权值为N组动态卷积核的候选输出进行加权求和,实现基于频率的权重赋值,从而为具备不同频率特性的声音信号进行自适应特征提取,使得神经网络的算法理论与频谱图的实际物理意义保持一致。

1.3　Conformer与特征融合模块

如图3(a)所示,与Transformer[16]相比,Conformer[18]增加了一层卷积模块,因此能够从时序数据中提取到额外的短时联系,并借助两个半步的前馈网络将其与多头自注意力模块捕捉到的全局联系相结合,其具有较好的特征表示能力。SED领域使用的CNN-Conformer混合模型往往将CNN与Conformer串联使用,连续地提取不同语义的特征信息[21-22]。然而,CNN专注于细化不同样本的相同特征,因此需使用批归一化,而Conformer则专注于细化同一样本的不同特征,因此需使用层归一化(Layer norm,

LN),这导致常规的CNN-Conformer混合模型往往会面临特征恶化的问题[24]。

本文提出使用与卷积神经网络并联的Conformer。如图1所示,对频谱图而不是经过卷积神经网络处理的特征图使用自注意力算法捕获长距离的特征联系,有效扩大模型的感受野,从而增强对不同持续时长声音信号的判别力,并使用特征融合模块将Conformer的输出与卷积神经网络的输出进行有效结合。特征融合模块的结构如图3(b)所示,定义Conformer部分输出的特征图为 Y_{global},卷积部分输出的特征图为 Y_{local},首先对全局特征 Y_{global} 使用基于批归一化的点卷积,使该特征图与局部特征 Y_{local} 具有同类分布,然后,将归一化后的全局特征与局部特征沿通道维度拼接,并使用全连接层进行降维选择尽可能地保留优质特征。

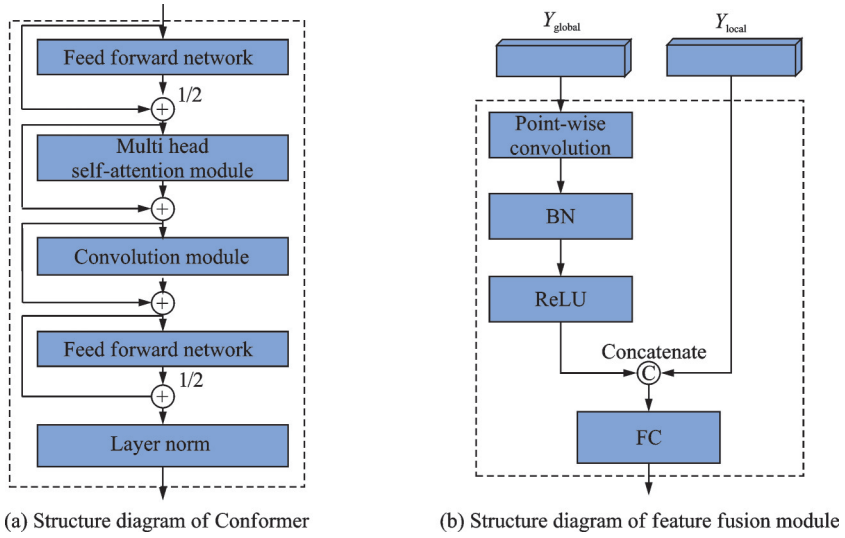

图3 Conformer与特征融合模块的结构图

Fig.3 Structure diagrams of Conformer and feature fusion module

1.4 基于多尺度注意力机制的双向门控单元

训练用于声音事件检测的深度神经网络往往需要大量同时具有声音事件类别与其起止时间的强标签数据,但该类数据由于标注困难而难以获得,仅有声音类别的弱标签数据则可以在一定程度上缓解这一困难,因此声音事件检测模型多使用弱监督方法进行训练[15,17,21]。由于两种数据均会被用于模型训练,因此模型需学习帧级与音频级两种不同时间尺度的事件预测,即强标签预测与弱标签预测。为此,本文提出一种基于多尺度注意力机制的双向门控单元(Bidirectional gated recurrent unit, BiGRU)[27],对融合后的高维特征信息进一步充分整合,学习时序特征顺序与逆序的上下文信息。具体结构如图4所示,

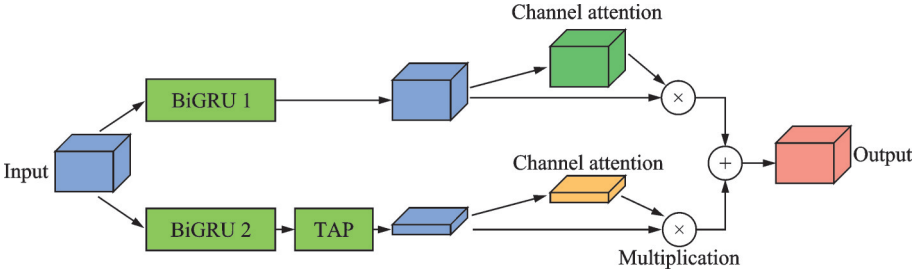

图4 本文提出的双向门控单元结构图

Fig.4 Structure diagram of the proposed BiGRU

使用基于挤压激活(Squeeze and excitation, SE)[28]的通道注意力模块学习帧级与音频级两种不同时间尺度上的互补特征信息,综合考虑声音事件在单位帧上以及声音信号整体上的存在情况,通过调整特征权重的方式为更倾向于发生声音事件的帧赋予较大值,对背景噪声则赋予较小值,从而增强对声音事件时变特征与不变特征的学习能力。

2　实验设置

2.1　数据集与预处理

本文使用的数据集为声音事件检测任务中广泛使用的DCASE2020任务4开发数据集。该数据集包括训练集、测试集和评估集,其中训练集包括2 595条强标签数据、1 578条弱标签数据以及11 412条无标签数据,强标签数据同时包括声音类别和起止时间,而弱标签数据仅有声音类别的信息,无标签数据无任何额外信息。为使用弱标签与无标签数据,模型使用DCASE官方提供的平均教师[29]框架进行半监督学习。所有声音数据的持续时间均为10 s,并可能包含若干个事件,如:Alarm Bell Ringing、Blender、Cat、Dog、Dishes、Electric Shaver/Toothbrush、Frying、Running Water、Speech和Vacuum Cleaner,等不同的声音事件可能会重叠,相同的声音事件也可能会是不连续的。

深度神经网络选择由声音信号转换得到的对数梅尔谱作为其输入。首先使用大小为2 048、帧移为256的汉明窗对采样率为16 k样本/s、持续时间为10 s的声音数据分帧,然后对帧信号使用快速傅里叶变换,最后经128个梅尔滤波器形成的滤波器组处理后得到梅尔谱图,取对数得到大小为625×128的对数梅尔谱,即神经网络的输入。

2.2　参数设置与训练过程

就卷积神经网络部分而言,令D^1为6,即该部分包括1层用于下采样的二维卷积模块和6层多维动态卷积模块,其中每个卷积模块的卷积核均为3×3,输出通道依次设定为32、64、128、256、256、256以及256,然后由批归一化、上下文门限(Context gating, CG)[30]激活函数和平均池化组成,平均池化的大小依次设定为2×2、2×2、1×2、1×2、1×2、1×2以及1×2;就Conformer部分而言,令D^2为3,即该部分包括3层线性堆叠的Conformer,取输出维度为128,其余参数设置遵循文献[18]。

深度神经网络使用Adam优化器进行训练,每个批次的训练数量取48,包括12条强标签数据、12条弱标签数据与24条无标签数据。训练过程分为热身阶段与适应阶段,其中热身阶段的训练周期为50次,学习率从1e−6开始沿指数曲线逐渐增加到0.001;适应阶段的训练周期为300次,学习率保持为0.001。

2.3　评估指标

为方便与其他方法进行性能对比,本文选择声音事件检测领域广泛使用的F_1分数作为主要的评估指标,其计算方式为

$$F_1\text{-Score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}=\frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{8}$$

式中:Precision和Recall分别表示声音事件检测的精度和召回率;TP、FP和FN则分别表示误检、漏检和正确检测的声音事件个数。

错误率(Error rate, ER)作为次要指标,其计算方式为

$$\text{ER}=\frac{\sum_i S(i)+\sum_i D(i)+\sum_i I(i)}{\sum_i N(i)} \tag{9}$$

式中:$S(i)$、$D(i)$、$I(i)$和$N(i)$分别表示第i个声音数据里替换错误、删除错误、插入错误和标注为存在的

声音事件数量。

3 实验结果与分析

3.1 模块分析

为证明本文所提出深度神经网络的有效性,选择由Kim等[19]提出的基于频率动态卷积的CRNN模型为基线模型并进行性能比对。

第1项实验研究所提出的多维动态卷积的有效性,结果如表1所示。卷积核共有卷积核组数、输入通道、输出通道和卷积核大小4个参数,而基线模型仅建立卷积核组数与频谱图的频率分布之间的联系,忽视卷积核的其余维度,限制了动态卷积的特征表示能力。从表1可见,增加对卷积核任一其他维度的注意力控制均可提高检测性能,任意两两组合形成的动态卷积均优于基线模型与单一维度的动态卷积,说明通过注意力机制建立频谱图的时频信息与卷积核之间的联系可以有效增强模型对时频特征的学习能力,从而提高模型的检测性能。但当其他3个维度均受注意力控制时,模型的性能将发生显著下降,这是因为频谱图的每一段频域均由与其对应的卷积核组进行特征提取,在较小的频率范围内增加过多的权重控制容易导致过拟合,使得模型学习了过多背景噪声在窄频带内的冗余信息,影响了对声音事件重要特征的学习。因此,本文提出的多维动态卷积选择对输入通道与卷积核大小这两个维度生成基于频谱图时频信息的注意力权重,从而为不同类别的声音信号建立自适应的特征提取器。将基线的卷积模块替换为本文提出的多维动态卷积模块后可见其性能明显提高,F_1分数超出其1.47%。

表1 卷积核不同维度的消融实验

Table 1 Ablation experiments for different dimensions of convolutional kernels

模型	卷积核组数	输入通道	输出通道	卷积核大小	F_1/%
Baseline	√	×	×	×	49.32
Ours	√	√	×	×	49.70
	√	×	√	×	50.30
	√	×	×	√	50.38
	√	√	√	×	50.41
	√	×	√	√	50.65
	√	√	×	√	**50.79**
	√	√	√	√	50.24

注:“√”指代卷积时增加对该维度的注意力控制;“×”指代不考虑该维度。

第2项实验研究本文所提出的Conformer与特征融合模块的有效性,结果如表2,3所示。从表2可见,为基线引入Conformer可以有效提升模型的检测性能,这是因为模型具有了更大的感受野,能够捕获长距离时序特征之间的联系,对不同持续时长的声音事件有了更好的判别力。但是,基于自注意力算法的模型往往需要足够的带标签数据才能发挥出较好的性能,这是由全连接层的特性决定的,而本实验所用的数据集只有少量带标签数据可用,引入过多的Conformer层容易导致过拟合,使得模型过于关注干扰噪声,稀释了对声音事件应有的关注,学习不到有效的特征,从而出现性能下降。从表3可见,Conformer层数为3层时表现出最好的性能,说明此时模型已经能比

表2 特征融合模块的重要性比对

Table 2 Importance comparison of the proposed feature fusion module

模型	F_1/%
Baseline	49.32
+ Conformer(有融合模块)	50.50
+ Conformer(无融合模块)	49.37

表3 使用不同层数Conformer的性能对比

Table 3 Performance comparison using different numbers of Conformer layers

模型	F_1/%
Baseline	49.32
+ Conformer(1层)	49.46
+ Conformer(2层)	49.57
+ Conformer(3层)	50.50
+ Conformer(4层)	48.61
+ Conformer(5层)	48.29

注:“+”指代在基线模型的基础上增加或替代模块,下同。

较好地建模频谱图的全局特征,超过3层后可训练的参数过多引起了过拟合反而使得模型性能显著下降。另一方面,从表3可见,去除特征融合模块后,模型在增加3层Conformer后网络变深但性能几乎没有得到提升,这是因为卷积与自注意力算法学习特征的方式不同,卷积使用批归一化使得不同样本的同一特征得以进行比较,而自注意力算法使用层归一化使得同一样本的不同特征得以进行比较。因此需引入额外的批归一化操作,使得由Conformer输出的全局特征与由卷积输出的局部特征具有同类分布,然后将这两种特征进行压缩选择,尽可能保留优质特征。表3显示引入具备融合模块的Conformer后检测性能相比基线高出1.18%。

第3项实验研究本文所提出的基于多尺度注意力机制的双向门控单元的有效性,结果如表4所示。由于数据集包含强标签数据与弱标签数据,用于声音事件检测的网络模型需同时学习帧级预测与音频级预测两种不同尺度的输出。从表4可见,为基线模型的BiGRU引入帧级或音频级的通道注意力均可提升检测性能,该注意力机制增强了模型对时序特征的感知性,即对有效特征赋较大值对无效特征赋较小值。另一方面,将基线模型的BiGRU替换为基于单一尺度注意力的并联BiGRU使得模型的检测性能发生显著下降,这是因为网络模型受单一的注意力控制时将会过于关注局部细节或声音信号的长期信息,导致顾此失彼,模型无法兼顾两种尺度的事件预测而导致特征劣化。由两种不同时间尺度的注意力机制所形成的互补组合则能更好地整合局部特征与全局特征,综合考虑声音事件的帧级分布与整体存在,通过调整权重的方式使得输出的时序特征中事件帧更为突出而背景帧则被抑制,可见相比基线模型其检测性能高出1.35%。

表4　使用不同尺度注意力机制的双向门控单元性能比对

Table 4　Performance comparison of BiGRU using attention module with different resolutions　%

模型	F_1
Baseline	49.32
＋单一BiGRU(仅有帧级注意力控制)	49.53
＋单一BiGRU(仅有音频级注意力控制)	50.21
＋并联BiGRU(帧级与音频级注意力控制)	**50.67**
＋并联BiGRU(帧级与帧级注意力控制)	49.40
＋并联BiGRU(音频级与音频级注意力控制)	48.05

第4项实验研究3个模块之间的联系,结果如表5所示。从表5可见,增加任一模块后模型的性能均优于基线;增加任意两个模块的组合后性能均得到了进一步的提升;同时增加3个模块的方法则取得了最高的F_1分数。这说明本文提出的3个模块之间满足互补性:多维动态卷积为不同类别的声音信号建立与其相适应的特征提取器;Conformer扩大了模型的感受野来增强对不同持续时长声音信号的判别力;多尺度双向门控单元从单位帧与整体两种时间尺度上综合学习声音事件的特征表示从而进一步提高检测性能。由此可以证明,本文所提出的3个模块均可以从不同角度增强基线模型对声音信号时频特征的学习能力,对声音事件检测任务而言均是有效的,F_1分数最高可提升3.25%。

表5　关于本文提出的3个模块的消融实验

Table 5　Ablation experiments for three proposed modules　%

模型	F_1
Baseline	49.32
＋多维动态卷积	50.79
＋ Conformer	50.50
＋多尺度BiGRU	50.67
＋多维动态卷积＋Conformer	50.92
＋多维动态卷积＋多尺度BiGRU	52.16
Conformer＋多尺度BiGRU	51.17
＋多维动态卷积＋Conformer＋多尺度BiGRU	**52.57**

3.2　与其他方法的对比分析

为了进一步证明本文所提出模型的有效性,将本模型与其他模型进行性能对比,结果如表6所示。首先,将本模型与基线模型进行详细的性能对比,从表6可见,本方

表6 本文所提模型与基线的详细性能对比

Table 6 Performance comparison between the proposed model and baseline in detail %

声音事件	F_1 (Ours)	F_1 (Base)	ER (Ours)	ER (Base)	Recall (Ours)	Recall (Base)	Precision (Ours)	Precision (Base)
Alarm Bell Ringing	**52.1**	44.9	**80**	93	**43.7**	37.8	**64.6**	55.3
Blender	**50.4**	50.0	**87**	99	**44.3**	49.4	**58.3**	50.6
Cat	**47.8**	47.2	**92**	98	**42.1**	43.8	**55.4**	51.3
Dog	**34.9**	29.4	**117**	138	**31.2**	28.8	**39.4**	30.1
Dishes	**36.1**	31.3	**140**	169	**39.6**	38.4	**33.2**	26.4
Electric Shaver / Toothbrush	**76.3**	70.6	**50**	65	**80.4**	78.3	**72.5**	64.3
Frying	**52.5**	47.8	**106**	128	**58.5**	58.5	**47.5**	40.3
Running Water	**38.4**	38.3	**103**	107	**32.1**	33.3	**47.7**	44.5
Speech	**61.8**	61.0	**77**	78	**62.5**	61.4	**61.2**	60.8
Vacuum Cleaner	**75.4**	72.7	**48**	52	**74.1**	69.0	**76.8**	76.9

法的检测性能显著优于基线，尤其是基线模型得分较低的。由于其具有较短的持续时间而容易和其他声音信号重叠混淆，Dog与Dishes类难以被检测出来，说明本方法能更好地处理复杂的声音数据。相比之下，本模型的平均F_1高于基线3.25%，平均ER低于基线0.13。

然后，将本模型与基于相同数据集的其他模型进行性能与参数对比，具体结果如表7所示。Gao等[2]提出一种基于Conformer的编码器-解码器结构来串行地学习局部特征与全局特征，并通过重构部分遮掩的频谱图来增强对短时特征的学习，本模型则将Conformer视作与卷积神经网络并行的编码器，专一学习时序数据的全局联系，相比之下本模型F_1分数超出其1.23%。Nam等[31]提出一种基于残差网络的CRNN模型，并使用卷积注意力模块来增强对声学特征的学习能力，本模型则使用多维动态卷积对不同频率特性的声音信号进行自适应特征提取，具备更好的可解释性，相比之下本模型F_1分数超出其1.97%。Wang等[32]搭建了基于多尺度卷积神经网络的CRNN模型，并设计特征选择模块将不同尺度的特征结合来扩大模型的感受野，但卷积操作受限于局部连接的特性仍不易捕捉到长距离的特征联系，相比之下本模型的F_1分数超出其3.77%。Chan等[33]设计了一种CNN-Conformer混合模型，使用Conformer替代CRNN模型中的循环神经网络来建模时序特征的上下文联系，本模型则使用基于多尺度注意力机制的Bi-GRU充分整合时序特征的短时与长期上下文信息，相比之下本模型F_1分数超出其4.07%。与DCASE2020竞赛中的前3名进行对比，本模型依次超出其4.27%、5.59%与6.17%。由此可见，本模型相比同数据集的其他模型具备更好的检测性能，本文所提出的深度神经网络具有明显优势。

表7 本文所提模型与其他模型的性能和可训练参数对比

Table 7 Performance and trainable parameters comparison between the proposed model and other models

模型	F_1/%	模型参数/10^6
本文模型	**52.57**	**14.26**
Gao等[2]	51.34	—
Nam等[31]	50.60	—
Baseline[19]	49.32	10.54
Wang等[32]	48.80	—
Chan等[33]	48.50	4.50
Ebbers等[34]	48.30	2.00
Mizayaki等[35]	46.98	2.00
Hao等[36]	46.40	2.00

动态卷积模块针对不同类别声音事件的自适应特征提取,基线模型本身也具备可观的性能,但与此同时也带来了不小的参数增量。从表7可见,基于动态卷积的CRNN模型Baseline[19]相比Hao等[26]提出的基于二维卷积的CRNN模型,F_1分数超出其2.92%,并增加了8.54×10^6可训练参数。这是因为动态卷积相比常规卷积具备了额外的若干组卷积核,因此能用更均衡、全面的特征提取器来处理复杂多变的数据。本模型则在基线模型的基础上进一步增强了其对时频特征的学习能力,F_1分数超出其3.25%,增加了3.72×10^6可训练参数。与Hu等[37]提出的一种用于实时监测儿童病患呼吸声音的模型相比,本模型多出了约2.76×10^6可训练参数。本模型在型号为Intel Xeon W-2133的CPU上对输入为10 s的声音信号进行推理所用的时间约为120 ms,与Grooby等[38]提出的一种用于实时评估新生儿心肺信号质量的模型相比,所用时间接近。现阶段的模型若要应用于实时声信号处理,还需进行额外的轻量化操作。

4　结束语

本文提出了一种适用于声音事件检测的混合卷积神经网络,同步提取与频谱图保持物理一致性的局部细节与长距离的特征联系,并提出一种多尺度的双向门控单元对融合后的高维特征表示进行充分整合,综合考虑声音事件的帧级分布与整体存在,突出重要特征并抑制无关特征。相关实验证明本模型可以有效识别复杂的声音信号,与同数据集的其他模型相比亦有较好性能。

参考文献:

[1] MESAROS A, HEITTOLA T, VIRTANEN T, et al. Sound event detection: A tutorial[J]. IEEE Signal Processing Magazine, 2021, 38(5): 67-83.

[2] GAO L J, MAO Q R, DONG M. On local temporal embedding for semi-supervised sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1687-1698.

[3] 曾金芳, 李友明, 杨恢先, 等. 基于多级残差网络的环境声音分类方法[J]. 数据采集与处理, 2021, 36(5): 960-968.
ZENG Jinfang, LI Youming, YANG Huixian, et al. Environmental sound classification method based on multilevel residual network[J]. Journal of Data Acquisition and Processing, 2021, 36(5): 960-968.

[4] WANG H, LIN X B, ZHANG J S. A lightweight CNN-Conformer model for automatic speaker verification[J]. IEEE Signal Processing Letters, 2024, 31: 56-60.

[5] NERI M, BATTISTI F, NERI A, et al. Sound event detection for human safety and security in noisy environments[J]. IEEE Access, 2022, 10: 134230-134240.

[6] FERNANDO T, SRIDHARAN S, DENMAN S, et al. Robust and interpretable temporal convolution network for event detection in lung sound recordings[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(7): 2898-2908.

[7] MARCHEGIANI L, NEWMAN P. Listening for Sirens: Locating and classifying acoustic alarms in city scenes[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 17087-17096.

[8] MENG J X, WANG X M, WANG J L, et al. A capsule network with pixel-based attention and BGRU for sound event detection[J]. Digital Signal Processing, 2022, 123: 103434.

[9] 袁文浩, 胡少东, 时云龙, 等. 一种用于语音增强的卷积门控循环网络[J]. 电子学报, 2020, 48(7): 1276-1283.
YUAN Wenhao, HU Shaodong, SHI Yunlong, et al. A convolutional gated recurrent network for speech enhancement[J]. Acta Electronica Sinica, 2020, 48(7): 1276-1283.

[10] YANG H Y, LUO L Y, WANG M, et al. Sound event detection using multi-scale dense convolutional recurrent neural network with lightweight attention[C]//Proceedings of 2023 3rd International Conference on Electronic Information Engineering and Computer (EIECT). Shenzhen, China: IEEE Press, 2023: 35-40.

[11] 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.

[12] GUO Y M, XU M X, WU Z Y, et al.Multi-scale convolutional recurrent neural network with ensemble method for weakly

labeled sound event detection[C]//Proceedings of 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). Cambridge, UK: IEEE Press, 2019: 1-5.

[13] VESPERINI F, GABRIELLI L, PRINCIPI E, et al. Polyphonic sound eventdetection by using capsule neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 310-322.

[14] SARA S, NICHOLAS F, GEOFFREY H. Dynamic routing between capsules[EB/OL]. (2017-10-26).https://arxiv.org/abs/1710.09829.

[15] KONG Q Q, XU Y, WANG W W, et al. Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2450-2460.

[16] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[EB/OL]. (2017-10-26). https://arxiv.org/abs/1706.03762.

[17] MIYAZAKI K, KOMATSU T, HAYASHI T, et al. Weakly-supervised sound event detection with self-attention[C]//Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE Press, 2020: 66-70.

[18] NIKI P, ZHANG Y, YU J H, et al. Conformer: Convolution-augmented Transformer for speech recognition[EB/OL]. (2020-05-16). https://arxiv.org/abs/2005.08100.

[19] KIM S, NAM H, KO B, et al. Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection[EB/OL]. (2022-03-29). https://arxiv.org/abs/2203.15296.

[20] WANG Q, DU J, WU H X, et al. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1251-1264.

[21] CHAN T, CHIN C. Lightweight convolutional iconformer for sound event detection[J]. IEEE Transactions on Artificial Intelligence, 2023, 4(4): 910-921.

[22] SHUL Y, CHOI J. CST-Former: Transformer with channel-spectro-temporal attention for sound event localization and detection[C]//Proceedings of ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE Press, 2024: 8686-8690.

[23] 李珠海, 郭武. 基于自注意力机制的音频对抗样本生成方法[J]. 数据采集与处理, 2024, 39(2): 416-423.
LI Zhuhai, GUO Wu. Audio adversarial examples generation method based on self-attention mechanism[J]. Journal of Data Acquisition and Processing, 2024, 39(2): 416-423.

[24] PENG Z L, GUO Z H, HUANG W, et al. Conformer: Local features coupling global representations for recognition and detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9454-9468.

[25] HAN M, LI R, ZHANG C K. LWCDNet: A lightweight fully convolution network for change detection in optical remote sensing imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[26] SHI C P, LIAO D L, ZHANG T Y, et al. Hyperspectral image classification based on expansion convolution network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[27] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures[C]//Proceedings of International Conference on Machine Learning. Lille, France: ACM, 2015: 2342-2350.

[28] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[29] ANTTI T, HARRI V. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[EB/OL]. (2017-03-06). https://arxiv.org/abs/1703.01780.

[30] MIECH A, LAPTEV I, SIVIC J. Learnable pooling with context gating for video classification[EB/OL]. (2017-06-21) [2020-12-29]. https://arxiv.org/pdf/1706.06905.

[31] NAM K, HONG K. Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function[J]. IEEE Access, 2021, 9: 7564-7575.

[32] WANG M Y, LI Y L, HU Y. Improved self-consistency training with selective feature fusion for sound event detection[C]//Proceedings of ICICSP 2023 6th International Conference on Information Communication and Signal Processing(ICICSP). Xi'an, China: IEEE Press, 2023: 460-464.

[33] CHAN T, CHIN C. Multi-branch convolutional macaron net for sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2972-2985.

[34] EBBERS J, HAEB U, REINHOLD. Convolutional recurrent neural networks for weakly labeled semi-supervised sound event detection in domestic environments[EB/OL]. (2020-06-18)[2020-11-18]. https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Ebbers_150.pdf.

[35] MIZAYAKI K, KOMATSU T. Convolution-augmented transformer for semi-supervised sound event detection[EB/OL]. (2020-06-12) [2020-11-12]. https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Miyazaki_108.pdf.

[36] HAO J Y, HOU Z W, PENG W.Cross-domain sound event detection: From synthesizedaudio to realaudio[EB/OL].(2020-06-15)[2020-11-15]. https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Hao_26.pdf.

[37] HU J H, TAO S L, YUAN G, et al. Supervised contrastive learning framework and hardware implementation of learned ResNet for real-time respiratory sound classification[J]. IEEE Transactions on Biomedical Circuits and Systems, 2025, 19(1): 185-195.

[38] GROOBY E, SITAULA C, FATTAHI D, et al. Real-time multi-level neonatal heart and lung sound quality assessment for telehealth applications[J]. IEEE Access, 2022, 10: 10934-10948.

作者简介:

赵明(1985-),女,副教授,研究方向:遥感图像处理,E-mail:zm_cynthia@163.com。

陈睿(1999-),**通信作者**,男,硕士研究生,研究方向:声音信号处理,E-mail:386398802@qq.com。

(编辑:刘彦东)