王恬, 朱晨阳, 李圣辰,等. 基于多尺度条带卷积注意的声音事件检测[J]. 智能计算机与应用,2025,15(2):168-174. DOI: 10.20169/j.issn.2095-2163.250226

基于多尺度条带卷积注意的声音事件检测

王 恬[1], 朱晨阳[1], 李圣辰[2], 邵 曦[1]
(1 南京邮电大学 通信与信息工程学院, 南京 210003; 2 西交利物浦大学 智能工程学院, 江苏 苏州 215123)

摘 要: 现有多尺度特征融合方法能够解决声音事件检测中声音事件时间尺度不一的问题,但对短时声音事件检测能力有时反而下降。本文在主流的卷积循环神经网络(CRNN)中增加了多尺度条带卷积注意模块,该模块通过多分支结构有效地捕捉了短时声音事件的不同上下文信息,也能匹配不同尺度的声音事件;每个分支使用两个深度条带卷积代替二维卷积,以匹配多次池化后短时声音事件呈现的线性特征。为了克服使用全局池化等方法的时频注意机制导致的短时声音事件特征提取不足的问题,本文引入十字交叉注意,在水平和垂直方向上聚合长程上下文信息,增强每个时频点的表征能力。使用 DCASE Challenge 2022 Task4 提供的 DESED 数据集进行了实验,实验结果表明本文提出的方法在测试集上相较于对比系统,显著提升了短时声音事件的检测能力。

关键词: 声音事件检测; 多尺度条带卷积注意; 十字交叉注意; 卷积循环神经网络
中图分类号: TN912.3;TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)02-0168-07

Sound event detection based on multi-scale strip-convolution attention

WANG Tian[1], ZHU Chenyang[1], LI Shengchen[2], SHAO Xi[1]
(1 School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2 School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, China)

Abstract: Existing multi-scale feature fusion methods can solve the problem of different time scales in sound event detection, but have poor ability to detect short time sound events. In this paper, a multi-scale striped convolution attention module is added to the mainstream CRNN model, which can effectively capture different context information of short-time sound events through multi-branch structure, and can also match different scale sound events. At the same time, each branch uses two deep strip convolution instead of two-dimensional convolution to match the linear characteristics of short-time sound events after multiple pooling. In order to overcome the problem of short time event feature extraction caused by time-frequency attention mechanism using global pooling, this paper introduces cross attention, which enhances the representation ability of each time-frequency point by aggregating long-range context information in horizontal and vertical directions. In this paper, the DESED data set provided by the DCASE Challenge 2022 Task4 is used for experiments. The results show that the proposed method significantly improves the detection capability of short-time sound events on the test set compared with the comparison system.

Key words: sound event detection; multi-scale strip-convolution attention; criss-cross attention; Convolutional Recurrent Neural Network

0 引 言

声音在生活中无处不在,可以通过分析声音得到需要的信息。例如,在黑暗的环境下监控时,声音可以代替图像和视频提供有效的信息。此外,诸如闹铃声等事件只能通过声音进行检测[1]。近年来,声音事件检测(Sound Event Detection, SED)技术获得了广泛的关注和研究,即检测音频片段中出现的事件种类,并返回识别出事件的开始和结束的时间[2]。声音事件检测在设备监控、生活辅助等多个领域具有较高的应用潜力。

基金项目: 国家科技创新 2030—"新一代人工智能"重大项目(2020AAA0106200); 国家自然科学基金(61936005,61872199 ,61872424,62001038), 姑苏领军人才青年人才创新项目(ZXL2022472)。

作者简介: 王 恬(1999—),女,硕士研究生,主要研究方向:音频分类与音频事件检测。
通信作者: 邵 曦(1976—),男,博士,教授,博士生导师,主要研究方向:多媒体信息系统分析与多媒体信息检索。Email:shaoxi@njupt.edu.cn。

收稿日期: 2023-08-07

神经网络在声音事件检测中的应用始于深度神经网络（Deep Neural Networks，DNN），即先从声音片段中提取声学特征，再将这些特征输入到神经网络中，进行高一级的抽象特征学习，最终利用这些抽象特征对声音事件进行分类和检测[3]。为了能够更好地学习局部的声学特征，Tokozume 等[4]将卷积神经网络（CNN）应用于声音事件检测中，利用有限范围内的音频上下文信息对每帧进行决策，但是由于卷积核大小固定，CNN 只能考虑固定时间的上下文信息。为了更好地利用音频的上下文信息，常将循环神经网络（Recurrent Neural Networks，RNN）用于声音事件检测的研究中[5]。由 RNN 组成的系统可以依据上下文信息做出逐帧的决策，但是 RNN 不能很好地提取时频谱图的频域信息[6]。因此，Adavanne 等[7]提出了结合 CNN 优秀的特征提取能力和 RNN 的时序建模能力的卷积循环神经网络（Convolution Recurrent Neural Networks，CRNN），该网络比单独使用 CNN 和 RNN 效果更好。近年来，在关于声音事件检测的研究和竞赛中，CRNN 是常用的一种网络[8]。

由于不同类型的声音事件具有不同的特征，包括持续时间、频率尺度等，使用单一尺度的卷积核难以匹配不同类型的声音事件[8]。为了解决这个问题，研究者们提出了各种方法，Zhang 等[9]通过上下采样和残差跳连融合 CNN 的低层和高层特征，将 CNN 提取到的特征送到 RNN 中进行进一步的时间特征学习；Koh 等[10]将特征金字塔的概念运用于 CRNN，将高层 RNN 输出的特征经过上采样与低层 RNN 输出的特征进行融合；Kim 等[11]使用了特征金字塔组件，但将 RNN 替换为 Transformer 编码器；Zhang 等[12]利用 4 组具有不同时频卷积核的并行 CNN，从输入的对数梅尔谱中提取更多的有用信息，适用于数据较少的声音事件检测系统中；Tang 等提出多尺度残差卷积循环神经网络（Multi-Scale Residual Convolutional Recurrent Neural Network，MS-RCRNN），该网络在 CRNN 的基础上堆叠了更深的 3×3 卷积层，并在基础残差块的基础上增加了一个卷积核大小为 5 的分支；Zheng 等[13]在 RCRNN 基础上增加选择核卷积模块，利用通道注意让网络学习不同感受野之间的联系，自适应地融合不同感受野的信息。这些方案在一定程度上提升了模型的整体性能，但忽略了短时声音事件容易出现特征提取不足的问题。

本文提出的多尺度条带卷积注意残差卷积循环神经网络（Multi-Scale Strip Convolution Attention Residual Convolutional Recurrent Neural Network，MSSCA-RCRNN）在一定程度上改进了短时声音事件容易出现特征提取不足的问题。MSSCA-RCRNN 即在 RCRNN 的高层 3×3 卷积后增加 MSSCA 模块，MSSCA 模块通过多个分支组合声音事件的不同上下文特征，同时也能匹配不同尺度的声音事件。由于短时声音事件在池化后，呈现更多的条状元素甚至点元素（短时、频窄），为了提取相关特征，本文采用两个深度条带卷积，作为网格卷积的有益补充。另一方面，一般的时间-频率注意机制会使用全局池化等方法，使得模型仍然关注时间较长的声音事件，而短时声音事件的特征表现不够明显。为此，本文引入了十字交叉注意机制，在水平和垂直方向上聚合上下文信息，建立不同时频点之间的联系，从而加强每个时频点的表征能力。

1　基于 MSSCA 的声音事件检测

1.1　模型框架

本文提出的 MSSCA-RCRNN 框架如图 1 所示。该网络以对数梅尔（Mel）为输入特征，共分为 3 部分：特征提取模块、时间定位模块以及事件决策模块。特征提取模块由两个普通的二维卷积块和 5 个 MSSCA 残差卷积块组成，MSSCA 残差卷积块采用多尺度条带卷积注意和残差结构，能够更好地提取音频的时频特征，并减少梯度消失问题。时间定位模块由两个双向门控循环神经网络（Gate Recurrent Units，GRU）构成，用于学习音频事件相关的时序信息。事件决策模块由两个全连接层组成，用于将特征转换为最终的类别标签。

图 1　MSSCA-RCRNN 框架图

Fig. 1　Framework of MSSCA-RCRNN

1.1.1　MSSCA 残差卷积块

MSSCA 残差卷积块的框架如图 2 所示，主要包括多尺度条带卷积模块、通道混合模块以及十字交叉注意模块。其中，多尺度条带卷积模块中普通二维卷积用于聚合局部信息，同时使用多分支结构能

够有效地考虑目标声音周围不同尺度的上下文信息,并能够匹配不同尺度的声音事件。鉴于短时声音事件在多次池化后呈现出较多的线性元素,二维卷积可能会捕捉过多的背景音,因此使用两个深度条带卷积代替深度二维卷积进行特征提取,增强短时声音事件的特征提取能力。由于深度卷积仅对每个通道进行单独的卷积操作,无法有效地跨通道运算,获得更完善的特征,因此在特征融合后加入了1×1 卷积建模通道之间的关系,并将其输出直接作为注意权重,以对输入特征进行重新加权。

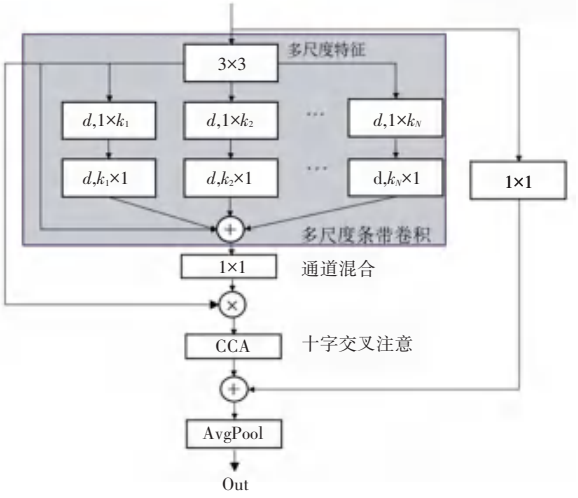

图 2 MSSCA 残差卷积块框架图

Fig. 2 Framework of MSSCA residual convolution block

输入特征图 F, 经过多尺度条带卷积模块后的特征图为 F_1, 计算公式如下:

$$F_1 = \sum_{i=0}^{N} \mathrm{Conv}_{\mathrm{Scale}_i}(F') \tag{1}$$

其中, F' 表示经过 3×3 卷积后的特征图; $\mathrm{Scale}_i, i \in \{0,1,\cdots,N\}$ 表示第 i 个分支,分支总个数为 $N+1$。

在 N 个分支中,使用两个深度条带卷积来近似标准深度卷积,深度条带卷积是轻量级的,能大量减少模型参数,模拟核大小为 $k_i \times k_i, i \in \{0,1,\cdots,N\}$ 的标准二维卷积,只需要一对 $k_i \times 1$ 和 $1 \times k_i$ 的卷积,很多特征中有条状的元素[14]。

F_1 经过通道混合模块和十字交叉注意模块后得到特征图 Att,计算公式如下:

$$\mathrm{Att} = \mathrm{CCA}(\mathrm{Conv}_{1\times1}(F_1) \otimes F') \tag{2}$$

其中,CCA 表示十字交叉注意, $\mathrm{Conv}_{1\times1}$ 用于建模不同通道之间的关系,通过该卷积的输出直接作为注意权重,对 F' 进行重新加权。

MSSCA 残差卷积块最终的输出特征图 Out 的计算公式如下:

$$\mathrm{Out} = \mathrm{AvgPool}(\mathrm{Att} + \mathrm{Conv}_{1\times1}(F)) \tag{3}$$

其中, $\mathrm{Conv}_{1\times1}$ 用于通道升维,使得 MSSCA 残差卷积块的输入维度和 Att 的维度一致。

1.1.2 十字交叉注意

为了避免常用的时间-频率注意机制使用全局池化等操作导致短时声音事件的特征丢失,本文使用了十字交叉注意(Criss-Cross Attention),聚合水平和垂直方向的长程上下文信息,建立不同时频点之间的联系,从而增强每个时频点的表征能力。十字交叉注意的架构如图 3 所示。

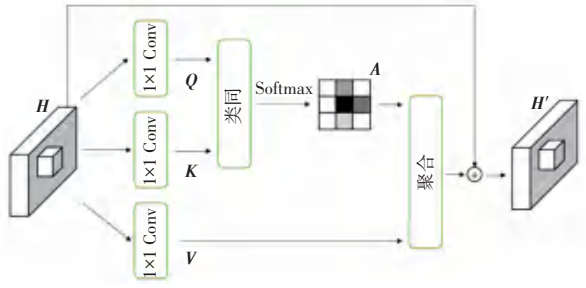

图 3 十字交叉注意框架

Fig. 3 Framework of Criss-cross Attention

十字交叉注意将特征图 H 分别经过 1×1 的卷积操作,得到 3 个映射特征 Q、K、V, Q 和 K 的通道数减少, V 的通道数不变,对 Q 和 K 进行类同操作,类同操作的公式如下:

$$d_{i,u} = \boldsymbol{Q}_u \boldsymbol{\Omega}_{i,u}^{\mathrm{T}} \tag{4}$$

其中, $\boldsymbol{Q}_u$ 是 Q 的空间维度的每个位置 u 对应的一个向量; $\boldsymbol{\Omega}_u$ 是在 K 中提取的与 u 在同一行同一列的特征向量; $\boldsymbol{\Omega}_{i,u}$ 是 $\boldsymbol{\Omega}_u$ 的第 i 个元素; $d_{i,u}$ 是特征 $\boldsymbol{Q}_u$ 与 $\boldsymbol{\Omega}_{i,u}, i = [1,\cdots,C+F-1]$ 的相关度。

在通道维度上应用 Softmax 计算注意特征图 A。

将 A 和 V 进行聚合操作,收集上下文特征。聚合操作公式如下:

$$H_u' = \sum_{i=0}^{T+F-1} A_{i,u} \boldsymbol{\Phi}_{i,u} + H_u \tag{5}$$

其中, $A_{i,u}$ 是特征图 A 在位置 u 第 i 个通道的标量值; $\boldsymbol{\Phi}_{i,u}$ 是 V 中与 u 在同一行同一列的特征向量; H_u 和 H_u' 分别表示特征图 H 和 H' 上位置 u 处的特征向量。

1.2 模型训练方法

由于训练数据中有大量无标签数据,研究者们提出了不同的训练方法,其中 Mean-teacher student 训练方法是最主流的。Mean-teacher student 训练方法由学生模型和教师模型组成,这两个模型的结构

完全相同，但权重更新方式不同。学生模型的权重使用梯度下降法训练得到，教师模型的权重是学生模型权重的指数移动平均（EMA）[15]。学生模型是最终被使用的模型，而教师模型是为了在训练中辅助学生模型。

在训练过程中，教师模型接收到与学生模型相同的输入，但添加了高斯噪声，强标签数据、弱标签数据和无标签数据均存在一致性损失（均方误差），包含强预测和弱预测两种一致性损失；而只有强标签数据和弱标签数据存在分类损失，对合成的强标签数据，计算帧级别损失（二进制交叉熵损失），对于弱标签数据，计算片段级别损失。因此，共有 4 种成分的损失：两种分类损失（强和弱）和两种一致性损失（强和弱），这些损失的组合公式如下[15]：

$$L = L_{\text{class}_w} + \lambda L_{\text{cons}_w} + L_{\text{class}_s} + \lambda L_{\text{cons}_s} \tag{6}$$

其中，L_{class_s} 和 L_{class_w} 分别表示强标签数据和弱标签数据的分类损失；L_{cons_w} 和 L_{cons_s} 分别表示强预测和弱预测的一致性损失；λ 为平衡一致性损失和分类损失的参数，其初始被设定为 0，随着迭代次数的增加逐渐变大，最终变为 1。

在训练初始阶段就考虑一致性损失，可能会对模型的拟合造成负面影响，甚至导致无法拟合[16]。因此需要对一致性损失项乘以一个系数 λ，控制其在总损失中占据的权重，以便平衡一致性损失和分类损失。

2　实验及结果分析

2.1　实验数据集

目前，声音事件检测的数据可以按照标签的不同分为 3 类。第一类是无标签数据，即未经过标记的原始数据；第二类是弱标签数据，仅包含声音事件的分类信息，缺少声音事件发生的时间信息；第三类是强标签数据，既包含声音事件的分类信息，又包含声音事件发生的具体起止时间。

本文采用 DCASE2022 Challenge Task4 提供的 DESED 数据集进行实验，该数据集包括 10 个事件类，每个音频样本持续的时间不超过 10 s。事件类别包括：闹钟/警报/铃声（Alarm/bell/ringing），搅拌机（Blender），猫（Cat），狗（Dog），盘子（Dishes），电动剃须刀/牙刷（Electric shaver/toothbrush），油炸（Frying），自来水（Running water），说话声（Speech），吸尘器（Vacuum cleaner）。实验按照 DCASE 2022 官方划定的数据集划分为训练集、验证集和评估集，训练集包含 1 578 个弱标签音频片段、14 412 个无标签音频片段以及 10 000 个合成的强标签音频片段，无标签数据集每个类的分布接近有标签数据集的分布；验证集中包含 1 168 个强标签标注的音频。所有的无标签音频和弱标签音频均来自 Audioset 的真实音频数据，而合成的强标签音频是使用 Scaper 软件生成的音频片段。

为了区分出长时声音事件和短时声音事件，本文统计了强标签数据集中每个声音事件类别中持续时间小于 0.3 s 的样本数量，见表 1。由于无标签数据集的分布与有标签数据集的分布相近，因此强标签数据集中的分布可以被近似看作整个数据集的分布。

表 1　强标签数据中各类声音事件小于 0.3 s 的样本数

Table 1　Number of samples of each type of sound event in strong label dataset that is less than 0.3 s

声音事件	小于 0.3 s 的样本数	声音事件	小于 0.3 s 的样本数
Speech	711	Cat	21
Dishes	1 505	Frying	0
Alarm/bell/ringing	249	Blender	0
Dog	447	Electric	0
Running water	29	Vacuum cleaner	0

本文将短时样本数多的声音事件，闹钟/警报/铃声、狗、盘子和说话声定为短时声音事件。

2.2　参数设置

在提取对数 Mel 特征时，本文设置采样窗口的长度为 2 048，帧移为 256，滤波器数量为 128，最大采样频率为 8 000 Hz。最后对于一个 10 s 的音频得到一个形状为 626 × 128 的特征矩阵。

模型从底层到高层，3×3 卷积的滤波器数量依次设置为 16，32，64，128，128，128，128，平局池化率设置为[[2，2]，[2，2]，[1，2]，[1，2]，[1，2]，[1，2]，[1，2]]，双向 GRU 的时间步为 128。

实验采用 Adam 作为优化器，学习率为 0.001，训练的迭代次数为 200，每个迭代批次的训练数据数量为 48，batch_size 设置为 24。为了平衡有标签数据和无标签数据的数量，每次迭代从强标签数据集、弱标签数据集和无标签数据集中分别选取 12、12、24 个音频样本。

2.3　评估指标

本文使用的评价指标包括 $F1$ - Score、错误率（Error Rate，ER）以及复音声音事件检测得分（Polyphonic Sound Detection Score，$PSDS$），分别从不同角度反映了检测系统的性能。$F1$ - Score 是精

确率和召回率的调和平均数,能够综合评估系统对声音事件的识别能力;错误率反映了系统检测错误的比例; *PSDS* 适用于评估系统在处理复杂声音场景时的性能,如多个声音事件同时发生的情况。

PSDS 还根据不同的应用程序设置了其他参数来进行调整,以满足不同用户体验需求。*PSDS* 的计算公式如下[18]:

$$PSDS = \frac{1}{e_{\max}} \int_0^{e_{\max}} r(e)\,\mathrm{d}e \tag{7}$$

其中, $e_{\max}$ 为最大有效假正例率 ($eFPR$), $r(e)$ 由一组与类别相关的受试者工作特征(Receiver Operating Characteristic, *ROC*) 函数简化得到的一条多音检测 *ROC* 函数。

本文采用与近年来声音场景和事件的检测与分类(Detection and Classification of Acoustic Scenes and Events, DCASE)比赛的任务 4 相同的两组参数设置方法,分别命名为 *PSDS*1 和 *PSDS*2, 将这两个指标之和作为系统整体性能的一个评价标准。*PSDS*1 注重评估声音事件的定位性能,*PSDS*2 注重评估声音事件的分类性能。

2.4 结果分析

2.4.1 不同多尺度特征融合模型性能对比

为了验证本文提出的多尺度条带卷积模块对短时声音事件的检测能力,本文将提出的模型与 CRNN、MS-RCRNN 以及 SK-RCRNN[13]进行对比实验。在实验时,除 CRNN 外,其余 3 种模型均采用两分支结构,并且分支的卷积核大小分别设置为 3 和 5;此外,这 4 种模型都使用了相同的训练方式。4 种模型对各类声音事件的检测结果见表 2, *PSDS* 得分、参数量以及计算复杂度见表 3。

表 2 4 种模型对各类声音事件的检测结果

Table 2 Detection results of various sound events by four models

声音事件	CRNN		MSSC-RCRNN		MS-RCRNN		SK-RCRNN[14]	
	*F*1/%	*ER*	*F*1/%	*ER*	*F*1/%	*ER*	*F*1/%	*ER*
Alarm	46.2	0.91	44.9	0.94	41.5	1.06	43.310	0.95
Speech	53.8	0.82	**59.0**	**0.74**	**55.1**	**0.80**	**57**	**0.78**
Dog	26.7	1.24	**32.8**	**1.21**	24.4	1.31	27.60	1.34
Dishes	23.9	1.33	**30.3**	**0.99**	**28.4**	**1.17**	**32.90**	**1.17**
Cat	34.5	1.30	42.4	1.14	47.8	1.01	43.610	0.95
Electric	50.0	1.02	49.7	1.31	45.3	1.26	51.90	1.17
Frying	23.8	1.69	41.5	1.53	43.9	1.28	43.20	1.34
Running	35.8	1.15	40.9	1.01	38.6	1.18	43.10	0.99
Blender	40.7	1.09	38.3	1.37	40.0	1.28	35.40	1.55
Vacuum	53.9	0.89	55.1	1.05	43.8	1.42	53.60	0.98
Overall	38.9	1.14	43.5	1.14	43.84	1.00	43.18	1.15

从表 2 中可以看出,3 个多尺度特征融合模型都在一定程度上提升了某些短时声音事件的检测效果,表明多尺度特征融合能够聚合上下文信息,从而增强对短时声音事件的特征提取能力; MSSC-RCRNN 模型在 Alarm、Speech 和 Dog 的检测能力均优于其他两个模型。尽管在对 Dishes 的检测效果上稍逊于 SK-RCRNN,但仍优于 MS-RCRNN,这说明使用条带卷积能够较好地提取出特征图中的条状元素,从而显著增强对短时声音的检测能力。

3 种模型的 *PSDS* 得分、参数量以及计算复杂度见表 3。

表 3 4 种模型的 *PSDS* 得分、参数量以及计算复杂度

Table 3 *PSDS*, number of parameters, and computational complexity for four types of models

模型	PSDS1	PSDS2	PSDS1+PSDS2	参数量/M	计算复杂度/G
CRNN	0.336	0.536	0.872		
MSCA-RCRNN	0.379	0.572	0.927	1.2	48.6
MS-RCRNN	0.362	0.559	0.921	3.2	165.3
SK-RCRNN[13]	0.355	0.532	0.887	1.3	54.1

从表 3 可以看出，MSSCA-RCRNN 的 *PSDS*1 和 *PSDS*2 得分均高于其他两种模型，说明多尺度条带卷积能够作为网格卷积的有益补充，在提升短时声音事件检测的同时，也能够增强模型整体的定位和分类能力，MSSCA-RCRNN 的参数量和计算量也是最小的。

2.4.2　不同注意机制性能对比

为了验证十字交叉注意对短时声音事件的提取能力，本文将其与已有的时间-频率注意进行了比较。其中，文献[18]引入时频挤压与激励块(SE_TFA)；文献[19]使用并行结构的时间-频率注意机制，并将通道维度压缩为 1(Parallel_TFA)；文献[20]也使用并行结构，但未将通道维度化为 1，而是直接对时间维度和频率维度进行了全局池化(TFA)。不同注意机制的 *PSDS* 得分、参数量以及计算复杂度见表 4。

从表 4 中可见，使用十字交叉注意的模型的性能最好，参数量和计算复杂度也较小。

为了验证十字交叉注意对短时声音事件的检测效果。无注意、十字交叉注意和 TFA[20] 对各类声音事件的检测结果见表 5。

表 4　不同注意机制的 *PSDS* 得分、参数量以及计算复杂度

Table 4　*PSDS*, number of parameters, and computational complexity for different attention mechanisms

注意力机制	*PSDS*1	*PSDS*2	*PSDS*1 + *PSDS*2	增加的参数量	计算复杂度(G)
SE_TFA[18]	0.365	0.582	0.927	+0	48.14
Parallel_TFA[19]	0.356	0.556	0.912	+581	48.18
TFA[20]	0.361	0.563	0.924	+1 162	48.26
SE_TFA[18]	0.379	0.572	0.951	+837 888	51.45
十字交叉注意	0.383	0.572	0.955	+97 183	49.15

表 5　十字交叉注意和 TFA 对各类声音事件的检测结果

Table 5　Detection results of various sound events for Cross-Cross Attention and TFA

声音事件	十字交叉注意		TFA[20]		无注意	
	*F*1/%	*ER*	*F*1/%	*ER*	*F*1/%	*ER*
Alarm	**47.1**	**0.96**	44.2	0.93	44.9	0.94
Speech	53.0	0.85	57.6	0.77	59.0	0.74
Dishes	**31.1**	**1.18**	27.6	1.10	30.3	0.99
Dog	**34.5**	**1.12**	32.8	1.17	32.8	1.21
Cat	43.6	0.95	42.6	1.18	42.4	1.14
Electric	51.0	1.12	53.1	1.17	49.7	1.31
Frying	44.3	1.36	41.5	1.35	41.5	1.53
Running water	44.4	1.05	41.0	1.05	40.9	1.01
Blender	38.4	1.47	43.6	1.21	38.3	1.37
Vacuum cleaner	54.7	0.99	54.80	1.08	53.1	1.05
Overall	44.3	1.04	43.86	1.10	43.5	1.14

结合表 4 和表 5 的实验结果可以看出，虽然 TFA[20] 能够提升模型的整体性能，但其对短时声音事件的检测能力却出现了下降，说明在时频注意机制中使用全局池化会导致模型过于专注于长时声音事件，而忽略了短时声音事件。而十字交叉注意能够进一步提升 Alarm、Dishes 和 Dog 3 种短时声音的检测能力，表明十字交叉注意机制通过在每个时频点上聚合水平和垂直的上下文信息增强了每个时频点的特征表达能力，确保不会忽视短时声音事件的特征，从而能够更好地捕捉这类事件。

3　结束语

本文针对多尺度特征融合不能有效提升短时声音事件检测能力的问题，提出了多尺度条带卷积注意-残差卷积循环神经网络(MSSCA-RCRNN)。该网络不仅考虑了多个尺度的信息，还能够捕捉短时

声音事件在特征图上的条状元素,通过引入十字交叉注意进一步提升了对短时声音事件的检测效果。通过与现有多尺度模型和时间-频率注意机制进行对比实验,结果表明提出网络使系统性能有一定的提升。接下来的工作中,考虑将全局和局部特征融合,以获得更加全面的声音特征表征。

参考文献

[1] CHAN T K, CHIN C S. A comprehensive review of polyphonic sound event detection[J]. IEEE Access, 2020(8): 103339-103373.

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

[3] RAVANELLI M, ELIZALDE B, NI K, et al. Audio concept classification with hierarchical deep neural networks[C]// Proceedings of the 22nd European Signal Processing Conference. Piscataway, NJ: IEEE, 2014: 606-610.

[4] TOKOZUME Y, HARADA T. Learning environmental sounds with end-to-end convolutional neural network[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 2721-2725.

[5] MA Junbo, WANG Ruili, JI Wanting, et al. Relational recurrent neural networks for polyphonic sound event detection[J]. Multimedia Tools and Applications, 2019, 78(20): 29509-29527.

[6] 杨利平,郝峻永,辜小花,等. 音频标记一致性约束 CRNN 声音事件检测[J]. 电子与信息学报,2022,44(3):1102-1110.

[7] ADAVANNE S, POLITIS A, NIKUNEN J, et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 13(1): 34-48.

[8] 侯元波. 基于模糊标签的音频标记与音频事件检测[D]. 北京: 北京邮电大学,2020.

[9] ZHANG Jingyang, DING Wenhao, KANG Jintao, et al. Multi-scale time-frequency attention for acoustic event detection[J]. arXiv preprint arXiv,1904.00063, 2019.

[10] KOH C Y, CHEN Y S, LIU Y W, et al. Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks[C]// Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2021: 376-380.

[11] KIM S J, CHUNG Y J. Multi-scale features for transformer model to improve the performance of sound event detection[J]. Applied Sciences, 2022, 12(5): 2626.

[12] ZHANG Keming, CAI Yuanwen, REN Yuan, et al. MTF-CRNN: Multiscale time-frequency convolutional recurrent neural network for sound event detection[J]. IEEE Access, 2020(8): 147337-147348.

[13] ZHENG Xu, SONG Yan, MCLOUGHLIN I, et al. An improved mean teacher based method for large scale weakly labeled semi-supervised sound event detection[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2021: 356-360.

[14] ZHANG Xulong, WANG Jianzong, CHENG Ning, et al. Susing: Su-net for singing voice synthesis[C]// Proceedings of the International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2022: 1-7.

[15] TURPAULT N, SERIZEL R, SALAMON J, et al. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis[C]// Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events. 2019: 253-257.

[16] 于东池. 居家环境下声音事件检测技术的研究[D]. 北京: 北方工业大学,2022.

[17] BILEN Ç, FERRONI G, TUVERI F, et al. A framework for the robust evaluation of sound event detection[C]//Proceedings of National Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 60-65.

[18] XIA W, KOISHIDA K. Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation[J]. arXiv preprint arXiv, 1908.01399, 2019.

[19] WANG Helin, ZOU Yuexian, CHONG Dading, et al. Environmental sound classification with parallel temporal-spectral attention[J]. arXiv preprint arXiv, 1912.06808, 2019.

[20] ZHANG Qiquan, Song Qi, NI Zhaoheng, et al. Time-frequency attention for monaural speech enhancement[C]// Proceedings of 2022 IEEE National Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2022: 7852-7856.