弱标签声音事件检测的空间-通道特征表征与自注意池化

杨利平[1],侯振威[1],辜小花[2],郝峻永[1]
(1. 重庆大学光电技术及系统教育部重点实验室,重庆 400044; 2. 重庆科技学院电气工程学院,重庆 401331)

摘 要: 深度神经网络声音事件检测方法需要大量标记声音事件类别和起止时间的强标签音频样本,然而强标签标注非常困难和耗时. 弱标签声音事件检测是解决这一困难的有效途径. 本文将弱标签声音事件检测作为多实例学习问题,并基于卷积循环神经网络提出弱标签声音事件检测的空间-通道特征表征与自注意池化方法. 该方法研究多实例弱标签声音事件检测的特征表征和帧级预测结果池化两个方面的内容. 在特征表征方面,为了增强卷积神经网络的特征表征能力,结合上下文门控和通道注意机制构建门控注意力结构并嵌入到卷积循环神经网络中,实现了音频样本特征的空间和通道特征选择;在预测结果池化方面,引入自注意思想设计音频帧预测结果的自注意池化方法,增强了音频样本中事件帧之间的相关度,使事件帧获得更大的权重. 本文方法通过对卷积循环神经网络特征表征和预测结果池化的革新,有效提升了模型的检测性能. 本文提出的方法在DCASE 2017任务4和DCASE 2018任务4数据集的评估集中分别取得了52.47%和31.00%的$F1$得分,性能优于当前绝大部分的弱标签声音事件检测方法. 实验结果表明:本文提出的空间-通道特征表征与自注意池化方法能显著改善弱标签声音事件检测的综合性能.

关键词: 特征表征;自注意池化;卷积循环神经网络;弱标签学习;声音事件检测
基金项目: 国家自然科学基金(No.61903054)
中图分类号: TP391.4;TP37 **文献标识码:** A **文章编号:** 0372-2112(2023)02-0297-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210035

Spatial-Channel Feature Representation and Self-attention Pooling for Weakly-Labeled Sound Event Detection

YANG Li-ping[1], HOU Zhen-wei[1], GU Xiao-hua[2], HAO Jun-yong[1]
(1. *Key Laboratory of Optoelectronic Technology and Systems*, *Ministry of Education*, *Chongqing University*, *Chongqing* 400044, *China*;
2. *School of Electrical Engineering*, *Chongqing University of Science & Technology*, *Chongqing* 401331, *China*)

Abstract: A large amount of strong labeled audio samples, which are annotated with detailed sound event categories and timestamps, is required for a deep neural network sound event detection (SED) model. However, obtaining strong label is very difficult and time-consuming. Weakly-labeled SED is an effective way to solve this problem. This paper approaches weakly-labeled SED as a multiple instance learning (MIL) problem and proposes a spatial-channel feature representation and self-attention pooling method for weakly-labeled SED based on convolutional recurrent neural network (CRNN). The proposed method studies the feature representation and the frame-level prediction pooling method for multi-instance weakly-labeled SED. In feature representation, in order to enhance the ability of CRNN, we design a gating attention structure by combining context gating and channel attention mechanism, and embed it into CRNN to realize the spatial and channel selection of audio sample features. In frame-level prediction pooling, we introduce the idea of self-attention and design a self-attention pooling (SAP) function to enhance the event frame correlation in the audio sample and assign great weights for event frames. The proposed method effectively improves the detection performance of SED model by innovating the feature representation of CRNN and the pooling method of frame-level predictions. The proposed method has achieved 52.47% and 31.00% $F1$ scores respectively in the evaluation set of DCASE 2017 task 4 and DCASE 2018 task 4 datasets, which outperforms most of the current weakly-labeled SED methods. Experimental results show that the proposed spatial-channel feature representation and self-attention pooling method can significantly improve the performance of weakly-labeled SED.

收稿日期:2020-12-29;修回日期:2021-04-07;责任编辑:王天慧

Key words: feature representation; self-attention pooling; convolutional recurrent neural network; weakly-labeled learning; sound event detection

Foundation Item(s): National Natural Science Foundation of China (No.61903054)

1 引 言

声音事件检测(Sound Event Detection,SED)旨在识别输入音频信号中所包含的目标声音事件并确定事件出现的起止时间.随着智能化设备在日常生活中的普及,声音事件检测的潜在应用也在不断增多.我们可以依据检测到的声音事件分析音频发生的场景[1],也可以通过目标声音事件(如枪声、尖叫声等)的检测实现公共安全监控[2,3].

目前绝大部分的声音事件检测方法都是基于深度学习的.卷积神经网络(Convolutional Neural Network,CNN)[4]能够学习特征图的局部特征,具有平移不变性.循环神经网络(Recurrent Neural Network,RNN)[5]可以建立音频各帧之间的关系,更好地学习声音事件的时间维度特征.卷积循环神经网络(Convolutional Recurrent Neural Network,CRNN)[6,7]则结合了CNN和RNN的优势,在声音事件检测上能够取得更好的效果.为了得到性能优良的深度学习声音事件检测模型,往往需要大量明确标注了事件类别和起止时间的训练样本.然而,声音事件样本的人工标注过程非常烦琐耗时且准确性不高,创建包含大量训练样本的数据集困难,制约了深度学习声音事件检测技术的发展.

为缓解训练数据与模型性能之间的矛盾,弱监督声音事件检测方法受到研究者的重视[8~10].弱监督声音事件检测指参与模型训练的样本仅给出事件类别标签而没有事件的起止时间.因此,弱监督声音事件检测也称为弱标签(weakly-labeled)声音事件检测.多实例学习(Multi-Instance Learning,MIL)[10]是当前解决弱标签声音事件检测问题的主流方法.该类方法将音频样本看作一个"包",将音频的每帧看作"包"中的一个实例,利用池化函数将实例的预测结果整合为"包"的预测结果.多实例学习具体可分为实例级法和特征级法(嵌入级法)[11].实例级方法将提取的声音事件特征送入分类器得到帧级预测标签,然后通过池化函数将帧级预测标签整合为弱标签预测.而特征级方法先将声音事件特征通过池化函数整合为音频级特征,再经过分类器得到弱标签预测.本文使用多实例学习的实例级方法对弱标签声音事件检测开展研究.

当前,基于深度神经网络的多实例弱标签声音事件检测方法研究主要集中在两个方面:一是强化深度神经网络对弱标签样本中声音事件特征的表征能力,二是设计有效的弱标签样本音频帧预测结果整合机制.在弱标签样本声音事件特征表征方面,文献[12]利用可学习门控线性单元(Gated Linear Unit,GLU)替换了CRNN网络中所有卷积层后的ReLU激活函数,实现了对声音事件检测弱标签样本时频谱图特征的选择.利用GLU单元控制谱图信息向更深卷积层的传递可看作是一种谱图的局部特征注意机制.文献[13]通过在GLU单元后增加特征选择结构实现对特征图的通道选择,并将时频谱图的局部和全局注意机制有机结合起来,增强了CRNN网络的特征表达能力.文献[14]则使用Transformer结构替换CRNN的中的RNN部分,解决模型无法并行训练的问题.在弱标签样本音频帧预测结果整合方面,将音频帧预测结果进行池化是最直接的处理方式.常用的池化函数有最大值池化[15]、平均池化[16]、Linear softmax池化[17]以及基于注意力机制的注意力池化[18]等.由于可以通过训练学习音频帧的权值,基于注意力机制的池化函数在深度神经网络声音事件检测中最受重视.然而在实际使用中注意力机制的学习效果并不理想.

从上述基于深度神经网络的多实例弱标签声音事件检测的研究可以看出,合理地设计特征提取网络中的注意力机制,以及改善池化函数中的注意力效果,是提升弱标签声音事件检测性能的两种有效途径.本文以CRNN多实例声音事件检测网络为基础,提出空间-通道特征表征与自注意池化的弱标签声音事件检测方法.该方法使用上下文门控单元(Context Gating,CG)[19]和Squeeze-and-Excitation(SE)[20]为CRNN引入注意力机制,使其具备选择重要的空间特征与通道特征的能力.然后利用自注意思想[21,22]设计一种自注意池化函数,在整合阶段为存在声音事件的音频帧赋予更大的权值.本文的主要贡献如下:

(1)本文将CG单元和SE模块与卷积层相结合构建一种门控注意力(Gating Attention,GA)结构.在门控注意力结构中,CG单元为每个通道的特征图生成对应的权重图,能够选择重要的空间特征,使模型关注更可能为声音事件的区域.SE模块则为所有特征通道建立联系,通过计算每个通道的权重值来衡量不同通道之间的重要程度,从而使网络选择更有用的特征通道,实现通道注意力机制.因此门控注意力结构通过对特征图中重要的空间特征和通道特征进行选择,使得网络能够提取有意义的事件特征,提高模型对声音事件的分类能力.

(2)本文结合自注意机制设计出一种自注意池化(Self Attention Pooling,SAP)函数用于音频帧预测结果的整合.自注意池化函数可以建立所有音频帧之间的关联,使得存在事件的音频帧之间的关联更加紧密,同时增大事件与背景帧之间的不相关度.此外,本文提出

的自注意池化函数使用多层感知机替代原本的缩放点积的计算形式,使自注意计算更符合弱标签声音事件检测的特点.

为了验证本文所提方法的有效性,本文在DCASE 2017任务4和DCASE 2018任务4数据集上进行实验.实验结果表明,本文方法能够显著提升弱标签声音事件检测的效果.

本文的其他各部分内容安排如下:第二部分介绍多实例深度声音事件检测网络;第三部分介绍声音事件的空间-通道特征表征;第四部分介绍弱标签声音事件检测的自注意池化机制;第五部分展示了本文的实验结果;第六部分总结了本文的工作内容.

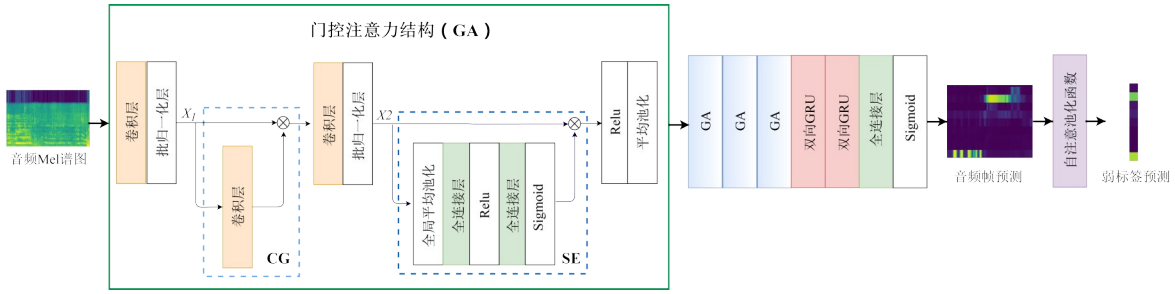
2　多实例深度声音事件检测网络

在弱标签声音事件检测中,多实例深度神经网络学习是一种代表性方法.一方面,深度神经网络可以有效地表征音频样本中的事件特征;另一方面多实例学习可充分整合音频样本各帧的预测结果,获得合理的弱标签预测.

一个通用的多实例深度神经网络弱标签声音事件检测框架如图1(*a*)所示,其包含特征表征和分类池化两部分.特征表征旨在对输入的音频样本(一般为样本的Mel谱图)进行特征表达,以利于鉴别样本中各音频帧的事件类别.CRNN网络是弱标签声音事件检测最主流的特征表征网络.由于级联了CNN和RNN,CRNN不仅能够学习音频样本谱图的时频特征,还可以建立样本音频帧之间的长时间关联关系.分类池化则利用全连接层对各音频帧进行分类,并通过对音频帧预测结果的池化获得音频样本的弱标签预测结果.

本文基于上述框架构建了如图1(*b*)所示的弱标签声音事件检测模型.其中CRNN的卷积部分由4个门控注意力(Gating Attention,GA)结构堆叠而成,实现声音谱图的空间特征和通道特征选择.循环神经网络部分由两层双向门控循环单元(Gated Recurrent Unit,GRU)[23]组成,用于建立声音谱图的帧间关联.本文模型使用自注意池化函数突出存在声音事件的音频帧,为其赋予更大的权值,实现弱标签音频样本帧级预测结果的整合.

(*a*)　弱标签声音事件检测框架



(*b*)　本文基于CRNN的弱标签声音事件检测模型

图1　弱标签声音事件检测框架和模型

3　声音事件的空间—通道特征表征

在弱标签声音事件检测中,音频通常包含多种重叠的声音事件,并且不同声音事件在时间和频率维度分布差异较大,再加上音频背景噪声的影响,极大地增加了声音事件检测的难度.为进一步增强模型对声音事件的表征能力,本文设计门控注意力结构替换原有的CNN部分.神经网络模型对音频中声音事件的表征能力在很大程度上影响着事件分类的准确性.而注意力机制可以使模型更加关注特征图中的重要部分,从而提取更多对分类有意义的特征.像GLU,CG和SE等注意力单元已经被应用在弱标签声音事件检测中,且成功提高了模型对声音事件的检测效果.因此,本文使用卷积层、CG和SE单元设计一种门控注意力结构,通过为CRNN引入空间和通道注意力机制增强模型对声音事件重要的空间特征和通道特征的选择能力.

从图1(*b*)可以看到,GA中第一个卷积层的输出首先通过CG单元实现空间特征选择.CG单元作为一种门控机制,通过计算出的权重值控制特征图中的每个元素能否向后续层传递.设CG单元的输入$\boldsymbol{X}_1\in\Re^{C\times H\times W}$,其中,$C$为特征图的通道数,$H$和$W$分别为特征图的高和宽.CG单元对输入特征图进行如下变换:

$$\boldsymbol{X}'_1=\sigma(\boldsymbol{W}*\boldsymbol{X}_1+b)\cdot\boldsymbol{X}_1 \tag{1}$$

其中，$\boldsymbol{W}$和b为卷积层的可学习参数；σ为Sigmoid激活函数，$*$表示卷积运算，$\cdot$表示逐元素相乘；$\boldsymbol{X'}_1$为CG单元的输出. 由式(1)可知，CG单元会为每个通道上的特征图生成对应的权重图，本质上是一种注意力机制. CG单元将计算出的注意力权重作用到输入特征图，实现对空间特征的选择，从而使卷积网络更加关注弱标签声音事件样本的局部，突出可能为声音事件的特征区域.

GA中的卷积层在提取声音事件特征的同时会增加特征通道的数量，但是对于不同的声音事件，并不是所有的特征通道都同等重要. 所以为了衡量特征图各通道的重要程度，本文将SE模块嵌入到第二个卷积层后，建立卷积网络的通道注意力机制. 设SE模块的输入为$\boldsymbol{X}_2 \in \Re^{C\times H\times W}$，$\boldsymbol{x}_2$为单通道的特征图. 先对特征图进行通道内的全局池化，即

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_2(i,j) \tag{2}$$

可得全局池化向量$\boldsymbol{z}=\{z_0,z_1,\cdots,z_c,\cdots,z_C\}$；然后使用两层连接层建立通道间的依赖关系，以获得表征各通道重要程度的权值向量$\boldsymbol{s}$，即

$$\boldsymbol{s}=\sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1\boldsymbol{z}\right)\right) \tag{3}$$

式(3)中，$\boldsymbol{W}_1 \in \Re^{(C/r)\times C}$和$\boldsymbol{W}_2 \in \Re^{C\times(C/r)}$表示全连接层的权重参数，$r$代表通道缩放率，$\sigma$和$\delta$分别表示Sigmoid函数和ReLU激活函数. 最后，特征图$\boldsymbol{X}_2$的各通道分别与对应权值相乘，实现基于注意力的通道选择，即

$$\boldsymbol{X'}_2=\boldsymbol{s}\cdot\boldsymbol{X}_2 \tag{4}$$

式(4)中，$\boldsymbol{X'}_2$表示SE模块的输出.

GA结构利用注意力机制选择特征图的空间特征和通道特征，增强模型对声音事件的分类能力. 此外，GA使用ReLU作为激活单元，并采用平均池化层对输出特征图进行降采样，以降低特征维度. 之后特征图被送入双向GRU层建立音频帧之间的关联，得到声音事件特征$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}$，$T$代表音频帧数.

4 弱标签声音事件检测的自注意池化

在弱标签声音事件检测的多实例学习框架中，池化函数的设计至关重要. 声音事件特征$\boldsymbol{X}$经过分类器得到帧级预测结果$\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T\}$. 池化函数的本质是提供权重$\boldsymbol{W}=\{w_1,w_2,\cdots,w_T\}$对$\boldsymbol{Y}$进行加权求和得到弱标签预测$\boldsymbol{y}$，且尽可能为存在声音事件的帧分配最大的权值，为背景帧分配最小的权值. 表1列举了常用的池化函数，其中$\mathcal{F}$表示全连接映射，t代表音频的第t帧. 最大值池化和平均池化是两种经典的池化函数. 最大值池化函数只选择声音事件出现概率最大的一帧，平均池化函数则为所有音频帧赋予相同的权重. 然而声音事件可能只出现在音频的一部分时间段内，所以这两种方式无法突出部分存在声音事件的音频帧. Linear softmax和注意力池化函数可以动态地实现权重变化，为存在事件的帧赋予较大权重，不存在事件的背景帧赋予较小的权重，但是在具体应用时难以保证权重的合理分配.

表1 四种常用池化函数

池化函数	定义	权重
最大值池化	$\boldsymbol{y}=\max\limits_t \boldsymbol{y}_t$	$w_t=\begin{cases}1, & \max\limits_t \boldsymbol{y}_t\\ 0, & \text{other}\end{cases}$
平均池化	$\boldsymbol{y}=\frac{1}{T}\sum\limits_{t=1}^{T}\boldsymbol{y}_t$	$w_t=1/T$
Linear softmax	$\boldsymbol{y}=\sum\boldsymbol{y}_t^2/\sum\boldsymbol{y}_t$	$w_t=y_t$
注意力池化	$\boldsymbol{y}=\sum w_t\boldsymbol{y}_t/\sum w_t$	$w_t=\mathcal{F}(\boldsymbol{X})$

为了在池化阶段使模型为事件帧分配大的权值，同时为背景帧分配小的权值，本文给出一种自注意池化函数. 该池化函数首先生成查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$，通过$\boldsymbol{Q}$和$\boldsymbol{K}$的相关性计算来构建各音频帧之间的关联，使得事件帧之间的关联性更加紧密，同时增大事件帧和背景帧之间的不相关度. 然后将计算出的权重矩阵作用回$\boldsymbol{V}$，得到自注意机制的输出$\boldsymbol{V'}$. 最后利用$\boldsymbol{V'}$生成注意力权重对帧级预测结果加权求和. 该过程通过模型训练可以动态地为音频帧生成合理的权重值. 标准的自注意机制一般使用缩放点积的计算形式，即

$$\boldsymbol{V'}=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

然而在矩阵相乘的过程中，由于相加求和运算的存在，那些含有较多相关帧的音频帧输出的值偏大，而那些含有较少相关帧的音频帧输出的值偏小. 对于持续时间很短的声音事件来说，其背景帧之间的相关数量远多于事件帧之间的相关数量，所以该方法不利于短事件的检测. 为此，我们使用如图2所示的多层感知机进行自注意权重的计算，即

$$\boldsymbol{V'}=\mathcal{B}\left(\boldsymbol{W}_{s2}\tanh\left(\boldsymbol{W}_{s1}[\boldsymbol{Q},\boldsymbol{K}]\right)\right)\cdot\boldsymbol{V} \tag{6}$$

式(6)中，$\boldsymbol{W}_{s1}$和$\boldsymbol{W}_{s2}$为全连接层参数，$\mathcal{B}$为边界限定函数，使$\boldsymbol{V'}\in[0,1]$. 从公式可以看出，$\boldsymbol{Q}$和$\boldsymbol{K}$被拼接在一起作为输入送入多层感知机中进行计算，然后生成与$\boldsymbol{V}$同样大小的权重矩阵并与其进行逐元素相乘，得到$\boldsymbol{V'}$. 这种计算方式利用神经网络灵活地为每帧学习合适的权重，避免池化函数忽略短事件的情况出现.

最终自注意池化函数形式如下式所示：

$$\boldsymbol{y}=\sum_t w_t\boldsymbol{y}_t \tag{7}$$

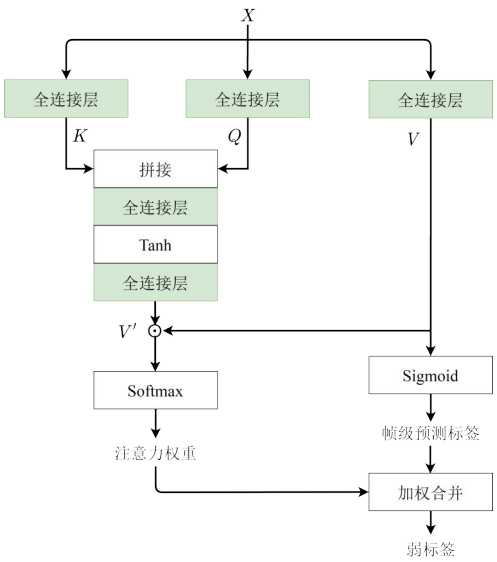


图2　自注意池化函数示意图

$$w_t=\frac{\exp\left(\boldsymbol{v}'_t\right)}{\sum_t \exp\left(\boldsymbol{v}'_t\right)} \tag{8}$$

式(8)中,$\boldsymbol{v}'_t$为$\boldsymbol{V}'$在第t帧的值. 自注意池化函数使用多层感知机灵活地学习每帧的权重,克服了缩放点积不利于短声音事件检测的缺陷. 自注意加权后的$\boldsymbol{V}'$使存在事件的音频帧之间相关度更高,作为注意力权重对帧级预测结果进行加权求和,进一步增强池化函数的整合效果.

5　实验与分析

为了验证本文所提方法的性能,我们将进行一系列实验,探究所提方法对声音事件检测结果的影响,证明其在弱标签声音事件检测中的有效性.

5.1　数据集与预处理

本文采用DCASE 2017任务4和DCASE 2018任务4数据集进行实验. 两个数据集都包含训练集、测试集和评估集. DCASE 2017数据集中的声音事件主要来自于汽车交通和警报两个大类,共17种声音事件类别. 训练集包含51 172个音频样本,且事件类别数量不平衡. 测试集包含488个音频样本,每种事件至少存在于30个音频样本中. 评估集包含1 103个音频样本,每种事件至少存在于60个音频样本中. 训练集只有音频的弱标签,测试集和评估集既有音频的弱标签又有强标签. 数据集中的音频持续时间均不超过10 s. DCASE 2018的训练集包含弱标签和无标签两类数据,本文仅使用弱标签数据进行模型训练. DCASE 2018数据集包含10种家庭环境中的声音事件:Speech, Dog, Cat, Alarm/Bell ringing, Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver/toothbrush. 训练集包含1 578个音频样本,测试集和评估集分别包含288和880个音频样本,同样的数据集中的音频持续时间均不超过10 s. 本文所有实验均使用只包含弱标签的训练集进行模型训练. 对于DCASE2017数据集,由于其事件类别严重不平衡,我们采取措施保证每批次训练数据的类别数量相同. 模型训练完成后使用包含强标签的测试集测试模型性能,同时给出模型在评估集上的结果.

模型训练时采用音频样本的对数梅尔谱图作为输入. 本文使用窗长为2 048的汉明窗对频率为44.1 KHz的音频样本进行分帧,帧移为511;然后进行快速傅里叶变换,再经过64个通道的梅尔滤波器组并对结果取对数得到大小为864×64的对数梅尔谱图.

5.2　实验过程和参数设置

本文的弱标签声音事件检测模型中4个GA结构的参数设置如表2所示,其中c表示输出通道数,s表示池化步长,r表示SE单元的通道缩放率,代号的下标表示处于第几个GA结构. 模型中两层双向GRU层输出维度均为128. 最后特征图被送入分类器中得到$\boldsymbol{V}$,其维度与类别数相同. 自注意池化函数中生成的$\boldsymbol{K}$和$\boldsymbol{Q}$的维度为64.

表2　GA结构参数设置

模块名称	参数设置	
GA	卷积层-1	c_1:32, c_2:64, c_3:64, c_4:128　k_1:3, k_2:3, k_3:3, k_4:3
	CG	c_1:32, c_2:64, c_3:64, c_4:128　k_1:1, k_2:1, k_3:1, k_4:4
	卷积层-2	c_1:32, c_2:64, c_3:64, c_4:128　k_1:3, k_2:3, k_3:3, k_4:3
	SE	c_1:32, c_2:64, c_3:64, c_4:128　r_1:4, r_2:4, r_3:4, r_4:4
	平均池化	s_1:(2, 2), s_2:(2, 2), s_3:(1, 2), s_4:(1, 4)

模型训练阶段采用Adam优化器,学习率为0.000 5,每批次训练的数据量为32,在DCASE2018数据集上的训练周期为150次,在DCASE2017数据集上的训练周期为100次. 训练过程使用二元交叉熵作为损失函数,其表达式如下式所示:

$$L=-\sum_{n=1}^{N}\left(\boldsymbol{y}_n\log\boldsymbol{p}_n+(1-\boldsymbol{y}_n)\log(1-\boldsymbol{p}_n)\right) \tag{9}$$

式(9)中,N表示样本数量,$\boldsymbol{y}_n$和$\boldsymbol{p}_n$分别表示第n个样本弱标签的真实值和预测值. 本文在后处理阶段采用文献[24]中的双阈值法对帧级预测结果进行处理,得到声音事件的时间预测. 在DCASE 2018数据集中,双阈值被设置为0.3和0.1,在DCASE2017数据集中,双阈值则被设置为0.5和0.3.

为了验证模型中门控注意力结构和自注意池化函数起到的效果,实验采用比赛官方提供的数据划分原则和评判标准[25],确保能公平地验证和比较本文模型与其他模型的性能. 本文采用$F1$得分评估模型在声音事件检测中的性能,其计算方式为

$$F1=\frac{2\times P\times R}{P+R}=\frac{2\times \mathrm{TP}}{2\times \mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{10}$$

式(10)中,FP表示误检的声音事件数,FN表示漏检的声音事件数,TP表示正确检测的声音事件数;P和R分别表示声音事件检测的精度和召回率. 根据比赛提供的标准,DCASE 2018采用基于事件的$F1$得分,DCASE2017采用基于1 s音频段的$F1$得分.

5.3 结果与分析

本文首先在DCASE 2018数据集上进行消融实验,分别验证门控注意力结构和自注意池化函数对弱标签声音事件检测的效果. 然后本文在DCASE 2018和DCASE 2017数据集上进行实验,给出本文所提方法在弱标签声音事件检测中的结果,并与其他研究结果进行对比. 最后本文给出不同池化函数在长声音事件和短声音事件上的检测结果,并进一步分析自注意池化函数的优势和特点. 本文列出的实验结果均为三次重复实验条件下单模型的最佳结果.

5.3.1 模型的消融实验和分析

为了验证本文提出的门控注意力结构和自注意池化函数的效果,本文分别在DCASE 2018数据集和DCASE 2017数据集上进行了弱标签声音事件检测的消融实验,结果如表3所示. 表中的第一组和第三组实验验证单独使用门控注意力结构的有效性,第一组和第二组验证单独使用自注意池化函数的有效性. 第三组至第七组实验验证本文在使用门控注意力结构下的自注意池化函数的有效性. 从DCASE 2018数据集中的前两组实验结果可以看出,CRNN-GA-Linear的$F1$得分在测试集与评估集上分别比CRNN-Linear高出15.75%和8.76%. DCASE 2017数据集中也表现出类似的结果,即模型使用门控注意力结构可以提高$F1$得分. 这表明门控注意力结构能够选择特征图中重要的空间特征和通道特征,从而提取到更多对分类有用的声音事件特征,提升模型的声音事件检测效果. 对比DCASE 2018和DCASE 2017数据集中的第一组和第三组实验结果可以看出,使用自注意池化函数的模型的检测效果无论是在测试集上还是在评估集上均优于Linear softmax池化函数. 这表明相比于Linear softmax池化,自注意池化函数更有利于提高模型的检测性能.

从表3中还可以看出,模型使用不同的池化函数时会有不同的检测性能. 以DCASE 2018数据集为例,最大值池化和平均池化在测试集上的$F1$得分为23.85%和17.62%,比Linear softmax池化分别低15.15%和23.18%,在评估集上分别低11.68%和15.05%. 这证明了最大值池化与平均池化做法的不合理,即很多声音事件不会只出现在音频的某一帧上,也不会贯穿整个音频,而是仅出现在一段区间内. 比较自注意池化函数、注意力池化函数以及Linear softmax池化函数的结果可以发现,本文的自注意池化在测试集的$F1$得分比Linear softmax池化低0.35%,比注意力池化高7.45%. 在评估集上比Linear softmax池化高0.93%,比注意力池化高6.8%. 这表明自注意池化函数在声音事件检测上的性能与Linear softmax池化函数不相上下,且优于注意力池化函数. 同时也表明自注意机制引入有利于提高池化函数的性能. 分析DCASE 2017数据集的结果同样可以发现,使用自注意池化函数在测试集与评估集上的检测效果均优于其他池化函数.

表3 模型在DCSAE 2018和DCASE 2017数据集上的消融实验结果

模型	DCASE2018		DCASE2017	
	测试集	评估集	测试集	评估集
CRNN-Linear	23.25%	21.31%	48.57%	48.23%
CRNN-SAP	29.98%	25.81%	49.49%	48.46%
CRNN-GA-Linear	39.00%	30.07%	50.00%	50.86%
CRNN-GA-Max	23.85%	18.39%	35.17%	35.96%
CRNN-GA-Average	17.62%	15.02%	48.56%	49.18%
CRNN-GA-Attention	31.20%	24.20%	50.66%	50.41%
CRNN-GA-SAP	38.65%	31.00%	52.91%	52.47%

最后,从消融实验的结果来看,本文提出的门控注意力结构和自注意池化函数均能提高弱标签声音事件检测模型的性能.

5.3.2 模型与其他方法的实验结果对比

为了能客观评价所提方法的效果,本文在DCASE 2018和DCASE 2017数据集上进行实验,与当前在弱标签声音事件检测任务中取得优异结果的其他方法进行对比. 遵循DCASE竞赛官方标准,DCASE 2018和DCASE 2017数据集的评判指标分别为基于事件和基于1秒音频段的$F1$得分.

表4对比了不同方法在DCASE 2018和DCASE 2017数据集上的实验结果. 文献[26]中的模型为DCASE 2018官方提供的基线模型,该模型使用标准的CRNN(包含3层卷积层和1层GRU层)训练数据,整体模型较为简单,并未引入其他功能模块. 文献[24]中的模型也是CRNN,包含5层卷积层和1层双向GRU层,并采用双阈值后处理法. 文献[27]采用同文献[26]一致的CRNN模型训练弱标签数据,主要区别在于前者使用带有余弦相似度惩罚项的损失函数用于提高不同事件的区分度. 文献[29]为DCASE 2017官方提供的基线模型,该模型为仅由2层全连接层组成的前馈神经网络(FNN). 文献[17]采用CRNN模型,其中包含十层卷积层和1层双向GRU层. 文献[12]的CRNN模型包含4个带有GLU单元的卷积块和2层双向GRU,每个卷积块含有2层卷积层. 文献[13]则在文献[12]中卷积块的基础上构建了特征选择结构,并在RNN部分之前使用

表4　DCSAE 2018和DCASE 2017数据集上不同方法的声音事件检测$F1$得分

模型	DCASE 2018		DCASE 2017	
	测试集	评估集	测试集	评估集
DCASE 2018 Baseline [26]	14.06%	10.80%	14.41%	17.23%
Dinkel [24]	36.00%	30.10%	47.05%	45.57%
Pellegrini [27]	34.75%	26.20%	41.23%	39.64%
DCASE 2017 Baseline [29]	0.19 %	0.28 %	13.80 %	21.31 %
Xu [12]	29.59%	20.46%	47.20 %	47.50 %
Wang [17]	28.09%	25.62%	46.80%	/
Yan [13]	31.26%	28.57%	51.30 %	55.10 %
Kao [28]	42.0%	29.5%	49.9%	49.4%
CRNN-GA-SAP	38.65%	31.00%	52.91%	52.47%

基于区域的注意力方法提取声音事件不同区域特征，而本文模型并未使用这一做法. 文献[28]使用改进后的Densenet作为特征提取器，并利用全局平均池化层预测帧级标签.

对于DCASE 2018数据集，从表4中可以看出，CRNN-GA-SAP在测试集和评估集上分别取得38.65%和31.00%的$F1$得分. 与官方基线模型相比，本文方法在测试集上高出24.59%，在评估集上高出20.2%，这表明引入注意力机制，适当增加CNN和RNN的层数有利于提高模型的检测性能. 另外，与文献[24]相比，本文方法在测试集和评估集上分别高出2.65%和0.9%，与文献[27]相比，本文方法在测试集和评估集上分别高出3.9%和4.8%. 与文献[28]相比，本文方法在测试集上$F1$得分低，但在评估集上$F1$得分高，整体表现结果不相上下. 这表明本文提出的声音事件空间-通道特征表征和自注意池化函数确实可以改善弱标签声音事件的检测效果.

对于DCASE 2017数据集，从表4中可以看出，CRNN-GA-SAP无论是在测试集上还是在评估集上的$F1$得分均远高于DCASE 2017官方基线模型. 这表明在弱标签声音事件检测中，CRNN对声音事件的特征表征能力要强于FNN. CRNN-GA-SAP在测试集上取得了52.91%的$F1$得分，高于文献[12]、文献[17]、文献[28]和文献[13]等方法. 在评估集中，CRNN-GA-SAP的$F1$得分为52.47%，仅低于文献[13]的55.10%，我们认为这可能是文献[13]中基于区域的注意力方法导致的. 从测试集和评估集的整体结果来看，本文方法对弱标签声音事件检测任务具有显著的性能改善.

从上述对比情况来看，本文方法在两个数据集中均能取得较大的优势，说明本文方法可以有效提升弱标签声音事件检测的性能，且对数据集没有依赖性，适应性较强.

5.3.3　池化函数对长短声音事件的影响

由于不同类别声音事件在音频中持续时间有长有短，这使得各声音事件检测的难度不同. 对不同声音事件的检测结果进行总结和分析，有利于更好地探究各种模型的检测效果和特点. 本文借助DCASE 2018数据集，重点探究Linear softmax、注意力和本文提出的自注意3种池化函数对不同持续时间声音事件的检测效果. 本文对DCASE 2018的评估集进行统计分析，根据每类的平均持续时间将10类声音事件划分为长事件和短事件. 其中长事件有Vacuum cleaner(8.37 s)，Running water(5.07 s)，Frying(7.89 s)，Electric shaver(7.97 s)和Blender(4.97 s)五种，短事件有Speech(1.49 s)，Dog(1.50 s)，Dishes(0.64 s)，Cat(1.60 s)和Alarm/Bell ringing(2.11 s)五种.

表5给出了使用3种池化函数的检测模型在10种声音事件上的检测结果，并且给出了长事件和短事件的检测结果. 从表中可以看出，Linear softmax池化函数对长事件的检测效果好，$F1$得分为38.04%，但是不利于短事件的检测，$F1$只有22.40%. 而注意力池化函数不管是在长事件还是在短事件上的检测效果都不理想. 自注意池化函数在长事件和短事件中的$F1$得

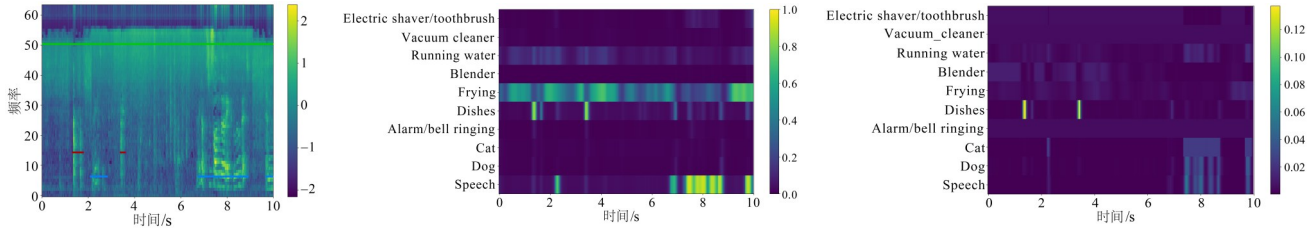
表5　不同模型对DCASE2018评估集中各类声音事件检测的$F1$得分

声音事件	CRNN-GA-Linear	CRNN-GA-Attention	CRNN-GA-SAP
长事件	38.04%	25.40%	28.66%
短事件	22.24%	22.60%	33.36%
Speech	10.90%	47.10%	39.10%
Dog	24.30%	14.70%	19.20%
Dishes	0.00%	4.30%	22.50%
Cat	27.20%	3.40%	34.70%
Alarm/Bell ringing	48.80%	43.50%	51.30%
Vacuum cleaner	63.30%	24.40%	15.30%
Running water	31.90%	26.50%	31.30%
Frying	41.90%	36.90%	48.20%
Electric shaver/toothbrush	23.50%	20.00%	18.00%
Blender	29.60%	19.20%	30.50%

分分别为28.66%和33.36%，虽然长事件的$F1$得分没有Linear softmax高，但是大大提高了短事件的检测效果．而且自注意池化函数对长短事件的检测效果较为均衡，对声音事件检测的鲁棒性要高于其他2种池化函数．

对表中给出的数据进一步观察可以发现，Linear softmax池化函数和注意力池化函数无法正确检测Dishes事件，其$F1$得分几乎为0．自注意池化函数反而有较好的表现结果．为了深入分析这这一现象，本文使用评估集中的一个音频样本进行测试，给出不同池化函数输出的可视化结果．该音频样本包含Speech，Dishes和Frying 3类声音事件．

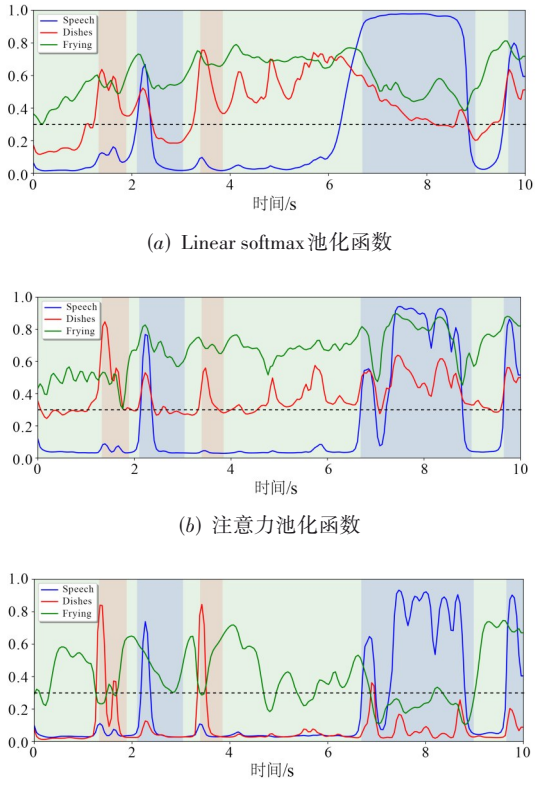
图3为自注意池化函数在该音频样本上的输出结果图．从图3(b)以看出，自注意池化函数正确预测出音频中包含的3类声音事件，而且对声音事件发生时间的预测也大致符合图3(a)中事件特征的位置．从图3(c)可以发现，自注意机制准确地为音频中出现Dishes事件的位置生成大的权值，这表明自注意池化函数通过加强事件帧之间的关联程度可以提高短事件的检测效果．



(a) Mel谱图　(b) 帧级预测结果　(c) 注意力权重

图3　自注意池化函数的输出可视化图(蓝色、红色和绿色线段代表Speech,Dishes和Frying的发生位置)

图4为3种池化函数的输出预测曲线，其中蓝色、红色和绿色曲线分别代表Speech，Dishes和Frying的预测，黑色虚线表示阈值．从图中可以看出，对于Speech和Frying事件，3种池化函数可以大致预测其出现的位置．但是从图4(a)和图4(b)可以发现，对于Dishes事件，Linear softmax与注意力池化函数无法识别该事件，将大部分未出现该事件的区域标记为发生Dishes事件．这说明该2种池化函数容易将持续时间较短的Dishes事件与其他声音事件相混淆，无法正确识别和检测．而图4(c)中的自注意池化函数通过自注意机制增强事件帧之间的关联程度并拉大与背景帧的差异，从而突出声音事件，提高模型的检测效果．



(a) Linear softmax池化函数

(b) 注意力池化函数

(c) 自注意池化函数

图4　3种池化函数的预测曲线图(蓝色、红色和绿色背景代表Speech、Dishes和Frying的真实发生时间)

综合模型在DCASE 2018和DCASE 2017数据集上的实验结果可以看出：本文提出的弱标签声音事件检测方法可以显著提高弱标签声音事件的检测效果，且$F1$得分优于绝大部分其他模型，并对持续时间不同的声音事件的检测具有较强的鲁棒性．

6　总结

本文创新性地设计了门控注意力结构和自注意池化函数，并将它们与多实例CRNN网络相结合构建弱标签声音事件检测方法．门控注意力结构实现了对特征图重要的空间特征和通道特征的选择；自注意池化函数通过利用自注意机制进一步加强事件帧的关联性，拉大事件帧与背景帧间的差异，从而为存在事件的音频帧赋予更大的权重．在DCASE 2017任务4和DCASE 2018任务4的数据集上的实验结果显示：本文提出的方法能够显著提升真实环境中弱标签声音事件

检测的综合性能,是目前单模型性能最佳的弱标签声音事件检测方法. 从实验中也发现,自注意池化函数对于长事件和短事件均具有较好的检测效果.

参考文献

[1] GIANNOULIS D, BENETOS E, et al. Detection and classification of acoustic scenes and events: An IEEE AASP challenge[C]//IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz: IEEE, 2013: 1-4.

[2] VALENZISE G, GEROSA L, et al. Scream and gunshot detection and localization for audio-surveillance systems [C]//2007 IEEE Conference on Advanced Video and Signal Based Surveillance. London: IEEE, 2007: 21-26.

[3] FOGGIA P, PETKOV N, SAGGESE A, et al. Reliable detection of audio events in highly noisy environments[J]. Pattern Recognition Letters, 2015, 65(C): 22-28.

[4] ZHANG H, MCLOUGHLIN I, SONG Y. Robust sound event recognition using convolutional neural networks[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Brisbane: IEEE, 2015: 559-563.

[5] PARASCANDOLO G, HUTTUNEN H, VIRTANEN T. Recurrent neural networks for polyphonic sound event detection in real life recordings[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai: IEEE, 2016: 6440-6444.

[6] 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.

[7] 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 W H, HU S D, SHI Y L, et al. A convolutional gated recurrent network for speech enhancement[J]. Acta Electronica Sinica, 2020, 48(7): 1276-1283. (in Chinese)

[8] CHOU S Y, YANG Y H, et al. Framecnn: A weakly supervised learning framework for frame-wise acoustic event detection and classification[DB/OL]. (2017) [2021]. https://www. semanticscholar. org/paper/fbe6d1324506755d901df17d6378d49713f15aea.

[9] KONG Q, XU Y, et al. Sound event detection and time-frequency segmentation from weakly labeled data[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27(4): 777-787.

[10] KUMAR A, RAJ B. Audio event detection using weakly labeled data[C]//Proceedings of the 24th ACM international Conference on Multimedia. Amsterdam: ACM, 2016: 1038-1047.

[11] LIN L W, WANG X D, et al. Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1466-1478.

[12] XU Y, KONG Q, Wang W, et al. Large-scale weakly supervised audio classification using gated convolutional neural network[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 121-125.

[13] YAN J, SONG Y, GUO W, et al. A region based attention method for weakly supervised sound event detection and classification[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 755-759.

[14] KONG Q, XU Y, WANG 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.

[15] SU T W, LIU J Y, YANG Y H. Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans: IEEE, 2017: 791-795.

[16] KUMAR A, KHADKEVICH M, FUGEN C. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 326-330.

[17] WANG Y, LI J, METZE F. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 31-35.

[18] KONG Q, XU Y, WANG W, et al. Audio set classification with attention model: A probabilistic perspective[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018: 316-320.

[19] 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.

[20] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(8): 2011-2023.

[21] LIN Z, FENG M, SANTOS C N, et al. A structured self-attentive sentence embedding[EB/OL]. (2017)[2021]. https://arxiv.org/abs/1703.03130.
[22] 张志昌, 曾扬扬, 庞雅丽. 融合语义角色和自注意力机制的中文文本蕴含识别[J]. 电子学报, 2020, 48(11): 2162-2169. ZHANG Z C, ZENG Y Y, PANG Y L. Chinese textual implication recognition combining semantic roles and self-attention mechanism[J]. Acta Electronica Sinica, 2020, 48(11): 2162-2169. (in Chinese)
[23] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [C]//International Conference on Machine Learning. Lille: ACM, 2015: 2342-2350.
[24] DINKEL H, YU K. Duration robust weakly supervised sound event detection[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 311-315.
[25] MESAROS A, HEITTOLA T, VIRTANEN T. Metrics for polyphonic sound event detection[J]. Applied Sciences, 2016, 6: 162.
[26] SERIZEL R, TURPAULT N, et al. Large-scale weakly labeled semi-supervised sound event detection in domestic environments[C]//Detection and Classification of Acoustic Scenes and Events. Surrey: IEEE, 2018: 19-23.
[27] PELLEGRINI T, CANCES L. Cosine-similarity penalty to discriminate sound classes in weakly-supervised sound event detection[J]. International Joint Conference on Neural Networks, 2019: 1-8.
[28] KAO C, SHI B, et al. A joint framework for audio tagging and weakly supervised acoustic event detection using DenseNet with global average pooling[C]//Interspeech. Shanghai: ACM, 2020: 846-850.
[29] MESAROS A, HEITTOLA T, DIMENT A, et al. DCASE 2017 challenge setup: Tasks, datasets and baseline system [C]//Detection and Classification of Acoustic Scenes and Events. Munich: IEEE, 2017: 85-92.

作者简介

杨利平 男, 1981 年生, 内蒙古鄂尔多斯人. 重庆大学副教授. 主要研究方向为机器学习, 模式识别, 以及图像、声音信号处理.
E-mail: yanglp@cqu.edu.cn

侯振威 男, 1996年生, 河北邢台人. 重庆大学硕士研究生. 主要研究方向为声音信号处理.