文章编号：1672-6987(2022)05-0121-05；**DOI**：10.16351/j.1672-6987.2022.05.016

基于自注意力路由胶囊网络的多音事件检测

李海涛，杨树国*

（青岛科技大学 数理学院，山东 青岛 266061）

摘　要：声音事件检测是目前计算机听觉领域中的重要问题，而多声音事件检测是其中一个极具挑战性的研究热点。基于最新提出的非迭代的自注意力路由方法和胶囊网络，本文提出了一种基于自注意力路由的多路径胶囊网络模型，将其用于多声音事件检测。由于自注意力路由方法是非迭代且高度并行的，大大加快了模型的训练速度；多路径基础胶囊层使用不同大小的非对称卷积核，不仅使模型能获得不同分辨率的信息，还能极大地保留时间信息，从而提高了模型的性能。本工作在 2017 年声音场景与事件检测分类挑战赛(Detection and Classification of Acoustic Scenes and Events，DCASE 2017) 挑战任务 4 数据集上对所提出的模型和方法进行了对比实验及性能评估。其中，音频标注子任务的 F 分数达到了 59.5%，音频事件检测的错误率降低到 0.72，检测效果有较大的提升。结果表明：本方法具有事件检测准确率高、速度快、泛化能力强等优点。

关键词：多声音事件检测；胶囊网络；DCASE 2017 挑战

中图分类号：TP 18　　**文献标志码**：A

引用格式：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.

Polyphonic Sound Event Detection Based on Self-Attention Routing Capsule Network

LI Haitao，YANG Shuguo

(College of Mathematics and Physics，Qingdao University of Science and Technology，Qingdao 266061，China)

Abstract：Sound event detection is currently an important issue in the field of computer hearing，and polyphonic sound event detection is one of the most challenging research hotspots. Based on the newly proposed non-iterative self-attention routing method and capsule network，this paper proposes a multi-path capsule network model based on self-attention routing，which is used for polyphonic event detection. Since the self-attention routing method is non-iterative and highly parallel，it greatly accelerates the training speed of the model；the multi-path primary capsule layer uses asymmetric convolution kernels of different sizes，which not only enables the model to obtain information of different resolutions，but also ex-

收稿日期：2021-09-08
基金项目：山东省自然科学基金项目(ZR2021QF040).
作者简介：李海涛(1997—)，男，硕士研究生.　*通信联系人.

tremely retains time information, thereby improving the performance of the model. This paper conducts comparative experiments and performance evaluation of the proposed models and methods on the data set of DCASE 2017 Task 4. The F score of the audio tagging subtask is 59.5%, and the error rate of the sound event detection is reduced to 0.72, which is a big improvement. The results show that the method in this paper has the advantages of high sound event detection accuracy, fast speed and strong generalization ability.

Key words: polyphonic sound event detection; capsule network; DCASE 2017 challenge

日常生活中,人们每天都会接触到很多不同的声音,如汽车的鸣笛声、孩子的叫喊声等等,这些声音中包含了丰富的信息,识别生活环境中发生的不同声音事件从而进行不同的处理是非常重要的。声音事件检测(sound events detection,SED)就是检测音频信号中不同的声音事件及其起止时间,为进一步分析和处理声音事件奠定基础。SED在音频监控[1]、城市声音分析[2]、设备监控[3]等诸多领域都有着广泛的应用。

一般来说,SED的任务大致分为两类:单音SED和多音SED。单音SED在任一时刻至多检测出一种声音事件,而多音SED系统可以检测出多个声音事件[4]。从用途上看,因为现实环境中包含多个声源的情况更加多见,所以多音SED应用更为广泛;不同的声音事件往往相互重叠,而从混叠的声音中提取出的特征可能与从单个声音中提取的任何特征都不匹配,导致无法提取出能够有效代表单个声音事件的特征[5],所以多音SED更加困难和复杂,也更具挑战性。

传统的多音事件检测的模型有隐马尔可夫模型[6]和高斯混合模型[7]等。近年来,数据集和计算资源可用性的提高推动了深度学习模型在声音事件检测和分类任务中的应用,包括前馈神经网络(FNN)[8]、卷积神经网络(CNNs)[9]和循环神经网络(RNNs)[10]等。基于CNN和RNN的方法在SED任务中取得了良好的性能,这得益于它们能够学习提取出的音频特征与目标向量之间的非线性关系。特别是在多音SED的情况下,CNN与RNN的结合(CRNN)具有CNN提供的局部位移不变性,并具有RNN层提供的短期和长期时间依赖进行建模的能力,两种体系结构的结合提高了检测性能和效果[4]。

2017年底,HINTON等[11]提出了胶囊网络的概念,它的引入是为了克服CNN的一些局限性,特别是最大池化造成的信息丢失。胶囊可以被认为是一组神经元,它们的输出代表同一实体的不同属性[11]。一层(低层)的胶囊通过变换矩阵对下一层(高层)的胶囊进行姿态预测,然后使用动态路由机制,通过迭代聚类的方法获得耦合系数,并将相关胶囊的信息传递给下一层。

基于胶囊的计算结构与路由机制相结合,胶囊网络可以识别数据特征之间的部分和整体关系,从而能够有效提高网络在重叠目标的检测任务上的表现[11]。从理论上讲,动态路由的引入可在不需要大量数据增强或专用域适应程序的情况下充分训练模型,能够极大地提高模型的泛化能力。文献[12]提出了用于多音事件检测任务的CapsNet,在网络的初始层应用了门控卷积层,并在最后的胶囊层中添加了并行的注意层。该算法在DCASE 2017任务4的弱标注数据集上进行了使用,取得了良好的性能。文献[13]将胶囊网络应用于多音事件检测中,并在三个公开的数据集上进行了评估,结果显示,基于CapsNet的算法不但优于CNN,而且也取得了良好的效果。

然而胶囊网络中的动态路由机制是通过迭代聚类的方法获得耦合系数,这使得网络的训练和推理过程变得缓慢。文献[14]用一种新的非迭代的、高度并行化的路由算法来代替动态路由,称为自注意力路由。本研究以文献[12]提出的CapsNet为基线系统,研究了自注意力路由算法以及多路径基础胶囊层结构对多音事件检测的影响,提出了自注意力路由和多路径基础胶囊层相结合的胶囊网络,并在DCASE 2017 task4数据集上对该模型进行评估。

1 模型与算法

1.1 胶囊网络

胶囊网络的概念是HINTON等[11]在2017年提出的,其主要思想是用向量神经元替代传统的标量神经元。胶囊是一种向量,它的维数与目标的各种性质有关,如位置、大小、方向等,其长度代表了目标的活动概率。胶囊网络主要包含了卷积层、基础胶囊层和数字胶囊层。其结构如图1所示。

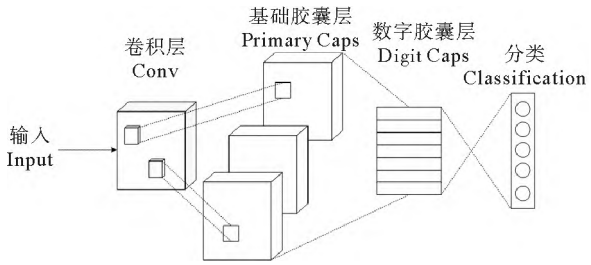

图 1　胶囊网络结构图

Fig.1　Capsule network structure

卷积层主要用来从输入中提取特征,其作用与卷积神经网络中的卷积层类似。低层胶囊通过动态路由机制来确定连接到高层胶囊的权重。动态路由算法的过程如图 2 所示。

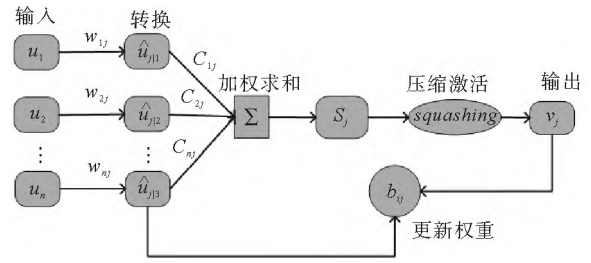

图 2　动态路由算法

Fig.2　Dynamic routing algorithm

假设低层胶囊为 i,高层胶囊为 j,则高层胶囊的输出 v_j 可由公式(1)～(3)计算得出:

$$\hat{u}_{j|i}=W_{ij}u_i\ , \tag{1}$$

$$s_j=\sum_i c_{ij}\hat{u}_{j|i}\ , \tag{2}$$

$$v_j=\frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|}\ 。 \tag{3}$$

其中 u_i 表示低层胶囊的输出,$\hat{u}_{j|i}$ 表示低层胶囊 i 对高层胶囊 j 的预测向量,W_{ij} 为相应的权重矩阵。将 v_j 的所有预测向量用一组耦合系数 c_{ij} 进行加权求和,并用一个非线性压缩函数(3)把向量的长度压缩在 0 到 1 之间,以表示目标存在的概率。耦合系数 c_{ij} 由动态路由算法确定:

$$c_{ij}=\frac{\exp(b_{ij})}{\sum_k\exp(b_{ik})}\ , \tag{4}$$

$$b_{ij}\leftarrow b_{ij}+\hat{u}_{j|i}\cdot v_j\ 。 \tag{5}$$

其中:$\hat{u}_{j|i}$ 和 v_j 之间的相似度越高(用内积表示),c_{ij} 就会越大。在每次正向传播中,b_{ij} 被初始化为 0,由方程(4)计算耦合系数 c_{ij} 的初始值,然后由网络的正向传播计算 v_j。b_{ij} 的值根据公式(5)进行更新,用于更新 c_{ij} 的值,并通过正向传播修正 s_j 的值,从而改变输出向量 v_j 的值,最后得到一组最优的耦合系数。

1.2　基于自注意力路由的多声音事件检测模型

为了提高胶囊网络的训练速度和推理速度,以及使模型充分利用原始特征中所包含的信息(尤其是时间信息),以进一步提高多声音事件检测的精度,本研究提出了基于自注意力路由的多声音事件检测模型(MpCaps-att)。该方法使用一种最近提出的非迭代且高度并行的自注意力路由算法和多路径基础胶囊层。

1.2.1　自注意力路由

自注意力路由是文献[14]提出的新型路由方法,具有非迭代且高度并行的特点,因此能大大加快网络的训练速度。自注意力路由过程如下:

首先,对于 l 层的胶囊 $u_n^l\in\mathrm{R}^{d^l}$(d^l 代表 l 层胶囊的维度),通过与权重矩阵相乘,获得对高层胶囊的预测向量,如公式(6)所示:

$$\hat{U}^l_{(n^l,n^{l+1},:)}=u_n^{Tl}\times W^l_{(n^l,n^{l+1},:,:)}\ 。 \tag{6}$$

其中,n^l 表示 l 层胶囊的数量,$W^l_{n^l,n^{l+1},d^l,d^{l+1}}$ 包含所有的权重矩阵,$\hat{U}^l_{n^l,n^{l+1},d^{l+1}}$ 包含所有 l 层胶囊的预测向量,则 $l+1$ 层胶囊 s_n^{l+1} 由公式(7)计算得出:

$$s_n^{l+1}=\hat{U}^{Tl}_{(:,n^{l+1},:)}\times(C^l_{(:,n^{l+1})}+B^l_{(:,n^{l+1})})\ , \tag{7}$$

其中,$B^l_{n^l,n^{l+1}}$ 是包含所有权重的对数先验矩阵,$C^l_{n^l,n^{l+1}}$ 是包含自注意力算法产生的所有耦合系数的矩阵。耦合系数通过自注意力张量 $A^l_{n^l,n^l,n^{l+1}}$ 计算,自注意力张量的计算公式:

$$A^l_{(:,:,n^{l+1})}=\frac{\hat{U}^l_{(:,n^{l+1},:)}\times U^{Tl}_{(:,n^{l+1},:)}}{\sqrt{d^l}}\ 。 \tag{8}$$

对于上层的每个胶囊 n^{l+1},都含有一个对称矩阵 $A^l_{:,:,n^{l+1}}$。耦合系数可通过公式(9)计算得出:

$$C^l_{(:,n^{l+1})}=\frac{\exp\left(\sum_{n^l}A^l_{(:,n^l,n^{l+1})}\right)}{\sum_{n^{l+1}}\exp\left(\sum_{n^l}A^l_{(:,n^l,n^{l+1})}\right)}\ 。 \tag{9}$$

最后将 $l+1$ 层胶囊的输出 s_n^{l+1} 代入到压缩函数中,将向量的长度压缩到 0 到 1 之间,以表示特定目标存在的概率,文献[14]中使用的压缩函数为

$$\mathrm{squash}(s_n^{l+1})=\left(1-\frac{1}{\mathrm{e}^{s_n^{l+1}}}\right)\frac{s_n^{l+1}}{\|s_n^{l+1}\|}\ 。 \tag{10}$$

1.2.2　多路径基础胶囊层

在声音事件检测任务中,时域信息的重要性比频域信息要高,所以应尽可能多的保留时间信息[4-5,15-16]。因此本研究提出了一种多路径基础胶囊层,如图 3 所示。

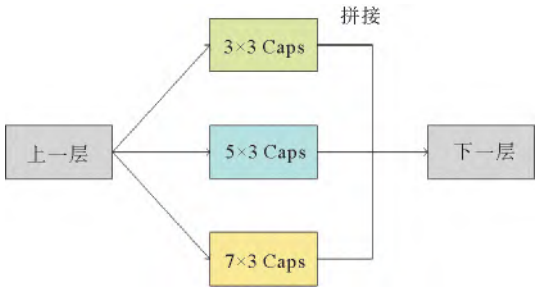

图3 多路径基础胶囊层

Fig.3 Multipath primary capsule layer

该结构由三层基础胶囊层组成,且三层基础胶囊层具有不同大小的卷积核。其中两层的卷积核尺寸为非对称的,且在时域上具有更大的卷积尺寸。其中,三层基础胶囊层的卷积核大小分别为(3, 3),(5, 3),(7, 3)。之后将三层基础胶囊层的输出进行拼接,送入高级胶囊层。一般来说,卷积核越大,获得的信息就越多,提取的特征就会更好。因此,在其中的两层基础胶囊层中,使用时域上尺寸更大的非对称卷积核,来获取更多的时间信息。不同的卷积核大小会提取出不同的特征,所以选择不同的卷积核大小就能获得不同分辨率的信息,使得模型能够充分利用特征信息。

1.2.3 基于自注意力路由的胶囊网络模型

本节提出了基于自注意力路由的胶囊网络模型,并用其进行多音事件检测,该模型包括卷积层、胶囊层和全连接层,如图4所示。

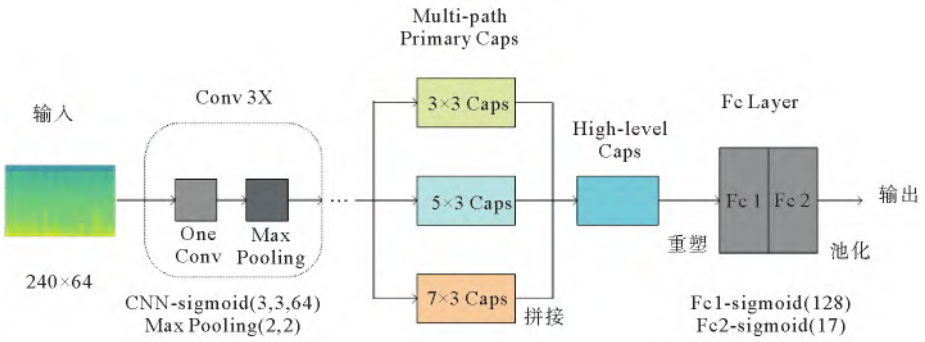

图4 基于自注意力路由的胶囊网络结构

Fig.4 Capsule network structure based on self-attention routing

图4中,模型的输入是对数Mel语谱图,是通过将每段音频进行重采样并进行短时傅里叶变换,然后和Mel滤波器组相乘并进行对数运算得出。3层卷积层用来从输入中提取局部特征,并使用最大池化来缩减时域和频域的维度。假设输入的特征向量的形状为 $T\times F$,其中,T 是样本中所含的帧数,F 为输入特征的频点数;卷积层的输出为 $T'\times F'\times Q$ 的张量,其中,Q 为特征图的数量,T' 和 F' 为经过一系列池化操作后的帧数和频带数。

本研究中使用的胶囊层由多路径基础胶囊层和高级胶囊层组成。多路径基础胶囊层的每个胶囊层是一个含有16通道的卷积层,每个通道由4维胶囊组成。特征被送入基础胶囊层中,经过卷积和 *squashing* 函数压缩后,将三层的输出进行拼接,然后特征压缩成形状为 $T'\times V\times U$ 的3维张量,其中,V 是从其它维度推断出的,U 是胶囊的维度,大小为4;然后将每一帧的胶囊送入高级胶囊层,来计算 K 个代表声音事件类别的8维高级胶囊,两层胶囊之间使用自注意力路由算法进行计算;最后,将得到形状为 $T'\times K\times 8$ 的张量。

胶囊层之后是两层全连接层,用来获取声音事件活动的概率。首先将胶囊层的输出重塑成形状为 $T'\times(K\times 8)$ 的张量,在经过两层全连接层后,张量的形状为 $T'\times K$,即 T' 个帧的每个声音事件的活动概率。由于使用的是弱标注的数据,训练集没有帧级别的标签可用,所以使用聚合函数将输出聚合成音频级的概率,即最后的输出形状为 $1\times K$。使用的聚合函数公式如式(11):

$$y_l=\frac{\sum_i y_i\exp(y_i)}{\sum_i\exp(y_i)}。\tag{11}$$

其中 $y_i\in[0,1]$ 是某个事件类型的帧级预测概率,$y_l\in[0,1]$ 音频级的聚合概率。

2 实验部分

2.1 数据集

本研究提出的方法是基于弱标注数据集的,其中弱标注数据是指只提供音频中的事件类型,而不

包含任何的时间信息。本研究使用 DCASE 2017 任务 4 提供的弱标记数据集进行评估,此数据集是 AudioSet[17]的一个子集,由 17 个声音事件组成,分为“警告”和“车辆”两类。每段音频的最长持续时间为 10 s,并且可能对应于多个可能重叠的声音事件。本工作在这个数据集上评估了 2 个任务:音频标注和声音事件检测。其中,音频标注旨在预测音频剪辑中包含的声音事件类型,声音事件检测还预测事件的开始时间和结束时间。对于音频标注子任务,使用精确率、召回率和 F 分数的微平均值来评估模型的性能。对于 SED,计算了一个 1 s 分辨率的基于分段的错误率。

2.2　实验设置

本研究使用对数 Mel 语谱图作为输入特征。在提取特征之前,将每个音频片段重新采样到 16 kHz。使用 64 ms 帧长度、20 ms 重叠和每帧 64 个 Mel 频率单元计算对数 Mel 特征。对于每个 10 s 的音频片段,将产生一个 240×64 的特征向量。

为减少过拟合的发生以及加快收敛的速度,本研究在每个卷积层和初级胶囊层之后使用批标准化。使用 Adam 优化器进行训练,固定学习率为 0.001 并且每两个 epoch 下降为原来的 0.9 倍。使用二元交叉熵作为损失函数,梯度通过大小为 44 的 mini-batch 进行计算。共训练 30 个 epoch。

验证集和评估集具有均衡的事件数,但训练集是不平衡的,这会导致分类的偏差。为了减轻这个问题带来的影响,本研究使用了文献[18]中提出的数据平衡技术,以确保每一个小批量中包含来自每个类的样本数量是相当的。对于本研究提出的系统,音频标注和声音事件检测的阈值分别设置为 $\tau_1=0.3$ 和 $\tau_2=0.6$。

2.3　实验结果

基于上述的弱标注数据集,下面检验前文提出的基于自注意力路由的胶囊网络模型的声音事件检测效果。本研究以文献[12]提出的 GCCaps 为基线系统,方案一将 GCCaps 中的动态路由算法更换为自注意力路由(记为 GCCaps-att);方案二在 GCCaps 的基础上使用多路径基础胶囊层(记为 GCCaps-mp);对本研究提出的方法进行对比性实验,具体结果见表 1 和表 2。

从表 1 的结果可以看出,自注意力路由和多路径基础胶囊层的加入能够提高音频标注任务的性能表现,分别比基线系统提高了 0.4%和 0.9%,而本研究提出的模型的 F 分数最高,相较于基线系统,提高了 1.4%。由表 2 可知,在声音事件检测子任务中,自注意力路由对于性能的提升更加明显。本研究提出的模型获得了最佳的表现,错误率为 0.72。

表 1　音频标注子任务的性能结果

Table 1　Performance results of audio tagging subtask　%

方法	F 分数	精确率	召回率
GCCaps	58.3	59.2	57.6
GCCaps-att	58.7	59.5	58.1
GCCaps-mp	59.2	59.7	58.9
MPCaps-att	59.7	60.1	59.5

表 2　声音事件检测子任务的性能结果

Table 2　Performance results of sound event detection subtask

方法	错误率
GCCaps	0.76
GCCaps-att	0.74
GCCaps-mp	0.75
MPCaps-att	0.72

表 1 和表 2 表明,本研究提出的自注意力路由和多路径基础胶囊层能够显著提高模型的性能,并且自注意力路由可以加快模型的训练过程,而多路径非对称的卷积结构能够使模型更充分地利用特征信息。

3　结　语

本研究提出了基于自注意力路由的胶囊网络模型,以实现弱标注数据下的多音事件检测。针对传统动态路由算法减缓网络运行速度的问题,采用了一种非迭代的自注意力路由算法,并且提出了一种多路径基础胶囊层结构,其中采用非对称的卷积核用来保留更多的时间信息,同时多路径的结构能够使模型获得不同分辨率的特征,从而使模型能够充分利用特征信息。实验结果也表明,本研究提出的模型具备更好的性能,模型在音频标注子任务上取得了 59.5%的 F 分数,在声音事件检测子任务中错误率仅为 0.72。未来的研究需要寻找更加高效的特征提取方法,为模型提取更全面的特征,以及研究最近提出的基于期望最大化算法(EM)的路由变体。

参　考　文　献

[1] VALENZISE G, GEROSA L, TAGLIASACCHI M, et al.

Scream and gunshot detection and localization for audio-surveillance systems[C]// IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007 Proceedings, 2007: 21-26.

[2] SALAMON J, BELLO J P. Deep convolutional neural networks and data augmentation for environmental sound classification [J]. IEEE Signal Processing Letters, 2017, 24(3): 279-283.

[3] CHAN T K, CHIN C S. Health stages diagnostics of underwater thruster using sound features with imbalanced dataset[J]. Neural Computing and Applications, 2019, 31(10): 5767-5782.

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

[6] DEGARA N, DAVIES M E P, PENA A, et al. Onset event decoding exploiting the rhythmic structure of polyphonic music [J]. IEEE Journal on Selected Topics in Signal Processing, 2011, 5(6): 1228-1239.

[7] HEITTOLA T, MESAROS A, ERONEN A, et al. Audio context recognition using audio event histograms[C]// European Signal Processing Conference, 2010: 1272-1276.

[8] MCLOUGHLIN I, ZHANG H, XIE Z, et al. Robust sound event classification using deep neural networks[J]. IEEE Transactions on Audio, Speech and Language Processing, 2015, 23(3): 540-552.

[9] PICZAK K J. Environmental sound classification with convolutional neural networks[C]// IEEE International Workshop on Machine Learning for Signal Processing, 2015: 9-14.

[10] ALEX GRAVES A M and G H. Speech recognition with deep recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 6645-6649.

[11] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]// Advances in Neural Information Processing Systems, 2017: 3857-3867.

[12] IQBAL T, XU Y, KONG Q, et al. Capsule routing for sound event detection[C]// European Signal Processing Conference, 2018: 2255-2259.

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

[14] MAZZIA V, SALVETTI F, CHIABERGE M. Efficient-CapsNet: Capsule network with self-attention routing[J]. Scientific Reports, 2021, 11(1): 1-13.

[15] XU Y, KONG Q, HUANG Q, et al. Convolutional gated recurrent neural network incorporating spatial features for audio tagging[C]// International Joint Conference on Neural Networks (IJCNN). IEEE, 2017: 3461-3466.

[16] LIANG K W, TSENG Y H, CHANG P C. Parallel capsule neural networks for sound event detection[J]. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC, 2019: 1933-1936.

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

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

(责任编辑 姜丰辉)