文章编号：1672-6987(2025)05-0152-07；**DOI**：10.16351/j.1672-6987.2025.05.019

基于深度胶囊网络融合模型的多声音事件检测

姜轻舟，杨树国*，王文武

（青岛科技大学 数理学院，山东 青岛 266061）

摘　要：传统的胶囊网络架构是基于动态路由机制实现的，需要大量迭代和向量计算来更新权值系数，并且胶囊之间不存在信息共享，导致信息冗余。针对这一缺陷，本工作提出了一种基于融合深度胶囊网络的多声音事件检测模型，在门控卷积和3D卷积下通过动态路由减少了特征重叠导致的信息冗余，并且对原始特征进行编码，将其用于特征信息补充，提高了训练次数模型的速度和准确性。本工作使用DCASE2017(Detection and Classification of Acoustic Scenes and Events 2017) Challenge Task 4数据集对模型进行评估，最终F1分数达到59.6%，声音事件检测错误率低至0.71。结果表明，所提出的方法可以显著提高训练速度和精度。

关键词：多声音事件检测；胶囊网络；融合网络；DCASE 2017挑战赛

引用格式：姜轻舟，杨树国，王文武. 基于深度胶囊网络融合模型的多声音事件检测[J]. 青岛科技大学学报(自然科学版)，2025，46(5)：152-158.

中图分类号：TN 912.2　　**文献标志码**：A

JIANG Qingzhou，YANG Shuguo，WANG Wenwu. Multi-Voice event detection based on fused deep capsule network fusion model[J]. Journal of Qingdao University of Science and Technology(Natural Science Edition)，2025，46(5)：152-158.

Multi-Voice Event Detection Based on Fused Deep Capsule Network Fusion Model

JIANG Qingzhou，YANG Shuguo，WANG Wenwu

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

Abstract：The traditional capsule network architecture is implemented based on dynamic routing mechanism，which requires a large number of iterations and vector calculations to update the weight coefficients，and there is no information sharing between capsules，leading to information redundancy. To address this shortcoming，this paper proposes a multi-sound event detection model based on fusion depth capsule network，which reduces the information redundancy caused by feature overlap by dynamic routing under gated convolution and 3D convolution，and encodes the original features and uses them for feature information supplementation to improve the speed and accuracy of training times models. In this paper，the model is evaluated using DCASE2017 (Detection and Classification of Acoustic Scenes and Events 2017) Challenge Task 4 dataset and the final F1 score reaches 59.6% with a low sound event detection error rate of 0.71. The results show that the proposed method can significantly improve the training speed and accuracy.

收稿日期：2024-11-19
基金项目：山东省自然科学基金项目(ZR2024QF112).
作者简介：姜轻舟(1998—)，男，硕士研究生.　*通信联系人.

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

声音事件检测(SED)[1]的目标为自动标记声音事件的类别以及它们发生的时间戳,即声音片段中的开始和结束时间。声音事件检测可以分为单声道声音事件检测(MSED)和多声道声音事件检测(PSED)[2]。单声道声音事件检测指的是对音频信号中单一声音事件的检测,而多声道声音事件检测的目的是检测音频信号中的多个声音事件。在实践中,因为声音事件经常同时发生,所以多声道声音事件检测比单声道声音事件检测有更大的应用范围。在现有公开的大规模数据集中,声音事件往往只标注了类别信息,而没有标注其在音频片段中出现的开始和结束时间,一般称为弱标签事件[3]。例如,路上的口哨声和声音很容易被检测到,但声音的开始和结束时间却不容易被人类注释者确定。然而,训练声音事件检测模型需要强标签,其中包括声音类别,以及它们在音频信号中的开始和结束时间。弱标签声音事件缺乏相应的时间戳,一般不能直接对其进行检测。因此,应该对原始事件进行相应的处理,或者改变检测模型中的网络结构,使其能够处理具有弱标签的输入信息。

传统的声音事件检测模型主要是基于经典的机器学习方法,如支持向量机、隐马尔科夫模型(hidden markov mode,HMM)和非负矩阵分解[4]。支持向量机在文献[5]中被用于声音事件检测。然而,HMM不能处理复杂的音频事件,特别是处理重叠的声音事件。基于非负矩阵分解(nonnegative matrix factorization,NMF)[6]的声音事件检测方法,可单独表示每一类声音事件,但在利用时间背景信息方面并不有效。近年来,深度学习的快速发展使得声音事件检测的新方法不断涌现。其中卷积神经网络(convolutional neural network,CNN)[8]和递归神经网络(recurrent neural network,RNN)[7,9]被广泛用于声音事件检测。然而,目前使用的CNN框架是基于标量操作的,例如卷积和具有恒定堆积的池子,它对高层特征和低层特征之间的位置关系不敏感。此外,适用于时间序列建模的递归神经网络(RNN)也被用于声音事件检测。例如,在文献[11]中,从原始立体声信号中构建了三个不同的通道数据,即右通道、平均通道和差分通道,并使用具有logmel能量的长短期记忆(long short-term memory,LSTM)网络(RNN的一个变种)作为输入特征;其对三个通道进行了声音事件检测,并使用了融合策略来实现性能的提高。尽管基于CNN的方法在音频标签和SED任务中都很成功,但它不能有效地捕捉音频片段中的长时间依赖性。为了解决这个问题,文献[12]提出了一个结合CNN和RNN的CRNN模型,用于多声道声音事件检测,其中用CNN提取频率不变的特征,而用双向的RNN实现分类。现有的大多数多声道声音事件检测方法都是基于CNN、RNN或CRNN的。CNN模型在解释部分和整体之间的位置关系方面有一个局限性[10]。为了解决这个局限性,HINTON等提出了胶囊网络[10],将几个具有相同位置特征的神经元包裹成一个多维向量,称为胶囊。通过反向传播得到的转换矩阵使高层的胶囊能够被低层的胶囊所预测。然后,对所有胶囊产生的预测进行聚类,并对耦合系数进行更新。最后,得到包含所有胶囊信息的预测向量。与传统的CNN相比,基于胶囊网络的声音事件检测实现了更好的性能。然而,传统的胶囊网络存在大计算量和低训练速度的缺点,这是因为在胶囊网络中使用了动态路由算法,需要迭代更新耦合系数,主要涉及矢量和矩阵运算,从而导致计算量增大。

本研究对文献[9]中基于3D卷积的动态路由算法的胶囊网络进行了改进,其中3D卷积减少了多层胶囊动态路由引起的信息冗余,然后通过使用跳跃连接(会跳跃神经网络中的某些层,并将一层的输出作为下一层的输入)进一步降低了网络复杂度,在保证模型精度的同时提高模型训练的速度;运用融合网络,首先通过深度胶囊网络进行特征提取和网络训练,然后通过对原始特征图进行编码,对原网络输出进行信息补充,进一步提高了模型精度。

1　胶囊网络

胶囊网络最早由HINTON等[10]提出,是一种可以替代传统卷积神经网络的新型网络结构。与传统的神经网络不同,胶囊网络的每个神经元都是一个多维向量,可以更好地表示目标的位置和方向等特征。胶囊网络主要包含卷积层、主胶囊层和数字胶囊层,如图1所示。

卷积层通过多个卷积核提取输入数据的特征,形成j个特征图。主胶囊层对应每个特征标量在这j个特征图中的位置,组成一个j维的特征向量,它是

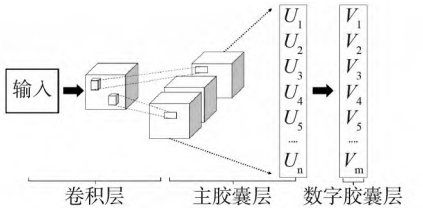


图1 胶囊网络结构

Fig. 1 Capsule network structure

胶囊网络的一个神经元。如图1所示,通过遍历所有的特征图,形成n个胶囊,其中每个胶囊是一个j维向量。n个胶囊通过动态路由算法,最终得到m个胶囊,形成数字胶囊层,即胶囊网络的输出。

从主胶囊层到数字胶囊层使用的动态路由算法是胶囊网络最重要的特征,如图2所示。

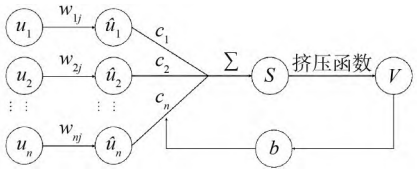


图2 动态路由

Fig. 2 Dynamic routing

动态路由的主要步骤如下:

1)主胶囊层的输出向量u_i通过转换矩阵w_{ij}得到向量$\hat{u}_{j|i}$:

$$\hat{u}_{j|i}=w_{ij}u_i。\tag{1}$$

其中$\hat{u}_{j|i}$可看作是低级胶囊i到高级胶囊j的预测向量。转换矩阵w_{ij}是编码低层特征和高层特征之间重要的空间和其他关系,由反向传播得到。

2)耦合系数c_{ij}由softmax函数得到

$$c_{ij}=\frac{e^{b_{ij}}}{\sum_k e^{b_{ik}}}。\tag{2}$$

其中b_{ij}为先验概率,初始化为0。

初级预测向量$\hat{u}_{j|i}$和耦合系数c_{ij}进行加权求和,得到预测向量s_j:

$$s_j=\sum_i c_{ij}\hat{u}_{j|i}。\tag{3}$$

初始预测向量s_j通过挤压函数,将值压缩到0~1之间,得到最终预测向量v_j:

$$v_j=\frac{\|s_j\|^2}{1+\|s_j\|^2}\cdot\frac{s_j}{\|s_j\|}。\tag{4}$$

3)先验概率b_{ij}:

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

最后经过规定迭代次数进行迭代,最终得到预测向量v_j。其中公式(5)的目的是使得相似度较高的低级胶囊预测向量与高级胶囊预测向量的权重变大,而相似度低的预测向量权重则不断减少,最终通过多次迭代取得最优结果。

2 融合特征

音频特征主要有过零率、短时能量、短时自相关系数、logmel谱图、语谱图和频谱图等,其中logmel谱图被广泛用于声音事件检测,并取得了最优的单特征实验结果。

2.1 特征选择

目前logmel谱特征被广泛用于声音事件检测,并取得了最先进的结果。过零率(ZCR)是指音频信号在每一帧中越过零值的次数(从正到负或从负到正)。这一特征已被广泛用于语音识别和音乐信息检索领域,是识别人声的一个关键特征。两种特征的可视化如图3、4所示。

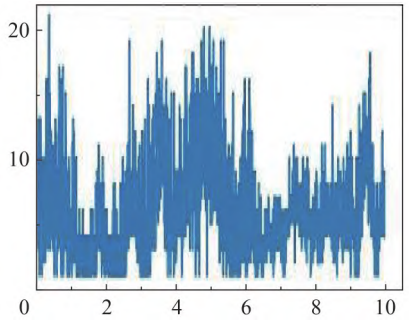


图3 过零率

Fig. 3 Zero crossing rate

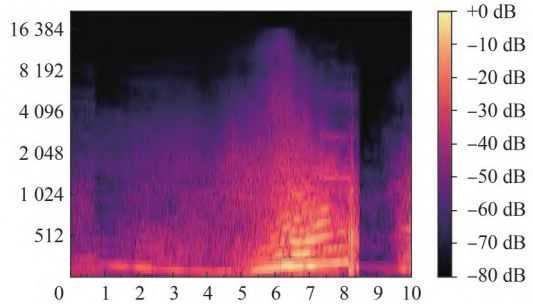


图4 Logmel谱图

Fig. 4 Logmel features

Logmel特征为目前声音事件检测常用的特征,单特征模型上精度最优;ZCR特征可有效区分白噪声。本研究通过实验发现,与只使用logmel相比,融合ZCR和logmel可以提高声音事件检测的性能。这主要是因为logmel在卷积操作后会有一些信息损失,见图5(a)。ZCR往往能很好地区分背景中的白噪声,加入ZCR可以帮助增强该特征,见图5(b)。如图5(b)所示,Conv2和Conv3卷积层的输出图中颜

色复杂度高于图5(a),说明加入ZCR后反应的特征信息要优于未加ZCR。

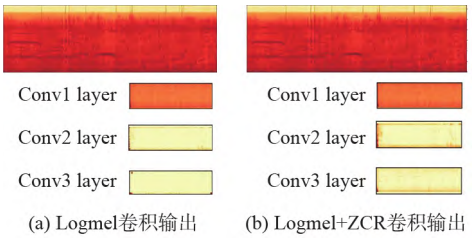


图5　卷积输出

Fig. 5　Convolutional output

2.2　融合方式

所用的特征融合方式为分别卷积融合,如图6所示。首先分别提取声音数据的过零率和Logmel谱,并用一维卷积核对过零率进行卷积,用二维卷积核对Logmel谱进行卷积,得到特征向量后进行级联融合。

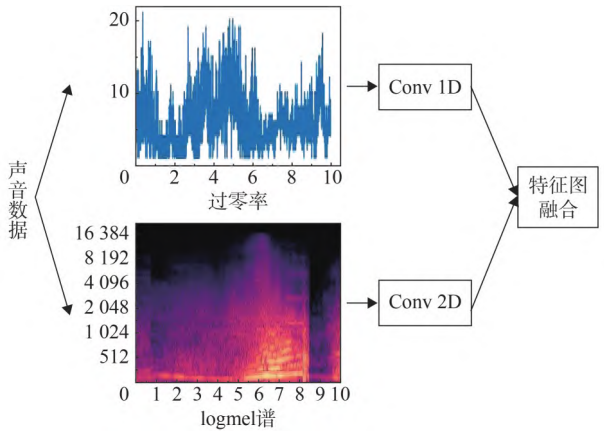


图6　特征融合

Fig. 6　Features fused

相较于对原始特征直接进行级联,此融合方法可通过卷积挖掘深层特征,并且分别通过不同卷积核对不同特征进行卷积处理,不会受到不同特征之间的相互影响,因此可以更好地捕获特征信息。

3　基于3D卷积的动态路由

与传统的CNN相比,胶囊网络可以得到更加准确的预测结果,但是训练速度慢,因为目前的动态路由只能以全连接层的形式实现。在传统的胶囊网络中,胶囊向量被展平并通过动态路由到分类胶囊,因此需要不断堆叠全连接层,这相当于在多层感知器(MLP)模型中堆叠全连接层[10],从而导致训练速度缓慢。为了解决这个问题,本工作提出了一种基于3D卷积的动态路由算法。

假设 $\Phi^l\in\mathbf{R}^{(w^l,w^l,c^l,n^l)}$ 是胶囊层 l 的输出,其中 w^l 包含特征图的高和宽,c^l 是3D胶囊向量的数量,n^l 是胶囊维度(即胶囊尺寸)。首先,重构 Φ^l 为单通道张量 $\tilde{\Phi}^l\in\mathbf{R}^{(w^l,w^l,c^l\times n^l,1)}$,并通过3D卷积核($c^{l+1}\times n^{l+1}$)进行卷积[11]。假设 φ_t^l 是胶囊层 l 通过卷积核 t($t\in[c^{l+1}\times n^{l+1}]$)卷积后的输出,卷积核步长为$(1,1,n^l)$,中间投票记为 S,维度为$(w^{l+1},w^{l+1},c^l,c^{l+1}\times n^{l+1})$,$S$的计算公式:

$$s_{i,j,k,m}=\sum_p\sum_q\sum_r\tilde{\Phi}^l(i-p,j-q,k-r)\cdot\varphi_t^l(p,q,r)。\tag{6}$$

保持 φ_t^l 的宽度为 n^l,可得到胶囊层 l 的投票。使用高和宽大于1的3D卷积核作为转换矩阵,得到低层胶囊的预测值。

重构 S 为 $\tilde{S}(w^{l+1},w^{l+1},n^{l+1},c^{l+1},c^l)$,其中 w^{l+1} 由以下公式得到:

$$w^{l+1}=\frac{w^l-V_{\text{Kernel\_size}}+2\times V_{\text{Padding}}}{V_{\text{Stride}}}+1。\tag{7}$$

与传统动态路由相比,此方法更好地使相邻胶囊共享信息,减少了信息冗余,减少参数个数为 $c\cdot(w^lw^{l+1})^2$。类似地,局部投票是通过3D卷积核将一个区域中的一个子集转化为一票来实现的。

初始化先验概率矩阵 B 为0,$B\in\mathbf{R}^{(w^{l+1},w^{l+1},c^{l+1})}$,使用Softmax_3D函数计算耦合系数 c:

$$c_{\text{pqrs}}=\frac{e^{b_{\text{pqrs}}}}{\sum_x\sum_y\sum_z e^{b_{\text{xyzs}}}}。\tag{8}$$

初始胶囊的输出 V 通过耦合系数 c 和 $\tilde{S}$ 的乘积得到:

$$V_{\text{pqr}}=\sum_s c_{\text{pqrs}}\cdot\tilde{S}_{\text{pqrs}}。\tag{9}$$

最终输出 $\hat{V}$ 通过挤压函数得到:

$$\hat{V}_{\text{pqrs}}=\frac{\|V_{\text{pqr}}\|^2}{1+\|V_{\text{pqr}}\|^2}\cdot\frac{V_{\text{pqr}}}{\|V_{\text{pqr}}\|}。\tag{10}$$

通过不断迭代更新先验概率 $b_{pqrs}\leftarrow b_{pqrs}+\hat{S}_{pqr}\cdot\tilde{V}_{pqrs}$ 得到最优输出 $\hat{V}$。

4　基于深度胶囊网络融合模型的多声音事件检测

所提出的基于融合深度胶囊网络的多声音事件检测模型如图7所示;其中网络A主要包括一个门控卷积层(包括卷积层和最大池化层)[12]、一个主胶囊层(包括3个跳跃连接的Caps单元、一个3D卷积Cap单元和一个最终的收集胶囊Collect caps)和一个全连接层fully connect(FC)caps层。网络

B包含编码器和解码器,直接对原始图像进行编码和解码,其输出作为网络A输出的补充特征信息。

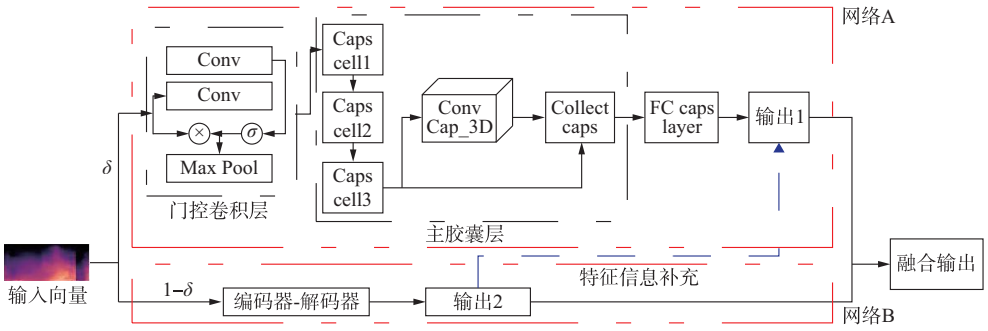


图7 基于融合深度胶囊网络的声音事件检测模型

Fig. 7 Sound event detection model based on fused deep capsule network

首先,输入向量为上节所提出的融合特征向量,分别送往网络A和网络B,对于网络A,融合后的特征输入到卷积层后,由门控卷积层和最大池化层进行局部特征提取,提取的特征向量将被送入初级胶囊层。每个Caps细胞层包括3个Convcaps层:第一个Convcaps层的输出跳接到最后一层的输出,相邻两个下层胶囊层的输出通过跳接后加入相应的元素进行连接。这种设计可以减少由于层的堆叠而导致的梯度消失问题。3个Convcaps层的迭代次数设置为1。然后,Convcaps层的输出被送到Convcaps3D层进行基于3D卷积的动态路由。路由迭代次数设置为3,此时特征谱图重构为单通道张量 $\tilde{\Phi}^l \in \mathbf{R}^{(w^l, w^l, c^l \times n^l, 1)}$,由步长为$(1,1,n^l)$的卷积核$(c^{l+1} \times n^{l+1})$进行卷积操作,得到初始预测矩阵,并与耦合系数矩阵加权求和得到初始输出。对于网络B,对原始融合特征图进行编码与解码,得到特征输出后与网络A的输出进行融合,融合公式:

$$V_{\text{output}} = \delta \times V_{\text{output1}} + (1-\delta) \times V_{\text{output2}}。 \tag{11}$$

其中V_{output1}是网络A的输出,V_{output2}是网络B的输出,权重系数δ初始化为0.8,由模型反向传播更新。

最后,将融合输出重构成维度为(a^l, n^l)的矩阵,其中$a^l = w^l \times w^l \times c^l$,n代表声音类别的数量。因此,相应声音事件的概率将通过全连接(FC)层获得。最终输出可通过以下函数聚合帧级概率获得:

$$y_l = \frac{\sum_i y_i \cdot \mathrm{e}^{y_i}}{\sum_i \mathrm{e}^{y_i}}。 \tag{12}$$

其中$y_i \in [0,1]$是为特定事件类型预测的帧级概率,$y_l \in [0,1]$是聚合的音频级概率。此时全连接层也可以用来去除下层胶囊之间的空间关系,保留胶囊层与全连接层之间的部分-整体关系。

5 实 验

5.1 数据集和实验设置

本工作使用来自DCASE 2017 Task 4的弱标记数据集进行评估和验证,在DCASE 2018之前的比赛中,基于RNN和CRNN的基线方法也经常使用该数据集。该数据集由来自各种场景的记录组成,包括城市、餐厅、森林小径等。对于每个录制的场景,分别捕获3~5 min,然后分段为最长10 s的音频。所有标签仅包含声音类别,但不包含它们的时间信息。对于这个数据集,本文实现了两个检测任务:检测音频片段中声音事件的发生,并提供声音事件的开始和偏移时间。对于第一个音频注释子任务,本文使用精度、召回率和F1分数的微平均来评估性能。对于第二个任务,本工作使用SEDEVAL工具箱[13]计算基于片段的等错误率(EER),并根据EER对其进行评估,时间分辨率为1 s。EER是指错误接受的比例等于错误拒绝的比例所代表的值,值越小代表模型结果越准确。

每个音频剪辑首先被重新采样到16 kHz,使用64 ms的帧长、20 ms的重叠和每帧64个Mel频率单元计算Logmel特征。因此,从每个10 s的剪辑中提取一个240×64的特征向量。在本文中,在每个门控卷积层和主胶囊层之后添加了一个批量归一化层[13]和一个Dropout层[14-15]以减少潜在的过度拟合。门控层的拒绝率(即拒绝单元的比例)设置为0.2,主胶囊层设置为0.5。

对于胶囊路由,3个Convcaps层的迭代次数设置为1,Convcaps3D层的迭代次数设置为3[9]。本工作使用二元交叉熵作为损失函数,Adam[16]作为优化器。对于梯度的计算,小批量大小设置为44;初始学习率设置为0.001。

胶囊网络经过30个epoch的训练,这里使用文献[17]中建议的数据平衡技术来确保每个小批量包含来自每个类别相当数量的样本。在推理过程中,选择了在验证集上达到最高准确度的5个模型(epoch),并对它们的预测进行平均。本工作中音频标注和声音事件检测的阈值分别设置为$\tau_1=0.3$和$\tau_2=0.6$。

5.2　实验结果

1)融合特征对比实验。

对于融合特征实验,本工作从是否进行融合和不同融合方式两方面进行了对比,所用网络结构为本工作提出的融合深度胶囊网络,结果见表1。

表1　特征对比实验

Table 1　Feature comparison experiment

特征	F1得分	准确率/%
Logmel	57.9	91.8
ZCRs	36.6	72.5
Logmel+ ZCRs(原始特征级联)	59.1	94.4
Logmel+ ZCRs(分别卷积融合)	59.6	95.2

由表1可见,使用融合特征可以使得模型精度更高,而相较于对原始特征进行级联,采用分别卷积并融合的方式进行特征融合可得到更高的模型精度。

2)网络结构对比实验。

对比实验以文献[18]提出的GCCaps为基线系统,将其与原始胶囊网络(Caps)、卷积递归神经网络(CRNN)、基于门控卷积的深度胶囊网络(DGCaps)以及本工作提出的融合深度胶囊网络(CDCaps)进行对比,结果见表2、3。

表2　音频标签子任务的训练准确性比较结果

Table 2　Training accuracy comparison results of audio tagging subtask

方法	F1得分	精确率/%	召回率/%
GCCaps	58.1	59.2	57.9
Caps	56.7	57.5	56.2
CRNN	57.9	59.0	57.5
DGCaps	58.8	60.2	58.3
CDCaps	59.6	61.5	59.1

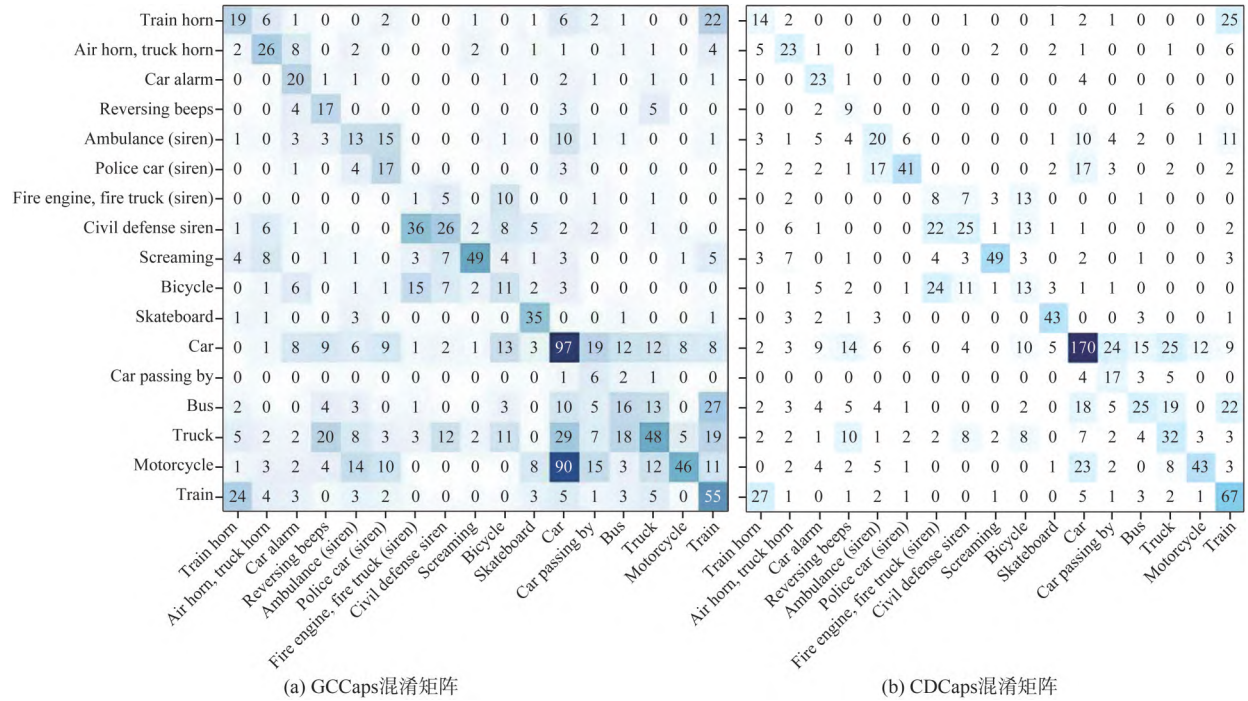
表3　声音事件检测子任务的等错误率比较结果

Table 3　Comparison results of EER of sound event detection subtask

方法	等错误率
GCCaps	0.75
Caps	0.84
CRNN	0.78
DGCaps	0.73
CDCaps	0.71

由表2、3可见,使用基于门控卷积的深度胶囊网络可以提高任务精度,相较于基线系统,F1得分提高了0.7%;而融合深度网络可以进一步提高模型性能,相较于基线系统,F1得分提高了1.5%;对于声音事件检测子任务,等错误率降低了0.04。

为了验证上述结果,本工作画出了基线系统(GCCaps)与融合深度胶囊网络(CDCaps)预测结果的混淆矩阵,见图8。



(a) GCCaps混淆矩阵　　(b) CDCaps混淆矩阵

图8　混淆矩阵

Fig. 8　Confusion matrix

图8横轴代表17个声音类别的真实值,纵轴代表预测值;结果表明,融合深度胶囊网络的分类结果优于基线系统。

6 结 语

提出了基于融合深度胶囊网络的声音事件检测模型,通过门控深度胶囊网络提取主要特征,其中3D卷积动态路由能更好地捕捉胶囊之间的相互联系,使相邻胶囊之间可以信息共享,大大减少了信息冗余,提高了训练速度和精度;利用编码器解码器进行特征补充,使得模型能更好地捕捉特征进行训练。

实验结果表明,对于音频标签子任务,本工作所提出的模型在F1得分上取得了最好的性能,为59.6%;对于声音事件检测子任务,取得了最低的等错误率,为0.71。未来该模型在网络融合上还需进一步研究出更好的融合方式,使特征信息可以更好地进行训练。

参 考 文 献

[1] THOMASSEN S, BENDIXEN A. Assessing the background decomposition of a complex auditory scene with event-related brain potentials[J]. Hearing Research, 2018, 370: 120-129.

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

[3] 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.

[4] SINGH S, JAISWAL U C. Audio classification using grasshopper-ride optimization algorithm-based support vector machine [J]. IET Circuits, 2021, 15(5): 342-354.

[5] JUNG S H, CHUNG Y J. Sound event detection using deep neural networks [J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2020, 18(5): 2587.

[6] CHAN T K, CHIN C S, LI Y. Semi-supervised NMF-CNN for sound event detection[J]. IEEE Access, 2021, 9: 130529-130542.

[7] JIN W Y, WANG X, ZHAN Y. Environmental sound classification algorithm based on region joint signal analysis feature and boosting ensemble learning [J]. Electronics, 2022, 11(22): 3743.

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

[9] GRAVES A, MOHAMED A, HINTON G E. Speech recognition with deep recurrent neural networks [J]. CoRR, 2013, 9: 1703-1709.

[10] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[J]. Advances in Neural Information Processing Systems, 2017, 9:3857-3867.

[11] RAJASEGARAN J, JAYASUNDARA V, JAYASEKARA S, et al. DGCaps: Going deeper with capsule networks [J]. CoRR, 2019, 1904: 122-154.

[12] ÇAKıR 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.

[13] HAN R Z, LIU Z L, PHILIP CHEN C L. Multi-scale 3D convolution feature-based Broad Learning System for Alzheimer's Disease diagnosis *via* MRI images [J]. Applied Soft Computing, 2022, 120: 108660.

[14] YUAN J, XIONG H C, XIAO Y, et al. Gated CNN: Integrating multi-scale feature layers for object detection [J]. Pattern Recognition, 2020, 105: 107131.

[15] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. CoRR, 2015, 1502: 1-14.

[16] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. CoRR, 2012, 1207: 566-571.

[17] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 12-53.

[18] IQBAL T, XU Y, KONG Q Q, et al. Capsule routing for sound event detection[J]. CoRR, 2018, 1806: 1-4.

(责任编辑 姜丰辉)