基于加性间距胶囊网络的家庭活动识别方法研究

郑启航[1],王章权[1,2],刘半藤[1,2],陈 阳[1],陈友荣[2]
(1. 常州大学信息科学与工程学院,江苏常州 213164; 2. 浙江树人大学信息科技学院,浙江杭州 310015)

摘 要: 本文研究基于音频的家庭活动识别方法,提出了一种基于加性间距胶囊神经网络识别模型,针对传统胶囊神经网络目标函数仅以输出胶囊模长作为约束的弊端,本文以几何学的视角,在胶囊神经网络结构中加入 Transition 层,使用 Transition 层对胶囊单元空间关系进行变基至一维空间,再使用加性间距 Softmax 作为目标函数,以同类特征变化小,非同类特征差异大作为优化策略构建基于胶囊向量空间关系的目标函数以提高模型分类能力,最后对方法进行试验,采用音频事件对家庭活动进行分类识别. 选择声学场景和事件检测与分类(Detection and Classification of Acoustic Scenes and Events,DCASE)2018 挑战任务 5 作为数据集,进行分类器构建和测试,最终平均 F1 分数达到 92.3%,优于其他主流方法.

关键词: 音频事件分类; 家庭活动识别; 胶囊网络; 加性间距
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2020)08-1580-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.08.017

Research on Family Activity Recognition Method Based on Additive Margin Capsule Network

ZHENG Qi-hang[1], WANG Zhang-quan[1,2], LIU Ban-teng[1,2], CHEN Yang[1], CHEN You-rong[2]
(1. *School of Information Science and Engineering, Changzhou University, Changzhou, Jiangsu* 213164, *China*;
2. *College of Information Science and Technology, Zhejiang Shuren University, Hangzhou, Zhejiang* 310015, *China*)

Abstract: We study the method of family activity recognition based on audio and propose a capsule neural network recognition model based on additive margin. In view of the drawbacks of the traditional capsule neural network objective function only with the output capsule mode length as the constraint, this paper adds a Transition layer to the capsule neural network structure from the perspective of geometry and uses the Transition layer to rebase the capsule unit spatial relationship to the one-dimensional. Then, using the additive margin Softmax as the objective function, the change of similar features is small, and the difference of non-similar features is used as the optimization strategy to construct the objective function based on the capsule vector space relationship to improve model classification ability. Finally, test this method by classified identified for audio events for family activities. Selecting Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge Task 5 as a dataset for classifier construction and testing, with a final average F1 score of 92.3%, which is superior to other mainstream methods.

Key words: classification of acoustic events; family activity recognition; capsule network; additive margin softmax

1 引言

经济的发展和医疗水平的提高使得人类寿命越来越长,发达国家甚至包括中国等新兴发展中国家均面临严重的人口老龄化问题,根据联合国的一份报告,2015 年至 2030 年间,60 岁以上的老年人数量预计将增长 56%,到 2050 年将达到近 21 亿[1]. 传统医疗保健的成本将按比例增长,因此对老年人健康状况、日常生活活动进行远程检测是十分必要的. 本文研究基于声学的声音事件分类方法,通过使用基于声学的监测方法,既不影响老人日常生活又保护老人隐私,对于提升养老看护系统的安全性、舒适性有巨大的意义.

同时,随着人工智能算法的兴起,基于深度学习的方法越来越多的用于音频事件识别中. 其中 Hershey 团

收稿日期:2019-07-05;修回日期:2020-05-18;责任编辑:马兰英
基金项目:浙江省公益技术研究项目(No. LGG19F010011);浙江树人大学省属高校基本科研业务费专项资金(No. 2020XZ009)

队首次使用多种基于卷积神经网络(Convolutional Neural Networks,CNN)架构的模型对音频信号进行分类,取得了较好的识别效果;Becker 团队建立了自己的音频数据集并使用分层相关传播的方法解释了 CNN 对音频信号进行分类的机理[2~5]. 然而传统的 CNN 模型主要对局部特征进行感知,缺乏对局部特征与整体特征之间空间联系的思考,导致模型对于时序相关数据识别效果较差. 众多学者进而研究 CNN 网络的变体模型,通过整合局部特征间相关性来弥补这个缺点,如 Keren 等人提出一种卷积递归神经网络(Convolutional Recurrent Neural Network,CRNN)模型,加强了对语音序列信号的特征提取,效果优于经典 CNN 模型[6];Chew 等人提出了长短期记忆神经网络(Long Short Term Memory networks,LSTM)结合 CNN 的混合架构神经网络,并联合多模型结果融合预测,取得了优良效果[7].

胶囊网络模型[8]由 Hinton 提出,使用向量作为神经元的基本单元(胶囊单元),可以表征多种特征组合而成的一个高信息量集合,相比传统神经单元,可承载更多的特征信息,并且胶囊单元间的特征传递是通过动态路由算法将底层与高层特征通过空间关系进行分配连接,使得特征在时序上、空间上紧密的联系在一起,确保模型能敏锐的区分不同的输入特征. 有学者对胶囊网络中特征整合过程进行了研究,如贾旭东等人使用注意力机制提升输入胶囊单元的特征的紧密性[9],提升了文本分类的准确率;任开旭等人利用概率矩阵分解模型将多个胶囊单元分解为物品内容特征向量提取多维语义信息后进行物品推荐[10],提升了预测精度. 而在输出目标函数构建过程中,以上工作与原始胶囊网络模型均采用基于 Margin Loss 的方式,即仅考虑胶囊单元的模长作为分类判别依据,却忽略了胶囊单元间的方向信息所包含的特征间相关性. 因此本文考虑加入其方向信息以构建输出目标函数,提出一种 Transition 层的映射方法,通过线性空间基变换将胶囊单元间空间关系映射至一维,再使用加性间距 Softmax 函数[11]计算胶囊单元间空间信息,通过空间夹角信息划分新的分类决策面构建输出目标函数,使相同类别胶囊单元间的方向性更加一致,提高模型分类识别能力.

2　加性间距胶囊网络模型

2.1　原始胶囊网络

胶囊网络中为解决底层特征与高层特征的空间联系问题,提出了使用权重矩阵方式控制底层胶囊单元与高层胶囊单元间的连接紧密程度,并且通过胶囊单元间方向一致性量化并反馈至权重矩阵,这一过程称为动态路由算法. 可有效避免特征的损失以及保证特征间的空间关系. 其动态路由总流程如算法 1 所示:

算法 1　动态路由算法

Input:$\hat{\boldsymbol{u}}_{j|i}, r, l$

Output:$\boldsymbol{v}_j$

对层 l 的所有胶囊单元 j 与层 $(l+1)$ 的所有胶囊单元 j 初始化:$\boldsymbol{b}_{ij} \leftarrow 0$;

for 迭代 r 次 do

对层 l 的所有胶囊单元 i:

$$\boldsymbol{c}_i \leftarrow \text{softmax}(\boldsymbol{b}_i);$$

对层 $(l+1)$ 的所有胶囊单元 j:

$$\boldsymbol{s}_j \leftarrow \sum_i \boldsymbol{c}_{ij}\hat{\boldsymbol{u}}_{j|i};$$

对层 $(l+1)$ 的所有胶囊单元 j:

$$\boldsymbol{v}_j \leftarrow \text{squash}(\boldsymbol{s}_j);$$

对层 l 的所有胶囊单元 j 与层 $(l+1)$ 的所有胶囊单元 j:

$$\boldsymbol{b}_{ij} \leftarrow \boldsymbol{b}_{ij} + \hat{\boldsymbol{u}}_{j|i} \cdot \boldsymbol{v}_j;$$

end

Return $\boldsymbol{v}_j$;

其中低层胶囊单元 $\boldsymbol{u}_i$ 与变换矩阵 $\boldsymbol{W}_{ij}$ 相乘后得到 $\hat{\boldsymbol{u}}_{j|i}$ 为预测向量. 设置初始权值矩阵 $\boldsymbol{b}_{ij}=0$,$\boldsymbol{b}_{ij}$ 表示低层胶囊单元 i 与高层胶囊单元 j 之间的连接权重,为了避免连接权重为 0 导致数值错误,对 $\boldsymbol{b}_{ij}$ 应用 softmax 函数得到 $\boldsymbol{c}_{ij}$,保证 $\boldsymbol{c}_{ij}\geqslant 0$,且 $\sum_j \boldsymbol{c}_{ij}=1$. 接着计算预测向量的加权和 $\boldsymbol{s}_j=\sum_i \boldsymbol{c}_{ij}\hat{\boldsymbol{u}}_{j|i}$,为保证输出向量的模长位于(0,1)之间,对 $\boldsymbol{s}_j$ 应用 squash 函数(公式(1))得到高层胶囊单元 $\boldsymbol{v}_j$. 最后根据底层胶囊与高层胶囊的方向一致特征来更新连接权重 $\boldsymbol{b}_{ij}=\boldsymbol{b}_{ij}+\hat{\boldsymbol{u}}_{j|i}\cdot\boldsymbol{v}_j$,为保证底层与高层胶囊单元的空间联系得到充分体现,继续迭代 r 次上述过程.

$$\boldsymbol{v}_j=\text{squash}(\boldsymbol{s}_j)=\frac{\|\boldsymbol{s}_j\|^2}{1+\|\boldsymbol{s}_j\|^2}\frac{\boldsymbol{s}_j}{\|\boldsymbol{s}_j\|} \tag{1}$$

图 1 为胶囊网络路由算法示意图(以三个胶囊单元输出为例). 其中绿线部分代表一个胶囊单元,每个胶囊单元包含着数个神经元,神经元代表单个特征,胶囊单元就代表一系列特征的集合. 在动态路由算法迭代 r 次之后,方向性相似的胶囊单元获得了来自底层胶囊单元更多的权重,此胶囊单元的模长更大,意味着对应分类的概率更大. 因此原始的胶囊网络目标函数如式(2)所示:

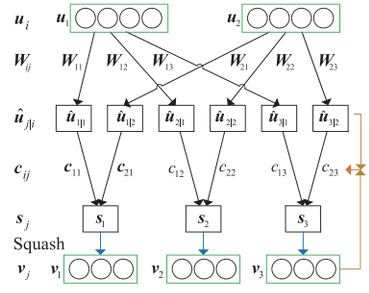

图1　胶囊网络动态路由示意图

$$L=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}T_{j==y_i}\max(0,m^{+}-\|\boldsymbol{v}_j\|)^2+\lambda(1-T_{j==y_i})\max(0,\|\boldsymbol{v}_j\|-m^{-})^2 \quad (2)$$

其中 $T_{j==y\_i}$ 代表分类 j 为正确类别时置 1，否则置 0. m^{+}，m^{-} 分别代表正确分类阈值和分类错误阈值.

2.1.1 加性间距损失函数

加性间距损失函数(Additive Margin Softmax)是一种改进的 Softmax 损失函数. 可以将分类的决策面扩大至设定值，使预测结果的类内向量的差异更小，类间向量差异更大，从而减少模型过拟合问题. 原始 Softmax 损失函数如式(3)所示.

$$\mathcal{L}_S=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i}}{\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{f}_i}}=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\boldsymbol{W}_{y_i}\boldsymbol{f}_i\cos(\theta_{y_i})}}{\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{W}_j\boldsymbol{f}_i\cos(\theta_j)}} \quad (3)$$

其中 $\boldsymbol{f}_i$ 代表输入至全连接层的第 i 个样本，$\boldsymbol{W}_j$ 代表全连接层第 j 列，y_i 代表样本正确类别索引，$\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i$ 第 i 个样本的目标概率.

加性间距 Softmax 通过替换原 Softmax 中的 cos 函数，获得到新的分类决策界. 它首先将 $\boldsymbol{W}$ 和 $\boldsymbol{f}_i$ 归一化至 1，得到 $x=\cos(\theta_{y_i})=\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i/\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i$，接着定义式(4)替换 cos 函数：

$$\psi(\theta)=\cos\theta-m=x-m \quad (4)$$

其中 m 为加性间距宽度参数，用于控制新的分类决策面. 由于将特征和权重都进行了归一化，会导致 x 值范围较小，因此加入新的超参数 s 作为尺度系数放大 x 的范围. 其直观的几何解释如图 2 所示. 原始的 Softmax 决策面被定义为 $\boldsymbol{P}_0$，且 $\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{P}_0=\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{P}_0$. 对于 AM-Softmax，决策面变成了一个边缘区域而不是单个向量，首先给定决策面参数 $m=(\boldsymbol{W}_1-\boldsymbol{W}_2)^{\mathrm{T}}\boldsymbol{P}_1=\cos(\theta_{W_1,P_1})-\cos(\theta_{W_2,P_2})$，那么类别 1 的决策面为 $\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{P}_1-m=\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{P}_1$. 如果进一步假设所有的类都具有相同的类内方差，并且类 2 的边界为 $\boldsymbol{P}_2$，那么得到 $\cos(\theta_{W_1,P_1})=\cos(\theta_{W_2,P_2})$，即加性间距宽度 $m=\cos(\theta_{W_1,P_1})-\cos(\theta_{W_1,P_2})$，代表此二分类的边缘区域的理想间距差值.

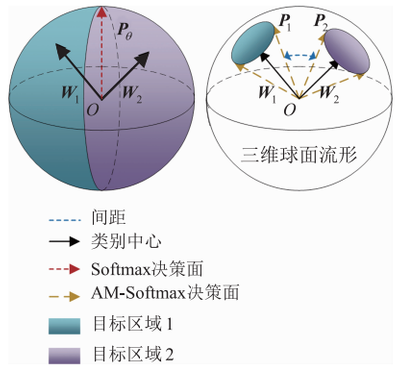

图2 原始softmax决策面与AM-softmax决策面对比

2.2 加性间距胶囊网络

本文提出基于加性间距的胶囊网络进行家庭活动识别，其网络模型如图 3 所示. 网络的输入样本为多通道音频中某一通道的音频对数 Mel 谱图，经过模型前向传播后输出音频分类预测标签. 提出的模型由两个卷积模块，一个池化层，一个初级胶囊层，一个高级胶囊层，一个 Transition 层和一个全连接层组成. 每个卷积模块包括一层卷积层、批归一化层、线性整流激活(Rectified Linear Unit，ReLU). 卷积层与池化层用于提取音频频率中的特征并减少总特征维度，胶囊网络层学习音频帧时序关系，Transition 层用于转换胶囊单元间方向关系矩阵至低维，全连接层用于计算方向夹角.

对数 Mel 谱图的第一维代表频率，第二维代表时间，因此首先使用 CNN 进行频率维度上的特征提取，经过第三层 CNN 后得到仅时间维度的特征. 接着重塑三维张量形状至 $P\times U$，即 P 个胶囊单元对应个音频帧，

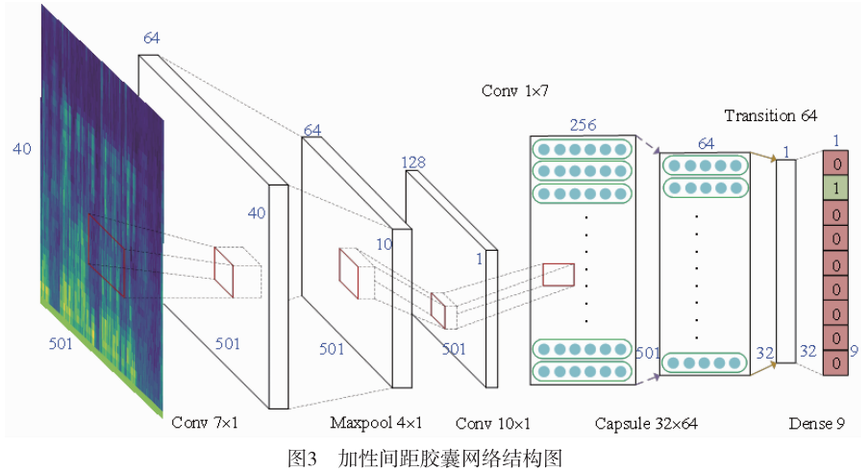

图3 加性间距胶囊网络结构图

其中包含 U 个特征元素. 此时胶囊单元间的空间关系,就代表着声音帧间的时序联系. 使用动态路由算法将底层特征整合并获得 Q 个长度为 V 的高层胶囊单元,高层胶囊单元代表着音频事件类别,底层胶囊单元与高层胶囊单元的方向差异作为反馈更新权重矩阵,既增加每个音频事件类别与相关声音帧间的权重,抑止与之不相关的声音帧间的权重,经过 r 次迭代后,权重矩阵即为声音帧与音频事件间关系的准确表征.

在传统的胶囊网络中,高层胶囊单元的模长代表着音频事件分类概率,其分类判别函数(公式(2))只简单的通过模长阈值来进行分类判别,这样就忽略胶囊单元向量的方向特性. 为使用其方向特性帮助模型做出更准确的分类判别,使用加性间距损失函数作为目标函数,他将同类目标的输出向量之间设定更小的方向夹角阈值,使同类目标输出向量的分布更加集中,以显著地减少模型过拟合,并且提升对于小样本的类别识别率. 但原始的加性间距函数并不直接适用与二维向量,因此本文提出使用名为 Transition 的层将胶囊网络输出的 $Q \times V$ 二维张量变基为一维向量,此向量代表胶囊单元间的空间关系,Transition 层如式(5)所示.

$$\boldsymbol{f} = \text{Transition}(\boldsymbol{v}) = \boldsymbol{v}\boldsymbol{t}^{\mathrm{T}} \tag{5}$$

其中向量 $\boldsymbol{f}$ 代表胶囊单元间空间关系向量,$\boldsymbol{v}$ 代表胶囊单元,t 为过渡矩阵.

将第 j 个空间关系向量 $\boldsymbol{f}_j$ 长度归一化后并与分类中心矩阵 $\boldsymbol{W}$(已归一化)进行相似计算,音频事件的余弦相似度为 $\cos(\theta_j) = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{f}_j$,因此第 j 类音频事件分类概率定义为:$y_j = \mathrm{e}^{\cos(\theta_j)} \Big/ \sum_{j=1}^{c} \mathrm{e}^{\cos(\theta_j)}$. 其中 $\boldsymbol{W}_j$ 代表对应音频事件的分类中心,c 代表总类别数. 此时利用所计算出的余弦相似度,使用加性间距宽度参数 m 构建新的决策面函数 $\psi(\theta) = \cos(\theta) - m$,为保证函数数值范围足够大,将其乘以尺度系数 s 得到 $\psi(\theta) = s \cdot (\cos(\theta) - m)$. 将 $\psi(\theta)$ 代替 $\cos(\theta)$ 后得到加性间距损失函数定义如式(6)所示.

$$\begin{aligned} \mathcal{L}_{AMS} &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\psi(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i)}}{\mathrm{e}^{\psi(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i)} + \sum_{j=1,j\neq y_i}^{c}\mathrm{e}^{s\cdot\cos\theta_j}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\psi(\boldsymbol{\theta}_{y_i})}}{\mathrm{e}^{\psi(\boldsymbol{\theta}_{y_i})} + \sum_{j=1,j\neq y_i}^{c}\mathrm{e}^{s\cdot\cos\theta_j}} \end{aligned} \tag{6}$$

其中 $\boldsymbol{\theta}_{y_i}$ 代表预测向量 $\boldsymbol{f}_i$ 与正确标签的分类中心向量 $\boldsymbol{W}_{y_i}^{\mathrm{T}}$ 间的夹角,利用加性间距宽度参数 m 对 θ_{y_i} 进行角度偏移得 $\psi(\theta_{y_i})$,即增大了相同类别向量间的损失值,使得相同类别向量逐渐向分类中心汇聚,从而减少过拟合,提升模型分类能力.

3　数据处理与实验设置

3.1　数据集

DCASE2018 挑战任务 5 数据集[12]收集了放置在厨房与客厅的 7 个麦克风阵列所记录的音频信号,要求使用多声道音频数据对家庭环境中日常活动进行事件分类,将由麦克风阵列获得的多声道音频片段分类到所提供的预定义类之一中,例如烹饪、看电视、工作等. 由于以上活动也可以由不同的声音事件组成,也可以被视为声学场景分类. 其传感器放置位置如图 4 所示. 它包含一个人在一周内的活动所产生的声音信号,一共 9 种活动类别. 它将每个活动分割成数个 10 秒的音频片段,并且忽略多个活动类的音频片段(例如活动间转换),因此每个音频片段都有单独的类别标签. 在 7 个麦克风阵列中,其中 4 个传感器采集的信号作为开发集,开发集中包含训练集与验证集,其他 3 个传感器采集的信号作为测试集.

在此数据集上已取得一些研究成果,如刘华平团队提出提取音频信号的对数 Mel 谱图特征、(Mel-frequency cepstral coefficients,MFCC)特征和 VGGish 特征,并进行特征融合学习,有效的提高了模型的准确率,最终验证集 F1 分数为 89.8%[13]. Tanabe 团队使用盲信号处理的前端模块进行盲源分离,后端采用一维卷积和 Visual Geometry Group 16 架构的神经网络,并使用数据增强来避免过拟合,验证集 F1 分数为 89.8%[14]. Inoue 等人提出使用音频混合以及片段洗牌的方式来进行数据增强,有效的抑止了 CNN 模型的过拟合,并且对样本较少的类别进行补充,消除了数据集类别样本不平衡的问题,最终验证集 F1 分数达到 90.0%[15].

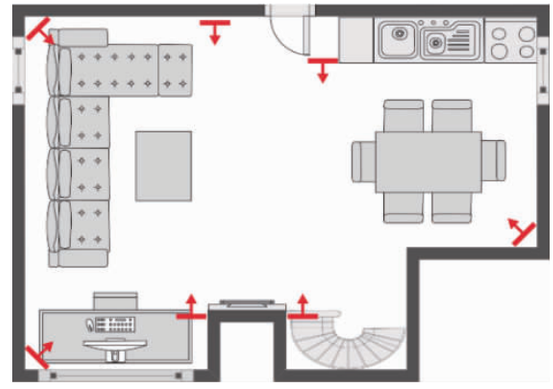

图4　传感器节点分布2D平面图

3.2　数据集预处理

本文对数据集中的所有的音频信号均采用对数 Mel 滤波的方式进行预处理,音频采样率为 16kHz. 首先进行帧长度为 64ms,步长为 20ms 的短时傅里叶变换. 再使用 40 个频带的 Mel 滤波器组进行滤波,最终得到 40×501 的特征矩阵. 图 5 为每个类别的典型对数 Mel 谱图.

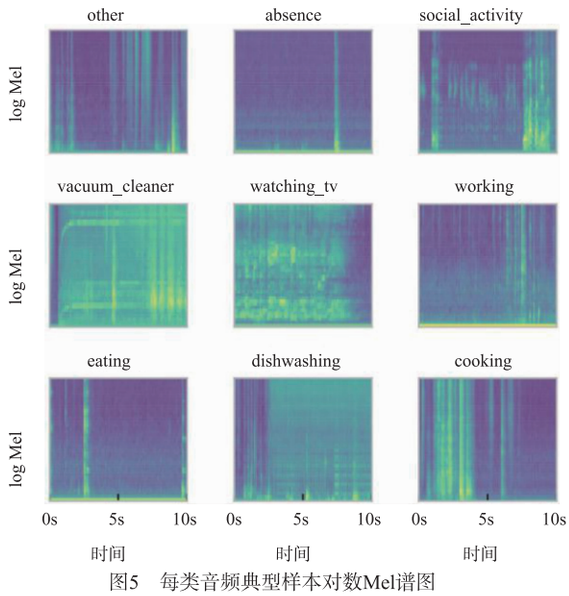

图5 每类音频典型样本对数Mel谱图

3.3 模型构建

使用预处理后的对数 Mel 谱图作为样本训练加性间距胶囊网络(Caps-ams),在对比算法的选取上,使用主流的CNN基线模型[16]、门控CNN模型(Gated Convolutional Neural Network,GCNN)[17]、胶囊网络模型(Caps),并在 CNN 与 GCNN 模型中加入了本文提出的输出目标函数,形成加性间距基线模型(基线-ams)、加性间距 GCNN 模型(GCNN-ams). 对于使用加性间距损失的模型,损失函数参数均设置为 $m=0.35$,$s=30$;其他模型损失函数参数均依据原论文设置. 训练参数均相同,优化算法为 Adam,初始学习率为 0.0001,批处理大小为 64,一共训练 100 轮.

3.4 实验分析

用不同模型对验证集数据进行分类,将得到的分类向量归一化后绘制在超球面中,结果如图 6 所示,不同颜色代表不同类别的分类向量,可见加入新的约束条件的模型,具有较大的分类向量类间间距,且分类效果要优于原始模型,而原始胶囊网络由于缺乏分类时的方向关系约束,其分类向量分布较不集中,当使用本文提出的 Caps-ams 模型,其分类向量分布明显更加集中,并能将不同类别的事件较好的区分开来,表明胶囊单元经 Transition 层映射能体现其空间关系,使得模型分类向量类内间距减小,类间间距增大,具备更好的分类决策面.

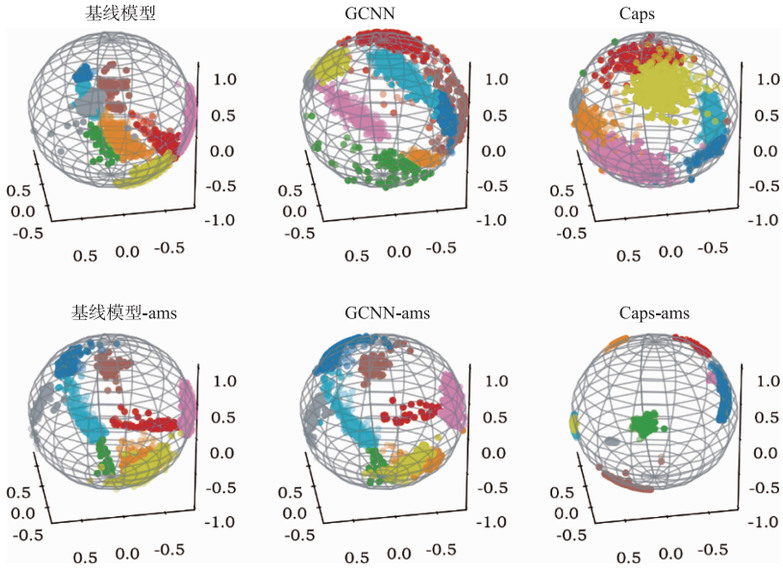

图6 模型中间向量 f 分布对比图

表 1 和表 2 分别是不同模型对验证集和测试集事件识别的 F1 分数,其中 Caps-ams 模型具有最高的 F1 分数,验证集 F1 平均分数为 92.3%,测试集 F1 平均分数为 88.8%,相比 Caps 模型,在验证集和测试集上的 F1 平均分数分别提高了 5% 和 2.2%. 同时采用了新约束条件的模型的分类效果,相比原始模型也得到了提升,基线-ams 模型、GCNN-ams 比基线模型与 GCNN 模型分别提升了 2.5%、2.1%.

然后对验证集和测试集发生误判的事件进行统计分析,每类中选取 2000 个样本进行预测,汇总当前所有预测结果中错误的个数,统计错误结果的真实标签个数与比例,结果如表 3 和表 4 所示,相比其他模型,本模型的误判次数最少,减少次 200 次以上. 再对误判的事件进行区分,选取了误判率最大的三个事件,从表中可

见,所有的模型误判最大的事件均为“工作”、“缺席”、“其他”三者间混淆,主要是因为三者部分音频片段信号过于相似导致,其他模型中缺乏更强的分类约束条件对于这些样本易产生过拟合问题,而本模型同时利用音频片段间的时序关系,和根据 Transition 层所映射的空间关系作为辅助分类条件做出进行分类识别,在一定程度上加强了模型的正则化,减少了对非同类相似样本的过拟合问题,从而提升了准确率.

表 1　验证集各模型各类别验证 F1 分数

模型种类	缺席/%	烹饪/%	洗碗/%	吃饭/%	其他/%	社交活动/%	真空吸尘/%	看电视/%	工作/%	均值/%
基线	85.4	95.1	76.7	83.6	44.7	93.9	99.31	99.6	82.0	84.5
基线-ams	88.6	96.1	78.3	87.6	48.7	95.5	99.7	99.9	89.0	84.1
GCNN	89	96.9	83.5	85.8	49.7	96.9	100	99.51	87.2	87.6
GCNN-ams	90.8	97.7	86.5	88.4	59.0	96.5	100	99.37	89.1	89.7
Caps	84.6	97.3	85.9	86.4	52.2	96.5	100	99.7	83.6	87.3
Caps-ams	**90.3**	**98.3**	**90.9**	**91.3**	**73.1**	**97.6**	**100**	**99.7**	**89.3**	**92.3**

表 2　测试集各模型各类别验证 F1 分数

模型种类	缺席/%	烹饪/%	洗碗/%	吃饭/%	其他/%	社交活动/%	真空吸尘/%	看电视/%	工作/%	均值/%
基线	87.7	93.0	77.2	81.2	35.0	96.6	95.8	99.9	81.4	83.1
基线-ams	93.2	85.1	76.3	76.3	46.2	97.1	96.1	99.9	86.7	84.1
GCNN	93.6	87.4	79.7	80.1	50.8	97.6	96.7	99.9	87.7	86.0
GCNN-ams	92.1	90.4	82.1	84.2	52.4	97.3	96.7	100	86.5	86.9
Caps	91.0	95.0	81.5	82.1	52.1	97.5	95.2	100	85.2	86.6
Caps-ams	**92.9**	**96.9**	**88.3**	**85.3**	**55.9**	**97.8**	**95.0**	**100**	**86.7**	**88.8**

表 3　验证集误判率统计

模型名称	误判事件 1 个数与比例	误判事件 2 个数与比例	误判事件 3 个数与比例	误判总数
基线	工作:483(35.0%)	缺席:304(22.0%)	其他:291(21.1%)	1382
基线-ams	工作:408(32.8%)	缺席:263(21.2%)	其他:281(22.6%)	1245
GCNN	工作:400(32.0%)	缺席:276(22.1%)	其他:265(21.2%)	1251
GCNN-ams	工作:361(31.5%)	缺席:251(21.9%)	其他:205(17.9%)	1149
Caps	工作:393(35.2%)	缺席:267(23.9%)	其他:186(16.7%)	1119
Caps-ams	工作:294(34.8%)	缺席:242(28.7%)	其他:175(20.8%)	845

表 4　测试集误判率统计

模型名称	误判事件 1 个数与比例	误判事件 2 个数与比例	误判事件 3 个数与比例	误判总数
基线	工作:418(30.3%)	缺席:371(26.9%)	其他:338(24.5%)	1941
基线-ams	工作:429(34.5%)	缺席:282(22.7%)	其他:324(26.1%)	1802
GCNN	工作:452(36.2%)	缺席:282(22.6%)	其他:242(19.4%)	1889
GCNN-ams	工作:368(32.1%)	缺席:278(24.2%)	其他:265(23.1%)	1741
Caps	工作:373(33.4%)	缺席:306(27.4%)	其他:231(20.7%)	1831
Caps-ams	工作:293(34.7%)	缺席:250(29.6%)	其他:177(21.0%)	1601

4　结论

在本文中提出了加性间距胶囊网络模型用于多通道音频事件分类任务,针对传统 CNN 对音频信号时序特征提取困难,提出使用胶囊单元表征每个音频帧;针对传统网络对局部与整体相对关系不敏感,提出使用动态路由算法进行音频帧与音频事件间相对关系的整合,从而准确反应出局部与整体间的关系;针对胶囊网络最终分类时丢弃了相对关系特性,提出使用 Transition 层映射胶囊单元间空间关系并结合加性间距损失

函数将空间关系用于分类时的辅助判别,从而减少模型过拟合问题. 通过实验得到模型在验证集与测试集上的 F1 得分分别为 92.3% 和 88.8%,证明所提出的模型具有更强的泛化能力. 未来计划研究改进胶囊单元间的路由算法,使其获得更准确空间关系. 并推广胶囊网络至自编码器中,使其可进行无监督的音频特征学习与音频事件聚类.

参考文献

[1] Nathan V, Paul S, Prioleau T, et al. A survey on smart homes for aging in place: Toward solutions to the specific needs of the elderly[J]. IEEE Signal Processing Magazine, 2018, 35(5): 111 – 119.

[2] Sophiya E, Jothilakshmi S. Large scale data based audio scene classification[J]. International Journal of Speech Technology, 2018, 21(4): 825 – 836.

[3] Ferguson E L, Ramakrishnan R, Williams S B, et al. Deep learning approach to passive monitoring of the underwater acoustic environment[J]. The Journal of the Acoustical Society of America, 2016, 140(4): 3351 – 3351.

[4] Kasnesis P, Tatlas N A, Mitilineos S A, et al. Acoustic sensor data flow for cultural heritage monitoring and safeguarding[J]. Sensors, 2019, 19(7): 1629.

[5] Lapuschkin S, Wäldchen S, Binder A, et al. Unmasking clever hans predictors and assessing what machines really learn[J]. Nature communications, 2019, 10(1): 1 – 8.

[6] Keren G, Schuller B. Convolutional RNN: An enhanced model for extracting features from sequential data[A]. 2016 International Joint Conference on Neural Networks [C]. Canada: IEEE, 2016. 3412 – 3419.

[7] Chew J, Sun Y, Jayasinghe L, et al. DCASE 2018 Challenge: Solution for task 5[R]. DCASE2018 Challenge, Tech. Rep, 2018.

[8] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[A]. Advances in neural information processing systems[C]. US: NIPS, 2017. 3856 – 3866.

[9] 任开旭,王玉龙,刘同存,李炜. 融合多维语义表示的概率矩阵分解模型[J]. 电子学报,2019,47(9):1848 – 1854.
REN Kai-xu, WANG Yu-long, LIU Tong-cun, LI Wei. A probabilistic matrix factorization model based on multidimensional semantic representation learning[J]. Acta Electronica Sinica, 2019, 47(9): 1848 – 1854. (in Chinese)

[10] 贾旭东,王莉. 基于多头注意力胶囊网络的文本分类模型[J]. 清华大学学报(自然科学版),2020,60(5):415 – 421.
JIA Xudong, WANG Li. Text classification model based on multi-head attention capsule networks[J]. Journal of Tsinghua University (Science and Technology), 2020, 60(5): 415 – 421. (in Chinese)

[11] Wang F, Cheng J, Liu W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926 – 930.

[12] Dekkers G, Lauwereins S, Thoen B, et al. The SINS database for detection of daily activities in a home environment using an acoustic sensor network[A]. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop [C]. Germany: DCASE. 2017. 32 – 36.

[13] Liu H, Wang F, Liu X, et al. An ensemble system for domestic activity recognition[R]. DCASE2018 Challenge, Tech. Rep, 2018.

[14] Tanabe R, Endo T, Nikaido Y, et al. Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling [R]. DCASE2018 Challenge, Tech. Rep, 2018.

[15] Inoue T, Vinayavekhin P, Wang S, et al. Domestic activities classification based on cnn using shuffling and mixing data augmentation[R]. DCASE2018 Challenge, Tech. Rep, 2018.

[16] Dekkers G, Vuegen L, van Waterschoot T, et al. DCASE 2018 Challenge-Task 5: Monitoring of domestic activities based on multi-channel acoustics[R]. DCASE2018 Challenge, Tech. Rep, 2018.

[17] Yuhan Shen, Kexin He, Weiqiang Zhang. Home activity monitoring based on gated convolutional neural networks and system fusion[R]. DCASE2018 Challenge, Tech. Rep, 2018.

作者简介

郑启航 男,1996 年出生,浙江绍兴人,硕士研究生. 主要研究方向为音频分类、半监督学习.

E-mail: 18000632@smail.cczu.edu.cn

刘半藤(通信作者) 男,1984 年出生,浙江余姚人,博士,副教授,硕士生导师,主要研究方向为信号处理、无损检测.

E-mail: hupo3@sina.com