基于注意力胶囊网络的家庭活动识别

王金甲[1,2]　纪绍男[1,2]　崔 琳[1,2]　夏 静[1,2]　杨 倩[1,2]

摘　要　本文提出了一种注意力胶囊网络的新框架利用录音识别家庭活动. 胶囊网络可以通过动态路由算法来选择基于每个声音事件的代表性频带. 为了进一步提高其能力, 我们在胶囊网络中加入注意力机制, 它通过加权来增加对重要时间帧的关注. 为了评估我们的方法, 我们在声学场景和事件的检测和分类 (Detection and Classification of Acoustic Scenes and Events, DCASE)2018 挑战任务 5 数据集上进行测试. 结果表明, F1 平均得分可达 92.1 %, 优于几个基线方法的 F1 得分.

关键词　DCASE 2018 挑战, 声音事件分类, 家庭活动识别, 胶囊网络, 注意力

引用格式　王金甲, 纪绍男, 崔琳, 夏静, 杨倩. 基于注意力胶囊网络的家庭活动识别. 自动化学报, 2019, **45**(11): 2199−2204

DOI　10.16383/j.aas.c180721

Domestic Activity Recognition Based on Attention Capsule Network

WANG Jin-Jia[1,2]　JI Shao-Nan[1,2]　CUI Lin[1,2]
XIA Jing[1,2]　YANG Qian[1,2]

Abstract　In this paper, a novel framework of attention capsule network is proposed, which uses sound recordings to identify domestic activities. The capsule network can select a representative frequency band based on each sound event by the dynamic routing algorithm. To further improve its ability, we add attention mechanism to the capsule network. It can increase the focus on significant time frames by weighting. To evaluate our approach, we test it on the dataset of task 5 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Challenge. The results show that the average F1 score can reach 92.1 %, outperforming several baselines.

Key words　DCASE 2018 challenge, sound event classification, domestic activity recognition, capsule network, attention

Citation　Wang Jin-Jia, Ji Shao-Nan, Cui Lin, Xia Jing, Yang Qian. Domestic activity recognition based on attention capsule network. *Acta Automatica Sinica*, 2019, **45**(11): 2199−2204

全球正在面临人口老龄化的问题, 预计到 2050 年, 64 岁及以上的人口将超过世界人口的 20 %. 据调查显示, 有 40 % 的老年人将独自居住在自己家中[1]. 这将导致许多社会问题, 例如疾病和卫生保健费用的增加、护理人员的短缺以及无法独立生活的人数增加. 因此, 开发环境智能辅助生活工具帮助老年人独立在家中生活是势在必行的[2]. 基于音频的家庭活动识别是一个新问题, 也是声音事件分类的一个新兴应用领域. 声音事件分类将语义标签与音频流相关联, 并识别产生它的事件. 用于家庭活动识别的声音事件分类系统能够预测对应的活动事件. 声音事件分类问题在基于人工智能 (Artificial intelligence, AI) 的机器人导航、智能驾驶、监测家庭活动及老年人生活等方面有重要应用[3].

传统的声音事件分类方法是从音频信号中提取预先设计的人工特征用于训练分类器[4]. 这种方法在很大程度上依赖于预先设计特征的能力, 而这需要大量信号处理方面的专业知识. 事实上, 鉴于现实生活中遇到的问题和特殊情况的高度多样性, 这种方法在许多问题中既没有效率也没有可持续性[5].

基于深度学习的声音事件分类方法采用端到端的深度神经网络实现自动特征提取和分类. 近年来, 基于卷积神经网络 (CNN) 和循环神经网络 (RNN) 等深度学习方法在声音事件分类方面显示出良好的性能, 并且卷积循环神经网络 (CRNN) 结合了 CNN 和 RNN 也已经获得了较先进的声音事件分类性能. 例如, Hershey 等通过将不同结构的 CNN 用于音频分类任务中, 发现以前应用于图像分类的 CNN 在音频分类任务中也表现良好, 并且更大的训练和标签集有助于达到更好的分类效果[6]. Parascandolo 等提出了一种基于双向长短时记忆 (Bi-LSTM) 循环神经网络用于复音声音事件检测, 并在来自不同日常环境的不同类别的音频样本上进行测试, 显示出了很好的效果[7]. Cakir 等提出了将卷积循环神经网络应用到复音声音事件检测任务中, 结果显示 CRNN 方法优于先前只用 CNN 和 RNN 的方法[8]. 徐勇等在 DCASE 2016 任务 4 弱监督音频标记问题中, 在卷积循环神经网络上加入注意力和定位方案[9]; 在 DCASE 2017 任务 4 弱监督声音事件检测问题中提出了门控卷积循环神经网络模型, 其中可学习的门控线性单元可以帮助选择对应于最终标签的最相关特征, 获得竞赛第一名的成绩[10].

DCASE 2018 挑战任务 5 是用于家庭环境中日常活动识别问题的多声道声音事件分类任务, 该任务的目标是将由麦克风阵列获取的多声道音频段分类为所提供的预定义类之一, 这些类是在家庭环境中进行的日常活动 (例如 "烹饪"). 这个任务的重点在于可以利用多声道音频系统来识别家庭活动, 多麦克风信号处理技术可以有效地提高音频分类的鲁棒性[11], 由于多个声音事件的并发性, 多声道音频分类是一项具有挑战性的任务. 该任务的基线系统使用了两个卷积层和一个全连接层的结构[12]. Kong 等使用了 AlexNetish 和 VGGish 的卷积神经网络, 更深网络层的 VGGish 模型有更好的性能, 这说明 VGG 模型不仅能够在大规模图像数据集上分类效果很好, 在音频数据集上的推广能力也非常出色[13]. 在此竞赛中并列第一名的两个团队是 Tanabe 团队和 Inoue 团队. Tanabe 等所提出的系统是基于盲信号处理的前端模块和基于机器学习的后端模块的组合方法. 为了避免过拟合, 前端模块采用盲去混响, 盲源分离等, 它们使用空间线索而无需机器学习. 后端模块采用基于一维卷积神经网络 (1DCNN) 的架构和基于 VGG16 的架构. 所有的网络概率输出进行集成[14]. Inoue 等提出了数据增强的前端模块和基于 CNN 分类方法的后端模块的组合方法. 首先, 它通过混洗和混合声音片段来增强输入数据, 这种数据增强方法有助于增加训练样本的变化, 并减少不平衡数据集的影响. 其次, 使用 CNN 深度学习模型作为分类器, CNN 模型输入是增强后数据的对数 Mel 语谱图[15].

收稿日期 2018-11-12　录用日期 2019-04-15
Manuscript received November 12, 2018; accepted April 15, 2019
国家自然科学基金 (61473339), 首批 "河北省青年拔尖人才" 项目 ([2013]17) 和京津冀基础研究合作专项 (F2019203583) 资助
Supported by National Natural Science Foundation of China (61473339), The First Batch of "Top Young Talents in Hebei Province" ([2013]17) and Basic Research Cooperation Projects of Beijing, Tianjin and Hebei (F2019203583)
本文责任编委 吴建鑫
Recommended by Associate Editor WU Jian-Xin
1. 燕山大学信息科学与工程学院 秦皇岛 066004　2. 河北省信息传输与信号处理重点实验室 秦皇岛 066004
1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004　2. Hebei Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao 066004

总的来说, CNN 是将局部特征提取进行处理, RNN 是对局部特征之间的时间依赖性进行建模, 尽管它们在很多方面取得了成功, 但是由于 CNN 网络对各个部件的朝向和空间上的相对关系并不敏感, 它只在乎有没有相应的特征, 所以 CNN 不能很好地反映部分和整体的关系. 加之各个特征的重叠性, 现有的深层学习技术仍然不足以将单个声音事件从它们的混合物中分离出来, 所以取得的效果并不是很理想. 而且 CNN 和 RNN 都不能很好地减少过拟合. 胶囊网络是 Hinton 在 2017 年提出的, 胶囊是一组神经元, 其表示特定类型的对象或对象部分的实例化参数[16]. 胶囊网络的一个主要优点是它提供了一种类似于人类感知系统的方法, 可以很简单地通过识别其部分来识别整体. 对于 DCASE 2018 任务 5, 我们使用胶囊路由机制的神经网络架构来完成.

在该网络中, 胶囊层为每个声音事件选择代表性的频带, 低级胶囊通过权值矩阵对高级胶囊所代表的事件类别进行预测, 如果该预测向量与高级胶囊层中某个胶囊的输出有较大点积值, 则通过反馈来增加胶囊与该高级胶囊的耦合系数, 并降低与其他胶囊的耦合系数从而可以准确地反映部分和整体的关系. 与最大池化实现的原始路由形式相比, 胶囊路由可以避免忽视除最显著特征之外的其他特征, 可有效地减少特征损失[16]. 另一个创新是在胶囊网络中的初级胶囊层后加入了注意力层, 它可以通过加权来提高对显著部分的关注度, 即可以自动选择音频事件类最相关的重要帧, 同时忽略不相关帧 (例如, 背景噪声段). 我们提出的注意力层通过对时间片的显著性选择实现了注意力机制, 从而减少了模型过拟合.

1 注意力胶囊网络模型

1.1 胶囊网络的动态路由

胶囊网络和标准神经网络的重要区别在于胶囊的激活是基于多个输入姿态预测之间的比较, 而在标准神经网络中, 它是基于单个输入活动向量和学习到的权重矢量之间的比较. 解决部分和整体关系问题的一种方法是找到高维投票的紧密聚类, 这个方法称为路由协议. 不同于 CNN 的输入输出形式, 也不同于 CNN 的池化操作, 胶囊层的输入输出均为向量形式, 并且采用了动态路由算法, 来对这些向量进行运算.

胶囊网络每一层有若干节点, 每个节点表示一个胶囊. 低级胶囊连接到更高级别胶囊的过程中, 连接权值会在学习中发生变化, 由此引起节点连接程度的变化, 因此称为动态路由. 通常, 在两层胶囊之间用动态路由算法对该网络进行训练. 以下是我们描述的动态路由算法[16].

如算法 1 所示, 已知前一层胶囊层的预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 为输入预测向量, 它是通过权重矩阵 $\boldsymbol{W}_{ij}$ 乘以前一层胶囊层的输出向量 $\boldsymbol{u}_i$ 计算得到的, 即 $\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i$. 设置初始权重 $b_{ij} = 0$, b_{ij} 表示第 i 个低级胶囊到第 j 个高级胶囊的连接权重. 迭代过程中, 首先对权重 b_{ij} 应用 softmax 函数得到 c_{ij} 并保证了 c_{ij} 均为非负数, 且 $\sum_j c_{ij}$=1; 其次用 $\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}$ 来计算前一层胶囊层的所有预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 的加权和; 再次对 $\boldsymbol{s}_j$ 应用 squashing 函数[15] 得出输出向量 $\boldsymbol{v}_j$; 最后根据公式 $b_{ij} = b_{ij}+\hat{\boldsymbol{u}}_{j|i}\cdot\boldsymbol{v}_j$ 更新相应的权重 b_{ij}. 重复这个过程直至收敛.

算法 1. 动态路由算法

Input: Prediction vectors $\hat{\boldsymbol{u}}_{j|i}$,layer l, max iterations r

Output: Layer $(l+1)$ capsules $\boldsymbol{v}_j$

1) Initialization: $b_{ij} = 0$

2) **For** r iterations **do**

3) $c_{ij} = soft\max(b_{ij})$

4) $\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j|i}$

5) $v_j = squash(s_j)$

6) $b_{ij} = b_{ij} + \hat{\boldsymbol{u}}_{j|i}\cdot\boldsymbol{v}_j$

7) End for

胶囊路由的概念图如图 1 所示, 圆圈为单个神经元, 虚线圈出的为一个胶囊. 胶囊可以代表实体, 左侧 L 层两个胶囊分别表示人的左右胳膊, 从实线箭头可以看出正确朝向的左胳膊对应右侧 $(L+1)$ 层胶囊的人体上半身构造, 而虚线箭头表示不能对应. 两个胶囊层之间通过识别局部的器官, 学习到局部和整体的关系, 然后找到正确的人体上半身结构.

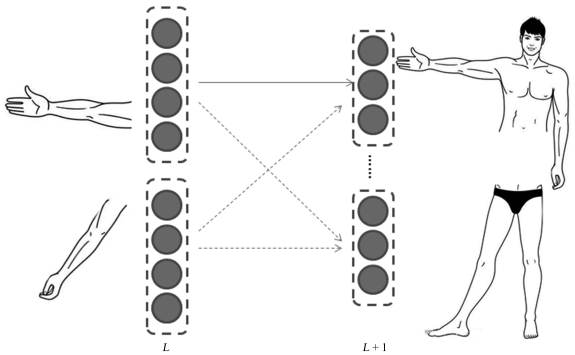

图 1 胶囊路由的概念图

Fig. 1 Conceptual diagram of capsule routing

1.2 注意力机制

注意力机制可以从大量信息中选择出对当前任务目标更关键的信息, 并抑制不相关的信息, 从而减少了过拟合问题. 图像处理中的注意力机制关注空间注意力, 我们提出的方法关注时间注意力. 注意力模块用 sigmoid 作为激活函数, 能在选择重要特征的同时抑制不相关的信息[9]. 它也可以帮助平滑训练集和测试集之间不匹配的问题. 第 t 帧的注意力因子 $z(t)$ 表示当前音频帧对音频类的重要程度. $z(t)$ 的输出值为 0 到 1 之间. 当 $z(t)$ 接近 1 时, 对应 t 时刻帧作为重要帧被选择, 当 $z(t)$ 接近 0 时, 对应 t 时刻帧作为不相关帧被忽略. 通过这种方法, 网络可以关注音频片段中的音频类事件帧, 忽略噪声帧. $z(t)$ 定义为:

$$z(t) = \sigma(w * x(t) + b) \tag{1}$$

其中, $x(t)$ 为输入特征, w 为权重矩阵, b 为偏置参数, σ 是 sigmoid 非线性激活函数. 通过训练网络来更新参数 w 和 b.

1.3 提出的网络模型

本节提出了注意力胶囊网络模型来进行家庭活动识别. 网络模型如图 2 所示, 首先将音频片段转变成对数 Mel 语谱图, 其次将对数 Mel 语谱图输入到提出的注意力胶囊神经网络模型, 最后模型输出是音频标签预测值.

提出的注意力胶囊网络模型由三个门控卷积模块, 一个初级胶囊层, 一个高级胶囊层, 一个注意力层和一个融合层组成. 每个门控卷积模块由两层门控卷积网络和最大池化组成, 每层门控卷积网络包括线性 (linear) 函数和 sigmoid 激活函数. 与传统的 CNN 相比, 门控卷积网络用门控线性单元 (GLUs) 取代了修正线性单元 (ReLU). 这个可学习的门能控制当前层传入下一层的信息量[10]. GLUs 能减少梯度消失现

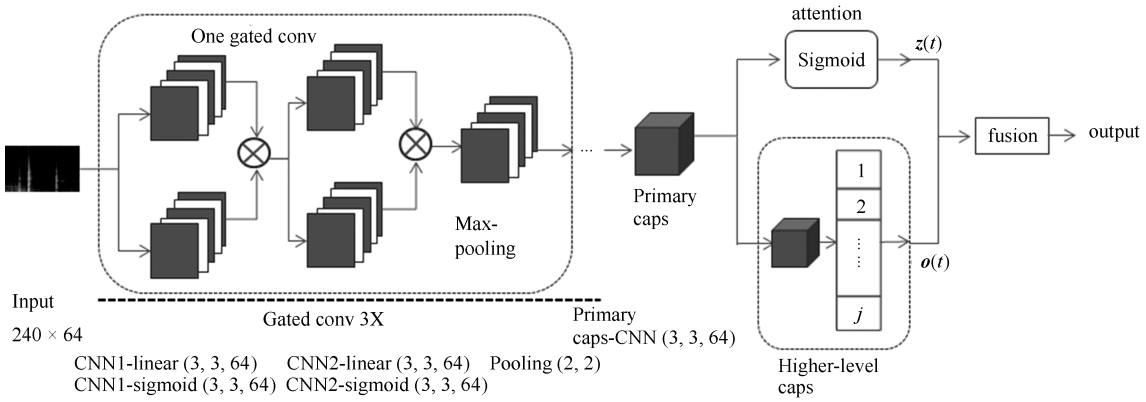

图 2　注意力胶囊网络模型

Fig. 2　Attention capsule network model

象[17], 这是通过用 sigmoid 激活函数保留了神经网络的非线性能力, 同时用线性 (linear) 函数为梯度提供线性路径来实现的. 最大池化操作能减少特征的空间维度.

经过三个门控卷积模块的输出特征被送入初级胶囊层. 初级胶囊层由卷积模块, 重塑模块和 squashing 模块组成. 输入特征先经过卷积层, 加入偏差之后, 又经过 ReLU 非线性激活函数, 然后重塑为一个 $T \times V \times U$ 的三维张量, 并用 squashing 函数压缩. T 是重塑前的时间维度, V 是从其他变量推测出的维度, $U = 4$ 是胶囊的大小. 也就是说初级胶囊层的输出有 T 个时间片, 每个时间片有 V 个胶囊, 每个胶囊是 $1 \times 1 \times U$ 的张量.

将每个时间片的 V 个胶囊输入高级胶囊层. 在初级胶囊层和高级胶囊层之间使用动态路由算法进行计算. 动态路由算法将 V 个代表音频帧的低级胶囊与 J 个代表事件类别的高级胶囊进行匹配. 当多个音频帧都预测到同一事件后, 则确定出音频事件的类别. 然后通过反馈来增加与该音频事件相关音频帧之间的权重, 并降低与该音频事件不相关音频帧的权重, 从而准确地学习到所有音频帧和音频事件之间的权重. 每一次训练, 路由算法的权重都会更新, 算法结束时保存最终权重. 用动态路由算法计算输出向量 $\boldsymbol{v}_j$, 再算出输出向量 $\boldsymbol{v}_j$ 的欧氏长度. 每个时刻 t 的所有 J 个类别的欧氏长度组成向量作为高级胶囊层的输出, 记为 $\boldsymbol{o}(t)$.

将每个时间片的 V 个胶囊输入注意力层. 注意力层可以让网络模型更专注地找出与音频事件类相关的输入音频的显著帧. 该层的 sigmoid 激活函数能够预测出每帧的重要性, 每个时刻 t 的注意力层输出为 $\boldsymbol{z}(t)$, $\boldsymbol{z}(t)$ 的值在 0 到 1 之间. 注意力层在抑制音频事件类不相关帧的同时选择显著帧. 时间注意力机制就是通过注意力层的输出来实现的.

最后是融合层, 将高级胶囊层的输出 $\boldsymbol{o}(t)$ 与注意力层的输出 $\boldsymbol{z}(t)$ 合并. 对时间片的显著帧选择实现时间注意力机制, 注意力因子大的时间片对应着类相关显著音频帧, 注意力因子小的时间片对应着类不相关的音频帧. 通过计算高级胶囊层的输出 $\boldsymbol{o}(t)$ 和注意力因子 $\boldsymbol{z}(t)$ 的加权和得到最终的预测输出 y_j. y_j 表示第 j 类音频类事件的预测值, 表达式如下:

$$y_j = \frac{\sum_{t=1}^{T} o_j(t) z_j(t)}{\sum_{t=1}^{T} z_j(t)} \tag{2}$$

其中, $o_j(t)$ 表示时刻 t 的第 j 个胶囊输出向量 $\boldsymbol{v}_j$ 的欧氏长度, $z_j(t)$ 表示时刻 t 的第 j 类注意力因子, $j = 1, \cdots, J$, $t = 1, \cdots, T$. $\boldsymbol{z}(t)$ 控制了 $\boldsymbol{o}(t)$ 传送信息中的显著音频帧. 选择一个概率阈值 τ, 当 $y_j > \tau$ 时, 输出是第 j 类音频活动事件.

2　实验

2.1　数据集

此次任务使用的是 DCASE 2018 任务 5 数据集, 它是 SINS 数据集的派生数据[18]. 对于这项任务, 在起居室和厨房混合区域使用了 7 个麦克风阵列组成网络收集音频, 每个麦克风阵列由 4 个线性排列的麦克风组成. 图 3 显示了声音录制环境的平面图以及使用的传感器节点的位置.

此数据集包含一个人一周住在度假屋中的连续录音, 这个连续录音被分成 10 s 的音频段, 包含多于一个活动类 (例如两个活动间的转换) 的音频段被忽略了, 这意味着每个音频段仅代表一个活动. 这些音频段和对应的类别标签作为单独的文件被提供. 每个音频段包含 4 个声道 (例如来自特定节点的 4 个麦克风声道). 这个 9 类任务的日常活动如表 1 所示, 表 1 中还包括开发集和评估集中每类活动的 10 s 片段的数量.

表 1　开发集和评估集音频数量

Table 1　Development set and evaluation set audio quantity

活动	开发集样本数	评估集样本数
缺席	18 860	21 112
烹饪	5 124	4 221
洗碗	1 424	1 477
吃饭	2 308	2 100
其他	2 060	1 960
社会活动	4 944	3 815
真空吸尘	972	868
看电视	18 648	21 116
工作	18 644	16 302
总计	72 984	71 971

2.2　特征提取

我们此次实验采用的特征提取方法是目前音频处理最常用的对数 Mel 滤波[19−20]. 在提取特征之前, 我们将每个剪辑的音频以 16 kHz 重新采样, 然后进行短时傅里叶变换得到语谱图; 其次我们生成一个 64 频带的 Mel 滤波器组; 将语谱

图和 Mel 滤波器组相乘, 并进行对数运算, 得到对数 Mel 语谱图. 即每个 10 s 音频样本产生一个 240 × 64 的特征向量. 图 4 是我们列举的每类活动的对数 Mel 语谱图.

2.3 实验设置

在训练阶段, 我们在预测标签和录音的真实标签之间应用对数交叉熵损失函数. 神经网络的权值可以通过反向传播计算的权值梯度来更新. 损失定义为:

$$E = -\sum_{n=1}^{N}(\boldsymbol{P}_n \log \boldsymbol{O}_n + (1-\boldsymbol{P}_n)\log(1-\boldsymbol{O}_n)) \quad (3)$$

其中, E 是对数交叉熵损失, $\boldsymbol{O}_n$ 和 $\boldsymbol{P}_n$ 表示样本索引 n 处的预测和真实类别标签向量, 批处理大小用 N 表示. 我们采用 Adam 作为随机优化方法, 初始学习率为 0.001, 以 0.9 的衰减率每两轮衰减一次学习率. 批处理的大小为 64, 总共训练了 30 轮.

2.4 实验结果

我们此次实验折叠了四次开发集数据, 三折数据集用于训练模型, 一折数据集用于预测结果, 然后计算四折结果的平均值. 重复该过程 10 次计算预测结果的平均值, 得到开发

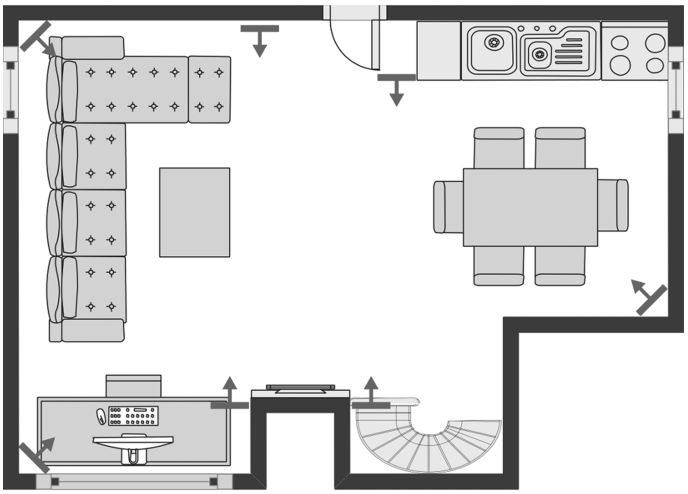

图 3 具有传感器节点的厨房和客厅混合的 2D 平面布置图

Fig. 3 2D floorplan of the combined kitchen and living room with the used sensor nodes

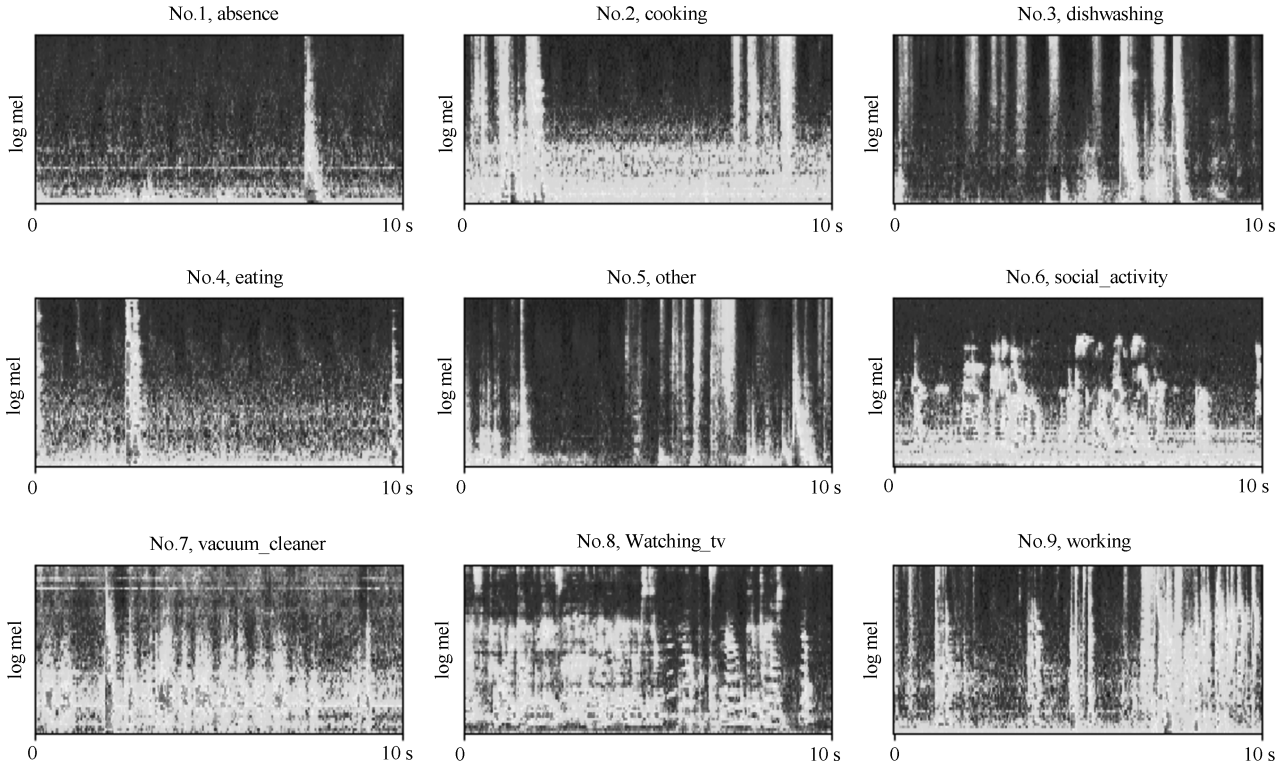

图 4 各类活动的对数 Mel 语谱图

Fig. 4 Logmel spectrum of various activities

集上模型的 F1 得分. 这样更好地避免了偶然性, 让实验结果更具有说服力. 最后我们在评估集上进行了测试, 得到了各模型的评估集 F1 得分.

表 2 显示了 5 个不同模型在开发集上各类活动的 F1 得分, 表 3 是评估集上各模型平均 F1 得分. 其中基线系统是简单的两层卷积结构[12]. GCRNN 是在卷积循环神经网络基础上加了门控线性单元. GCRNN-att 是 GCRNN 后端加上了前文提到的注意力模块. Caps 是指没有加入注意力模块的胶囊网络模型. Caps-att 是我们提出的模型.

表 2 开发集上各模型的 F1 得分

Table 2 F1 scores of each model on development dataset

活动	基线系统	GCRNN	GCRNN-att	Caps	Caps-att
缺席	85.4 %	85.8 %	86.9 %	87.5 %	91.3 %
烹饪	95.1 %	93.7 %	96.9 %	93.8 %	95.8 %
洗碗	76.7 %	78.3 %	81.1 %	67.3 %	82.7 %
吃饭	83.6 %	83.3 %	87.8 %	82.8 %	90.5 %
其他	44.8 %	39.1 %	41.5 %	38.0 %	55.4 %
社会活动	93.9 %	84.7 %	98.8 %	89.8 %	96.8 %
真空吸尘	99.3 %	99.9 %	100.0 %	99.5 %	99.6 %
看电视	99.6 %	98.7 %	99.8 %	100.0 %	99.9 %
工作	82.0 %	84.1 %	84.4 %	84.3 %	87.6 %
平均值	84.5 %	86.9 %	87.8 %	87.3 %	92.1 %

从表 2 的结果可以明显看出, 我们的模型相比于其他 4 个模型在 9 类活动中有 5 类活动的 F1 得分都是最高的, 其中缺席类的 F1 得分比其他 4 个模型高出 5 % 左右, 其他类的得分比另外 4 个系统高出 10 % 左右. 可以看出对于不是具体相关活动的类别, 我们的模型能很好地减少过拟合现象.

从实验结果可以看出, 我们模型在开发集和评估集上 F1 得分的平均值都要高于其他 4 个模型. 胶囊网络模型 (Caps) 在开发集和评估集的 F1 得分明显高于基线系统, 分别高出 2.8 % 和 1.6 %. 这说明胶囊网络在音频分类问题中的效果是要明显好于这种浅层的 CNN 结构. Caps 在开发集和评估集的 F1 得分也高于 GCRNN, 分别高出 0.4 % 和 0.1 %. 这说明相比于 GCRNN 这种较深的网络结构, 胶囊网络在分类效果上也有较好的表现. GCRNN-att 较 GCRNN 在开发集和评估集 F1 得分分别提高了 0.9 % 和 0.7 %; Caps-att 较 Caps 在开发集和评估集 F1 得分分别提高了 4.8 % 和 2.2 %, 这说明注意力机制成功抑制了音频事件类不相关帧, 选择了显著帧.

3 结论

在本文中, 我们提出了注意力胶囊网络模型用于多声道音频分类任务. 针对 CNN 对局部特征间相对关系不敏感, 提出采用胶囊网络学习局部特征与整体间的相对关系; 针对最大池化路由造成的特征损失问题, 提出采用动态路由避免忽视不显著局部特征, 得到初级胶囊层与高级胶囊层间的权重系数, 更加准确反映出部分与整体的关系; 针对音频剪辑所有帧对音频类贡献程度不同, 提出时间注意力机制赋予帧不同权重, 减少模型过拟合问题. 通过实验可以看出, 相比于一般的卷积网络和卷积循环网络等方法, 提出的网络模型具有更好的学习能力, 模型在开发集和评估集上的 F1 得分分别为 92.1 % 和 88.8 %. 我们下一步的研究计划包括将注意力胶囊网络推广到注意力矩阵胶囊网络, 将注意力胶囊网络用于弱标签半监督音频事件检测以及将注意力胶囊网络用于其他的类别区分度低的海量数据问题上.

表 3 评估集上各模型 F1 得分

Table 3 F1 scores of each model on evaluation dataset

模型	F1 得分
基线系统	85.0 %
GCRNN	86.5 %
GCRNN-att	86.9 %
Caps	86.6 %
Caps-att	88.8 %

致谢

本文作者衷心感谢英国萨里大学的 Wang Wen-Wu, Xu Yong, Huang Qiang, Kong Qiu-Qiang 以及 Turab Iqbal 五位学者对本文实验和写作的热情帮助.

References

1 Rafferty J, Nugent C D, Liu J, Chen L. From activity recognition to intention recognition for assisted living within smart homes. *IEEE Transactions on Human-Machine Systems*, 2017, **47**(3): 368−379

2 Erden F, Velipasalar S, Alkar A Z, Cetin A E. Sensors in assisted living: a survey of signal and image processing methods. *IEEE Signal Processing Magazine*, 2016, **33**(2): 36−44

3 Phan H, Hertel L, Maass M, Koch P, Mazur R, Mertins A. Improved audio scene classification based on label-tree embeddings and convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, **25**(6): 1278−1290

4 Zhu Yu, Zhao Jiang-Kun, Wang Yi-Ning, Zheng Bing-Bing. A review of human action recognition based on deep learning. *Acta Automatica Sinica*, 2016, **42**(6): 848−857
(朱煜, 赵江坤, 王逸宁, 郑兵兵. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, **42**(6): 848−857)

5 Fonseca E, Gong R, Serra X. A simple fusion of deep and shallow learning for acoustic scene classification. In: Proceedings of the 15th Sound and Music Computing Conference. Limassol, Cyprus, 2018

6 Hershey S, Chaudhuri S, Ellis D P W, Gemmeke J F, Jansen A, Moore R C, et al. CNN architectures for large-scale audio classification. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. Seoul, South Korea: IEEE, 2017. 131−135

7 Parascandolo G, Huttunen H, Virtanen T. Recurrent neural networks for polyphonic sound event detection in real life recordings. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016. 6440−6444

8 Cakir E, Parascandolo G, Heittola T, Huttunen H, Virtanen T. Convolutional recurrent neural networks for polyphonic

sound event detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 2017, **25**(6): 1291−1303

9 Xu Y, Kong Q Q, Huang Q, Wang W W, Plumbley M. D. Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. In: Proceedings of Interspeech 2017. Stockholm, Sweden: ISCA, 2017. 3083−3087

10 Xu Y, Kong Q Q, Wang W, Plumbley M D. Large-scale weakly supervised audio classification using gated convolutional neural network. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Alberta, Canada: IEEE, 2018. 121−125

11 Barker J, Marxer R, Vincent E, Watanabe S. Multimicrophone speech recognition in everyday environments. *Computer Speech & Language*, 2017, **26**: 386−387

12 Dekkers G, Vuegen L, Waterschoot T V, Vanrumste B, Karsmakers P. Dcase 2018 challenge–task 5: monitoring of domestic activities based on multi-channel acoustics. [Online], available: https://arxiv.org/pdf/1807.11246.pdf, August 1, 2018

13 Kong Q Q, Iqbal T, Xu Y, Wang W W, Plumbley M D. Dcase 2018 challenge surrey cross-task convolutional neural network baseline. [Online], available: https://arxiv.org/pdf/1808.00773.pdf, September 29, 2018

14 Tanabe R, Endo T, Nikaido Y, Ichige T, Nguyen P, Kawaguchi Y, et al. [Online], available: http://dcase.community/documents/challenge2018/technical_reports/DCASE2018_Tanabe_55.pdf, September 15, 2018

15 Inoue T, Vinayavekhin P, Wang S, Wood D, Greco N, Tachibana R. [Online], available: http://dcase.community/documents/challenge2018/technical_reports/DCASE2018_Inoue_14.pdf, September 15, 2018

16 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. In: Proceedings of the 2017 Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 2017. 3856−3866

17 Dauphin Y N, Fan A, Auli M, Grangier D. Language modeling with gated convolutional networks. In: Proceedings of the 2016 International Conference on Machine Learning. New York, USA: ACM, 2016. 933−941

18 Dekkers G, Lauwereins S, Thoen B, Adhana M W, Brouckxon H, Waterschoot T V, et al. The sins database for detection of daily activities in a home environment using an acoustic sensor network. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop. Munich, Germany: DCASE, 2017. 32−36

19 Kong Q Q, Xu Y, Wang W W, Plumbley M D. A joint separation-classification model for sound event detection of weakly labelled data. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Alberta, Canada: IEEE, 2018. 321−325

20 Kong Q Q, Xu Y, Sobieraj I, Wang W W, Plumbley M D (2019). Sound Event Detection and Time–Frequency Segmentation from Weakly Labelled Data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, **27**(4): 777−787

王金甲 燕山大学信息科学与工程学院教授. 主要研究方向为信号处理和模式识别. 本文通信作者. E-mail: wjj@ysu.edu.cn
(**WANG Jin-Jia** Professor at the School of Information Science and Engineering, Yanshan University. His research interest covers signal processing and pattern recognition. Corresponding author of this paper.)

纪绍男 燕山大学信息科学与工程学院硕士研究生. 主要研究方向为信号与信息处理. E-mail: jsn1533915375@163.com
(**JI Shao-Nan** Master student at the School of Information Science and Engineering, Yanshan University. His research interest covers signal and information processing.)

崔 琳 燕山大学信息科学与工程学院硕士研究生. 主要研究方向为信息处理. E-mail: 15733598690@163.com
(**CUI Lin** Master student at the School of Information Science and Engineering, Yanshan University. Her research interest is information processing.)

夏 静 燕山大学信息科学与工程学院硕士研究生. 主要研究方向为信号处理. E-mail: xiajing_527@sina.com
(**XIA Jing** Master student at the School of Information Science and Engineering, Yanshan University. Her research interest is signal processing.)

杨 倩 燕山大学信息科学与工程学院硕士研究生. 主要研究方向为模式识别. E-mail: yqlhp@sina.cn
(**YANG Qian** Master student at the School of Information Science and Engineering, Yanshan University. Her research interest is pattern recognition.)

勘误声明

本刊 2019 年第 45 卷第 8 期 1536−1547 页所刊 “基于多层忆阻脉冲神经网络的强化学习及应用” 一文中, 第 4 节实验与分析部分: 图 9 (c) 和 9 (d) 位置颠倒, 图 9 (c) 应为 50 次迭代后的结果, 图 9 (d) 应为训练前的结果.

特此更正, 并对由此带来的困扰表示歉意.

《自动化学报》编辑部