计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190200365

环境辅助的多任务混合声音事件检测方法

高利剑　毛启容
江苏大学计算机科学与通信工程学院　江苏 镇江 212013
(2221608013@ujs.edu.cn)

摘　要　在混合声音事件检测任务中,不同事件的声音信号相互混杂,从混合语音信号中提取的全局特征无法很好地表达每种单独的事件,导致当声音事件数量增加或者环境变化时,声音事件检测性能急剧下降。目前已存在的方法尚未考虑环境变化对检测性能的影响。鉴于此,文中提出了一种基于多任务学习的环境辅助的声音事件检测模型(Environment-Assisted Multi-Task,EAMT),该模型主要包含场景分类器和事件检测器两大核心部分,其中场景分类器用于学习环境上下文特征,该特征作为事件检测的额外信息与声音事件特征融合,并通过多任务学习方式来辅助声音事件检测,以此提高模型对环境变化的鲁棒性及多目标事件检测的性能。基于声音事件检测领域的主流公开数据集 Freesound 以及通用性能评估指标 F1 分数,将所提模型与基准模型(Deep Neural Network,DNN)及主流模型(Convolutional Recurrent Neural Network,CRNN)进行对比,共设置了 3 组对比实验。实验结果表明:1)相比单一任务的模型,基于多任务学习的 EAMT 模型的场景分类效果和事件检测性能均有所提升,且环境上下文特征的引入进一步提升了声音事件检测的性能;2)EAMT 模型对环境变化具有更强的鲁棒性,在环境发生变化时,EAMT 模型事件检测的 F1 分数高出其他模型 2%～5%;3)在目标声音事件数量增加时,相比其他模型,EAMT 模型的表现依旧突出,在 F1 指标上取得了 2%～10%的提升。

关键词:声音事件检测;环境辅助;多任务学习;特征融合;环境鲁棒性
中图法分类号　TP391

Environment-assisted Multi-task Learning for Polyphonic Acoustic Event Detection

GAO Li-jian and MAO Qi-rong
School of Computer Science and Communication Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China

Abstract　Polyphonic Acoustic Event Detection (AED) is a challenging task as the sounds are mixed with the signals from different events,and the overall features extracted from the mixture can not represent each event well,leading to suboptimal AED performance especially when the number of sound events increases or environment changes. Existing methods do not consider the impact of environmental changes on detection performance. Therefore,an Environment-Assisted Multi-Task learning (EAMT) method for AED was proposed. EAMT model mainly consists of two core parts:environment classifier and sound event detector,where the environment classifier is used to learn environment context features. As additional information of event detection,the environment context features are fused with sound event features to assist sound event detection by muli-task learning,so as to improve the robustness of EAMT model to environmental changes and the performance of polyphonic event detection. Based on Freesound dataset,one of the mainstream open data set in the field of AED,and general performance evaluation metrics F1 score,three sets of comparative experiments were set up to compare the proposed method with DNN(baseline) and CRNN,which is one of the most popular methods. The experimental results show that:compared with the single task model,EAMT model improves the performance of environment classification and event detection,and the introduction of environment context features further improves the performance of acoustic event detection. EAMT model has stronger robustness than DNN and CRNN as the F1 score of EAMT is 2% to 5% higher than other models when environment changes. When the number of target events increases,EAMT model still performs prominently,and compared with other models,EAMT model achieves an improvement of about 2% to 10% in F1 score.

Keywords　Acoustic event detection,Environment-assisted,Multi-task learning,Features fusion,Environmental robustness

到稿日期:2019-02-26　返修日期:2019-05-19　　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金通用联合重点项目(U1836220);国家自然科学基金面上项目(61672267)
This work was supported by the Key Projects of the National Natural Science Foundation of China(1836220) and National Nature Science Foundation of China (61672267).
通信作者:毛启容(mao_qr@ujs.edu.cn)

1 引言

声音事件是声信号中被标记为一个独特概念的音频片段[1]。声音事件检测(Acoustic Event Detection, AED)的目的是识别连续声音信号中出现的声音事件。声音事件检测可用于多种应用[2],如声学监控[3]、环境背景检测[4]及自动音频索引[5]等。通常来说,声音事件检测可以分为两类:单音(monophonic)检测和混合声音(polyphonic)检测。其中,单音检测旨在检测某一段语音信号属于哪个最有可能发生的事件,而多音检测处理多种声音同时发生的情况,目的是检测出每个声音事件发生的始末时间[1]。

现实生活中产生的语音信号大多是混合的,不同事件的声音信号是混合在一起的,从混合的语音信号中提取的全局特征无法很好地表达每个独立的事件,因此混合声音事件检测通常比单音事件检测要复杂得多。此外,在任何特定时刻发生的事件数量是未知的,并且可能很复杂,因此混合声音事件检测具有很高的挑战性。

传统的混合声音事件检测方法主要是模仿语音识别和模板匹配,例如使用混合高斯模型和隐马尔可夫模型对梅尔频率倒谱系数(MFCCs)建模,或者利用非负矩阵分解来表示每一种事件,然后将它们与声音词典进行模板匹配[6-7]。除了这几个具有代表性的方法外,利用手工特征进行建模与分析的传统方法,在过去的几年内的确取得了初步效果[8-9],但手工特征远不能很好地区分不同的声音事件,效率很低,并且传统方法的检测效率严重依赖手工特征的选择,这就使得深度学习模型在声音事件检测领域全面领先。

近年来,人们引入了深度神经网络(DNN)在声音事件检测中学习更深层的特征表示,并取得了很好的效果[10]。DNN通过将原始特征映射到高维空间后,再通过降低维数来去除冗余特征,以更好地学习特征表示,从而显著提高了分类和检测性能[11]。此外,文献[1]提出了一种多标签学习的DNN模型来作为"声音场景和事件的检测和分类(Detection and Classification of Acoustic Scenes and Events, DCASE)"比赛的基准模型,这是一个由IEEE举办的官方比赛。而后,基于长短时记忆网络(Long Short-Term Memory, LSTM)的方法[12]也被引入声音事件检测领域,并取得了更好的效果。与DNN不同,LSTM可以直接对音频中自然呈现的序列信息进行建模,同时学习时序信息和深度特征。研究表明,单向和双向LSTM网络均在声音事件检测领域取得了不错的效果[13-15]。

尽管通过深度学习学到的深层特征可使声音事件检测得到更好的效果,但是当目标事件数量很大时(如大于10[16]),这种全局特征很难表达出每种事件的差异,导致检测性能明显降低。此外,目前已存在的方法都未曾考虑环境上下文信息,而环境上下文信息中包含了许多有利于引导声音事件检测的有用信息,如声音事件发生的背景、声音事件在不同背景中的表现不同(相比街道上的说话声,会议室的声音存在回声)等。录制于不同场景下的同一种声音事件具有不同的数据分布及环境噪声,这就导致当环境发生变化时,现有方法对于声音事件的检测性能将会极大降低。因此,如何训练一个同时容忍环境变化和多个事件检测能力的模型是至关重要的。

在机器学习领域,多任务学习方法能够通过在相关任务间共享表示信息,来有效地协调并促进多个任务的学习[17-18]。在深度学习中,早期多任务学习被分为两种:隐层参数的硬共享网络和软共享。其中:1)参数的硬共享机制是神经网络的多任务学习中最常见的一种方式,一般来讲,它可以应用到所有任务的所有隐层上,而保留任务相关的输出层,越多任务同时学习就需要捕捉越多任务的同一个表示,因此硬共享机制降低了过拟合的风险;2)参数的软共享机制与硬共享相反,每个任务都有各自独立的模型和参数,参数间的相似性仅依靠正则项来规约,这样独立的模型的联合学习方式更容易收敛和保留特定任务的特异性。近年来,多任务学习模型逐渐朝着混合共享的方式(硬共享和软共享混合)发展。如文献[19]提出了一种自底向上的方法,从瘦网络(thin network)开始,使用对相似任务自动分组的指标,贪心地动态加宽网络。文献[20]则提出了层次性的共享型网络。

鉴于上述问题以及多任务学习的协作学习与参数共享的能力,本文提出了一种环境辅助的多任务学习模型(EAMT)来解决环境变化和多事件检测的挑战。该模型通过多任务学习的方法,来协调和促进场景识别任务和声音事件检测任务的学习,并将学到的具有场景鉴别能力的环境上下文特征作为额外信息来辅助声音事件的检测。因此,对于来自不同环境下录制的声音事件,一个统一的模型就能够同时检测这些声音事件。本文主要的贡献如下:

1)通过多任务学习框架,EAMT模型能够学到具有场景鉴别能力的环境上下文特征来辅助声音事件的检测,该特征包含声音事件发生的背景环境以及该环境下与声音事件相关的隐藏信息,从而提升了整体性能。

2)提出的EAMT是一个容忍环境变化的声音事件检测模型。与已有的方法相比,该模型能够有效检测来自不同场景的同类声音事件。同时,通过训练不同场景下的不同事件,所提模型能够检测出更多种类的声音事件。

本文第2节给出了EAMT模型的结构和算法;第3节给出了实验设计和性能评价;最后总结全文并对下一步工作进行了展望。

2 环境辅助的混合声音事件检测

为了详细地阐述环境辅助的多任务学习模型EAMT,首先介绍EAMT模型结构,再介绍多任务学习训练方法。

2.1 EAMT模型结构

2.1.1 输入层

输入层包含数据预处理过程及特征提取过程。MFCCs是目前最主流的手工特征之一,作为EAMT模型的输入。

为了得到MFCCs特征,首先对语音信号做预处理,预处理包括预加重、分帧及加窗。其中,预加重的目的是增加语音的高频分辨率;由于语音信号具有时变特性,对语音信号的分析需要建立在"短时"(即一帧,通常取40 ms)的基础上,即需要对长语音进行分帧操作,为了保持帧间的连贯性,还需要取

20 ms 的重叠部分作为帧移；加窗的目的是对帧信号的端点做平滑处理。以原始语音 s 为例，预处理后即可得到 M 帧信号：

$$y=[y(1),y(2),\cdots,y(m),\cdots,y(M)] \tag{1}$$

在完成预处理后提取 MFCCs 特征。MFCCs 特征是根据人耳听觉机理的研究发现的，人耳对不同频率的声波有不同的听觉敏感度，其提取过程分为以下几个步骤：快速傅里叶变换、三角带通滤波、离散余弦变换及差分参数提取。其中，快速傅里叶变换将时域信号转换成频域信号，从而得到频谱上的能量分布；三角带通滤波则计算三角带通滤波组输出的对数能量，经过离散余弦变换后即可得到 MFCC 系数；最后分别提取该系数的一阶差分和二阶差分。

至此，可以得到第 t 帧的 MFCCs 特征 $X(t)$。由于 EAMT 模型的建模对象为序列化数据，因此以滑动窗口方式采集训练样本，当窗口长度为 n 时，即一次送入网络中训练的数据为 n 帧信号时的 MFCC 特征 X，也就是说每一次训练模型时，需要采集连续的 n 帧信号作为输入，即：

$$X=[X(t-n),X(t-n+1),\cdots,X(t)] \tag{2}$$

2.1.2 硬共享层

硬共享层的目的是从原始特征 $X(t)$ 学到更精确的全局特征，在降低特征维度的同时去除冗余特征。在深度学习中，全连接层能够自由地将浅层特征映射到不同维度的特征空间，通过先升维后降维的方式筛选出更有价值的特征向量。因此，硬共享层由两个全连接层组成，每层分别有 512 和 128 个神经元，均由 Relu 激活。最终得到隐层特征 F，用于同时接入场景分类器和事件检测器，其中 F 为 $n*128$ 维向量：

$$F=[F(t-n),F(t-n+1),\cdots,F(t)] \tag{3}$$

2.1.3 场景分类器

场景分类器是一个由多层全连接层组成的神经网络，该部分旨在学习一个具有场景鉴别能力的环境特征，并通过特征融合层为事件检测提供额外的有效信息，从而实现容忍环境变化的事件检测能力。通过将硬共享层的输出 F 映射到高维空间并做降维处理，可得到潜在特征 $D(t)$，再经过一个简单的多层感知机（Multi-layer Perceptorn，MLP），通过 softmax 激活即可得到 t 时刻预测的 one-hot 形式的环境场景标签。

具体而言，该部分中的深度神经网络（DNN）负责将共享特征 F 映射到高维空间以学习深度特征 $D(t)$（128 维向量），而 MLP 负责将高维特征降至低纬度并送入分类层。其中，DNN 包含 3 层，维度分别为 256，512，128，MLP 为 3 层全连接层，分别含有 128 个、64 个及 J 个神经元（J 表示场景类别数）。除最后一层由 softmax 激活外，其余各层由 Relu 激活。图 1 以 3 种场景为例，若当前数据被预测为来自第一个场景，则场景标签 $\hat{y}_{\text{env}}(t)=[0,0,1]$。

为了让潜在特征 $D(t)$ 具备场景鉴别能力，引入交叉熵衡量预测的场景标签和真实标签间的距离，以此作为场景分类的目标函数 J_{env}：

$$J_{\text{env}}(\theta_1)=-\frac{1}{I}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}1\{y^{(i)}=j\}\log\frac{e^{\theta_j^{\mathrm{T}}x(i)}}{\sum_{l=1}^{k}e^{\theta_l^{\mathrm{T}}x(i)}}\right] \tag{4}$$

其中，y 表示真实标签，I 表示样本数。

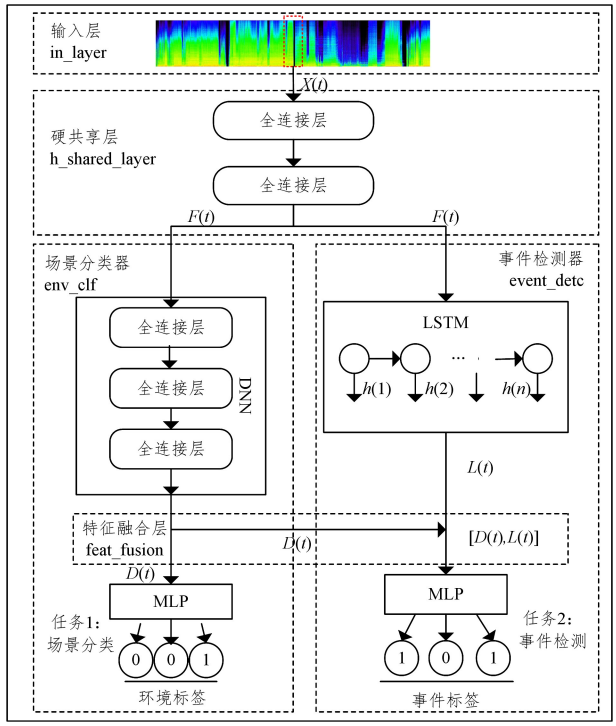

图 1 EAMT 模型框架图

Fig. 1 Framework of EAMT model

2.1.4 声音事件检测器

EAMT 模型中的声音事件检测器主要包含 3 个重要部分（见图 1）：LSTM 网络、特征融合层及 MLP 多标签事件检测器。

从图 1 可以看出，LSTM 网络用于同时学习深度事件特征和语音信号的时序信息，即事件特征 $L(t)$，该 LSTM 网络的输入为硬共享层得到的隐层特征 F。为了将从场景分类器中学到的环境特征 $D(t)$ 作为额外信息传递给声音事件检测器，特征融合层将 $L(t)$ 和从场景分类器中学到的环境特征 $D(t)$ 串联，串联后的融合特征 $[D(t),L(t)]$ 既包含了声音事件的信息又融入了环境因素，以此作为 MLP 的输入来预测声音事件标签。由于每个声音事件发生的状态是相互独立的，因此 MLP 的输出层应当由多个 sigmoid 单元组成，即得到多标签（multi-label）形式的检测结果 $\hat{y}_e(t)$：

$$\hat{y}_e(t)=[\hat{y}_e^1,\hat{y}_e^2,\cdots,\hat{y}_e^c] \tag{5}$$

其中，$c\leqslant C$，C 表示声音事件的种类数。

具体而言，该部分中 LSTM 网络共有两层，每层各有 256 个单元和 128 个单元，激活函数均为 tanh。此处的 LSTM 网络是一个“多对一”的结构，因此输出向量 $L(t)$ 的维度为 128 维。MLP 为 3 层全连接层，分别含有 128 个、64 个及 C 个神经元。除最后一层由 sigmoid 激活外，其余各层由 Relu 激活。图 1 以 3 类声音事件为例，若当前时刻检测到第一个和第三个事件正在发生，而没有检测到第二个事件，那么检测结果为 $\hat{y}_e(t)=[1,0,1]$。

最后，引入联合的二分类交叉熵衡量每类声音事件的预测标签和真实标签的距离，作为声音事件检测的目标函数 J_e：

$$J_e(\theta_2)=-\frac{1}{I}\sum_{i=1}^{I}\sum_{c=1}^{C}y_e^{(c,i)}\log(\hat{y}_e^{(c,i)})+(1-y_e^{(c,i)})\log(1-\hat{y}_e^{(c,i)}) \tag{6}$$

其中，$y_e^{(c,i)}$ 表示第 i 个样本中第 c 类事件的真实标签。

2.2 多任务学习方法

多任务学习的核心目标是训练一个容忍环境变化同时具有多个事件检测能力的模型。为了达到这一目的，首先将场景分类任务作为辅助任务，将声音事件检测作为主要任务。其中辅助任务旨在学到一种具有场景鉴别能力的环境特征，主要任务要求能够在辅助任务的协作下提升声音事件检测的性能。

为了具备这种协作学习的能力，所提模型的核心思路是在保留硬共享层的基础上，额外增加软共享层，即 2.1.4 节提出的特征融合层。通过将学到的环境特征直接与深层事件特征相融合，使得本文的 EAMT 模型在事件检测过程中提供额外的有用因素。进一步地，通过联合场景分类任务及事件检测任务的目标函数，EMAT 模型能够学会两个任务间直接的联系：

$$L_{\text{coop}}(\theta_1,\theta_2)=\lambda J_{\text{env}}(\theta_1)+(1-\lambda)J_e(\theta_2) \tag{7}$$

其中，λ 为辅助任务所占权重，应当小于 0.5。

除此之外，为了防止两个任务过于依赖彼此而失去各自的特异性，额外增加一项正则项来约束彼此同化，该正则项为环境特征 $D(t)$ 与深层事件特征 $L(t)$ 之间的 KL 散度(Kullback-Leibler divergence)，通过该约束来促使两个特征概率分布最大化：

$$L_{\text{regular}}(\phi)=-D_{\text{KL}}(p_{\theta_1}(D(t))\parallel p_{\theta_2}(L(t))) \tag{8}$$

最终，EAMT 模型的目标函数为：

$$\begin{aligned}L_{\text{EAMT}}(\theta_1,\theta_2,\phi)&=L_{\text{coop}}(\theta_1,\theta_2)+\alpha L_{\text{regular}}(\phi)\\&=\lambda J_{\text{env}}(\theta_1)+(1-\lambda)J_e(\theta_2)-\\&\quad\alpha D_{\text{KL}}(p_{\theta_1}(D(t))\parallel p_{\theta_2}(L(t)))\end{aligned} \tag{9}$$

通过优化该目标函数，EAMT 模型能够在声音事件的检测中同时考虑环境的变化，因此对于不同环境下的同一类声音事件，该模型也具有检测能力。此外，对于不同环境下不同种类的声音事件，EAMT 模型能够作为一个统一模型分别检测出不同的声音事件。

3 实验

3.1 数据集

由于本文实验需要动态增加声音事件的种类及构造不同的环境背景，因此实验选用一个公开的通用数据库，即 Freesound 数据库[21]，该数据库包含了 28 种不同的声音，但每个样本都只包含一种声音。在混合声音事件检测领域，由于目前还没有公开的数据集包含不同数量的目标事件，因此为了满足实验需求，我们根据这个数据集构造了 4 种子数据集：Freesound-env，Freesound-6，Freesound-9 及 Freesound-12。

1)Freesound-env：该数据集包含两种不同环境(街道和室内)中的相同 3 类事件(说话声、笑声及鼓掌声)。该数据集的具体构造方式如下：从原始 Freesound 数据库中选出以上 3 种声音，每种声音选出 10 个语音样本，根据式(10)从每类声音中随机选出一个来混合成新的语音信号 S：

$$S=aS_1+bS_2+cS_3+0.2*Noise \tag{10}$$

其中，a,b,c 均随机取值为 0 或 1，$Noise$ 为街道噪声或室内噪声(分别用 01 或 10 表示)，0.2 为噪声强度系数。因此，新生成的混合信号 S 对应的标签即为[$Noise,a,b,c$]。

2)Freesound-6：该数据集包含两个指定的环境(街道和室内)，每种场景选取 3 种不同事件，构造方式如式(10)。

3)Freesound-9：该数据集包含 3 种不同环境(街道、室内及会议厅)，每种环境包含了不同的 3 类事件，即一共有 9 种不同的声音事件，构造方式如式(10)。

4)Freesound-12：相比 Freesound-9，该数据集多了一种环境(咖啡厅)和该环境下的 3 类事件，因此一共包含了 12 种声音事件，构造方式如式(10)。

3.2 实验设置

实验中，基准模型采用文献[1]提出的多标签的深度神经网络模型(DNN)。该 DNN 模型已成为 DCASE 2016 比赛中最为主流的模型，并且被引入 DCASE2017 中作为比赛的基准模型。本文实验不仅与基准模型做对比，同时也与声音事件检测领域的主流模型之一 CRNN 做比较，该模型同样在 DCASE 2017 比赛中取得了突出的效果[22]。其中，DNN 模型包含 4 层全连接层，前 3 层每层包含 50 个神经元，均由 Relu 激活，最后一层由 K 个神经元组成(K 表示事件类别数)，由 sigmoid 激活，为了防止过拟合，每层都连接了一个丢失率为 0.5 的 Dropout 层；CRNN 模型与文献[22]一致，包含 8 层卷积层、4 层池化层、2 层双向长短时记忆网络(Bi-directional Long Short-Term Memory，BLSTM)以及最后的 3 层全连接层。

实验采用 F1 分数作为评价指标来衡量声音事件检测的性能，计算方式为：

$$F1=\frac{2\cdot TP}{2TP+FP+FN} \tag{11}$$

其中，TP，FP 及 FN 分别表示预测为正类的正样本数、预测为正类的负样本数及预测为负类的负样本数。

首先，为了验证所提 EAMT 模型的多任务学习的性能，我们在 Freesound-6 数据集上设置了 3 组对比实验(具体层如图 1 所示)：1)只保留场景分类部分(in_layer+h_shared_layer+env_clf)，对比完整的 EAMT 模型，比较环境分类的 F1 分数；2)只保留声音事件检测部分(in_layer+h_shared_layer+event_detc)，对比完整的 EAMT 模型，比较事件检测的 F1 分数；3)仅去掉特征融合层，而保留其他部分(in_layer+h_shared_layer+env_clf+event_detc)，对比完整的 EAMT 模型，比较整体的 F1 分数。其次，为了验证 EAMT 模型对环境的鲁棒性，与 DNN 及 CRNN 进行对比，我们在 Freesoundenv上验证了事件检测的性能。最后，为了验证 EAMT 模型检测多个声音事件的能力，我们在 Freesound-6，Freesound-9 及 Freesound-12 上做了对比实验。

所有的实验都在 Ubuntu 16.04 LST 操作系统下完成，开发语言为 Python 3.6.2，深度学习框架为 Keras 2.2.0 及 Tensorflow 1.8.0，GPU 为 NVIDIA TITAN X。我们利用网

格搜索(Grid Search)方法确定了式(9)中的超参数 λ 和 α 分别为 0.2 和 0.5。批量训练的数量设置为 200，迭代次数设置为 300。

3.3 性能评估

本节从 3 个角度分析了 EAMT 在声音事件检测中对 F1 的影响：1)多任务学习的结构；2)环境的变化；3)声音事件的种类、数量。

1)不同的多任务学习结构对 F1 分数的影响。为了验证多任务学习方法的有效性，基于 Freesound-6 数据集，首先给出了不同情况下多任务学习对 F1 分数的影响。如表 1 所列，相比于单一任务的模型，完整的 EAMT 模型的场景分类效果和事件检测能力均有所提升，这就说明了多任务学习能够互相协作促进彼此的性能。具体地，当联合了场景分类(env_clf)和声音事件检测(event_detc)时，场景分类的 $F1$ 提高了 0.12%，事件检测的 $F1$ 提高了 0.62%。此外，将环境特征作为额外信息来引导事件检测，进一步提升了事件检测的性能。

表 1 在 Fressound-6 数据集上不同情况下多任务学习对 $F1$ 分数的影响

Table 1 Effect of multitask learning on $F1$ scores in different situations on Freesound-6

(单位：%)

模型	环境分类 $F1$	事件检测 $F1$
in__layer+h_shared_layer+env_clf	99.20	—
in__layer+h_shared_layer+event_detc	—	83.50
in_layer+h_shared_layer+env_clf+event_detc	99.26	83.84
EAMT in_layer+h_shared_layer+env_clf+feat_fusion+event_detc	**99.32**	**84.12**

表 2 在 Fressound-env 数据集上环境变化对于不同模型 $F1$ 分数的影响

Tabel 2 Effect of environmental changes on $F1$ scores of different models on Freesound-env

模型	事件检测 $F1$/%
DNN	82.24
CRNN	85.76
EAMT	**87.01**

2)环境变化对声音事件检测 $F1$ 的影响。为了验证 EAMT 模型对不同环境下的声音事件检测能力，基于 Freesound-env 数据集，对比了 EAMT 模型、DNN 及 CRNN 模型在声音事件检测方面对环境的鲁棒性，结果如表 2 所列。可以看到，EAMT 模型取得了最好的结果。由于 DNN 模型和 CRNN 模型均未针对环境场景做适应性优化，导致数据集中的环境上下文信息并没有起到作用；此外，两种环境下同一类事件的背景噪声不同、数据差异较大，导致其性能不如 EAMT 模型。这说明 EAMT 模型对于环境变化的容忍性明显优于其他模型。

3)声音事件数量对声音事件检测 $F1$ 的影响。为了验证 EAMT 模型对大量类别的声音事件的检测能力，在 Fressound-6，Fressound-9，Fressound-12 这 3 组数据集上，对比了 EAMT 模型与其他两种模型，验证了在声音事件数量增加时检测性能的影响，结果如表 3 所列。由于声音事件种类增加时，用于区分不同声音事件的全局特征更难学习，导致检测性能随着种类的增加而逐渐降低。尽管如此，在所有的情况下，所提模型的 $F1$ 值都高于 DNN 模型和 CRNN 模型，这说明 EAMT 模型能够利用环境特征的辅助适应大量事件的检测场景，如在 12 个声音事件时，其 $F1$ 值仍比 DNN 高出了 10.77%。

表 3 在 Fressound-6，Fressound-9，Fressound-12 数据集上声音事件数量的增加对于不同模型 $F1$ 分数的影响

Table 3 Effect of increase of number of sound events on F1 scores of different models on Fressound-6，Fressound-9 and Fressound-12

事件数量	模型	事件检测 F1/%
6	DNN	74.14
	CRNN	78.58
	EAMT	**84.12**
9	DNN	69.26
	CRNN	74.48
	EAMT	**80.59**
12	DNN	63.77
	CRNN	72.86
	EAMT	**74.54**

结束语 混合声音事件检测的目标是识别连续声信号中出现的声音事件。本文提出了一种新的方法(EAMT 模型)来学习一种具有场景鉴别能力的环境特征，同时通过多任务学习的方法利用该特征来辅助声音事件的检测。与目前的主流方法相比，所提方法在容忍环境变化和检测大量声音事件两个方面，取得了非常好的效果。尽管如此，该方法仍未能很好地解决在大量声音事件检测的场景下无法保持很高准确率的问题；此外，也未能解决同一个环境下大量声音事件检测困难的挑战。因此，未来我们的主要研究方向是进一步扩大环境信息对事件检测的辅助能力，同时增加同一环境下大量声音事件的检测能力。进一步地，本文提出的针对环境容忍的多任务学习模型是一种可用于跨域分类或检测任务的通用架构，也适用于其他跨域的任务，因此未来的研究工作将该方法应用到其他模型和任务中。

参 考 文 献

[1] CAKIR E，HEITTOLA T，HUTTUNEN H，et al. Polyphonic sound event detection using multi label deep neural networks [C]// Proceedings of the 6th International Joint Conference on Neural Networks. Killarney，Ireland，2015：1-7.

[2] ZHANG A Y，NI C J. Research on background model adaptive method of audio monitoring system based on audio event detection and classification[J]. Computer Science，2016，43(9)：310-314.

[3] ZHANG D，ELLISD. Detecting sound events in basketballvideo archive[R]. Department of Electrical Engineering，Columbia University，New York，2001.

[4] CHU S，NARAYANAN S，KUO C l. Where am I? Scene Recognition for Mobile Robots using Audio Features[C]// Proceedings of the 7th International Conference on Multimedia and Expo. Toronto，Canada，2006：885-888.

[5] HARMAA, MCKINNEY M F, SKOWRONEK J. Automatic surveillance of the acoustic activity in our living environment [C]// Proceedings of the 6th IEEE International Conference on Multimedia and Expo. Amsterdam, Netherlands, 2005: 634-637.

[6] INNAMI S, KASAH. NMF-based environmental sound source separation using time-variant gain features [J]. Computers & Mathematics with Applications, 2012, 64(5): 1333-1342.

[7] DESSEINA, CONT A, LEMAITRE G. Real-time detection of overlapping sound events with non-negative matrix factorization [M]. Matrix Information Geometry, 2013: 341-371.

[8] MESARO A, HEITTOLA T, ERONEN A, et al. Acoustic event detection in real life recordings [C]// Proceedings of the 18th Signal Processing Conference. Aalborg, Denmark, 2010: 1267-1271.

[9] HEITTOLA T, MESAROS A, VIRTANEN T, et al. Supervised model training for overlapping sound eventsbased on unsupervised source separation [C]// Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 8677-8681.

[10] MUN S, SHON S, KIM W, et al. Deep neural networkbottleneck features for acoustic event recognition [C]// Proceedings of the 16th INTERSPEECH. SanFrancisco, USA, 2016: 2954-2957.

[11] GENCOGLUO, VIRTANEN T, HUTTUNEN H. Recognition-of acoustic events using deep neural networks [C]// Proceedings of the 22nd European Signal Processing Conference. Lisbon, Portugal, 2014: 506-510.

[12] PARASCANDOLO G, HUTTUNEN H, VIRTANEN T. Recurrent neural networks for polyphonic sound event detection in real life recordings [C]// Proceedings of the 9th International Conference on Acoustics, Speech, and Signal Processing. Shanghai, China, 2016: 6440-6444.

[13] WANG Y, METZE F. A transfer learning based featureextractor for polyphonic sound event detection using connectionist temporal classification [C]// Proceedings of the 19th INTERSPEECH. Hyderabad, India, 2018: 3097-3101.

[14] XIA X, TOGNERI R, SOHEL F, et al. Frame-wise dynamic threshold based polyphonic acoustic eventdetection [C]// Proceedings of the 19th INTERSPEECH. Hyderabad, India, 2018.

[15] ZHRER M, PERNKOPF F. Virtual adversarial trainingand data augmentation for acoustic event detection withgated recurrent neural networks [C]// Proceedings of the 19th INTERSPEECH. Hyderabad, India, 2018: 493-497.

[16] MCLOUGHLIN I, ZHANGH, XIE Z, et al. Robust sound event classification using deep neuralnetworks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(3): 540-552.

[17] DO V H, CHEN N F, LIM B P, et al. Multitask Learning for Phone Recognition of Underresourced Languages Using Mismatched Transcription [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2018, 26(3): 501-514.

[18] TAN Z, MAK M W, MAK K W. DNN-Based Score Calibration With Multitask Learning for Noise Robust Speaker Verification [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2018, 26(4): 700-712.

[19] LU Y, KUMAR A, ZHAI S, et al. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification [C]// Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1131-1140.

[20] SØGAARD A, GOLDBERG Y. Deep multi-task learning with low level tasks supervised at lower layers [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 231-235.

[21] FONT F, ROMA G, SERRA X. Freesound technical demo [C]// Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain, 2013: 411-412.

[22] ADAVANNE S, VIRTANEN T. A report on sound event detection with different binaural features [R]. Technical Report, DCASE2017 Challenge, 2017.

GAO Li-jian, born in 1993, postgraduate. His main research interests include multimedia intelligent analysis.

MAO Qi-rong, born in 1975, professor, Ph.D supervisor, is member of China Computer Federation (CCF). Her main research interests include multimedia intelligent analysis and emotional computing.