计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900111

弱标签环境下基于多尺度注意力融合的声音识别检测

郑伟哲[1] 仇 鹏[2] 韦 娟[2]
1 西安电子科技大学电子工程学院 西安 710071
2 西安电子科技大学通信工程学院 西安 710071
(634877973@qq.com)

摘 要 目前大多数声音识别检测的研究都是基于强标签数据集的,但在真实环境的声音识别与检测任务中,音频标签不完整并且含有大量噪声,使得获取强标签音频数据比较困难,进而影响对声音的准确识别与检测。为此,在卷积循环神经网络模型的基础上,提出了一种多尺度注意力融合机制。该机制使用注意力门控单元,在降低声音时频图特征中噪声影响的同时,能够更多地利用有效特征。同时,通过结合多个尺寸的卷积核进行特征融合,进一步提升对声音特征的有效提取。此外,采用一种结合帧检测结果的加权法对声音信号进行识别。最后,在弱标签环境下,从 AudioSet 数据库中选取一个包含 17 种城市交通工具声音的弱标签数据集进行检测识别,所提模型对测试集声音识别结果的 $F1$ 值为 58.9%,检测结果的 $F1$ 值为 43.7%。结果表明,在弱标签城市交通工具声数据集下,网络模型相比传统的声音识别检测模型具有更高的识别检测精度;同时,重要性加权识别方法、多尺度注意力融合方法均可提升模型对声音识别检测的精度。

关键词:弱标签;多尺度;注意力;声音识别;声音检测
中图法分类号 TP391

Sound Recognition and Detection Based on Multi-scale Attention Fusion in Weak Label Environment

ZHENG Wei-zhe[1], QIU Peng[2] and WEI Juan[2]
1 School of Electronic Engineering, Xidian University, Xi'an 710071, China
2 School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract At present, most of the research on sound recognition and detection is based on the datasets with strong labels. However, in real-world sound recognition and detection tasks, it is difficult to obtain strong label audio data due to incomplete audio labels with a large amount of noise, which in turn affects the accuracy of sound identification and detection. To this end, a multi-scale attention fusion mechanism is proposed based on the convolutional cyclic neural network model. This mechanism uses the attention gating unit to make more use of the effective features while reducing the effects of noise in the sound time-frequency map features. At the same time, feature fusion is performed by combining convolution kernels of multiple sizes to further improve the effective extraction of sound features. In addition, the sound signal is identified by a weighting method that combines the results of the frame detection. Finally, using the proposed model, a weak labeled data set containing 17 kinds of urban vehicle sounds is selected from the AudioSet database for detection and identification in the weak label environment. For the test set, the F1 value of sound recognition result is 58.9%, and the F1 value of detection result is 43.7%. The simulation experiments show that the CRNN baseline model used in this paper is more accurate than the traditional sound recognition detection model under the weakly labeled city vehicle sound datasets. And the methods involved in the paper, such as the importance weighted recognition method and multi-scale attention fusion method, can improve the accuracy of the model for sound recognition and detection.

Keywords Weak label, Multi-scale, Attention, Sound recognition, Sound detection

1 引言

传统的声音识别检测方法的一个基本假设是训练集中样本标签是准确和完整的。但是,在真实应用环境中,这个假设很难成立,因为给出音频中所有事件的标签是一项繁琐、费力的工作。因此,在实际环境中往往只给出少部分标签,而不给出完整的标签列表,我们称该环境为弱标签环境[1-3]。弱标签环境下,声音的识别与检测在信息检索、公共场所异常声音监

到稿日期:2019-09-16 返修日期:2020-01-07 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(51675425);陕西省重点研发计划(2018SF-365)
This work was supported by the National Natural Science Foundation of China (51675425) and Key Research Program of Shaanxi Province, China (2018SF-365).
通信作者:韦娟(weijuan@xidian.edu.cn)

测和工业应用中有广泛的应用[4-8]，是人工智能领域的一个重要研究方向。弱标签环境的核心任务是对某一段音频进行特征提取，然后对其进行识别与检测，判断其属于哪一类音频，并找出包含特定声音事件的起始点和结束点。

随着人工智能的快速发展[9-10]，深度学习被用于弱标签环境下的声音识别与检测。文献[7,11]将时频图(Time-Log-mel)作为原始特征输入卷积神经网络（Convolutional Neural Network，CNN），相比直接输入音频频谱图为特征，这种操作；对声音识别与检测的效果有所提升。文献[12-13]以时频图为特征，研究了卷积循环神经网络（Convolutional Recurrent Neural Network，CRNN)在弱标签环境下对声音识别的性能，发现 CRNN 对声音的识别效果相比 CNN 模型有了较大提升。文献[14]以 CRNN 为基础模型，研究了一种线性门控单元，在弱标签环境下在声音的识别与检测方面取得了较好的效果。

对于一段音频，时频图同时包含时域和频域的信息，因此时频图特征更适合用于对声音的识别与检测。现有文献也是将整个时频图作为输入进行学习，但在弱标签环境下，训练集音频中出现的声音事件一般占比较小，噪声占比较大，难以提取有效的深层特征。为此，本文提出了一种多尺度注意力融合方法，使用注意力门控单元控制神经网络中每一层信息的流动，使得网络在减小噪声影响的同时还能更多地利用有效特征。此外，在卷积层中通过使用不同大小的卷积核进行卷积融合，使神经网络学习到的特征更多样化，进而提升声音识别与检测的效果。

2 网络结构

CNN 适用于提取时频图中频域的深层特征，循环神经网络(Recurrent Neural Network，RNN)可以充分利用时频图中时域的相关性，而本文使用的 CRNN 将二者的优势相结合，如图 1 所示。将声音时频图特征输入 CNN，通过多次的卷积和池化得到高级特征，再将该特征输入 RNN，最终通过前馈神经网络（Feedforward Neural Network，FNN)得到 RNN 中每一个时间戳的预测结果。

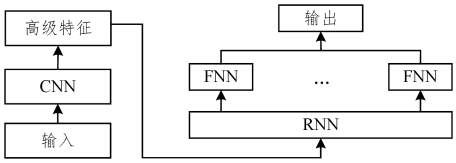

图 1 CRNN 模型结构图

Fig. 1 CRNN model structure

CRNN 模型的整个流程为：

$$X=\mathrm{CNN}(I) \tag{1}$$

$$h_t=\mathrm{RNN}(x_t),t=1,2,3,\cdots,N \tag{2}$$

$$p_t=\mathrm{FNN}(h_t),t=1,2,3,\cdots,N \tag{3}$$

其中，I 是输入 CNN 的特征，X 是 CNN 的输出，x_t 为 RNN 中每一个时间戳的输入，h_t 为 RNN 中每一个时间戳的输出，p_t 为 h_t 通过 FNN 后的输出，N 为 RNN 输入时间戳的数量。

3 多尺度注意力融合和声音识别检测

3.1 多尺度注意力融合

在弱标签环境下，训练集音频中的声音事件出现的时长均未知。音频中声音事件占比较小的音频在对应到时频图中后，其相应特征只占小部分，而剩余部分为噪声。由于噪声占比过大，将时频图输入 CNN 将难以学习到有效的特征。为此，提出将门控注意力机制与多尺度卷积进行融合，来提高模型的识别检测效果。

门控注意力机制是指在 CNN 中对卷积操作后的特征图分别进行两个不同激活函数的运算，一个为 sigmoid 激活函数，一个为 relu 激活函数，输出为二者逐元素相乘。sigmoid 可看作一种门控单元，用来控制信息的流动。对于通过 relu 的特征图，当门控单元激活值接近 1 时，相应特征部分将会向下一层传播更多的信息；当激活值接近 0 时，相应特征部分会被忽略。换言之，CNN 网络将会忽略无关信息而更关注特征图中的重要信息，使网络学习到更有效的特征。从特征选择的角度来看，这可被视为一种局部注意力机制[8]。门控单元的定义为：

$$Y=\mathrm{relu}(w*x+b)\circ\sigma(v*x+c) \tag{4}$$

其中，relu 是非线性激活函数；σ 是 sigmoid 非线性激活函数；$\circ$是逐元素相乘；$*$ 是卷积操作；w 和 v 是卷积核的权重参数；b 和 c 为偏置参数；x 在第一层为时频图，在其他层则代表特征图。

借鉴 inception 结构[15]的创新思想，在门控注意力机制的基础上使用一种多尺度卷积融合方法，融合过程如图 2 所示。

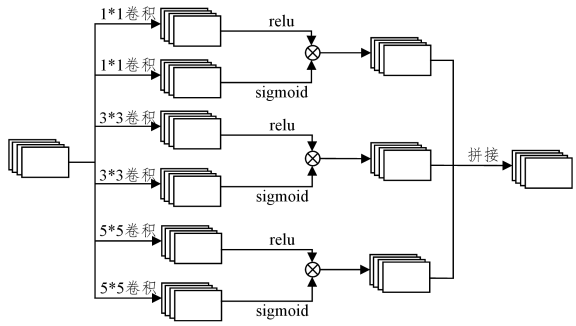

图 2 多尺度注意力融合机制

Fig. 2 Multi-scale attention fusion mechanism

对 CNN 中的每一卷积层使用 $1*1$ 的卷积核，得到输出 Y_1：

$$Y_1=\mathrm{relu}(w_1*x+b_1)\circ\sigma(v_1*x+c_1) \tag{5}$$

其中，w_1 和 v_1 是 $1*1$ 卷积核的权重参数，b_1 和 c_1 为 $1*1$ 卷积核的偏置。

对 CNN 中每一卷积层分别使用 $3*3$ 的卷积核，得到输出 Y_3：

$$Y_3=\mathrm{relu}(w_3*x+b_3)\circ\sigma(v_3*x+c_3) \tag{6}$$

其中，w_3 和 v_3 是 $3*3$ 卷积核的权重参数，b_3 和 c_3 为 $3*3$ 卷积核的偏置。

对 CNN 中每一卷积层分别使用 $5*5$ 的卷积核，得到输出 Y_5：

$$Y_5=\mathrm{relu}(w_5*x+b_5)\circ\sigma(v_5*x+c_5) \tag{7}$$

其中，w_5 和 v_5 是 $5*5$ 卷积核的权重参数，b_5 和 c_5 为 $5*5$ 卷积核的偏置。

对 Y_1，Y_3 和 Y_5 进行融合，得到最终输出：

$$Y=concat(Y_1,Y_3,Y_5) \tag{8}$$

其中，$concat(\ )$为特征图拼接操作。

不同尺寸的卷积核可以学习到不同尺度的深层特征，而

门控单元可以筛选出音频的重要特征，将二者结合可以得到更加丰富、有效的特征。

3.2 声音事件的检测与识别

为了对声音事件进行检测，首先将音频分帧，得到时频图。通过 CNN 提取时频图的高级特征并将其输入 RNN，最后将 RNN 的输出输入至 FNN，并经过 sigmoid 激活函数得到每一帧的检测结果。声音事件检测与识别的整体模型结构如图 3 所示。

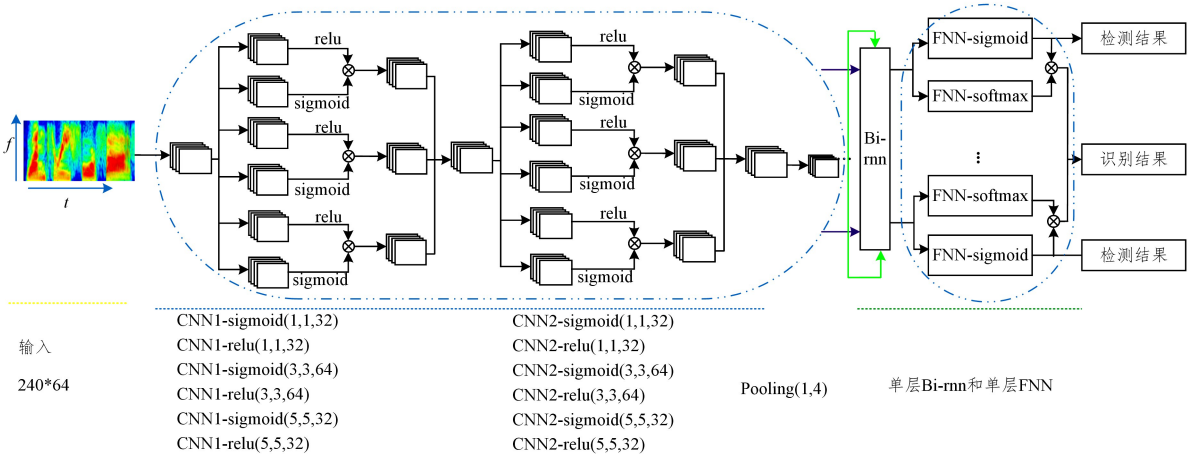

图 3 时间注意力机制

Fig. 3 Time attention mechanism

为了更好地识别声音，首先将每一帧 t 的识别结果 P_t 进行平均，得到识别结果 O：

$$O=\frac{1}{T}\sum_{t=1}^{T}P_t \tag{9}$$

其次，采用一种结合帧检测结果的加权法进行识别。将 FNN 的输出分别输入 sigmoid 和 softmax 两个激活函数，输出分别为 P_t 和 Z_t，如图 3 所示。将两者逐元素相乘，输出 O_t'：

$$O_t'=Z_t\circ P_t, t=1,2,3,\cdots,T \tag{10}$$

最后，对 O_t' 进行平均，得到最终的识别结果。

FNN 经过 sigmoid 的输出为相应帧的识别结果；经过 softmax 可以得到每一帧对不同类识别的重要性系数，为每一类声音中重要性高的帧分配较大的权重，为噪声所在帧分配更小的权重。最后，对每一帧的识别结果进行加权，以提升识别精度。

4 仿真与结果分析

4.1 环境配置

实验框架为 tensorflow，keras，编程语言为 python3.5。神经网络部分以 tensorflow 为底层框架，keras 为顶层框架，来实现模型与算法。整体模型如图 3 所示，RNN 部分使用双向门控循环单元（Bi-directional Gate Recurrent Unit，GRU）。

4.2 数据与特征提取

实验选取的数据集是 AudioSet 的一个子集[16]，其主要包含城市内的交通工具声和警报声，共计 51 660 个样本，17 类声音。数据的标签分布如图 4 所示。数据集中的样本大部分是长度为 10 s 的音频，少数音频时长少于 10 s。有 51 172 个音频只存在弱标签；仅有 488 个音频带有强标签，即拥有完整的标签列表。

首先对数据集进行预处理，对所有音频进行采样、分帧。采样率为 16 000 Hz；分帧时，滑窗长度为 1 024 个采样点，每两帧之间重叠 360 个采样点，最终将 10s 的音频分成 240 帧，在分帧长度不够 240 帧的音频尾部填 0 使其分为 240 帧。然后，对于每一帧，提取 64 维的 log-mel 特征。最终，每个音频可以提取得到 240 * 64 的时频图特征。

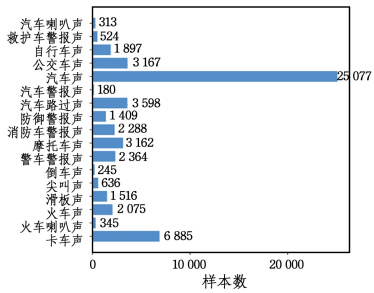

图 4 数据标签分布图

Fig. 4 Data label distribution

4.3 结果与分析

训练时，权重参数使用 glort 方法进行初始化，并使用 adam 算法进行优化。每批次的大小为 44，共训练 40 轮，每轮迭代 100 次，学习率为 0.001，损失函数使用二值交叉熵。

对于识别任务，将弱标签音频中 70% 的样本作为训练集，15% 的样本作为验证集，另 15% 的样本作为测试集。对于检测任务，直接使用 488 个具有强标签的音频作为测试集。

使用 Precision，Recall，F1 和 Accuracy 等指标对识别结果进行评估。其中，Precision 是指在被预测为正的所有样本中实际为正样本的概率；Recall 是指在实际为正的样本中被预测为正样本的概率；而 F1 是 Precision 和 Recall 的调和均值，属于综合评价指标；Accuracy 是所有预测正确的样本与总样本的比值。不同模型的识别结果如表 1 所列。

表 1 17 种弱标签声音的识别结果

Table 1 17 kinds weak labeled data set sound recognition results

分类算法	F1	Recall	Precision	Accuracy
CNN+RNN	0.570	0.571	0.570	0.926
CNN+RNN[14]	0.580	0.575	0.585	0.930
CNN+RNN（多尺度注意力融合）	**0.589**	**0.582**	**0.597**	0.930

同时，使用 F1，Recall 和 Precision 指标对检测结果进行评估，不同模型的检测结果如表 2 所列。

表 2 17 种弱标签声音的检测结果

Table 2 17 kinds weak labeled data set sound detection results

分类算法	F1	Recall	precision
CNN+RNN	0.420	0.528	0.349
CNN+RNN(加门控单元)[14]	0.427	0.526	0.360
CNN+RNN(多尺度注意力融合)	**0.437**	**0.545**	**0.365**

表 1 和表 2 表明，在识别结果方面，所提模型的所有指标均优于 CNN+RNN 模型，与文献[8]提出的模型相比，除了准确率相同外，其余指标均为最优；对于检测结果，所有指标均优于 CNN+RNN 模型和文献[8]提出的模型。以上结果说明，在弱标签环境下，所提方法对最终的识别与检测效果都有所提升。

但是，所提模型由于将门控注意力机制与多尺度卷积进行融合，相比其他算法，计算复杂度有所提高。以双核 3.2GHz 处理器为例，每进行一次预测，CNN+RNN 网络需要 1.775 s，文献[14]网络需要 3.549 s，而所提网络模型需要 10.649s。

结束语 针对弱标签环境下的声音识别与检测问题，本文在 CRNN 模型的基础上提出了多尺度注意力融合方法。门控注意力机制可以减小时频图中噪声部分的影响，并且能更多地利用有效特征；多尺度卷积融合则考虑了不同大小的卷积核对应不同的感受野，通过不同尺寸卷积核卷积后的特征图将更加具有多样性，将二者结合可以学习到更加丰富、有效的特征。在音频识别方面，使用了一种结合帧检测结果的加权法进行最终识别。在城市交通工具和警报声的弱标签数据集中，所提模型的识别与检测效果比 CNN+RNN 模型和文献[14]提出的模型更好。本文将时频图作为原始特征输入到网络中进行识别检测，而特征的选取对网络模型及识别检测结果有较大的影响，未来将研究不同的声音特征对模型识别检测精度的影响。

参 考 文 献

[1] KUMAR A，RAJ B. Audio event detection using weakly labeled data[C]// Proceedings of the 2016 ACM on Multimedia Conference. ACM，2016：1038-1047.

[2] TSENG S Y，LI J，WANG Y，et al. Multiple Instance Deep Learning for Weakly Supervised Small-Footprint Audio Event Detection[C]// Proc. Interspeech. 2018：1-5.

[3] CHOU S Y，JANG J S，YANG Y H. Frame CNN：A weakly supervised learning framework for frame-wise acoustic event detection and classification [R]. DCASE2017 Challenge，2017.

[4] DIMITROV S，BRITZ J，BRANDHERM B，et al. Analyzing sounds of home environment for device recognition[C]// European Conference on Ambient Intelligence. Cham：Springer，2014：1-16.

[5] BOGDANOV D，WACK N，GóMEZ E，et al. Essentia：an open-source library for sound and music analysis[C]// Proceedings of the 21st ACM international conference on Multimedia. ACM，2013：855-858.

[6] ANWAR，M Z，KALEEM Z，et al. Machine learning inspired sound-based amateur drone detection for public safety applications [J]. IEEE Transactions on Vehicular Technology，2019(68)：2526-2534.

[7] PARASCANDOLO G，HEITTOLA T，HUTTUNEN H，et al. Convolutional recurrent neural networks for polyphonic sound event detection [J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2017，2(6)：1291-1303.

[8] ZHOU Q，FENG Z R，BENETOS E. Adaptive Noise Reduction for Sound Event Detection Using Subband-Weighted NMF [J]. sensors，2019，19(14)：3206.

[9] CAKIR E，VIRTANEN T. End-to-End Polyphonic Sound Event Detection Using Convolutional Recurrent Neural Networks with Learned Time-Frequency Representation Input[C]// 2018 International Joint Conference on Neural Networks (IJCNN). 2018.

[10] XIA X，TOGNERI R，SOHEL F，et al. Random Forest Classification based Acoustic Event Detection Utilizing Contextual-Information and Bottleneck Features [J]. Pattern Recognition，2018(81)：1-13.

[11] CHOI K，FAZEKAS G，SANDLER M. Automatic tagging using deep convolutional neural networks [J]. arXiv：1606.00298.

[12] XU Y，KONG Q，HUANG Q，et al. Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging[C]// Proc. Interspeech. 2017：3083-3087.

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

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

[15] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2015：1-9.

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

ZHENG Wei-zhe，born in 1998，postgraduate. His main research interests include sound and image recognition.

WEI Juan，born in 1973，Ph.D，associate professor. Her main research interests include sound localization and recognition.