NCMMSC

会议论文集

第十六届
全国人机语音通讯学术会议

NATIONAL CONFERENCE ON MAN-MACHINE SPEECH COMMUNICATION

2021年10月15-18日　江苏 · 徐州

主办单位

中国中文信息学会　中国计算机学会

协办单位

中国声学学会语言、听觉与音乐声学分会　中国语言学会语音学分会　中国电子学会信号处理分会

承办单位

江苏师范大学　北京工业大学

基于点积自注意力卷积神经网络的歌声检测

桂文明[1,2]　曾岳[1]　夏泽宇[1]

(1. 金陵科技学院软件工程学院，江苏南京 211169；2. 南京邮电大学 宽带无线通信与传感网技术教育部重点实验室，江苏南京 2100031.)

摘　要：传统的歌声检测过程往往包含了复杂的特征工程，而基于深度神经网络统一框架的算法则可以利用其强大的学习能力学习到特征，从而忽略特征工程。但是，这些学习到的特征通常得不到重要性区分，在网络中所占权重相同。针对这一问题，提出在卷积神经网络中嵌入点积自注意力模块的算法，该算法通过学习得到各个特征的注意力分布，调整注意力权重，使得卷积神经元在“观察”这些特征时能区分轻重，从而提升网络的整体性能。在实验部分，通过在两个公开数据集下测试，并和基准模型进行对比，证明了该算法对提升歌声检测水平切实有效。

关键词：歌声检测；卷积神经网络；余弦注意力；点积自注意力

中图分类号：TP391　　**文献标识码：**A　　**DOI:**10.16798/j.issn.1003-0530. ****.**.***

Singing voice detection algorithm based on a scaled dot-product attention embedded convolutional neural network

GUI Wenming[1,2]　ZENG Yue[1]　XIA Zeyu[1]

(1. School of Software Engineering, Jinling Institute of Technology, Nanjing 211169, Jiangsu, China；2. Key Lab of Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications), Ministry of Education, Nanjing 210003, Jiangsu, China)

Abstract: The complicated feature engineering usually plays a significantly important role in the conventional singing voice detection algorithm, while it could be neglected in those algorithms based on the deep neural network because they can learn the features through their strong learning capability. However, the learned features are treated equally in the network despite their different importance for the result. To address this problem, a scaled dot-product attention embedded convolutional neural network was proposed, in which attention distribution for the feature maps was achieved by learning, and then the weights of the feature maps were adjusted so that the convolutional neurons could distinctively “observe” the features in terms of importance, resulting in the overall performance improvements. In the experimental section, compared to the base line model, with the experiments on the two public datasets, the results proved the effectiveness of this algorithm.

Key words: singing voice detection; convolutional neural network; cosine attention; scaled dot-product attention

1　引言

歌声检测（Singing Voice Detection, SVD）严格来说是检测音乐中存在的歌声区域的过程。但在实践中，我们往往是检测帧级别的音乐片段是否含有歌声。歌声检测是音乐信息检索（Music Information Retrieval, MIR）领域的重要基础性工作，歌手识别[1]，旋律提取[2]，歌声分离[3]，歌词对齐[4]等研究方向都把歌声检测作为前驱工作或者增强技术。歌声检测的主要难点是鉴别歌声和乐器的声音。要在混合了乐器和歌声的

收稿日期：****-**-**；修回日期：****-**-**

基金项目：国家自然科学基金(61872199)；南京邮电大学宽带无线通信与传感网技术教育部重点实验室开放研究基金资助；江苏省教育厅高校优秀中青年教师和校长境外研修项目；金陵科技学院和澳大利亚昆士兰科技大学中外合作办学高水平示范性建设工程

音乐片段中判断是否含歌声，对机器来说目前仍是颇具挑战性的工作。特别是一些乐器在制作过程中，发声原理就是根据人的发声过程来模拟制作的，鉴别它们和人声更是困难的事情。当前歌声检测工作主要集中在流行音乐领域，最近 Krause 等研究了歌剧中的歌声检测[5]。

一般地，歌声检测通过特征提取和分类两个步骤来完成。特征提取是根据歌声区别于乐器的特性，对音乐信号进行变换、降维等加工处理的过程。早期的特征提取技术大多来自于语音识别，在提取的特征中，最常用的特征是经过短时傅里叶变换后的时频图，其变形包括梅尔时频图和对数梅尔时频图等。其他特征还包括梅尔频率倒谱系数 MFCCs（Mel Frequency Cepstral Coefficients）、线性预测系数 LPCs（Linear Predictive Coefficients）、过零率 ZCR（Zero Cross Rate）等等。后来研究人员开始注重挖掘歌声的音乐特性，并组合多个包含传统语音特性和音乐固有特性的特征，一同作为歌声检测的最终特征。Regnier 等认为歌声的特征包括和声（Harmonicity）、共振峰（Formants）、颤音（Vibrato）和震音（Tremolo）[6]，并提出一种提取颤音和震音参数来进行歌声检测的方法；Mauch 等提出了从基音波动和旋律信息中提取歌声特征的方法[7]。Lehner 等提出一种基于基音波动的动谱特征（Fluctogram）[8]，并辅以两个可靠因子：谱平坦因子（Spectral Flatness）和谱收缩因子（Spectral Contraction）来鉴别歌声和乐器，并组合 MFCCs 来进行歌声检测。Zhang 等在他们的歌声检测算法中组合了包括 Chroma 在内的 8 中不同的特征[9]。上述深度挖掘音乐信号信息，提取歌声特征的复杂过程，是歌声检测过程的特征工程。

分类步骤根据特征工程提取的特征，对音乐片段进行分类。分类方法包括两种：基于传统分类器的方法和基于深度神经网络 DNN（Deep Neural Network）分类器的方法。传统的分类器包括隐马尔可夫模型 HMM（Hidden Markov Model）、支持向量机 SVM（Support Vector Machine）、随机森林 RF（Random Forest）等；DNN 分类器包括各种卷积神经网络 CNN（Convolutional Neural Network）[10]分类器、各种循环神经网络 RNN（Recurrent Neural Network）[11, 12]分类器、以及各种混合 CNN 和 RNN 的分类器，比如 LRCN（Long-Term Recurrent Convolutional Network）[9]分类器。

特征工程，再加上 DNN 强大的分类能力使得基于 DNN 的算法占据当前最高歌声检测水平的行列[8, 9]。事实上，DNN 的分类能力强大来自于它的特征学习能力。因此，在歌声检测算法领域，有一种新的趋势是把特征工程和分类两个独立的过程统一到一个 DNN 框架，一步完成（以下称统一 DNN 框架）。Schlüter 等在 CNN 中[10, 13]，Leglaive 等在 RNN 中[12]都没有特征工程步骤，网络输入是简单的时频图特征。其中文献[13]输入网络的是梅尔时频图，而文献[10, 12]输入网络的是对数梅尔时频图。对于统一 CNN 框架算法[13]，文献[14]在 Jamendo 数据集下做了第三方的比较，其性能（准确率为 86.8%）和经过复杂特征工程的 RF 算法（准确率为 87.9%），以及经过歌声分离预处理的 RNN 算法相比（准确率为 87.5%）差距很小，说明统一 DNN 算法是可行的。

本文探讨的是如何在统一 CNN 框架下提升歌声检测算法性能的问题。在统一 CNN 算法中，学习到的特征，可以认为存在于各层次的特征图中，是通过卷积核“观察”上一层特征图得到的。然而对于一般 CNN 学习到的特征，CNN 在“观察”他们时往往不进行重要性区分，在网络中所占权重相同。针对这一问题，本文提出在卷积神经网络中嵌入点积自注意力模块的算法，该算法通过学习得到各个特征的注意力分布，调整注意力权重，使得卷积神经元在“观察”这些特征时能区分轻重，从而提升网络的整体性能。

2 相关研究

2.1 注意力机制背景

Graves 等首次提出神经网络中的注意力机制[15]，他在设计神经图灵机的内存时提出一种基于内容的寻址机制，不同于传统的基于确定性的地址和存储内容的读写，这种读写机制是对工作内存的模糊(blurry)操作。例如在读操作过程中，某时刻的读操作可以用如下公式表示：

$$\boldsymbol{r} = \mathrm{softmax}(\mathrm{g}(\boldsymbol{k},\boldsymbol{M}))\boldsymbol{M} \tag{1}$$

其中 $\boldsymbol{M} = (\boldsymbol{m_1}, \boldsymbol{m_2}, ...\boldsymbol{m_n})^T$ 是工作内存矩阵，包含 n 个行向量，$\boldsymbol{k}$ 是读写头产生的关键字，而

$g(u,v)=\frac{u\cdot v}{||u||\cdot||v||}$ 是一个余弦函数，用来度量 $\boldsymbol{k}$ 和工作内存的行向量 $\boldsymbol{m}_i, i\in[1..n]$ 之间的相似性。此时，读出的向量 $\boldsymbol{r}$ 是内存向量的加权求和。这种机制就是余弦注意力机制。

自从 Graves 等的注意力机制在 RNN 实施后，对于 DNN 中的注意力机制研究迅速升温，Bahdanau 等提出了应用于机器翻译和文字对齐的加性注意力模型[16]，Luong 等提出了一系列全局和局部注意力的模型，同样应用于机器翻译[17]。

2.2 点积自注意力网络

以上提及的注意力模型目标都是为某一外部向量求解注意力分布，Vaswani 等提出著名的自注意力网络（或称点积自注意力网络），则是关注自身内部向量的关系[18]。他把向量键值对 $<\boldsymbol{k},\boldsymbol{v}>$ 和查询向量 $\boldsymbol{q}$ 通过点积注意力(Scaled dot-product, Sdp)映射为另一个向量，其矩阵形式的映射过程可用如下公式表示：

$$\mathrm{sdp}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{2}$$

其中 d_k 是关键字向量 $\boldsymbol{k}$ 的维度，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别是多个查询向量 $\boldsymbol{q}$ 和多个向量键值对 $<\boldsymbol{k},\boldsymbol{v}>$ 组成的矩阵。虽然点积自注意力和余弦注意力机制都包含了点积过程，但是二者的性质差别很大，一是关注点不一样，自注意力模型关注的是内部向量之间的关系；二是计算注意力分布时，自注意力模型不是度量真实向量之间的相似性，而是度量和中间向量 $\boldsymbol{k}$ 的相似性，向量 $\boldsymbol{k}$ 是真实向量 $\boldsymbol{v}$ 的抽象，这就为实际操作带来了更多的灵活性。

2.3 基于 CNN 的歌声检测模型

图 1 构造的 CNN 网络是典型歌声检测算法模型之一，文献[14]使用的 CNN 模型和文献[10, 13]是都是基于此模型的，该模型包含 4 个卷积层和 3 个全连接层，本文把该模型当做基准的对比模型。

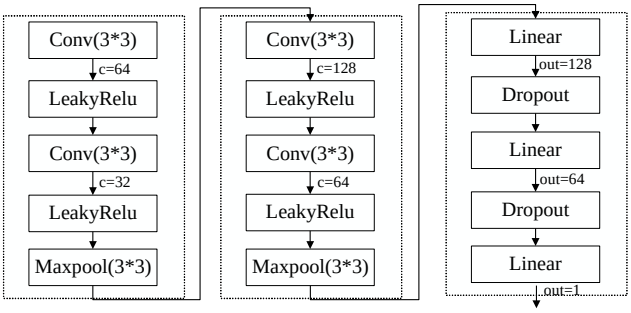


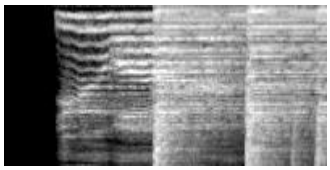
图 1 基于 CNN 的歌声检测网络结构图

Fig.1 The network structure of singing voice detection based on CNN

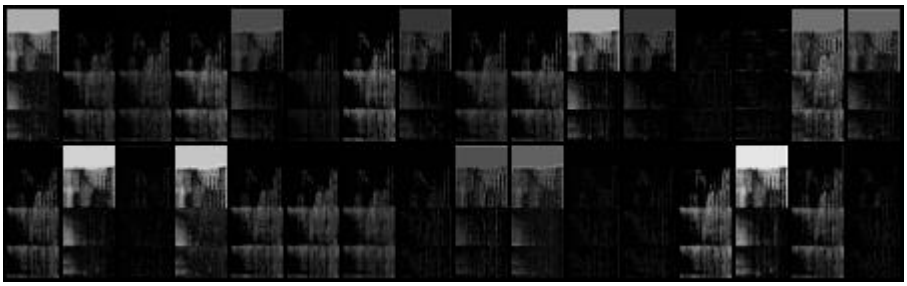
上图中，每个卷积层后面有一个 LeakyRelu 增强非线性，每两个卷积层分别紧跟一个最大值池化层缩小特征图组成一个卷积组模块，卷积层的输出通道数按顺序分别是 64, 32, 128, 64，卷积层的输出特征图进入全连接层前平铺成向量，然后进入三个线性变换层逐层降低向量的维度，每个线性变换紧跟一个 Dropout 层防止过拟合，线性变换的输出长度分别是 128, 64, 1，最后输出的一维向量用于二分类。

3 算法模型

3.1 算法动机



(a) 音乐 You are 的对数梅尔时频图片段



(b) CNN 的 64 通道特征图

图 2 音乐 You are 的对数梅尔时频图输入片段和 CNN 中间层各通道输出特征图

Fig.2 An image from the log-mel spectrogram of the music "you are"
and the corresponding output feature maps from a CNN median layer

在一般基于 CNN 的方案中，每一层次的特征都是不加重要性区分地进入下一层，下一层“观察”他们时，把他们的权重看成相同的，然而这些由 CNN 学习到的特征往往对歌声检测起着不同的作用，有些特征显著而重要，需要得到重点“关注”，而有些特征因其作用微弱需要得到抑制。

例如图 2 中，图 2（b）是图 2（a）输入的对数梅尔时频图片段在一个 CNN 中间层的 64 通道输出的特征图。CNN 的下一层在“观察”该 64 个特征图时，按照一般 CNN 的处理方式，其对应的权重是相等的，但是从图 2（b）中用肉眼显然能看出，他们的权重不应该是相等的，比如其中的黑色特征图因信息量少其权重应该小。因此，如果给这些通道加入权重指标，或者说“观察”时的注意力分布值，让 CNN 把“注意力集中”在关键特征图上，其鉴别特征能力应能得到提高。

3.2 点积自注意力卷积神经网络模型

本算法的目标是通过自注意力模型，将 CNN 中学习到的特征根据注意力分布进行重估，然后再进入下一层网络，其模型可通过图 3 描述。图中 CNN 网络的 n-1 层的特征图 $\boldsymbol{F}=(\boldsymbol{F}_1,\boldsymbol{F}_2...,\boldsymbol{F}_m)$ 经过 Sdp 模块进行了注意力重估，转换成 $\boldsymbol{F}'=(\boldsymbol{F}'_1,\boldsymbol{F}'_2...,\boldsymbol{F}'_m)$ 再进入到第 n 层网络。

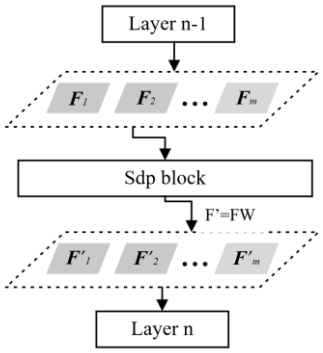


图 3 特征图注意力重估

Fig.3 The attention is adjusted for the feature maps

大多数自注意力网络应用于机器翻译，构建于 RNN，据作者所知，到目前为止，没有自注意力网络在歌声检测框架中的应用。为了适用于 CNN 模型的歌声检测，对自注意力网络需要做如下调整：

（1）$\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$ 的表达含义变化。原自注意力网络的 $\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$ 在进行点积前都进行了线性变换，而本文的 $\boldsymbol{v}$ 保持不变，代表原特征图，网络学习到的自注意力分布和 $\boldsymbol{v}$ 相乘可得到新的特征图；而 $\boldsymbol{q},\boldsymbol{k}$，它们分别经过线性变换，长度由 h*w 变为 h，代表着特征图的抽象，其中 $\boldsymbol{k}$ 是特征图 $\boldsymbol{v}$ 的键值，$\boldsymbol{q}$ 是特征图查询向量，$\boldsymbol{q},\boldsymbol{k}$ 在线性变换后都经过一个 Relu 单元，以增强非线性特性，如图 4。

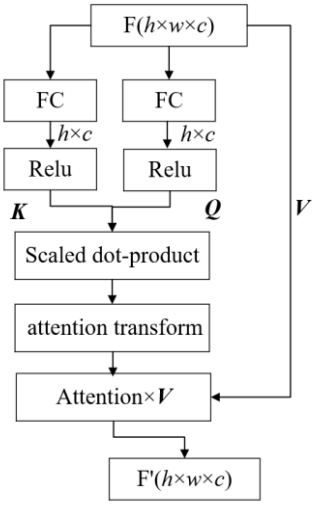


图 4 本文点积注意力模块结构

Fig.4 The module structure of the scaled dot-product attention

（2）$\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$ 的长度不等。原自注意力网络处理对象是词向量，输入与输出向量长度相等，三者保持长度不变，有利于各层次网络处理。而本算法的处理对象是对数梅尔时频图，且 CNN 网络的特征图大小在网络中发生变化，因此，这三者的向量长度应自适应变化。本算法中 $\boldsymbol{q},\boldsymbol{k}$ 的长度保持和特征图的高度相同，而高度是时频图中的频率个数；$\boldsymbol{v}$ 的长度是特征图的高度和宽度之积，是特征图展开的向量。

（3）增加一个注意力分布变换（attention transform）机制。本算法在得到注意力分布后，还增加了一个变换机制，从分布矩阵形式变换为特征图的权重向量，降低了注意力重估的复杂度。变换过程可以用以下公式表示：

$$\boldsymbol{w} = mean(\boldsymbol{R}(1-\boldsymbol{E}), \dim = 1) \tag{3}$$

其中 $\boldsymbol{R}$ 和 $\boldsymbol{E}$ 分别是自注意力分布矩阵和单位对角矩阵，$\boldsymbol{R}(1-\boldsymbol{E})$ 则将注意力分布矩阵的对角线置零，不计算查询向量对自身的注意。在特征图的频率维度（dim=1）上取均值得到各特征图的注意力权重 $\boldsymbol{w}$ (attention)。最后得到加权注意力后的特征图 $\boldsymbol{F}' = \boldsymbol{w}\boldsymbol{V}$ 。

为了检验点积自注意力模块的有效性，本文对图 1 基准模型进行改造，嵌入点积自注意力模块。嵌入的方法是在两个卷积组模块（图 1 中最大值池化层）后分别进行嵌入，这样就在基准模型中一共增加了两个点积自注意力模块。增加的模块分别对其输入的特征图进行注意力权重重估，并把重估后的特征图送入到网络的下一层。由此，我们可以通过嵌入前后的实验结果来检验模块的有效性。此外，对于 CNN 中输出 1 维改成输出 2 维（图 1 最后一个线性变换），用以应用交叉熵损失函数，由于嵌入前后均使用同样的交叉熵损失函数，因此，不会改变实验结果的公平性。

3.3 对数梅尔时频图输入

本算法重点关注对特征的注意力重估，因此忽略复杂的特征工程。本算法的输入是歌声检测中的常用基本特征，即对数梅尔时频图(log mel-spectrogram)，首先对音频文件进行短时傅里叶变换，计算时分别对音频数据序列左右分别填充窗口长度的一半，此处音频文件采样率为 22050Hz，窗长为 1024，帧移为 315，时频的时间分辨率是 14.3ms；然后对傅里叶变换后的时频图进行梅尔刻度规范化，考虑到音乐的频率范围，此处频率区间取[27.5, 8000]Hz，梅尔频率数量取 80 个；最后对梅尔时频图的幅值取对数，便可得到对数梅尔时频图。得到的对数梅尔时频图实际上是一个二维的矩阵, 矩阵的行表示梅尔频率序号，矩阵的列对应音乐的进行时间。

输入到 CNN 网络时，从对数梅尔时频图矩阵的起始列位置开始提取图像，图像大小为 80*115，提取下一个图像时设置跳数为 5 列，直到该对数梅尔时频图到达末尾为止。图像的宽度决定了输入到卷积神经网络的每个图像所属的音乐时长，为 1.6s。跳数决定了算法的检测精度，为 71.5ms。

3.4 网络设置

本文算法中点积自注意力卷积神经网络的实验部分基于 Pytorch 平台，通过 Homura[19]封装网络输入和训练组件进行实现。模型采用 Adam 作为训练优化器，并在算法中通过早停机制和限制最大训练轮数来控制训练的终止。本文算法中的早停次数设置为 10 次，最大训练轮数设置为 50 轮。歌声检测可看成是一个二分类问题，因此，损失函数使用二分类交叉熵损失函数。又因为流行音乐中一般歌声占比大，纯乐器部分占比小，所以损失函数最终采用加权二分类交叉熵损失函数。在算法的实验中，权重设置为数据集中的歌声和非歌声的样本数量比例。设数据集中有 N 个样本，样本预测为歌声的概率为 x_i，样本所属标签为 y_i，其权重为 w_i 则算法的加权二分类交叉熵损失函数为：

$$l = -\frac{1}{N}\sum_{i} w_i[y_i \log x_i + (1-y_i)\log(1-x_i)] \tag{4}$$

其中 $i \in [1, N]$。

4 实验

4.1 数据集

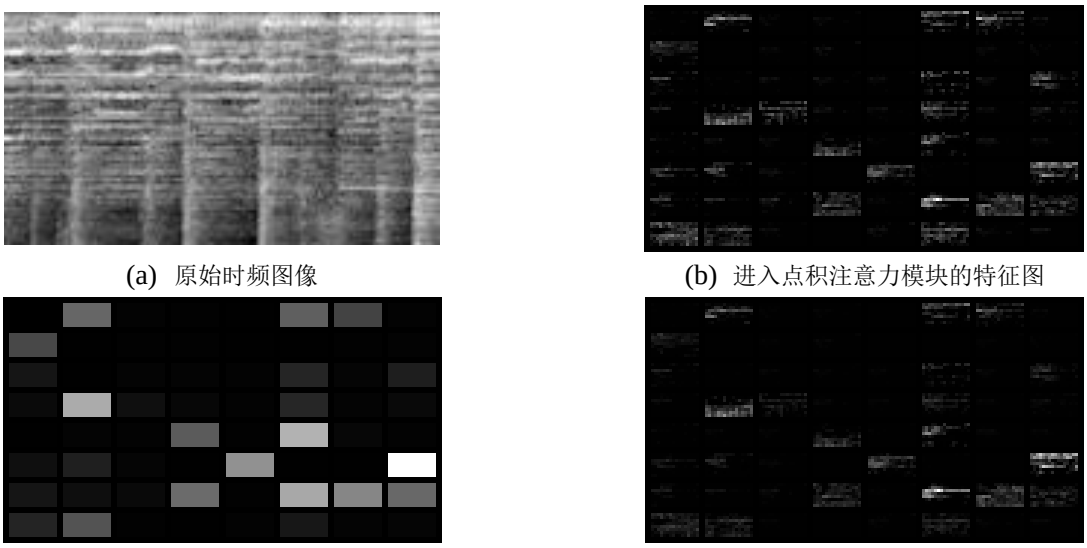
为了验证算法的有效性，本文一共选择了 2 个公开数据集进行实验，一个是数据集 RWC 中的流行歌曲（简称 RWC），另一是数据集 Jamendo（简称 JMD），两个数据集都有歌声和非歌声的标注信息。其中 RWC 包含 100 首流行歌曲，共 407 分钟；JMD 包含 93 首歌曲，共 371 分钟。文献[14]对于 JMD 的训练、验证和测试集的划分有一个固定的样板，在本实验中保持不变。对于 RWC，其他文献没有公开的划分方式，本文的划分方式是把数据集文件名中结尾为 0-4 的文件划为训练集，把结尾为 5 和 6 的文件划为验证集，而把结尾为 7-9 的文件划为测试集。可以说这种划分方式是准随机的方法，这可以保证实验结果对比的公正性。RWC 和 JMD 的歌声和非歌声的样本数量比分别为 1.12 和 1.55，这个比例将在加权二分类交叉损失函数中使用。

4.2 基准系统实现

本实验是为了比较在 CNN 中嵌入点积自注意力模块前后的网络性能变化，其中 2.3 小节中描述的基于 CNN 的歌声检测系统是用于比较的基准系统。对于该系统的实现，文献[14]有公开的在 keras 框架的实现代码，但是为了保持各条件不变的情况下对比实验结果，因此把文献[14]的代码移植到和本文算法实现相同的 Pytorch 框架,这样网络的训练、验证和测试方法可以保持一致。

4.3 特征图的注意力分布

深度神经网络的不可解释性是它的一个主要弱点。虽然我们无法通过分析完全解释本文点积自注意力模块通过注意力分布调整达到提升 CNN 歌声检测效果的来龙去脉，但是，通过展示网络的特征图注意力分布可以加深对过程的理解。图 5 是音乐 You are 的一个对数梅尔时频图片段（如图 5(a)），进入本文构造的基于点积自注意力的卷积神经网络，在第二个嵌入点积自注意模块前后的特征图的比较。从图 5(b)中看到经过一系列网络“观察”处理后，64 个通道的特征图中分别出现了“辨别是否含有歌声”的纹理特征，然而，这些特征图在辨别能力方面各有差异，点积自注意力网络通过注意力学习，给这些特征图赋以一定的注意力权重（图 5(c) 的灰度小图片表示了对应特征图的权重）。在一般 CNN 中没有点积自注意力模块，这些权重全是相同的，图 5(c)中的灰度小图片应呈现全白色。经过注意力重估后，点积自注意力模块输出的特征图如图 5(d)，从前后特征图对比来看，施加了注意力分布后，64 个特征图中有部分偏白的纹理特征变得暗淡或者消失了（如图 5(d)中的最左列的部分小特征图），这说明部分歌声辨别贡献小的特征图被抑制或剔除了。



(a) 原始时频图像　(b) 进入点积注意力模块的特征图

(c) 点积自注意力权重分布　(d) 点积自注意力权重重估后的特征图

图 5 音乐 You are 的一个对数梅尔时频图片段在点积自注意力权重重估前后的特征图比较

Fig.5 Comparison between the input feature maps and the output feature maps of a scaled dot-product attention module, with an input image from the log-mel spectrogram of the music “you are”

4.4 实验结果

表 1 本文算法和基于 CNN 的算法实验结果对标

Tab.1 The experimental result comparison between the proposed algorithm and the traditional CNN based algorithm

Dataset	μ±σ (%)	Accuracy	F-measure	Precision	Recall	FP	FN
JMD	CNN	86.54±0.49	86.57±0.15	80.93±2.17	93.16±2.86	19.22±3.36	6.84±2.86
	Sdp-CNN	88.36±0.49	87.93±0.47	85±0.98	91.08±0.82	14.01±1.15	8.92±0.82
	μ 提升	1.82	1.36	4.07	-2.08	-5.21	2.08
RWC	CNN	88.95±0.51	90.65±0.52	91.77±0.6	89.57±1.5	11.98±1.14	10.43±1.5
	Sdp-CNN	90.93±0.19	92.28±0.17	94.05±0.42	90.57±0.52	8.54±0.68	9.43±0.52
	μ 提升	1.98	1.63	2.28	1.00	-3.44	-1.00

实验结果的评估标准包括准确率(Accuracy)、F 值(F-measure)、精确率(Precision)、召回率(Recall)、假正例(FP)和假负例(FN)比率，其中 F 值是精确率和召回率的综合。因 DNN 训练收敛有一定的随机性，因此，使各算法分别在 JMD 和 RWC 下分执行 3 次，计算各指标百分数的均值和方差(μ±σ)，实验结果见表 1。表中可看出，本文对于 CNN 算法的实现在 JMD 下的准确率和文献[14]结果基本一致，这在一定程度上反应本实验中算法实现的正确性。表中各指标的方差都在可接受范围之内，不存在太大的波动。在 JDM 和 RWC 两个数据集下，本算法相对基于 CNN 的算法，准确率和 F 值均有提升，这说明在 CNN 中嵌入点积自注意力模块对歌声检测的有效性。二者的精确率都有相对其他指标最大幅度的提升，带来的是假正例比率的下降。在 JMD 下，虽然精确率提升幅度大，但是假负例比例上升，有可能是 JMD 中有部分乐器的泛音特征和歌声类似造成的。总的来说，在 RWC 下的结果要好于 JMD，这是因为 RWC 的歌曲都是流行歌曲，其训练、验证、测试集的分布相对 JMD 来说更趋一致，网络相对容易学习到更有效的特征。

5 结论

本文提出一种新型歌声检测算法，在卷积神经网络中嵌入了点积自注意力模块，该模块使得卷积网络学习到的特征在网络中的注意力分布不再是相同的，这种注意力重估机制使得各特征得到网络不同的对待，从而提升整体网络性能。通过实验证实，在原基准 CNN 模型中嵌入点积自注意力模块，歌声检测的准确率

和 F-值等指标均有一定的提升。

参考文献

[1] HSIEH T H, CHENG K H, FAN Zhecheng, et al. Addressing the confounds of accompaniments in singer identification[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 1-5.

[2] KUM S, NAM J. Joint detection and classification of singing voice melody using convolutional recurrent neural networks[J]. Applied Sciences, 2019, 9(7): 1324.

[3] PRÉTET L, HENNEQUIN R, ROYO-LETELIER J, et al. Singing voice separation: A study on training data[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 506-510.

[4] GUPTA C, YıLMAZ E, LI Haizhou. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? [J]. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 496-500.

[5] KRAUSE M, MÜLLER M, WEIß C. Singing voice detection in opera recordings: A case study on robustness and generalization[J]. Electronics, 2021, 10(10): 1214.

[6] REGNIER L, PEETERS G. Singing voice detection in music tracks using direct voice vibrato detection [C]//2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009: 1685-1688.

[7] MAUCH M, FUJIHARA H, YOSHII K, et al. Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music [C]// 2011 12th International Society for Music Information Retrieval Conference (ISMIR). Florida, USA. 2011: 233-238.

[8] LEHNER B, SCHLÜTER J, WIDMER G. Online, loudness-invariant vocal detection in mixed music signals[J]. IEEE/ACM Transactions On Audio, Speech, And Language Processing, 2018, 26(8): 1369-1380.

[9] ZHANG Xulong, YU Yi, GAO Yongwei, et al. Research on singing voice detection based on a long-term recurrent convolutional network with vocal separation and temporal smoothing[J]. Electronics, 2020, 9(9): 1458.

[10] SCHLÜTER J. Learning to pinpoint singing voice from weakly labeled examples [C]// 2016 International Society for Music Information Retrieval (ISMIR). New York, USA, 2016: 44-50.

[11] LEHNER B, WIDMER G, BÖCK S. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks[C]//2015 23rd European Signal Processing Conference (EUSIPCO). Nice, France. IEEE, 2015: 21-25.

[12] LEGLAIVE S, HENNEQUIN R, BADEAU R. Singing voice detection with deep recurrent neural networks[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015: 121-125.

[13] SCHLÜTER J, GRILL T. Exploring data augmentation for improved singing voice detection with neural networks [C]// International Society for Music Information Retrieval (ISMIR). Malaga, Spain, 2015: 121-126.

[14] LEE K, CHOI K, NAM J. Revisiting singing voice detection: A quantitative review and the future outlook [C]// International Society for Music Information Retrieval Conference. Paris (France), Paris (France). International Society for Music Information Retrieval, 2018.

[15] GRAVES A, WAYNE G, DANIHELKA I. Neural turing machines [EB/OL]. 2014.

[16] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. 2014.

[17] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015:1412-1421.

[18] VASWANI A, SHAZEER N, PARMAR N , et al. Attention is all you need [C]// Advances in neural information processing systems, 2017: 5998-6008.

[19] HOMURA. Homura package [EB/OL]. 2019.

桂文明，男，1974 年生，江西鹰潭人。金陵科技学院，副教授，博士，英国伦敦大学玛丽女王学院数字音乐研究中心访问学者，主要研究方向为音乐人工智能、音乐信号处理。

E-mail：guiwenming@126.com

曾岳，男，1972 年生，湖北洪湖人，金陵科技学院教授，博士，主要研究方向为智能信息处理。

E-mail：53901444@qq.com

夏泽宇，男，2001 年生，安徽马鞍山人，金陵科技学院学生，本科在读，主要研究方向为人工智能应用。

E-mail：xiazeyu_2011@126.com