DOI:10.16515/j.cnki.32-1722/n.2021.01.004

一种基于深度残差卷积神经网络的歌声检测算法

桂文明[1],吕家伟[1],敖志强[2]

(1.金陵科技学院软件工程学院,江苏 南京 211169;2.南昌航空大学软件学院,江西 南昌 330063)

摘 要:歌声检测是音乐人工智能领域重要的基础性工作,也是很多相关研究的必备技术或者增强技术。提出一种基于深度残差卷积神经网络的歌声检测算法,该算法在仅仅输入简单朴素特征的情况下,通过多层次卷积神经网络,能学习到比浅层卷积神经网络更多的、更有效的歌声特征,从而提高算法的整体性能。根据2种基本的残差网络结构,设计了6种不同深度的卷积神经网络,通过与基线系统的实验结果进行比较,证明了新算法的性能优于基于浅层卷积神经网络算法的性能。同时,新算法的网络深度可调性为应用增加了灵活性。

关键词:歌声检测;残差网络;深度神经网络;卷积神经网络;循环神经网络

中图分类号:TN912.2 **文献标识码**:A **文章编号**:1672-755X(2021)01-0019-05

A Singing Voice Detection Algorithm Based on Deep Residual Convolutional Neural Network

GUI Wen-ming[1], LYU Jia-wei[1], AO Zhi-qiang[2]

(1. Jinling Institute of Technology, Nanjing 211169, China; 2. Nanchang Hangkong University, Nanchang 330063, China)

Abstract: Singing voice detection is an important segment in the field of musical artificial intelligence, and it is also a necessary technology or enhancement technology for various related studies. In this paper, we propose an algorithm based on deep residual convolutional neural network. The multi-level convolutional neural network can learn more valid singing features than a shallow convolutional neural network when only simple and plain features are input, thereby improving the overall performance of the algorithm. In this paper, based on two basic residual network structures, six convolutional neural networks with different depths are designed. By comparing with the experimental results of the baseline system, it is proved that the performance of the algorithm in this paper is ahead of the algorithm based on shallow convolutional neural network. At the same time, the network depth adjustability of this algorithm also adds more flexibility to its application in practice.

Key words: singing voice detection; residual network; deep neural network; convolutional neural network; recurrent neural network

歌声检测(singing voice detection,SVD)是判断数字形式的一小段音乐中是否含有人的歌声的过程,其检测精度一般为50~200 ms。在每一小段音乐中,除了歌声,一般还含有乐器的声音,要在混有乐器和

收稿日期:2020-08-06

基金项目:江苏省教育厅高校优秀中青年教师和校长境外研修项目(2018-191);金陵科技学院博士科研启动基金(jitb-201509)

作者简介:桂文明(1974—),男,江西鹰潭人,副教授,博士,主要从事音乐人工智能研究。

人声的音乐片段中判断是否含有歌声，虽然对人来说轻而易举，但对机器来说仍然颇具挑战。歌声检测是音乐人工智能领域重要的基础性工作，很多其他研究比如歌手识别、歌声分离、歌词对齐等都需要歌声检测作为必备技术或者增强技术。例如，在歌手识别过程中，首先对音乐进行歌声检测就是必备技术，只有检测到歌声才能通过歌手鉴别过程进行歌手识别；而在歌词对齐过程中，如果能准确地进行歌声检测，必然增强歌词对齐的准确性。

歌声检测是对每一小段音频的二分类过程。可以把这段音频记为 x，假定分类函数为 f，这小段音频若含有歌声则记为1，若不含歌声则记为0，则可以用 $y=f(x)$ 的形式来表示歌声检测问题，其中 y 的值为0或1。

1 歌声检测算法的一般框架

歌声检测的过程一般分为预处理、特征提取、分类和后处理等几部分。

输入的音频文件一般是物理样本级的，例如 wav、mp3 等文件。预处理主要包括对音频信号去噪和对信号进行分频等，也包括利用歌声分离技术在一定程度上先把歌声提取出来再进行处理。

特征提取和用分类器对特征信息进行分类是歌声检测的2个重要步骤。特征提取是从音频信号中提取能区别含或不含歌声的鉴别信息，该鉴别信息称为特征。对歌声检测来说，较简单的特征是短时傅里叶变换后的时频图；此外，还有在时频图基础上提取的诸多特征，包括线性预测系数(linear predictive coefficient，LPC)、感知线性预测系数(perceptual linear predictive coefficient，PLPC)、过零率(zero cross rate，ZCR)、梅尔频率倒谱系数(Mel frequency cepstral coefficients，MFCCs)、动谱(fluctogram)特征、谱平坦(spectral flatness)因子、谱收缩(spectral contraction)因子等。

分类器采取机器学习等方法对特征信息进行分类，主要的分类方法包括支持向量机(support vector machine，SVM)、隐马尔可夫模型(hidden Markov model，HMM)、随机森林(random forest，RF)等，也包括近年来出现的深度神经网络(deep neural network，DNN)等。一些采用卷积神经网络(convolutional neural network，CNN)[1-4]和循环神经网络(recurrent neural network，RNN)[5-6]的分类方法在某种程度上提高了歌声检测的准确率，但是仍有提升空间。

后处理主要是对分类的结果利用光滑处理等技术进行微调，从而达到提高歌声检测准确率的目的。

在现有的歌声检测算法中，研究者们通过精心设计某种特征，并选择某种分类器进行分类，从而达到歌声检测的目的。当单一特征不能满足要求时，研究者们自然会想到组合多种特征[3,7]，因此歌声检测算法的发展历史就是研究者们寻找和设计特征的历史。这种人工设计特征的方式存在明显弊端：一是需要人工研究人声和乐器在频谱信息或者其他信息中的不同点，研究过程的周期较长；二是研究过程中研究者往往需要用人类肉眼去发现，导致特征不可靠；三是特征提取和分类方法的提出是2个独立的阶段，在特征提取后需要额外研究分类方法，再次导致整个研究周期延长，增加了整个检测方案的复杂度。事实上，DNN 不仅可以充当歌声检测框架的分类器，还可以通过多层次的学习，对歌声进行多层次的特征提取[8]。因此，找到合适的 DNN 框架，则可以克服上述弊端。采用适当的 DNN 框架，只需要对音乐文件进行简单朴素的特征提取，不需要进行复杂的特征提取，把复杂的特征提取和分类两个阶段合二为一，通过对输入的简单特征进行多层次特征再学习和分类，完成歌声检测工作。在歌声检测研究领域，当前这方面的工作并不多，大部分基于 DNN 框架的工作还是先进行复杂的与特征有关的工作，然后再输入 DNN 分类器来实现。仅输入简单朴素的特征如梅尔时频图并采用 DNN 框架的工作，目前只有 Schlüter 的 CNN 方案[1-2]。然而在该方案中，CNN 的深度有限，仅有14层，受限于浅层深度，该方案的网络学习能力有限，导致学习到的歌声特征有限，被称为浅层 CNN(shallower convolutional neural network，SCNN)。而且，Schlüter 的浅层方案无法通过堆叠卷积层来增加深度，这是因为它会导致梯度消失，使得堆叠的网络无法进行训练或达到饱和状态。本文提出一种网络深度更深、深度灵活可调、检测结果优于浅层 CNN 的检测算法，并通过实验验证了本算法的性能比浅层 CNN 有所提升。

2　基于深度残差卷积神经网络的歌声检测算法

2.1　残差网络

残差网络（residual network，ResNet）来源于图像分类领域，它在很大程度上解决了梯度爆炸和梯度消失问题，使得网络可以构建得很深而不会退化[9]。残差网络由残差结构叠加组成，残差网络的一般构造如图 1 所示。残差网络不是直接学习堆叠网络的潜在映射 $H(x)$，而是通过增加恒等映射（identity mapping）后拟合一个残差映射（residual mapping）$F(x)=H(x)-x$。残差映射相比潜在映射更容易优化，从而解决了深度增加后产生的梯度问题和网络退化问题。而且残差结构是无侵入式结构，可以叠加到其他网络中，以提升网络的深度和性能。残差网络一般用在 CNN 中，本文的歌声检测算法就是基于 CNN 构造了一个深度残差 CNN。

图 1 中的 $F(x)$包括 2 个权值层和一个 Relu。事实上在构造 ResNet 过程中，$F(x)$可根据需要采取不同的结构。图 2 是用于构建基于深度 CNN 的 ResNet 的两种典型的残差结构：基于基本块（basic block）和基于瓶颈块（bottleneck block）的残差结构。本文设计的深度残差卷积神经网络就是根据这两种结构和目标网络层数进行选择和构建的。深度残差卷积神经网络通过不同的卷积层对音乐信号进行多层次特征自动提取，不必依赖人工设计的特征，降低了特征设计的复杂度。同时，从残差网络的输出信息中可推导出分类结果，不必设计额外的分类器。本算法集特征提取和分类功能于一体，简化了歌声检测的步骤。再者，这两种结构在解决梯度问题的基础上，使得歌声检测的网络可以变得更深，从而使得自动提取的歌声特征更多，有利于歌声的鉴别。

图 2 中，n,i,j 分别表示经过 1×1 或 3×3 卷积后的特征图数量，也就是通道数量。图 2(a)与图 2(b)的不同在于后者具有 3 个卷积层，特征图数量也发生了变化，但二者的输出特征图数量都是相同的。本文中图 2(a)用于构建深度为 18 和 34 的网络，图 2(b)用于构建深度为 50、101、152 和 200 的网络。

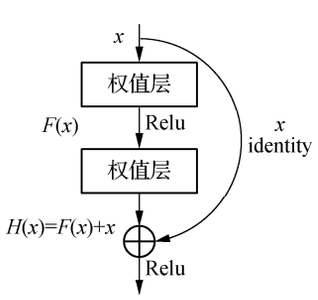
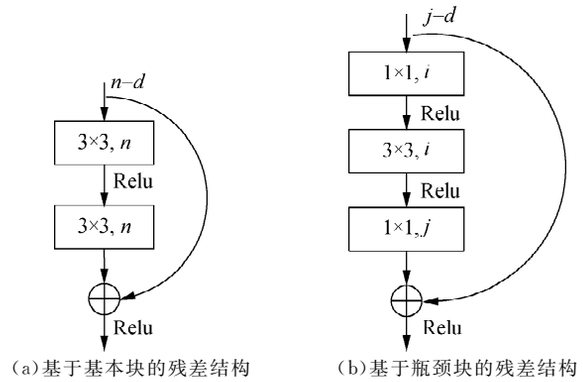


图 1　残差网络的一般结构



(a)基于基本块的残差结构　　(b)基于瓶颈块的残差结构

图 2　两种典型的残差结构

2.2　深度残差卷积神经网络设计

如前所述，通过图 2 所示的 2 种残差结构，本文设计了 6 种深度不一的残差 CNN 网络用于歌声检测，其中 18、34、50、101 和 152 是典型的深度，200 是本文为测试更深网络的检测性能而设计的深度。6 种网络的详细架构见图 3。所有网络的第一层都是一个卷积核大小为 7×7、步数（stride）为 2 的卷积层，之后再经过 3×3 的最大值池化层，步数同样为 2，这使得进入卷积层的图片大小缩小了一半。

由图 2(a)构造的残差 CNN 中（深度为 18 和 34），在 conv2_X、conv3_X、conv4_X、conv5_X 四个层次上，对应的特征图数量分别为 $n=[64,128,256,512]$，而在由图 3(b)构造的残差 CNN 中（深度为 50、101、152 和 200），对应的特征图数量分别为 $i=[64,128,256,512]$，$j=[256,512,1\ 024,2\ 048]$。此外，构造不同深度的 CNN 过程中，还有一个重要参数（即规模参数），该参数规定了 conv2_X、conv3_X、conv4_X、conv5_X 四个层次残差结构的堆叠个数，并最终确定网络的总层数。深度为 18、34、50、101、152 和 200 的 CNN 对应的规模参数分别为[2,2,2,2]，[3,4,6,3]，[3,4,6,3]，[3,4,23,3]，[3,8,36,3]和[3,12,48,3]，

layer name	18-layer	34-layer	50-layer	101-layer	152-layer	200-layer
conv1	7×7, stride 2					
	3×3 max pool, stride 2					
conv2_X	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_X	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8	[1×1, 128; 3×3, 128; 1×1, 512] ×12
conv4_X	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1 024] ×6	[1×1, 256; 3×3, 256; 1×1, 1 024] ×23	[1×1, 256; 3×3, 256; 1×1, 1 024] ×36	[1×1, 256; 3×3, 256; 1×1, 1 024] ×48
conv5_X	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2 048] ×3	[1×1, 512; 3×3, 512; 1×1, 2 048] ×3	[1×1, 512; 3×3, 512; 1×1, 2 048] ×3	[1×1, 512; 3×3, 512; 1×1, 2 048] ×3
	average pool, 2-d fc					

图 3 本文设计的 6 种深度的残差 CNN 的详细结构

其中，虽然深度为 34 和 50 的规模参数一致，但是因为所构成网络的残差结构不一样，最终的网络层数并不一样。

所有网络的最后一层都经过一个均值池化层，然后通过一个全卷积层，把不同长度的输入信息转换为长度为 2 的输出信息，分别对应的是非歌声和歌声，可用于导出分类。

2.3 算法实现

本文深度残差 CNN 的构建采用 Pytorch 方法，并借助 Homura 包[10]进行实现。首先将每个音频文件转换成一个包含对数梅尔时频图(log Mel-spectrogram)的文件。计算过程是先计算音频信号的时频图(spectrogram)，音频采样率 $f_s=22\ 050$ Hz，帧长 $l=1\ 024$，帧移 $h=315$；然后把时频图转换成梅尔时频图，转换时，梅尔频率数量为 80 个，频率区间取[27.5，8 000]Hz，梅尔频率数量对应时频谱图的行数 H；最后对梅尔时频图中的幅值取对数，就可以得到对数梅尔时频图。一个对数梅尔时频图等同于一个数据矩阵 $\boldsymbol{A}(H,L)$，L 由音频的长度确定。然后，从对数梅尔时频图的起始列位置开始提取大小为 80×115 的图像数据，并把数据送入到深度残差 CNN 中进行训练、验证或测试，其中读取图片时每跳(hop)为 5。深度残差 CNN 的输入即对数梅尔时频图是音乐人工智能领域常用且简单的一种特征。

如前所述，每张图片的网络输出为 2 个输出值，对应着非歌声和歌声的分类信息，如果对应一个批次的图片，网络输出就是一个 2 维的向量，本文在实验中将批次大小设定为 64。在网络输出的结果上，本文应用二分类交叉熵损失函数对网络进行训练，优化器采用 Adam，训练过程中使用早停机制(early stopping)和最大轮数(epoch)结束训练，其中早停机制的次数(patience)为 10，最大轮数设为 50。

3 实验及分析

3.1 实验数据集

本文选择公开数据集 RWC(real world computing)中的流行歌曲(以下简称为“RWC 数据集”)作为实验和分析对象。该实验数据集包含 100 首流行歌曲，共 407 min。采用准随机的方法把 RWC 数据集分成训练、验证和测试 3 个数据集，以保证实验结果比较公正。将所有 RWC 数据集中结尾为 0～4 的文件划分为训练集，结尾为 5 和 6 的文件划分为验证集，而结尾为 7～9 的文件划分为测试集。

本文选择文献[11]中的系统作为用于比较的基线系统，该文献实现了基于浅层 CNN 的歌声检测，并公开了代码。该浅层 CNN 包含 4 个卷积层和 3 个全连接层，是目前歌声检测准确率最高的框架之一，网络输入是对数梅尔时频图。运行该系统的歌声检测代码，并在 RWC 数据集上进行实验。

3.2　实验结果与分析

使用本文提出的歌声检测算法，在 RWC 数据集上进行训练、验证和测试，并与基线系统进行对比，结果如表 1 所示。其中，F 值(F-measure)是一个综合查全率(recall)和查准率(precision)的折中指标，FP 是将负类预测为正类的误报率，FN 是将正类预测为负类的漏报率，$\mu\pm\sigma$ 是 6 种不同深度的实验结果的均值和方差。从实验结果看，本算法所有均值指标都优于 SCNN，准确率(accuracy)、F 值、查全率、FN 的所有指标均优于 SCNN 的相应指标，仅查准率和 FP 中各有 2 个指标比 SCNN 差。实验结果有力地证实了本算法的有效性和先进性。如果选择深度为 101 的网络，则准确率较 SCNN 提升了 3.42 个百分点。

表 1　本文算法与 SCNN 的实验结果比较

算法类型	accuracy/%	F-measure	precision/%	recall/%	FP/%	FN/%
SCNN	87.94	89.73	91.46	88.07	12.25	11.93
ResNet18	90.39	91.90	92.47	91.34	11.02	8.66
ResNet34	90.89	92.23	93.86	90.67	8.79	9.33
ResNet50	90.28	91.94	91.06	92.83	13.51	7.17
ResNet101	91.36	92.67	93.93	91.44	8.75	8.56
ResNet152	90.92	92.49	91.35	93.67	13.15	6.33
ResNet200	90.20	91.79	91.82	91.76	12.11	8.24
ResNet $\mu\pm\sigma$	90.77±0.46	92.25±0.35	92.53±1.24	91.99±1.1	11.04±2.09	8.01±1.1

4　结　语

本文提出一种基于深度残差 CNN 的歌声检测算法，残差结构使得网络的深度可以扩张至 200 层甚至更多。通过更深的网络，本算法能在仅仅输入简单朴素特征的情况下，学习到歌声的更多有效特征，从而提升算法的性能。实验结果显示在 RWC 数据集下，本算法的准确率可比 SCNN 高 3.42 个百分点。

参考文献：

[1] SCHLÜTER J. Learning to pinpoint singing voice from weakly labeled examples[C]. New York：ISMIR，2016：44－50

[2] SCHLÜTER J，GRILL T. Exploring data augmentation for improved singing voice detection with neural networks[C]. Malaga：ISMIR，2015：121－126

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

[4] HUANG H M，CHEN W K，LIU C H，et al. Singing voice detection based on convolutional neural networks[C]. Taipei：7th International Symposium on Next Generation Electronics(ISNE)，2018：1－4

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

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

[7] LEHNER B，WIDMER G，SONNLEITNER R. On the reduction of false positives in singing voice detection[C]. Florence：IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP)，2014：7480－7484

[8] ZEILER M D，FERGUS R. Visualizing and understanding convolutional networks[C]. Zurich：European Conference on Computer Vision，2014：818－833

[9] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]. Las Vegas：Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770－778

[10] Homura. Homura package[EB/OL]. (2019－08－04)[2020－07－23]. https：//github. com/moskomule/homura

[11] LEE K，CHOI K，NAM J. Revisiting singing voice detection：a quantitative review and the future outlook[J]. 2018，arXiv preprint：1806. 01180

(责任编辑：湛　江)