北京林业大学
Beijing Forestry University

知山知水
树木树人

学校代码：10022

北京林业大学

Beijing Forestry University

本科毕业论文(设计)

音乐流派分类研究

Research on Music Genre Classification

胡汉林

学　　院　信息学院

专　　业　计算机科学与技术（创新实验班）

指导教师　徐艳艳　副教授

2020 年 6 月 12 日

音乐流派分类研究

计算机科学与技术（创新实验班）16　胡汉林
指导教师　徐艳艳

摘要

从早期的分类存储到现在的基于流派的听歌偏好推荐，音乐流派分类的应用价值正在不断提高，应用领域正在不断拓宽。本文提出了一种使用音乐梅尔谱图作为输入、多种类多尺度的重复卷积神经网络作为特征提取、全连接神经网络作为分类器得到每种流派分类概率的音乐流派分类方法。其中特征提取所使用的网络结构由重复卷积神经网络和残差学习改进得到，采用了两种不同种类和两种不同尺度的池化层。通过在 GTZAN 数据集上做实验对比了多种已有网络结构，在 1000 首 10 种音乐流派分类任务中得到了 86%的准确率，远高于人工分类的 70%准确率，同时也高于仅适用卷积神经网络或卷积循环网络的识别准确率。所提出模型对于音乐类别上相似的流派如 classical 和 jazz 也能做出很好的区分，达到了 100%无误分的识别率。除此之外，本文还对不同音乐流派的梅尔谱图特征进行了分析和对比，讨论了各种网络模型在不同音乐流派识别中的优劣。相比其他所对比的方法，本文提出的基于多种类多尺度重复卷积神经网络的音乐流派分类方法拥有更高的实际应用价值。

关键词：音乐信息检索，音乐流派分类，卷积神经网络，残差学习

Research on Music Genre Classification

Computer Science and Technology(Innovation Class)16 Hu Han-lin
Supervisor Xu Yan-yan

Abstract

From the early use of storage to the current genre-based listening preference recommendation, the application of music genre classification is constantly improving, and the application field is constantly expanding. In this study, a music genre classification method is proposed, which uses music Mel spectrogram as input, multiple types and multi-scale duplicated convolutional neural networks as feature extraction, and fully connected neural networks as classifiers to predict the classification probability of each genre. The network structure used for feature extraction is improved by duplicated convolutional neural networks and residual learning. Two different types and scales of pooling layers are used. Through experiments on the GTZAN data set, the existing network structures in various fields are compared with the CNN and CRNN model proposed in this paper. The accuracy of 86% on 1000 songs of 10 music genre classification tasks is reached, which is much higher than the 70% accuracy of manual classification, and higher than other network model structures. The proposed model can also make a good distinction between similar genres in music categories such as classical and jazz, reaching 100% accuracy, without the occurrence of misclassification. Moreover, the Mel spectrogram characteristics of different music genres are analyzed and compared, and the advantages and disadvantages of various network models for the recognition of different music genres are discussed. Compared with other traditional methods, the music genre classification method based on multiple types of multi-scale repetitive convolutional neural networks proposed in this paper has a higher practical application value.

Key words：music information retrieval, music genre classification, convolutional neural network, residual learning

目录

1 绪论

1.1 研究背景和目的

音乐信息检索(Music information retrieval, MIR)是一个多领域的研究方向，它结合了信号处理，机器学习和音乐理论等学科。音乐流派识别(Music Genre Recognition, MGR)是MIR 中的一个重要领域。近些年来，多媒体和音乐的广泛传播让合理地对音乐进行标注和推荐成为了一项至关重要的任务。但传统的依赖人工进行分类标注和整理的任务的低效和不准确性使得它变得不切实际，研究表明大学生只能在 10 种音乐流派的分类任务中达到70%的准确率 [1]。通过流派对音乐进行分类和推荐是一种简单有效的方法。

1.2 国内外研究现状

最早的音乐流派分类相关研究工作[1]是 2002 年由 Tzanetakis 和 Cook 共同发表的，使用了音质、节奏和音高内容作为特征，借助高斯混合模型(Gaussian Mixture Model, GMM)和 K-近邻算法进行分类。借助手工设计特征进行分类的传统音乐流派分类还使用了其他机器学习方法，其中包括在语音识别领域被广泛利用的隐马尔可夫模型[2] (Hidden Markov Models, HMMs)，以及不同距离矩阵的支持向量机[3] (Support Vector Machines, SVMs)对于流派分类的影响。在 Lily 和 Rauber 于 2005 年发表的工作[4]中，还研究了心理声学特征(Psycho-acoustic features)对音乐流派分类的贡献，以及短时傅里叶变换(Short-time Fourier transform, STFT)在巴克尺度中的重要性。

随着最近深度神经网络(Deep Neural Networks, DNN)的成功，很多相关技术被应用在语音识别和其他声音处理的领域上[5]。单一从时间域上对音频信息进行处理并输入到神经网络中的效果并不好，一个常见的替代方案是使用频谱图，因为它同时包含了时域和频域上的信息。卷积神经网络(Convolutional Neural Networks, CNN)在语音识别和图像识别上取得了成功，CNN 假定特征处于不同的层次结构中，并且可以通过卷积核提取。在有监督训练中，通过学习分层的特征可以完成不同的任务。通过深度学习的方法可以有效的避开使用人工设计的特征，CNN 在图像分类的任务[6]上被证明可以取得良好的效果，向 CNN 中输入 3 通道(RGB)矩阵表示的图像，从而完成了图像分类的任务。类似的，在 2016 年音乐流派分类的研究中[7]，频谱图在训练时可以被作为图像输入到卷积神经网络中，CNN 通过输入训练完成对于流派分类的预测。

使用传统人工设计特征如 MFCC、节奏特征等进行 10 类别音乐流派分类时只能达到61%的准确率[1]，使用主成分分析(Principal Component Analysis, PCA)方法对 MFCC 特征进行白化后的频谱图作为输入，卷积深度置信网络在 5 分类任务中达到了 70%的准确率，但这样的结果仍然低于人工分类的准确率，这充分说明了使用传统人工设计特征的局限性。

单一使用 CNN 的音乐流派分类并没有很好的考虑到音乐信号作为一个序列模型的上下文关联性，而循环神经网络(Recurrent Neural Networks, RNNs)的出现恰好解决了这个问题，RNN 通过上一时刻的信息和当前时刻的输入决定输出的信息，这使得它可以捕捉到序列在时间上的上下文关联性。将 CNN 与 RNN 结合形成的卷积循环网络(Convolutional recurrent neural network, CRNN)可以被认为是将最后一层卷积替换为 RNN 的 CNN 模型。在 CRNN 中，CNN 和 RNN 分别承担了特征提取和时间关联的任务，使用 RNN 来汇集这些特征可以考虑到模型的全局性，同时保留的部分 CNN 使得局部特征的提取得到保证。这样的 CRNN 模型最早在文档分类的工作中被提出，随后被使用在图像分类[8]和音乐自动转录[9]上。在 2017 年 Choi 的研究[10]中表明 CRNN 非常适合音乐流派分类的工作，因为RNN 在汇总局部特征上比 CNN 静态的使用平均权重和降采样更加灵活，这样的灵活性在

对音乐的情感分类时可以有很大的帮助，与此同时在进行乐器分类中 CNN 善于提取局部特征的特点又可以起到很大的帮助。在 Million Song Dataset 上的实验[10]表明，CRNN 在音乐流派分类上的准确率相比使用 CNN 模型的准确率有所提升。但由于 RNN 时间序列的特点，当前时间步的计算需要等待上一时间步的计算完成后才可以开始，无法像 CNN 一样进行并行计算，从而导致计算的时间和计算量都更大，模型的训练时间变长。以及 RNN 需要输入较长的音乐片段才能体现出全局特征的优点，实验中使用长达 29 秒的片段在实际应用中也是不现实的。

后续的研究针对基于 CNN 的音乐流派分类模型采取了其他的改进方案，Zhang 的工作[11]采用了基于 k-最大池化的 CNN 模型来完成音乐的语义建模，这样的方法可以通过添加更多的神经网络层数来产生鲁棒性更强的音乐表征。在 Dieleman 的研究[12]中，使用端到端学习的方式直接训练 CNN 学习音乐的原始信号，并将其与梅尔谱图作为特征的方法进行对比，结果表明虽然直接学习音乐原始信号的效果不如梅尔谱图，但证明了这种方法的可行性。此方法可以在最大程度上减少音乐流派分类过程中对先验知识的依赖，为后续研究提供了一个可行的方向。

在机器学习领域，迁移学习(Transfer Learning)的定义是在训练目标任务时，重新利用在源任务上已经训练好的参数，进而达到类似领域之间知识的迁移。这样做的动机通常是因为目标任务的训练数据量不足，通过迁移学习，模型在目标任务中需要学习的参数量大大减少，这样就可以完成在少量数据上更加高效的学习。一个很好的例子是在计算机视觉领域上，网络模型从大量的图片中学习到物体或模式检测的模板，用在少量数据的图像识别任务上[13]。迁移学习在 MIR 上也被广泛使用，Oord 团队提出了 k-均值(k-means)和多层感知机(Multi-layer Perceptron, MLP)或线性回归结合的迁移学习方法[14]，在原始数据集上训练各类标签分类任务，并将网络参数迁移到音乐流派分类的小数据集上使用。

在 2017 年的文献[15]中作者提出首先使用五层的卷积网络 vggnet 对 Million Song Dataset (MSD) [16]进行训练，MSD 中共有 244224 条音乐片段，分别带有流派、年代、乐器和情感等标签。在绝大多数的迁移学习训练中，会使用源任务最后一层卷积层的输出作为目标任务的特征提取参数，在本研究中，作者还尝试使用了多个中间层相组合的方式，试图找出适合不同目标任务的最佳特征。使用中间层作为特征的问题是，中间层特征的维度较大，会增加分类时的计算量。为此，作者对一到四层的卷积层输出使用了平均池化来概括特征，在减小维度的同时保证全局的数据特点。在后续的分类中，作者使用 SVM 来专注于比较特征组合的优劣，同时根据其他研究中的对比，在训练集较小的情况下，SVM 的分类效果会好于 K-近邻方法。在 GTZAN 上测试的结果表明，结合使用全部五层特征组合的效果最好。在最新的研究中[17]，作者使用了 CNN、LSTM 和 MLP 组合的神经网络结构，以及迁移学习特征在 GTZAN 上对音乐进行分类。

首先，在 MSD 数据集上训练 CNN-LSTM 网络进行特征提取。其次，使用 MLP 对迁移学习特征进行分类。最后，针对每一首音乐的所有片段根据投票机制进行判别，具体的做法是统计所有片段的流派分类频率，选取最大频率作为预测结果，如果有相同频率分类则选择 softmax 预测概率最大分类，这样的做法使得分类的准确率提高了近 10%，达到了 94.5%，是当前的最佳模型。

国内研究方面在音乐流派分类领域的研究开展时间较短，研究的分类效果一般，其中很多使用了基于传统的分类方法。天津大学学者徐星将音质特征作为特征，并在 2012 年提出了基于最小一范数的分类方法[18]，与传统的 svm 方式相比，准确率有所提高。喻晓雯等人[19]选择了音高和相对节奏来表示音乐信号，也得到了相对好的分类结果。

在涉及深度学习的研究中，2017 年南京大学的学者[20]将梅尔多频系数提出作为特征，并且提出了基于深度置信网络的识别方法。也有采用了三种不同频谱作为特征，分别是傅

利叶转换、常数 Q 变换和梅尔变换，通过改进的 CNN 模型分类[21]，以及优化的 CRNN 完成分类[22]，两种方法分类准确率为国内研究中最高。

1.3 研究内容

本课题的主要研究目标是使用深度学习方法实现音乐流派分类，使用设计的深度神经网络模型对给定音乐进行流派识别。

需要重点解决的问题分别是特征选择、神经网络结构设计和在公开数据集上的实验。在特征选择上，将尝试对于音乐的梅尔频谱图特征进行分析作为特征进行分析。在神经网络的结构设计上，会完成一个基于重复卷积神经网络模型的改进方案，并对模型训练和测试。

1.4 本文结构

论文文章结构如下，在第 2 章中将介绍涉及的深度学习和语音处理相关背景知识，列出实验中对音乐提取的频谱图特征；第 3 章中详细介绍多个用于音乐流派识别的神经网络结构，其中包括本文提出的改进的重复卷积模型。第 4 章主要说明实验所用的音乐数据集，实验参数设置以及与其他音乐流派识别模型的效果对比。最后对全文进行总结并讨论未来可能的改进。

2 背景知识介绍/基础知识

2.1 过拟合解决方案

2.1.1 正则化

在计算损失函数时，需要额外加入一项，如式(2.1)所示：

$$\frac{1}{2}\lambda\sum_i w_i^2 \tag{2.1}$$

其中w指神经网络权值，这种方法的作用是惩罚过高的权值，我们希望权重分布在所有模型参数中，而不仅仅是少数几个参数中。同样，直观上，较小的权重将对应于较不复杂的模型，从而避免过度拟合。本研究中系数λ被设置为 0.01。

2.1.1 神经元丢弃 dropout

这是另一种通过在训练中随机关闭一些神经元（将权值暂时设置为 0）的一种简单有效防止过拟合的方案[23]。在每次迭代中使用不同的神经元组合作为分类器对样本预测，这可以让模型具有更强的泛化能力，可以摆脱对某个特定集合内神经元的依赖性。

2.2 残差网络结构和重复卷积模型

在使用深度学习方法的音乐流派分类中，神经网络的结构是影响分类准确率的最大因素，Yang 的研究[24]对原有的结合残差学习[25] (Residual Learning)的 CNN 模型[26]结构进行了改进。残差学习的核心思想是，假设神经网络需要学习的目标是$H(x)$，那么除了将拟合$H(x)$作为目标之外，残差学习还将拟合一个残差目标，如式(2.1)所示：

$$F(x) = H(x) - x \tag{2.2}$$

原始的函数将变为式(2.3)：

$$H(x) = F(x) + x \tag{2.3}$$

因此他们会将网络中间卷积层的输出与网络最后一层卷积层的输出相加获得最终的输出。在他们的研究中证明，残差学习更适于优化深层网络，并且可以在图像识别任务上通过增加网络层数来提高准确率。但单纯的相加会导致网络损失一部分学习到的信息，所以在 Yang 的研究中，采用了将卷积层输出加入池化层(Pooling Layers)并将池化层输出拼接。作者还采用了不同的池化层来试图保存不同类型的特征，从而提高分类的准确率。这种重复卷积层(Duplicated Convolutional Layers)和不同池化层结合的结构能在参数量不大幅增加的条件下提高分类效果。

2.3 梅尔谱图

作为音乐流派分类中的一个重要环节，选用合适的特征是分类是否成功和提高分类准确率的关键。一种常见的传统方法是从原始歌曲中提取出手工设计的特征，然而这种过程依赖于特定领域和工程上的专业知识。Sarkar 等人通过经验模式分解(Empirical mode decomposition, EMD)的方法来捕获不同流派音乐之间的局部特征差别，然后从分解后的音乐中的计算音高特征(如 Mel-frequency cepstral coefficients, MFCC 等)来提高分类的表现。这些人工设计的特征有很多缺点：首先，它很难为一个特定任务设计特征；其次，这种方法缺乏泛化能力，不同任务的需要分别计算不同的特征；最后，该模型缺乏可扩展性，因为系统的性能改进不依赖于统一的框架，例如，通常需要使用不同的特征或分类器才能在不同的任务上实现更好的分类精度。

生理学研究表明人类的听觉系统以分层方式工作，首先人耳将连续的声音波形分解成不同的频率，从而在低频上拥有更好的精度，从低到高的听觉结构的神经元逐渐提取具有更复杂的光谱时域接受场（Spectro-temporal receptive filed, STRF）的更复杂的光谱时域特征[27]。通过将频谱图作为输入，并将相应的流派作为标签，CNN 将学习在频域和时域提

取特征的滤波器。 如果这些学习的滤波器模仿了人类听觉系统中的 STRF，则它们可以提取出有用的特征，用于音乐流派分类。由于音乐信号通常在时域中是高维的，因此无法使用适合音乐信号完整频谱图的 CNN。为了解决这个问题，研究人员提出了一种分治的策略[9]：将音乐的频谱信号特征分割为连续的 3 秒片段，并分别对每个片段进行预测，最终结果将结合所有音乐片段预测结果的结合来判断。这样做的原因是人类的预测的准确率在 3 秒以上长度时没有明显提高，同时在卷积深度置信网络对音乐流派的分类中[28]，3 秒时长的分割也取得了最佳的效果。

为了进一步的减少输入频谱的维度，使用了梅尔倒谱图作为 CNN 的输入，梅尔倒谱图(Mel-frequency Cepstrum, MFC)是在自动语音识别中广泛使用的一种特征[29]，它模仿人类的听觉系统在时频域上对声音表征，可以看作是频谱图在频域上进行平滑的结果，它在低频上拥有较高的准确度，在高频上的准确度较低。通常来说可以通过将音乐信号的幅度谱在频域上进行映射到梅尔尺度上来获得。

3 研究方法

3.1 音乐流派识别的过程

音乐的流派与音乐的情感相同是一种高层级的标签。作为一个分类问题，典型的自动音乐流派分类主要包含三个步骤：

（1）信号转换

从原音乐信号中提取出例如节奏、音高和数据特征；

（2）特征提取

使用一些技术来从中选取有用的特征或是整合特征；

（3）分类器

训练一个基于挑选出特征的机器学习分类器来实现对于输入音乐流派的自动分类。

图 3.1 以本文中模型为例介绍这一过程，图中三个虚线框分别代表了主要的三个步骤，其中的模型结构不限于图内所示。

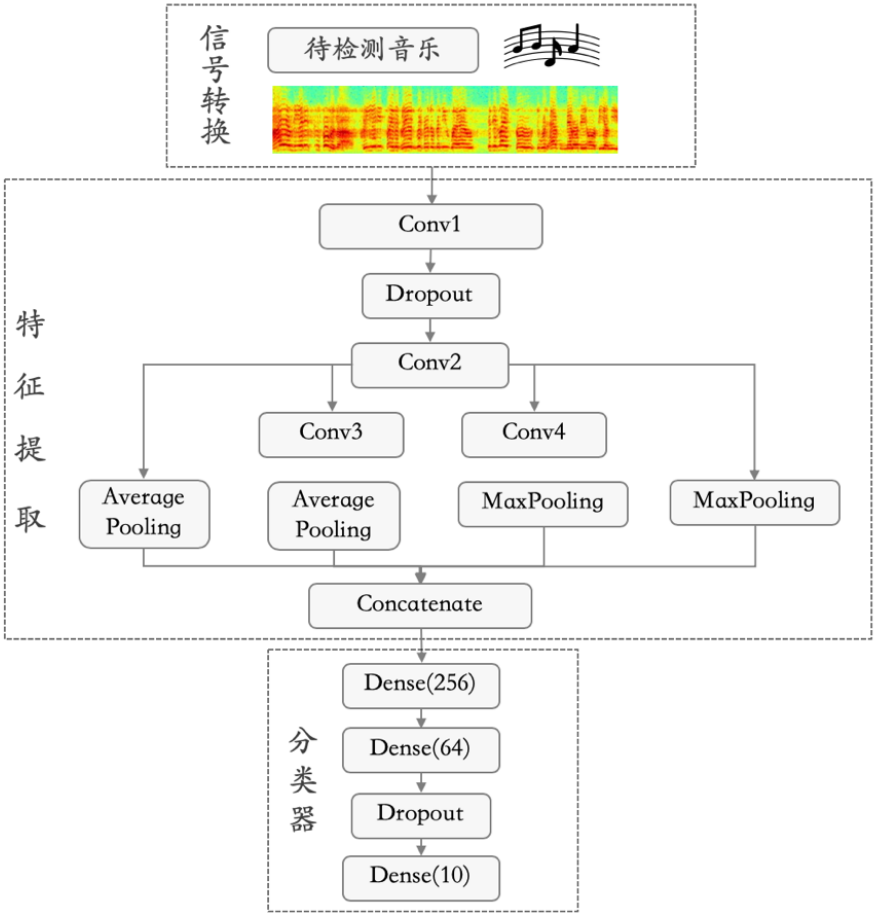


图 3.1　识别过程示例

Fig.3.1　Illustration of genre classification

3.2 数据预处理

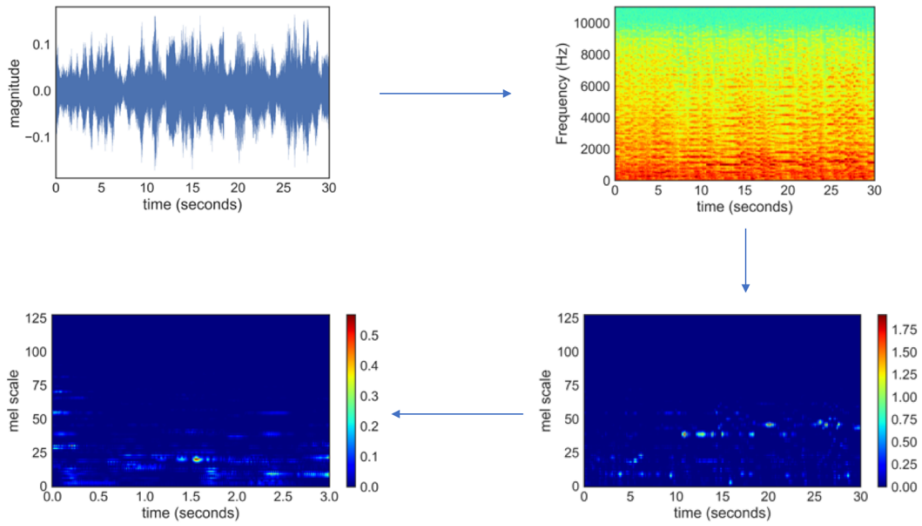


图 3.2　音乐数据预处理

Fig.3.2　Preprocess of music data

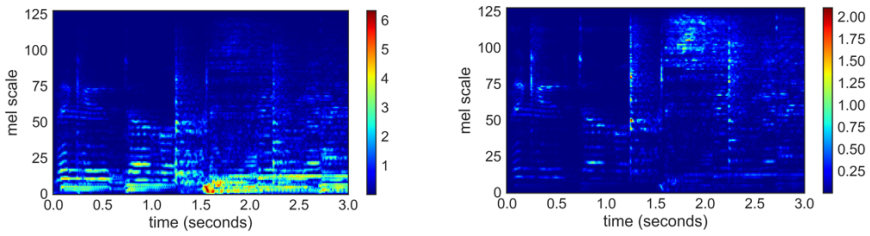
图 3.2 展示了一首 30s 长度的音乐样本从原始音频经过预处理到可作为神经网络模型输入的梅尔谱图的全过程。首先将以 wav 形式存储的音乐片段使用 librosa 读入为幅度谱，为保证训练数据的一致性，对于超过 30 秒的音乐部分忽略不计，根据统计，所有训练数据中最长片段有 661794 帧，根据 22050 赫兹的采样频率计算，忽略部分最多只有 294 帧，对于模型的训练影响几乎可以忽略不计。类似的，对于不足 30 秒的音频在结尾进行补 0 操作，补齐至 66150 帧。

第二步将幅度谱转换为频率谱，考虑到当下大部分音乐由人声组成，而人声语音的能量又主要集中在低频部分，这就导致了频谱图整体向 y 轴负半轴方向倾斜。为了解决这个问题，使用一阶预加重对频率谱进行变换，有助于提高语音信号相较于低频分量的高频分量的幅度。

假设当前时刻为t_d，当前帧的语音信号为$X[t_d]$，变换后的语音信号$X'[t_d]$计算方法见式 (3.1)

$$X'[t_d] = X[t_d] - \alpha X[t_d - 1] \tag{3.1}$$

其中α为预加重系数，一般取值区间为 $[0.95, 0.99]$，预加重系数越大低频的削弱和高频的增强就越大。这种方法可以去除声门激励和人说话时口鼻辐射的影响，进而提高信噪比。经过预加重处理的频谱图对比如图 3.3 所示，图(a)是转换前，图(b)是转换后



(a)　　**(b)**

图 3.3　预加重

Fig.3.3　Pre-emphasis

频谱图的纵坐标是频率，在梅尔轴上，听觉是等距离的、等差的，但频率轴上不是，因此要对频率再进行一步变换为梅尔尺度。频率和梅尔尺度的转换方法见式 (3.2)

$$mel(f) = 2595 * \log_{10}(1 + f/700) \tag{3.2}$$

最后，为保证输入维度在训练时可以接受，也就是可以在保证显存足够情况下进行 gpu 加速的批训练，将 30 秒的音乐截取为 10 个 3 秒长的片段作为训练样本。

数据预处理及特征转换中所涉及的参数在表 3.1 中列出，

表 3.1　参数设置
Table 3.1　Hyper parameters

音乐时长	30 s
切分片段时长	3 s
采样频率	22050 Hz
短时傅里叶变换帧长	512
帧移	256（50%重叠）
梅尔频度	128
重预加重系数	0.97

3.3 基线模型

3.3.1 卷积神经网络模型

基线卷积神经网络模型由两个二维卷积和一个全连接神经网络分类器组成，网络模型结构图及每层网络输入输出大小如图 3.4 所示。网络的输入是 128 x 128 大小的梅尔谱图，分别代表了 128 帧和 128 梅尔频度，这样可以在保证计算量较小的情况下尽可能增加网络深度。第一层卷积层拥有 64 个卷积核，大小为 (3, 3)，第二层卷积层拥有相同数量卷积核，大小变为 (3,5)，激活函数选用线性整流函数[30] (Rectified Linear Unit, ReLU)，其计算方法见式 (3.3)

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{3.3}$$

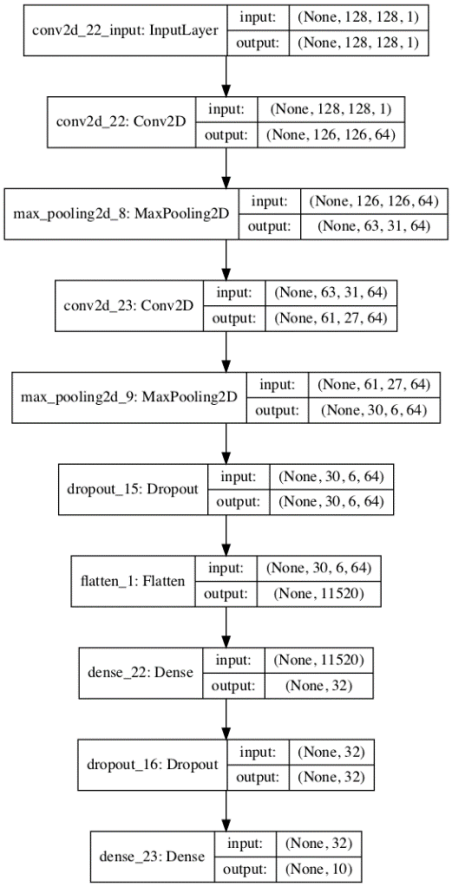


图 3.4　基线卷积神经网络模型

Fig.3.4　A CNN model baseline

两层卷积层后均接有大小为(2,4)的最大值池化层，来进一步缩减特征大小。池化是一种最普遍的对卷积层提取出特征降采样的方法，如本文中使用的为 2 x 4 大小的池化窗口，即只保留窗口内的最大值，并按照默认 1 的步长在整个图上移动。经过卷积层和最大值池化层后提取出的频谱图特征经过扁平层 (flatten)变换为一维，作为输入送入后续的全连接神经网络分类器进行分类。两层的全连接分类器分别有 32 和 10 个隐藏层节点，其中第二层的 10 个节点分别对应了 10 种音乐流派，经过 softmax 激活函数后得到每种类别标签的概率作为结果的判定依据。

接下来通过交叉熵计算损失函数

$$L = -\sum_{C=1}^{M} y_{o,c} * \log p_{o,c} \tag{3.4}$$

其中M是分类数；$y_{o,c}$是二进制标签，当样本o的正确分类为c时为 1，其他情况下为 0；$p_{o,c}$是模型对样本o的预测概率。计算得出的误差用于反向传播计算梯度，进而更新网络参数。网络持续迭代到损失收敛到最小值。

3.3.2 长短时记忆神经网络模型

长短时记忆神经网络模型如图 3.5 所示，在基线卷积神经网络模型的基础上，使用 LSTM 代替了卷积层，尝试通过 RNN 的特性建立音乐的前后句关联性。使用 LSTM 的神经元个数与输入大小相同为 128 个，只输出最后一个时间的输出，所以经过 LSTM 后网络的输入输出大小相同。后续的网络结构与基线卷积神经网络模型相同，使用全连接神经网络作为分类器。

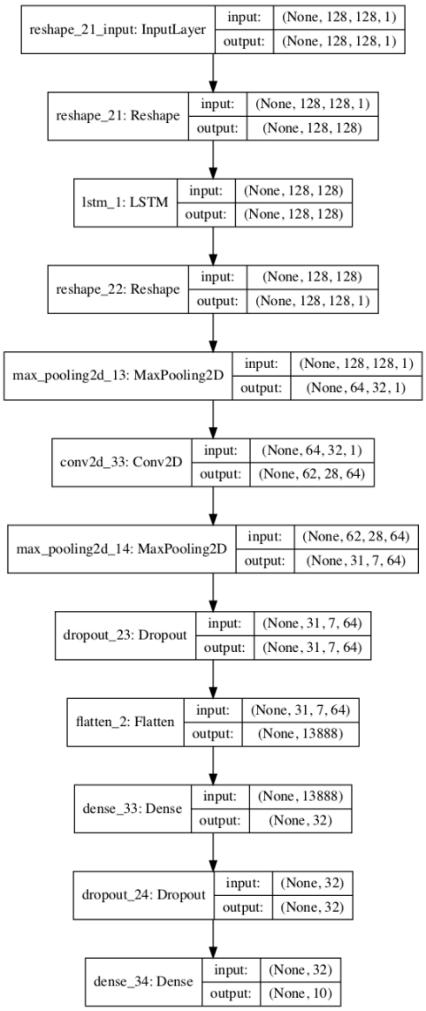


图 3.5　长短时记忆神经网络模型
Fig.3.5　An LSTM model

3.4 改进的重复卷积模型

改进的重复卷积神经网络模型如图 3.6 所示。这种模型结构相较基线模型最大的改进在于增加了一个由多层卷积组成的重复卷积网络模块，并代替了残差网络模块。同时使用了多尺度对称结构的池化层对特征降采样。

传统的残差网络模块仅采用了卷积层后特征以相加方式组合，从而增加网络深度的同时，能够避免训练时出现梯度消失和梯度爆炸的问题，比没有残差相加的原始网络更容易优化。然而它在训练速度上还有一定的缺陷，并且相加操作可能会带来一定程度上的信息损失，本结构将通过拼接池化层输出的方式，将残差网络模块替换为重复卷积模块。

在过去的研究中多使用单一一种的池化层或是依次使用多种池化层，在改进的重复卷积网络结构中同时采用了最大值池化(MaxPooling)和平均值池化(AveragePooling)，两种池化用局部最大值和局部平均值取代了多个特征点，所以依次可以完成对于卷积层输出的特征概括和总体概括。同时随着卷积层数的增加，相同大小的特征区域概括能力增强，在经过池化层后得到的特征也就拥有了不同层级的意义。不同流派音乐需要在不同的层级上进行区分，如 classical 音乐和 hip hop 音乐可以从高层的节奏特征来区分，但 rock 音乐和 pop 音乐可能就需要从低层的曲调或音色来判断。所以本结构在第一层卷积后除了顺序的卷积层堆叠外还模仿残差网络直接将输出送入最后一层池化层与第三层卷积池化结果进行拼接。

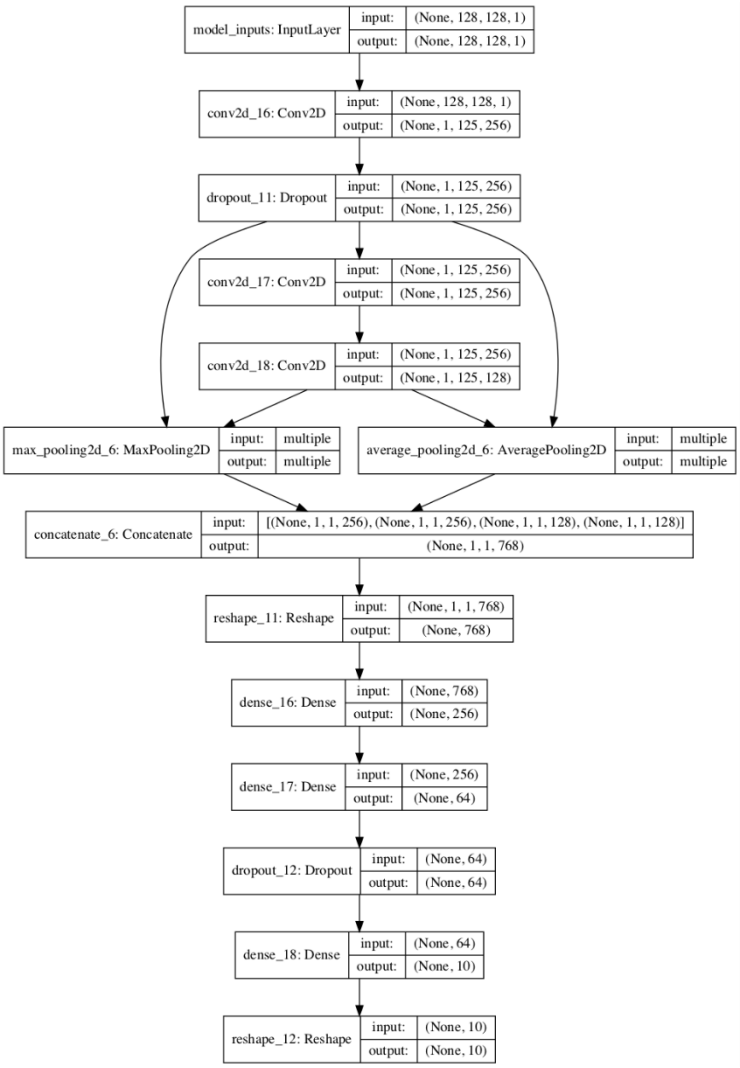


图 3.6　改进的重复卷积网络模型

Fig.3.6　The refined duplicated convolutional model

改进的重复卷积模型在输入上与基线模型相同，在卷积层部分更换为三层不同尺寸大小的卷积，在分类器上使用了隐藏层神经元数量更多，层数更深的全连接网络，输出的激活函数与基线卷积神经网络相同。

网络各部分参数在表 3.2 中列出

表 3.2　网络参数

Table 3.2　model parameters

第 1 层卷积	(256, 128, 4)
最大值池化层 1	(1, 125)
dropout1	0.2
第 2 层卷积	(256, 1, 4)
第 3 层卷积	(128, 1, 4)
平均值池化	(1, 125)
dropout2	0.1
全连接层	768, 256, 64, 10

注：卷积层参数分别代表卷积核个数，卷积核大小，全连接层分别为四层的隐藏层节点个数

所有卷积层均使用了 ReLU 作为激活函数，同时为了防止模型过拟合，加入了 dropout 层，对第 1 层卷积加入了 l2 正则化的限制。

4 实验及结果分析

本章将对实验使用的数据集、参数和实验结果进行描述。

4.1 数据集

4.1.1 数据集 GTZAN

GTZAN 是由 Tzanetakis 和 Cook 共同收集制作的音乐数据集[1]，这个数据集被广泛使用音乐流派分类上作为基准与其他方法进行对比。它包含了 1000 首的音乐片段，每段时长 30 秒。分为 Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae 和 Rock 十类，基本均匀分布，所有录音均以 22050Hz 的单声道 wav 文件分类存储。

在实验中，数据的训练集、验证集、测试集被分割为 8 : 1 : 1 的比例。在实验中，每种流派的 100 首音乐各自被随机分为 80 首训练集、10 首验证集和 10 首测试集，以防止不均衡的数据划分出现。同时，为了消除数据集划分对模型准确率的影响，在对数据提取频谱特征前，随机打乱了数据集内每类音乐的顺序，由此得到了两种数据集划分，分别命名为划分 1 和划分 2，对于每种网络模型结构，分别在划分 1 和划分 2 上做训练和验证并得到结果。数据的随机打乱通过 numpy 对每类数据进行随机排序实现，numpy 是一个 python 上用于科学计算的基础程序包。同时通过设置不同的随机数种子，使程序同时保证了随机性和可重复性。

4.1.2 不同音乐流派频谱图

本文介绍的音乐流派识别方法均基于将音乐的梅尔谱图视为图片，进而采用类似图像识别的方法来对音乐分类。然而图像识别中的样本不同于横纵轴为时频域信息的梅尔谱图，它很重要的一部分是空间、位置信息。卷积神经网络使用每个小的滤波器来识别或构建一个小特征，再通过多个滤波器的堆叠来完整识别或构建整张图片，其中非常流行且典型的 VGG[31]网络就是如此。本条中将对使用该类方法对音乐进行分类的可行性进行分析。

在图 4.1 中展示了四张来自两种不同音乐流派的音乐的梅尔谱图，第一行中的两首为 classical 流派，第二行为两首 pop 流派的音乐。不难看出两种音乐流派的差别很大，同时类内又有一定的相似性。根据第二章中的介绍可以知道，梅尔谱图是原始音乐频谱图变换后的结果，其纵轴的梅尔尺度可以线性反应听觉对于频率变化的感知，也就是说，梅尔谱图的分布特征区别可以间接反映出不同流派音乐的风格差别。从结构上来看，classical 音乐的分布较为均匀，而对比来说 pop 音乐的分布更接近山峰状，说明节奏和音高的起伏比 classical 音乐大。从纵轴梅尔尺度的值域来看，classical 音乐从低频到高频的分布较为均匀，pop 音乐主要集中在低频段，由此我们可以结合实际情况推断出，pop 音乐中存在着比 classical 音乐中更多的人声片段，因为人声主要集中在低频段，所以得到的梅尔谱图也在低频段较为密集。

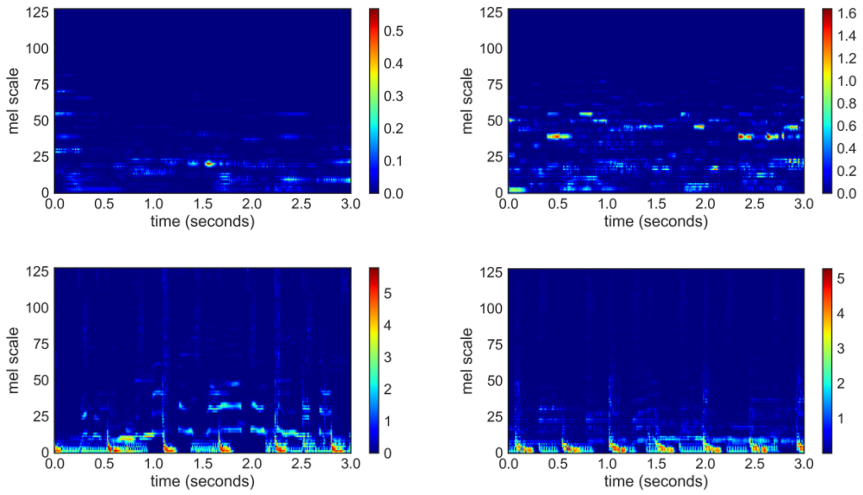


图 4.1　两种音乐流派的梅尔谱图样例

Fig.4.1　Sample spectrograms of two different genres

综合以上几点，虽然使用图像识别方法在音频信号的处理上结合了较少领域内知识，但优点在于它可以通过少量的需要领域内知识的人为处理（使用梅尔谱图特征），结合深度神经网络的优势，让算法代替人类对语音特征进行挖掘和学习，这样的便利性是其他方法无法代替的。

4.2 实验设置

所使用的远程服务器具备 62G CPU 内存，配备有 GTX 1080TI GPU，用于加速深度学习的训练，同时还有至少 3T 以上的存储空间用于备份和存储数据。

所有代码在 Ubuntu 16.04 LTS, python 3.6.9，tensorflow-gpu 1.14.0，keras-gpu 2.2.4 环境下完成，可正常运行，同时使用了在语音领域上被广泛使用的 python 包 librosa[32] 0.7.1 对语音原始信号进行处理。

在实验参数设置上，学习率使用 Adam 优化器更新[33]，每个模型训练 10 个 epoch，批训练的 batch size 选用 200。模型最后一层输出是每个截取片段经过 softmax 激活函数的 10 分类概率，最终结果使用所有片段概率求和的方式判定，分类概率和最大的被选为预测结果。

4.3 结果分析

基线模型和改进的重复卷积模型混淆矩阵分别见图 4.2，图 4.3，其 x 轴为预测标签，y 轴为真实标签。从混淆矩阵中可以直观的分析出模型在不同音乐流派中的分类效果区别，这样的差异可能由模型特性和不同流派音乐的特性共同导致。比如在基线模型和改进的重复卷积模型的混淆矩阵中，country 和 rock 音乐的识别率都不高，平均分别为 70%和 50%，远低于模型的平均准确率。

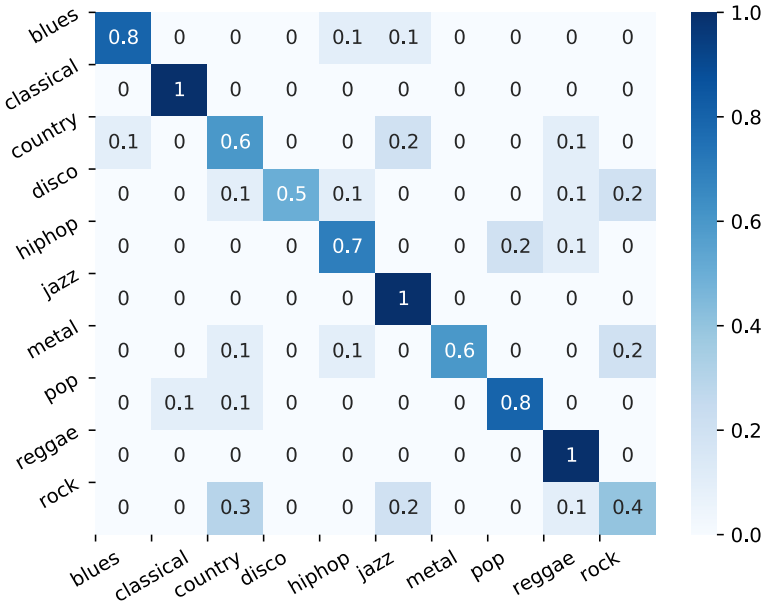


图 4.2　基线模型混淆矩阵

Fig.4.2　The confusion matrix of the baseline model

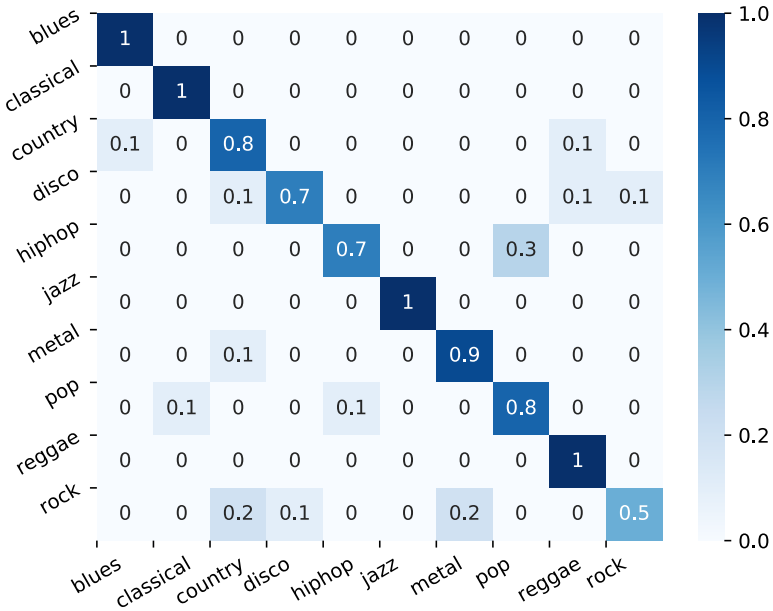


图 4.3　改进的重复卷积模型混淆矩阵

Fig.4.3　The confusion matrix of the refined duplicated convolutional network

有研究表明[34] country 音乐中的一些典型特征（如节拍，韵律等）需要较长时间（大于 3 秒）来判别。然而本文中的检测方式，为保证输入梅尔谱图特征维度可以接受，网络模型的输入均采用了 3 秒长度，也就导致了没有足够长的分割片段来识别出类似的特征。这个缺陷未来可以尝试通过在梅尔频度上减少对高频段的采样或在时间域上进行降采样来解决。另外一种可能是 country 音乐拥有与其他流派音乐较多相似的特征，从而导致了神经网络较难将他们与其他流派音乐完全区别开。

在杜威等人的研究[35]中使用支持向量机对包含 MFCC 和其他特征在内的 180 维特征对 GTZAN 分类，其中 country，disco 以及 rock 被归为同一类，拥有类似的特征。这也可以从多个流派音乐多被误识为 country 类中看出，其中 rock 被识别为 country 的概率在两种模型中分别为 20%和 30%。在甄超等人的多模态音乐流派分类研究中[36]也有类似的结果，摇滚音乐在基于声学特征和基于音乐标签的分类中均是分类准确率最低的音乐流派，其中基于声学特征的分类方法中，摇滚乐比预测准确率第二低的爵士乐还要低 10%以上。这些研究都证实了 rock 音乐在音乐流派分类任务中的困难程度。

从两个混淆矩阵的对比中可以看出，尽管两种模型的准确率相差 12%，但仍在 classical，jazz 和 reggae 三类音乐中达到了 100%的识别率。有研究[37]指出 rock 和 blues 两种流派的音乐存在音乐类别之间的相似性，他们共同被誉为“黑人的流行音乐”。类似的联系也存在于 classical 和 jazz 音乐中，然而在本文提出的方法中，classical 和 jazz 音乐被很好的区分开，这体现出了模型在挑选局部关键特征的优越性。

表 4.1　实验结果

Table.4.1　Results of different networks on GTZAN

网络结构	参数量	数据集划分	准确率
基线模型	493, 226	划分 1	74%
重复卷积模型 1	738, 890	划分 1	80%
LSTM 模型	708, 970	划分 1	77%
改进的重复卷积模型	730, 698	划分 1	84%
基线模型	493, 226	划分 2	80%
重复卷积模型 1	738, 890	划分 2	82%
LSTM 模型	708, 970	划分 2	81%
改进的重复卷积模型	730, 698	划分 2	**86%**

注：参数量均指可训练参数，数据集划分 1 和划分 2 对应实验设置中的两种不同划分

实验证明不同的数据集划分也会对模型准确率产生影响。如表 4.1 所示，在不同的两种数据集划分方案中分别进行了四个模型的训练，可以看出从划分 1 到划分 2 四种模型均有 2%至 6%的准确率提升，这证实了不同数据划分是会对模型的训练效果产生显著影响的。猜测是由于数据集中存在这某些具有极强代表性的样本，通过对该样本的学习，模型的泛化能力可以得到显著的提升。

与此同时，结构最简单的基线模型提高为 6%，相反，最复杂的改进的重复卷积模型准确率提升仅为 2%，是四种模型中最小的。这一方面是因为本研究中提出的改进的重复卷积模型原有准确率较高，训练数据的改变很难再带来提升。另一方面也可以理解为是这种多尺度更加复杂的重复卷积模型对于频谱图特征提取的能力更强，对数据的要求相对较低，有更好的鲁棒性。

基于 LSTM 的音乐流派分类模型准确率不高，在两个数据集划分中仅为 77%和 81%，这是在预计之中的，因为 LSTM 作为 RNN 模型，其优势在于挖掘上下文关联性，而实验中所有模型的输入维度均相同并且较小，这样的特征输入使得 LSTM 无法发挥出它的优势，未来可以设计更大维度输入的实验来进一步验证其模型的学习能力。

最后从参数量上来分析，三种改进模型参数量都显著大于基线卷积模型，同时两种重复卷积模型的参数量又略大于 LSTM 模型。可见在目前阶段，模型最终的准确率会宿舍可训练参数量的上升而上升，而这样的参数量在训练过程中也是可以接受的，证明了模型还有进一步提升的空间。

5 结论

5.1 总结

随着大量音乐的出现，对音乐进行快速准确分类的需求正在逐步攀升，与此同时，音乐流派识别在音乐信息检索中的地位也在不断提升。大型国外音乐公司 Spotify 就拥有一个专门的团队来对总计 6000 万首歌曲进行分类，分类种数更是高达 1000 个子类型[38]。本文在原有研究的基础上提出了一种基于多种类多尺度的重复卷积网络模型对音乐进行风格流派识别，其结构上继承了卷积神经网络和残差学习的优点，并且可以更好的提取多尺度的特征用于后续的分类。

为验证所提出模型的有效性，在公开数据集 GTZAN 上与几种常见且效果较好的网络结构进行了测试和对比。实验过程中提取了音乐的梅尔谱图作为网络的输入，通过特征提取后均送入全连接神经网络进行分类，最终得到预测的音乐流派。实验结果表明，本文提出的模型结构在准确率上高于所复现的其他模型结构，同时受不同训练数据变化的影响较小。后续还针对不同流派的识别效果以及原因进行了分析，证明了模型在一些特定流派识别上的优势。尽管在参数量方面略高于其他模型，但综合考虑到流派识别的准确性和模型的泛化能力，所提出模型仍有很高的实际应用价值。训练出的音乐流派识别模型只需在数据的并行处理和标签分类结果的记录上稍加改动便可以投入实际的应用和生产中。

5.2 展望

考虑到模型的最终准确率和模型的结构等因素，本文提出的基于多种类多尺度的重复卷积网络模型仍有很多可以改动的空间，比如网络的输入现在仍采用了一定程度上人工处理的梅尔谱图，在设备计算能力不断上升的今天，可以尝试加入原始频谱进行端到端学习，从而减小信号转换过程中信息的丢失。在实际应用的方面，也需要在训练数据中加入更多的音乐风格流派标签来适应不断变换的曲库和当今社会的音乐环境。

致谢

大学四年的生活即将随着论文的结束落下帷幕，我希望能在这个特殊的时间点，借助如此宝贵的机会，向从入学直到完成论文以来，给予过我帮助和鼓励的老师和共同在北林学习的同学们表示感谢。

首先要感谢指导我完成这篇论文的导师徐艳艳副教授。徐老师在学业上悉心的教导和在学术上的认真钻研一直以来是我学习的动力和榜样，不但如此，徐老师对我们学生给予了朋友一般的关怀和鼓励。从大二带我进入北京林业大学人工智能实验室以来，徐老师一直在学业上不断督促着我的前行和成长，指导我完成了第一个主持的校级大创项目，带我走入了人工智能学习的大门。

同时还要特别感谢的就是作为我论文和在实验室的指导老师柯登峰博士，柯老师在研究中付出的努力和苦心钻研的态度是我对于一名学者定义的标杆。柯老师在指导我进行研究和实验时总是能不失耐心的指出问题并给我解惑，没有柯老师的帮助，我在深度学习的入门之路将会坎坷许多。

当然还要感谢的是在北京林业大学信息学院的每一名老师和同学，是老师讲授的一门门基础课加在一起给我打下来的未来研究坚实有力的基础，也是身边不断涌现出的优秀同学让我保持谦虚，保持高涨的学习动力。在老师之中特别要感谢的是班主任王春玲老师，作为计创班班主任，王老师会针对我们每个人的特长和不足给予帮助，将这个特殊的班集体凝聚在了一起，爆发出更强的力量。在同学之中我特别需要感谢丁可，刘天禹学长，二位学长耐心回答了我很多学业上的问题，为我未来的学习方向给予指导，同时二位优秀的学长也一直是我四年大学生活学习中的榜样。

最后还要感谢与我共同学习，清晨到图书馆晨读，熬夜备战考试的室友们，是你我并肩奋战的这股力量让我能在知识的海洋中远航。大学四年的生活短暂又令我不舍，这里塑造了我从学生到学者的第一步，我相信，走出北林校门后的我，能用自己的努力和成就给母校带来回报和骄傲。

参考文献

[1] Tzanetakis G, Cook P. Musical genre classification of audio signals[J]. IEEE Transactions on speech and audio processing, 2002, 10(5): 293-302.
[2] Scaringella N, Zoia G. On the Modeling of Time Information for Automatic Genre Recognition Systems in Audio Signals[C]//ISMIR. 2005: 666-671.
[3] Mandel M I, Ellis D P W. Song-level features and support vector machines for music classification[J]. 2005.
[4] Lidy T, Rauber A. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification[C]//ISMIR. 2005: 34-41.
[5] Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2014, 22(10): 1533-1545.
[6] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[7] Nanni L, Costa Y M G, Lumini A, et al. Combining visual and acoustic features for music genre classification[J]. Expert Systems with Applications, 2016, 45: 108-117.
[8] Zuo Z, Shuai B, Wang G, et al. Convolutional recurrent neural networks: Learning spatial dependencies for image representation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 18-26.
[9] Sigtia S, Benetos E, Dixon S. An end-to-end neural network for polyphonic piano music transcription[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(5): 927-939.
[10] Choi K, Fazekas G, Sandler M, et al. Convolutional recurrent neural networks for music classification[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2392-2396.
[11] Zhang P, Zheng X, Zhang W, et al. A deep neural network for modeling music[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015: 379-386.
[12] Dieleman S, Schrauwen B. End-to-end learning for music audio[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 6964-6968.
[13] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1717-1724.
[14] Van Den Oord A, Dieleman S, Schrauwen B. Transfer learning by supervised pre-training for audio-based music classification[C]//Conference of the International Society for Music Information Retrieval (ISMIR 2014). 2014.
[15] Choi K, Fazekas G, Sandler M, et al. Transfer learning for music classification and regression tasks[C]//18th International Society for Music Information Retrieval Conference, ISMIR 2017. International Society for Music Information Retrieval, 2017: 141-149.
[16] Bertin-Mahieux T, Ellis D P W, Whitman B, et al. The million song dataset[J]. 2011.
[17] Ghosal D, Kolekar M H. Music Genre Recognition Using Deep Neural Networks and Transfer Learning[C]//Interspeech. 2018, 2018: 2087-2091.
[18] 徐星.基于最小一范数的稀疏表示音乐流派与乐器分类算法研究 [D]. PhD thesis. 天津大学, 2012.
[19] 喻晓雯,张楠,张勇.音乐作品风格流派的神经网络识别方法研究[J].计算机工程与应用, 2011,47(27):246﹣248.
[20] 王芳.基于深度学习的音乐流派及中国传统乐器识别分类研究[D]. Master’s thesis.南理工大学, 2017.
[21] 黄琦星. 基于卷积神经网络的音乐流派分类模型研究[D].吉林大学,2019.
[22] 冯楚祎. 基于深度学习的音乐自动标注方法研究[D].北京邮电大学,2019.
[23] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.
[24] Yang H, Zhang W Q. Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks[J]. Proc. Interspeech 2019, 2019: 3382-3386.
[25] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[26] Zhang W, Lei W, Xu X, et al. Improved Music Genre Classification with Convolutional Neural Networks[C]//INTERSPEECH. 2016: 3304-3308.
[27] Theunissen F E, Elie J E. Neural processing of natural sounds[J]. Nature Reviews Neuroscience, 2014, 15(6): 355-366.
[28] Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//Advances in neural information processing systems. 2009: 1096-1104.
[29] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE transactions on acoustics, speech, and signal processing, 1980, 28(4): 357-366.
[30] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807-814.
[31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[32] McFee B, Raffel C, Liang D, et al. librosa: Audio and music signal analysis in python[C]//Proceedings of the 14th python in science conference. 2015, 8.

[33] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[34] Dong M. Convolutional neural network achieves human-level accuracy in music genre classification[J]. arXiv preprint arXiv:1802.09697, 2018.
[35] 杜威,林浒,孙建伟,于波,姚恺丰.一种基于分层结构的音乐自动分类方法[J].小型微型计算机系统,2018,39(05):888-892.
[36] 甄超,宋爽,许洁萍.多模态音乐流派分类研究[J].计算机科学与探索,2011,5(01):50-58.
[37] 孙辉,许洁萍,刘彬彬.基于多核学习支持向量机的音乐流派分类[J].计算机应用,2015,35(06):1753-1756.
[38] 何丽,袁斌.利用长短期记忆网络进行音乐流派的分类[J].计算机技术与发展,2019,29(11):190-194.