一种基于改进CRNN的轻量化乐谱识别方法

蒋凌云,鞠金恒,徐 佳,肖 甫
(南京邮电大学计算机学院、软件学院、网络空间安全学院,江苏南京 210003)

摘 要: 基于深度学习的乐谱识别方法提高了识别精度,但存在模型训练单次迭代耗时长、总迭代轮数多的问题. 本文提出了一种改进卷积循环神经网络的轻量化乐谱识别方法CRNN-lite(lightweight Convolutional Recurrent Neural Networks),该方法在卷积层引入残差式深度可分离卷积,减少计算量并加速特征图的提取;在循环层使用双向简单循环单元,采用并行计算避免了串行计算的强依赖问题;在转录层调节交叉熵函数参数,针对性地学习不均衡样本数据. 实验结果表明,该方法提高训练速度,单次迭代耗时为基准网络的43%,在失真图像数据上符号错误率为1.12%,序列错误率为14.5%,错误率指标均优于对比方案.

关键词: 光学乐谱识别;序列识别;卷积循环神经网络;深度可分离卷积;简单循环单元;不均衡样本学习
基金项目: 国家自然科学基金(No.62372250);江苏省333高层次人才培养工程项目(No.BRA2020065)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3167-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230031

A Lightweight Music Recognition Method Based on Improved CRNN

JIANG Ling-yun, JU Jin-heng, XU Jia, XIAO Fu
(*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*)

Abstract: The deep learning-based music score recognition method has improved recognition accuracy, while there is a dilemma of long single iteration time and multiple total iterations in model training. This work proposes CRNN-lite (lightweight Convolutional Recurrent Neural Networks) for music score recognition. CRNN-lite introduced residual depth separable convolution into the convolution layer, which reduced the computation and speeds up the feature map extraction. The bidirectional simple recurrent unit was used in the recurrent layer, and the strong dependence on serial computation was avoided by parallel computation. The parameters of the cross entropy function were adjusted at the transcription layer to learn unbalanced sample data. The results show that the proposed method improves the training speed, the single iteration time is 43% of the benchmark network, the symbol error rate is 1.12% and the sequence error rate is 14.5% on the distorted image data. The error rate indexes are better than the comparison scheme.

Key words: optical music recognition; sequence recognition; convolutional recurrent neural networks; depthwise separable convolution; simple recurrent unit; unbalanced sample learning
Foundation Item(s): National Natural Science Foundation of China (No.62372250); The Research Foundation of Jiangsu for "333 High Level Talents Training Project" (No.BRA2020065)

1 引言

光学乐谱识别(Optical Music Recognition,OMR)[1]研究如何利用计算机将纸质乐谱自动转换成数字乐谱,对于构建数字乐谱库和存档音乐遗产具有重要意义. 深度学习方法采用端到端的方式训练模型,避免了传统OMR方法冗余的步骤,提高音符识别效果. 基于端到端的深度学习乐谱识别方法主要分为目标检测和序列识别两大类.

基于目标检测的方法使用锚框定位音符,对乐谱中的每个独立音符区域进行检测分类来识别[2~4],但复杂音符的原语(符头、符干、符尾)过于紧凑,具有重合区域的音符置信度不高. 此外,音符需要与谱线和上下文绑定,在重组语义阶段会衍生其他问题. 乐谱识别中的序列识别方法是将输入的乐谱图像映射成符号序列并进行相应的分析处理. 卷积循环神经网络(Convolutional Recurrent Neural Networks,CRNN)[5]是序列识别

收稿日期:2023-01-06;修回日期:2023-08-25;责任编辑:朱梅玉

的代表方法，CRNN结合了卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Networks，RNN）及连接时序分类（Connectionist Temporal Classification，CTC），具备CNN图像处理、RNN序列分类及CTC自动对齐的优势，可以有效应用在OMR领域．文献[6]提出了CRNN-baseline，该方法将CNN和双向长短记忆网络（Bi-directional Long Short-Term Memory，Bi-LSTM）相结合构成了卷积循环神经网络，对单声部乐谱图像的音符进行识别，使用链式时序分类CTC自动对齐音符和标签．在原始数据集上达到了0.8%的符号错误率，但是模型收敛速度慢，耗时久，训练困难．文献[7]提出了残差循环卷积循环神经网络（Residual Recurrent CRNN，R2-CRNN），该方法对卷积层进行改进，使用循环残差块提取特征，提取更多的图像特征从而提高乐谱识别的精度．这也导致卷积层的计算增多，增加了模型的复杂度．文献[8]提出了一种基于多尺度残差式卷积神经网络和双向简单循环单元的端到端的光学乐谱识别方法．虽然该方法对循环层改进提升了速度，但是对卷积层的改进大幅度增加了计算量，导致该层运算速度降低．基于序列的方法本身存在训练耗时久，学习效率低的问题．多数研究为了提升识别效果，往往使用深层网络对模型进行改良，虽然这些改良在实际表现中取得了不错的效果，但是随着网络层数加深，网络参数和计算量的激增，导致模型训练的速度更为缓慢[9]．MobileNet V1[10]采用深度可分离卷积代替传统卷积，在保证网络精度的前提下减少网络参数．在MobileNet V1基础上，MobileNet V2[11]使用倒残差结构和线性单元进一步提高了模型性能．MobileNet V3[12]引入了网络结构搜索．ShuffleNet V1[13]使用通道重排解决逐点组卷积带来的信息流通不畅问题．ShuffleNet V2[14]提出了新的block结构提升计算效率．

在轻量级网络模型的启发下，针对乐谱识别方法存在的问题，本文提出了一种轻量化乐谱识别方法CRNN-lite，该方法优化了网络结构，使用残差结构增强了梯度传播，并针对不平衡的音符数据进行改良，在降低网络参数的同时，加快模型的运算速度，缓解模型训练过程中单次迭代耗时长，总迭代轮数多问题，提高了失真乐谱的识别精度．

2 CRNN-lite的乐谱识别方法

2.1 系统框架

针对CRNN训练时间长和识别精度低的问题，CRNN-lite在卷积层引入深度可分离卷积的基础上使用残差连接[15]提取特征，降低计算量的同时提高网络的学习能力；在循环层引入简单循环单元（Simple Recurrent Unit，SRU）[16]，减少门电路单元，通过并行计算加快模型训练速度；在转录层CTC方法引入Focal Loss[17]，将其改进为Focal CTC，均衡学习样本，提高模型的精度．

CRNN-lite的乐谱识别模型结构如图1所示．

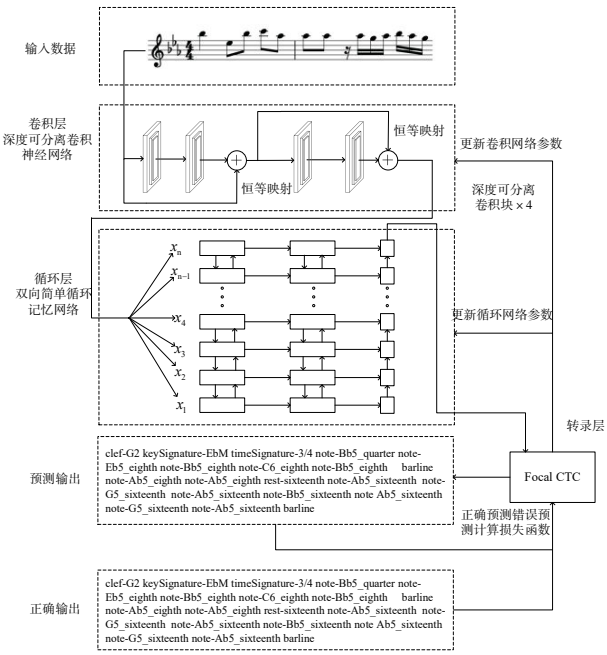


图1 CRNN-lite的乐谱识别模型

具体乐谱识别流程如下：

（1）原始图片输入：输入不定长度的乐谱图片；

（2）卷积层特征提取：将乐谱图片转化为单通道的灰度图．使用4个深度可分离卷积块对灰度图进行卷积操作，并对卷积过程中的中间结果进行残差连接，输出特征图（详见2.2节）；

（3）循环层特征序列分类：调整特征图对应矩阵，得到维度数较低的特征序列．将特征序列矩阵送入SRU组成的双向循环神经网络，输出序列的预测分布矩阵，每一帧对应一维向量，向量的长度等于所有预测音符种类的总和，向量的每个元素为每种音符的概率分布（详见2.3节）；

（4）转录层对齐输出序列：将概率分布矩阵中概率最高的元素重新组合成一维序列矩阵，利用CTC对齐相同语义的帧，去除多余的无效帧；

（5）输出预测序列：将处理后的序列矩阵转换为音符编码序列输出；

（6）更新模型参数：将正确的序列矩阵与预测得到的序列矩阵作为Focal Loss损失函数的参数，根据样本均衡程度调节损失函数值继而反向传播，更新CRNN网络参数（详见2.4节）．

2.2 基于残差式深度可分离卷积的特征提取

对于乐谱图像而言，卷积层提取的特征图将转化

为特征序列,进而传入循环层进行学习. 因此本文在卷积层进行改良,既加快训练的速度,也保证提取特征的有效性.

本文结合了深度可分离卷积和残差网络的思想,改进卷积层为残差式深度可分离卷积网络. 区别于MobileNets中深度可分离卷积采用的Rule6激活函数,本文定义的深度可分离卷积采用LeakyRule激活函数,具体结构如图2所示. 残差式深度可分离卷积网络包括四个深度可分离卷积块,每个卷积块对输入进行卷积操作分别输出C_1、C_2、C_3、C_4,其中原始图片除了作为第一个卷积块的输入外,还通过恒等映射与第二个卷积块的输出C_2进行加运算,得出结果M_1将作为第三个卷积块的输入,并再次恒等映射与第四个卷积块的输出C_4进行加运算,得到M_2并作为最终输出. 具体的网络结构如图3所示.

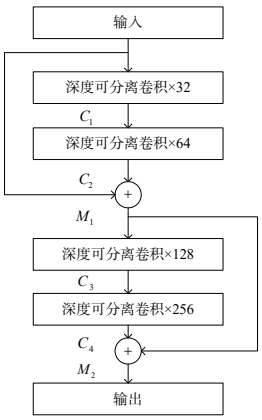


图2　深度可分离卷积结构

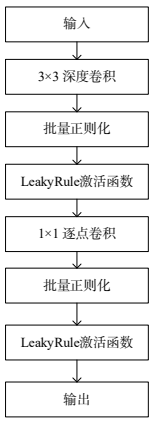


图3　残差式深度可分离卷积网络结构

2.3　基于SRU的音符分类预测

为了提高训练和推理速度,SRU对门控单元(Gated Recurrent Unit,GRU)进行改进,解除了网络对隐藏状态的强依赖,通过预先求得门控参数矩阵,实现了网络的并行计算. SRU在t时刻的结构如图4所示,x_t表示输入,h_t表示输出状态,c_t表示内部状态,σ为Sigmoid激活函数,v_f和v_r是对c_{t-1}进行映射的参数向量,g_t是关于x_t的线性变换:$g_t=\omega x_t$,ω为其参数矩阵. f_t表示遗忘门,r_t表示重置门.

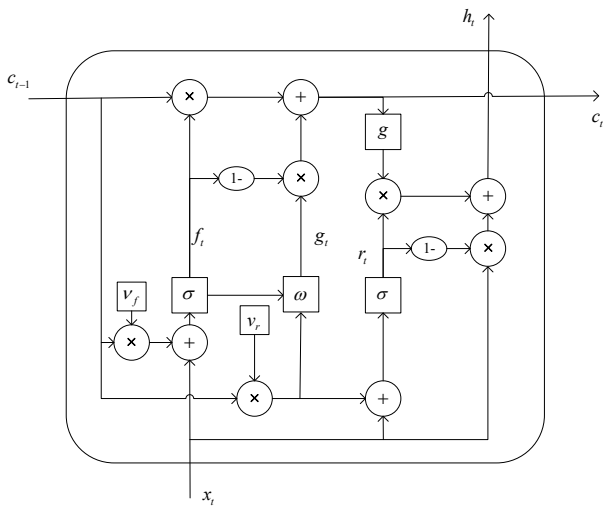


图4　SRU结构图

为了减轻递归的程度,它的两个门控单元f_t和r_t不再依赖于上一时刻的隐藏状态h_{t-1},而是依赖于上一时刻的内部状态c_{t-1},遗忘门f_t计算见式(1),重置门r_t计算见式(2),b_f和b_r分别为f_t和r_t的偏置单元. c_t综合了过去状态的信息和当前输入的信息,用Hadamard乘积代替矩阵乘积减少计算量,见式(3):

$$f_t=\sigma(w_f x_t+v_f\odot c_{t-1}+b_f) \tag{1}$$

$$r_t=\sigma(w_r x_t+v_r\odot c_{t-1}+b_r) \tag{2}$$

$$c_t=f_t\odot c_{t-1}+(1-f_t)\odot g_t \tag{3}$$

h_t采用了跳跃连接的方法,计算方法见式(4),直接将输入x_t纳入计算,其目的在于优化梯度传播,在网络深度增加时,不会因为传播距离过远而使得梯度消失.

$$h_t=r_t\odot c_t+(1-r_t)\odot x_t \tag{4}$$

SRU循环网络结构如图5所示,模型共包含两层双向SRU,每层双向SRU共512个隐藏单元,首先按时间顺序传递信息,最后一个输出单元再按时间逆序传递信息. 两层双向SRU输出的结果最终通过点乘计算进行预测分类.

2.4　基于Focal Loss的均衡样本学习

在乐谱中频率高的符号在训练过程中对模型的影响较大,而频率低的符号在训练过程中往往被忽略. Focal Loss[17]克服了由于数据集不平衡导致的过拟合和欠拟合问题. 基于焦点理论和交叉熵,Focal Loss损失函数定义见式(5):

$$L_{Focal\_Loss}(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

其中p_t为预测概率大小,α_t和γ为可调节因子,

$L_{Focal_Loss}(p_t)$代表预测概率为p_t时的损失函数. 式(5)中p_t反映了预测序列与真实序列的接近程度,p_t越大说明越接近真实序列,即分类越准确. p_t也反映了分类的难易程度,p_t越大说明分类的置信度越高,代表样本越易分类;p_t越小说明分类的置信度越低,代表样本越难分类. α_t用于调节正负样本损失之间的比例,抑制正负样本的数量失衡,γ用于控制简单/难区分样本数量失衡. $(1-p_t)^{\gamma}$为该损失函数的调制因数,对于分类准确的样本$p_t \to 1$,调制因数趋近于0. 对于分类不准确的样本$p_t \to 0$,调制因数趋近于1. 在α_t和γ的共同作用下,对于分类不准确的样本,Focal Loss损失没有改变,对于分类准确的样本,损失会变小. 整体而言,Focal Loss增加了分类不准确样本在损失函数中的权重,使得损失函数倾向于难分的样本,有助于提高难分样本的准确度.

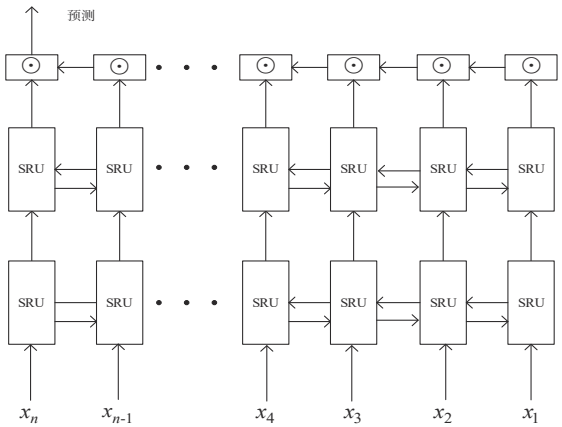


图5 SRU循环网络结构

3 实验设计与结果分析

3.1 数据集及实验环境

本文选用公开数据集PrIMus(Printed Images of Music staves)[18],该数据集包含87 678张完整的乐谱图像和相同数量的合成失真图像. 每张图像对应一种semantic编码序列,semantic编码蕴含了音符的内在语义(如表1所示). 语义格式在乐谱图片中的呈现如图6和图7所示,其中bB大调签名被标记为音乐意义上的keySignature,两个符号b合并为一个序列元素,而非单独识别. 图6所示乐谱图像所对应的semantic编码标签序列如图8所示.

本实验所用到的计算机配置如表2所示,使用基于Tensorflow链接库实现本文的方法.

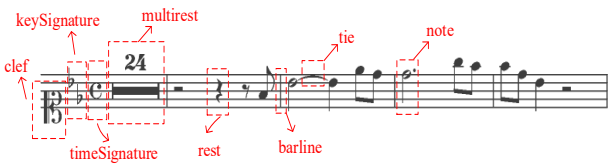


图6 乐谱原始图像上语义标签对应的乐谱符号

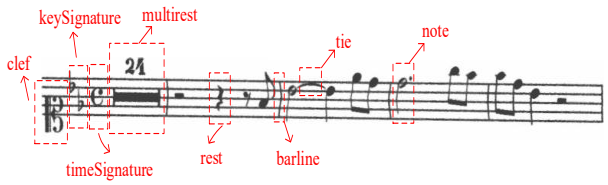


图7 图6所示乐谱失真图像上语义标签对应的乐谱符号

clef-C1, keySignature-BbM, timeSignature-C, multirest-24, barline, rest-half, rest-quarter, rest-eighth, note-F4_eighth, barline, note-Bb4_half, tie, note-Bb4_quarter, note-Eb5_eighth, note-D5_eighth, barline, note-D5_half., note-G5_eighth, note-F5_eighth, barline, note-F5_eighth, note-D5_eighth, note-Bb4_quarter, rest-half, barline

图8 图6所示乐谱图像所对应的semantic编码标签序列

在实验过程中,使用Adam自适应学习率算法优化深度学习模型参数,初始学习率设置为0.001,批处理大小设置为16,每经过1 000次迭代算法在验证集上进行符号错误率的评估以验证模型的精度,整个过程共64 000次迭代.

3.2 实验指标

模型的训练速度和推理速度往往与模型的复杂度呈正相关,因此引入网络参数量(Params)和计算量(FLoating Point Operations,FLOPs)增加可解释性. 同时为了验证模型的学习能力,根据已有的相关研究结果提出两个评价指标来对模型的学习能力进行评估:

(1)符号错误率(Symbol Error Rate,ER_{sym}) 通过模型预测生成的符号中,非正确符号数量占序列总符号的比例,其定义见式(6):

$$ER_{sym}=\frac{\sum_{i=1}^{n}(S_i+D_i+I_i)}{\sum_{i=1}^{n}N_i} \tag{6}$$

其中,n为预测序列的总数. S_i表示第i个序列需要人为修改的符号数量;D_i表示第i个序列需要人为删除的符号数量;I_i表示第i个序列需要人为添加的符号数量. N_i表示第i个序列预期符号总数.

(2)序列错误率(Sequence Error Rate,ER_{seq}) 通过模型预测生成的标签序列中,非正确序列个数所占总序列个数的比例(非正确序列是指该序列中至少有一

表1 semantic语义格式

乐谱曲调调号	乐谱拍号	音符	休止符	多小节休止符	延音符号	倚音、装饰音	小节线
keySignature	timeSignature	note	rest	multirest	tie	gracenote	barline

表2　计算机配置

中央处理器/GHz	Intel Xeon E5-2683 V3 @ 2.00
显卡/GB	NVIDIA Tesla K80, 12
内存/GB	32
操作系统	Ubuntu 16.04

个符号为非正确符号),具体定义见式(7),其中η为指示函数,当$(S_i+D_i+I_i)>0$时其值为1,否则η为0.

$$\mathrm{ER}_{\mathrm{seq}}=\frac{\sum_{i=1}^{n}\eta_{(S_i+D_i+I_i)>0}}{n} \tag{7}$$

3.3　实验对比与分析

3.3.1　深度可分离卷积的有效性

本文改动卷积层的目的是降低神经元数量,减少计算参数. 因此,本节选用参数量Params和计算量FLOPs作为卷积网络的衡量指标. 为了证明本文网络的轻量性,选用了基准网络、主流网络Resnet中参数量最小的Resnet18[15]和文献[7]改进的卷积网络作为实验的对比网络. 表3展示了输入乐谱图像大小为768×128×1时各个网络的Params和FLOPs.

表3表明,使用残差网络的其他网络参数量和计算量均大于基准网络,而本文采用深度可分离卷积改进的网络与基准网络相较,无论在参数量还是计算量都大大降低. 实验结果表明,本文采用的卷积网络结构更加轻量化.

表3　各模型参数量和运算量

模型	Params/MB	FLOPs/G
CRNN-baseline[6]	0.38	2.8
R2-CRNN[7]	2.04	17.7
Resnet18[15]	11.19	4.38
CRNN-lite	**0.24**	**1.7**

深度可分离卷积除了更加轻量化,还通过残差结构提高了特征提取效果. 表4比较了两种卷积和残差结构组合下的符号错误率和序列错误率. 结果显示,无论是原始卷积和残差结构的组合,还是使用深度可分离卷积和残差结构的组合,都可以显著降低符号错误率和序列错误率. 因此,卷积和残差结构的组合是一种有效的手段,可以在符号识别和序列识别任务中提高模型的性能表现. 虽然原始卷积和残差结构的组合在错误率指标上略好于深度可分离卷积和残差结构的组合,但是深度可分离卷积和残差结构的组合更加轻量化.

表4　残差结构的有效性

原始卷积	深度可分离卷积	残差结构	符号错误率/%	序列错误率/%
√			3.4	38.3
√		√	1.72	20.9
	√		3.68	37.9
	√	√	1.79	22.3

除此以外,为了更好说明所设计的深度可分离卷积网络中残差结构的有效性,还提取了四个卷积块的输出C_1、C_2,C_3,C_4和残差连接后的输出M_1,M_2所代表的特征图,如图9所示,M_1特征图相较于C_2锐化了音符轮廓,M_2特征图相较于C_4对音符最终像素点进行了调整. 这说明残差结构可以有效保留浅层特征、提高特征表达能力,在音符边缘锐化和音符像素点微调方面具有一定的有效性. 通过残差结构,网络可以更好地学习和提取音符的特征信息,使其在训练过程中能够更准确地识别和定位音符,从而提高模型的性能表现.

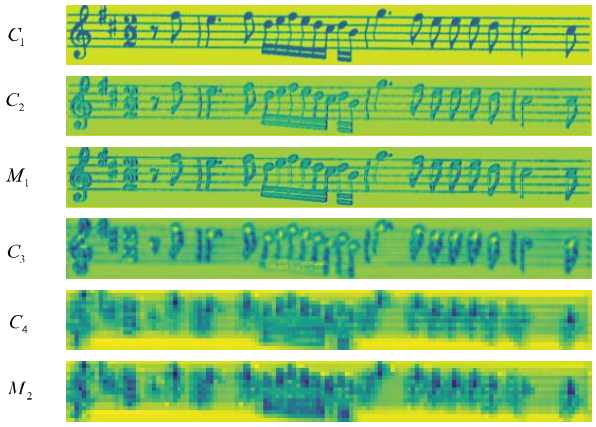


图9　各层特征图效果

3.3.2　SRU的有效性

本文选取SRU的目的是通过并行计算提高计算速度,并利用SRU中类似残差网络的结构提高模型精度. 文献[16]已经证明SRU在各类数据集中的速度优于其他RNN网络,因此本小节的实验仅证明SRU在OMR任务中的表现优于其他网络. 本节对CRNN循环层的网络分别使用LSTM、GRU、SRU进行实验,对比三种网络的损失函数、符号错误率、序列错误率在相同时间内的收敛情况,实验结果如图10~12所示. 图10表明,在前2 000轮的迭代中,三种网络损失值的下降趋势都比较陡峭,尽管三种模型的收敛速度比较快,但是采用了SRU的模型下降趋势对比其他两个模型更明显,下降幅度更大;在2 000轮以后,三种网络损失值趋于平缓,并在损失值10左右开始震荡下降,且后续的训练过程中一直保持这种趋势. 从曲线上看,在震荡过程中,采用SRU的模型损失值对比其他模型相对更小. 可见SRU在收敛速度上更具备优势.

从图11符号错误率及图12序列错误率来看,尽管迭代多轮以后错误率都比较低,但是使用SRU的模型符号错误率和序列错误率均低于对比模型. 这表明相同迭代轮次下,SRU在OMR任务中的表现优于其他

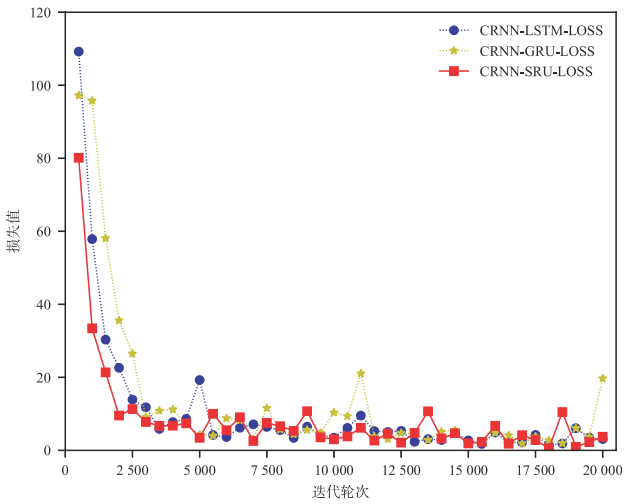


图10 损失函数变化曲线

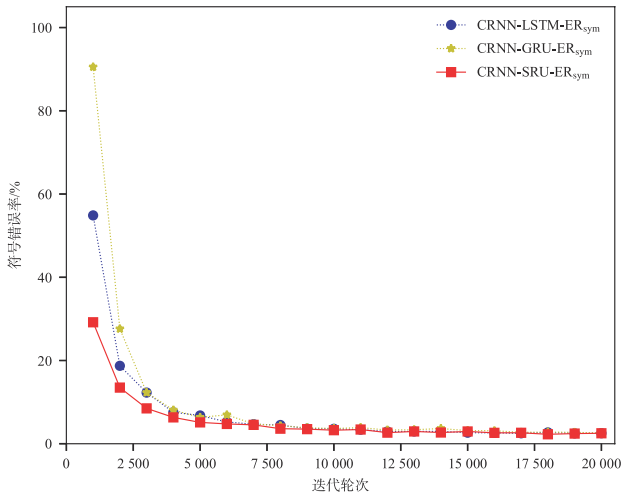


图11 符号错误率变化曲线

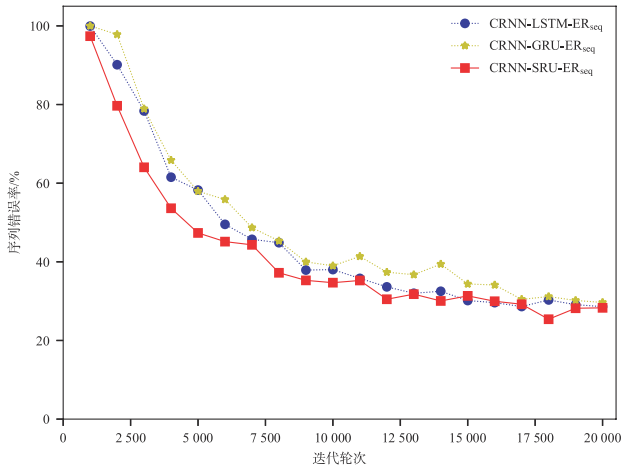


图12 序列错误率变化曲线

RNN网络,在音符识别精度更具优势. 结合SRU本身并行计算的特性,迭代速度也优于LSTM和GRU. 因此可以得出结论,SRU在OMR任务中可以大幅度提高运算效率.

3.3.3 Focal Loss的有效性

Focal CTC针对低频率音符进行识别,从而提高模型识别率. 通过控制Focal Loss的调节因子,获取OMR任务中Focal CTC均衡学习的最佳参数. 为检测Focal CTC低频率样本识别的有效性,从测试集中随机选取了8 760张乐谱图片,分别使用CTC及使用最佳Focal Loss的Focal CTC对低频率音符识别. 表5展示部分低频音符标签的识别准确率,实验结果表明,使用Focal CTC的对低频率音符识别准确率有明显提高.

表5 低频率音符标签符号准确率

标签	出现次数/次	CTC/%	Focal CTC/%
gracenote-B4_thirty_second	9	45	56
gracenote-D5_thirty_second	21	42	85
timeSignature-9/8	31	81	93
rest-thirty_second	28	85	96
note-C5_sixty_fourth	5	60	80

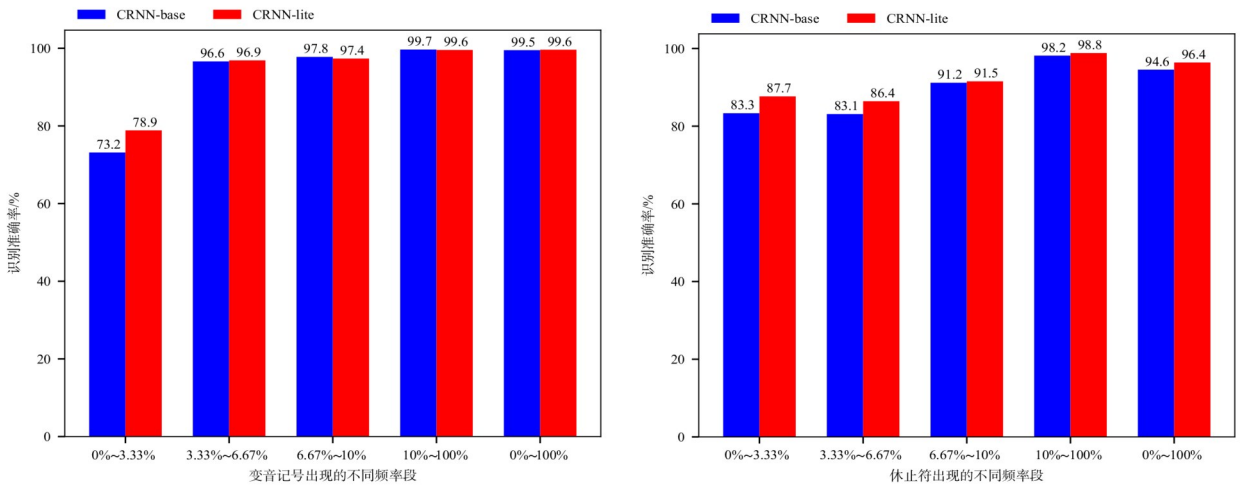
3.3.4 变音记号和休止符的识别效果

为了更好地验证CRNN-lite对复杂音符的识别效果,本文针对变音记号和休止符的识别效果进行测试. 由于音符通常不是单独存在,而是与其他音符组合形成音乐中的语义信息,因此在对音符进行semantic编码时,即使在乐谱中某些音符的纹理图像相同,但它们与别的音符组合成的编码标签也可能有几十种甚至上百种. 此外,每种编码标签出现的次数也不尽相同. 本节统计了数据集中所有编码标签出现的次数,并按照升序排序,划分前10%出现次数越少的编码标签为低频,并在此基础上,对这部分编码标签再次划分为三等分. 使用CRNN-lite和CRNN-baseline分别对不同频率分布的编码标签中变音记号和休止符的错误率进行了测试. 测试结果如图13所示,相较于CRNN-base,CRNN-lite在低频0~6.67%上的变音记号和休止符准确率都有所提升.

3.3.5 综合实验对比

为了全面评估模型的性能,本文对比了所有条件下符号错误率、序列错误率和单次迭代耗时情况. 除了CRNN-baseline和R2-CRNN,对比实验还加入了轻量化的模型进行比较. 对比模型使用轻量化网络替换了CRNN-baseline的卷积层,选用MobileNetv3[12]的small版本作为骨架网络的CRNN-MobileNetv3 small. 此外,本文还使用了ShuffleNetv2[14]输出维度缩放倍数为0.5和1.0的两个版本对CRNN-baseline进行微调,分别为CRNN-ShuffleNetv2 0.5x和CRNN-ShuffleNetv2 1.0x.

表6列举了它们与CRNN-lite在原始数据集和失真数据集下的表现. 在原始数据集上训练和测试,CRNN-



(*a*) 不同频率下变音记号的准确率　　(*b*) 不同频率下休止符的准确率

图13　变音及休止符不同频率准确率对比

lite的符号错误率为0.67%,序列错误率为9.6%,而CRNN-baseline的符号错误率为0.8%,序列错误率为12.5%,CRNN-lite在错误率指标上均优于CRNN-baseline. 尽管相同情况下CRNN-lite的符号错误率略高于R2-CRNN的0.56%,但序列错误率低于R2-CRNN的13.4%,这说明CRNN-lite完整识别单行乐谱的正确率更高. 在失真数据集上训练,利用原始数据集测试,CRNN-lite的符号错误率和序列错误率上的表现分别为0.81%和12.7%,低于CRNN-baseline的3.3%和44.6%及R2-CRNN的1.06%和25.4%;利用失真数据集测试,CRNN-lite的符号错误率和序列错误率分别为1.12%和14.5%,低于CRNN-baseline的3.4%和38.3%及R2-CRNN的1.5%和35.8%,这说明CRNN-lite在失真数据集上表现优于其他对比网络. 因为失真训练集与现实场景中拍摄的乐谱图像相似,而CRNN-lite对失真训练集更具优势,这表明CRNN-lite更适应现实中的乐谱识别任务. 此外,在失真数据集下训练的CRNN-lite同样可以有效识别原始数据集,在原始数据集上测试符号错误率和序列错误率分别为0.67%和10.1%,这说明CRNN-lite具备一定的泛化性,能够适应不同类型乐谱上的识别任务.

表6　评估各类数据集上的模型表现

训练数据集	识别方法	原始数据集(测试集)		失真数据集(测试集)		耗时/s
		符号错误率/%	序列错误率/%	符号错误率/%	序列错误率/%	
原始数据集(训练集)	CRNN-baseline[6]	0.8	12.5	59.7	97.9	1.644
	R2-CRNN[7]	0.56	13.4	38.19	91	2.342
	CRNN-MobileNetv3 small	0.95	12.2	49.02	95.5	0.73
	CRNN-ShuffleNetv2 1.0x	0.86	10.9	37.29	92.1	0.65
	CRNN-ShuffleNetv2 0.5x	0.89	12.8	40.7	95.8	0.55
	CRNN-lite	**0.67**	**9.6**	**47.16**	**93.5**	**0.723**
失真数据集(训练集)	CRNN-baseline[6]	3.3	44.6	3.4	38.3	1.632
	R2-CRNN[7]	1.06	25.4	1.5	35.8	2.351
	CRNN-MobileNetv3 small	1.13	17.6	1.91	21.6	0.70
	CRNN-ShuffleNetv2 1.0x	1.15	17.1	1.84	21.2	0.61
	CRNN-ShuffleNetv2 0.5x	1.42	19.4	2.24	25.5	0.53
	CRNN-lite	**0.81**	**12.7**	**1.12**	**14.5**	**0.719**

除了模型的错误率更优以外,CRNN-lite每次迭代的平均耗时也远低于CRNN-baseline及R2-CRNN,相同情况下单次迭代时间仅需0.72 s左右,耗时为CRNN-baseline的43%,R2-CRNN的30%,训练耗时短,训练效率更高. 尽管与使用主流轻量化网络的CRNN(CRNN-MobileNetv3 small、CRNN-ShuffleNetv2 0.5和CRNN-ShuffleNetv2 1.0x)相比,CRNN-lite在速度方面并不占上风. 但是,除在原始训练集上训练并在失真数据集上测试的错误率指标外,CRNN-lite在其他所有错误率指标上都优于对比网络. 失真数据集对原始数据集上训

练的模型干扰太多,因此不能以此指标来说明其他轻量化网络的优势. 除此之外的错误率指标能够很好地说明本文方法在轻量化模型中的识别效果更好. 综合来看,CRNN-lite并不会因为轻量化而损失过多精度,也不会为了提升准确率使得模型臃肿,在具备轻量化属性的同时在准确率方面也有所提升.

4 结论

本文提出了一种轻量化乐谱识别方法CRNN-lite,在精简网络结构的基础上,采用并行计算提高模型的训练速度. 此外,还针对不均衡的音符样本调节网络参数,通过残差结构增强梯度传播. 实验表明,CRNN-lite在失真乐谱的识别任务中准确率优于其他对比网络. 未来工作中,将针对训练数量少的复杂音符识别率的提高技术开展进一步研究,并扩展对简谱识别的支持.

参考文献

[1] CALVO-ZARAGOZA J, HAJIČ J, PACHA A. Understanding optical music recognition[J]. ACM Computing Surveys, 2020, 53(4): 1-35.

[2] PACHA A. Incremental Supervised Staff Detection[C]// Proceedings of the 2nd International Workshop on Reading Music Systems. Alicante: WoRMS, 2019: 16-20.

[3] TUGGENER L, ELEZI I, SCHMIDHUBER J, et al. DeepScores-a dataset for segmentation, detection and classification of tiny objects[C]//2018 24th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2018: 3704-3709.

[4] JIA X, SONG Y Q, MA S C, et al. Printed score detection based on deep learning[C]//2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS). Piscataway: IEEE, 2021: 173-177.

[5] 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). Piscataway: IEEE, 2017: 2392-2396.

[6] BARÓ A, RIBA P, CALVO-ZARAGOZA J, et al. From optical music recognition to handwritten music recognition: A baseline[J]. Pattern Recognition Letters, 2019, 123: 1-8.

[7] LIU A Z, ZHANG L P, MEI Y Q, et al. Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores[C]//Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding. New York: ACM, 2021: 23-27.

[8] 吴琼, 李锵, 关欣. 基于多尺度残差式卷积神经网络与双向简单循环单元的光学乐谱识别方法[J]. 激光与光电子学进展, 2020, 57(8): 67-76.
WU Q, LI Q, GUAN X. Optical music recognition method combining multi-scale residual convolutional neural network and bi-directional simple recurrent units[J]. Laser & Optoelectronics Progress, 2020, 57(8): 67-76. (in Chinese)

[9] 袁海英, 成君鹏, 曾智勇, 等. Mobile_BLNet:基于Big-Little Net的轻量级卷积神经网络优化设计[J]. 电子学报, 2023, 51(1): 180-191.
YUAN H Y, CHENG J P, ZENG Z Y, et al. Mobile_BLNet: Optimization design of lightweight convolutional neural network based on big-little net[J]. Acta Electronica Sinica, 2023, 51(1): 180-191. (in Chinese)

[10] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17) [2021-12-28]. https://arxiv.org/abs/1704.04861.

[11] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2018: 4510-4520.

[12] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1314-1324.

[13] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.

[14] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design[C]//European Conference on Computer Vision. Cham: Springer, 2018: 122-138.

[15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[16] BALLAKUR A A, ARYA A. Empirical evaluation of gated recurrent neural network architectures in aviation delay prediction[C]//2020 5th International Conference on Computing, Communication and Security (ICCCS). Piscataway: IEEE, 2020: 1-7.

[17] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Con-

ference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2999-3007.

[18] CALVO-ZARAGOZA J. The printed images of music staves (PrIMuS) dataset[EB/OL]. (2018-04-11)[2022-10-08]. https://grfia.dlsi.ua.es/primus.

作者简介

蒋凌云　女,1978年4月出生,湖南永州人. 现为南京邮电大学副教授,硕士生导师. 主要研究方向为信息网络与智能信息处理.
E-mail: jianglingyun@njupt.edu.cn

鞠金恒　男,1997年8月出生,江苏南京人. 现为南京邮电大学硕士. 主要研究方向为图像识别与时间序列处理.
E-mail: 1558332432@qq.com

徐　佳(通迅作者)　男,1980年10月年出生,江苏常州人. 现为南京邮电大学教授,博士生导师. 主要研究方向为无线充电网络、智能信息处理. 中国电子学会会员编号:E190011107M.
E-mail: xujia@njupt.edu.cn

肖　甫　男,1980年10月出生,湖南省邵阳市. 现为南京邮电大学教授,博士生导师. 主要研究方向为物联网感知与计算、数据中心网络、智能信息处理. 中国电子学会会员编号:E190029628M.
E-mail: xiaof@njupt.edu.cn