基于注意力机制藏文乌金体古籍文字识别研究

童　攀[1,2,3],龙炳鑫[1,2,3],拥　措[1,2,3*]
(1. 西藏大学 信息科学技术学院,西藏 拉萨 850000;
2. 西藏大学 藏文信息技术人工智能西藏自治区重点实验室,西藏 拉萨 850000;
3. 西藏大学 藏文信息技术教育部工程研究中心,西藏 拉萨 850000)

摘　要:藏文乌金体古籍文字识别是古籍文字识别领域的一个难题。针对藏文乌金体古籍中存在的文字粘连和背景复杂问题,提出一种基于注意力机制的藏文乌金体古籍文字识别方法。该方法主要包含两部分,编码器部分采用卷积神经网络(CNN)与双向长短期记忆(Bi-LSTM)获得图像文本的特征序列和序列标注,解码器部分使用注意力机制计算注意力权重并与循环神经网络(RNN)相结合得出识别结果。采用实验室的 616 张藏文乌金体古籍作为实验数据集以及藏文字丁准确率作为实验评测指标。采用两种文字识别模型作为基线模型,从模型大小和识别率进行对比,文中识别模型在模型大小和识别效果上都优于其他两个模型,文中模型大小 41.2 MB,相比基线模型中最小的优化了 36 MB,字丁识别准确率 90.55%,相比基线模型中最好的结果提高了 7.94 百分点。表明所提出的基于注意力机制的藏文乌金体古籍识别模型,显著提高了藏文乌金体古籍中的粘连文字和背景复杂图像的识别效果。

关键词:藏文古籍;文字识别;乌金体;注意力机制;字丁准确率
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1673-629X(2023)10-0163-06
doi:10.3969/j.issn.1673-629X.2023.10.025

Research on Tibetan Ujin Ancient Book Character Recognition Based on Attention Mechanism

TONG Pan[1,2,3], LONG Bing-xin[1,2,3], YONG Cuo[1,2,3*]
(1. School of Information Science and Technology, Tibet University, Lhasa 850000, China;
2. Key Laboratory of Tibetan Information Technology and Artificial Intelligence of Tibet Autonomous Region, Tibet University, Lhasa 850000, China;
3. Engineering Research Center of Tibetan Information Technology of Ministry of Education, Tibet University, Lhasa 850000, China)

Abstract: The Tibetan Ujin ancient book character recognition is a difficult problem in the field of ancient book character recognition. Aiming at the problems of text adhesion and complex background in Tibetan Ujin ancient book character recognition, we propose an attention mechanism based recognition method for Tibetan Ujin ancient books, which consists of two parts. The encoder adopts the convolutional neural network (CNN) and Bi-LSTM to obtain the feature sequence and sequence annotation of image text. The decoder uses the attention mechanism to calculate the attention weight and obtains the recognition result by combining the method of recurrent neural network (RNN). 616 Tibetan Ujin ancient books in the laboratory are used as the experimental data set and the accuracy rate of Tibetan characters is used as the experimental evaluation index. Two text recognition models are used as the baseline model. Compared with the model size and recognition rate, the proposed recognition model is superior to the other two models in terms of model size and recognition effect. The size of proposed recognition model is 41.2 MB, which is optimized by 36 MB compared with the smallest baseline model. The recognition accuracy of character block is 90.55%, which is 7.94% higher than the best result in the baseline model. It is showed that the proposed recognition model of Tibetan Ujin ancient books based on attention mechanism significantly improves the recognition effect of text adhesiont and complex background images in Tibetan Ujin ancient books.

收稿日期:2022-10-20　　**修回日期**:2023-02-22
基金项目:国家重点研发计划重点专项(2017YFB1402202);西藏自治区科技创新基地自主研发项目(XZ2021HR002G)
作者简介:童　攀(1996-),男,硕士研究生,研究方向为文字识别;通信作者:拥　措(1974-),女,教授,博导,研究方向为藏语自然语言处理、古籍智能信息处理。

Key words: ancient books in Tibetan; text recognition; the sharply body; mechanism of attention; accuracy of character

0 引 言

藏文乌金体古籍文字识别是计算机视觉领域的一个难题,同时也是国内外文献资源数字化领域的一个重要研究方向。藏文乌金体古籍是藏族文化的重要组成成果,同时也是中华宝贵文化遗产的一部分,藏文古籍的数字化,对研究藏族文化教育,藏学研究、传承优秀传统文化等方面都发挥着极其重要的作用。目前,多数藏文乌金体古籍识别算法在清晰的藏文乌金体古籍图像中能取得较好的识别效果,而对于藏文乌金体古籍中存在的文字粘连和背景复杂的图像,其识别效果有待进一步提高。

国内外关于藏文古籍识别的研究相对稀少。20 世纪 90 年代日本情报处理学会为了研究藏文佛教典籍,设立了藏文字符识别项目,1996 年完成了识别系统[1]。该系统并没有解决藏文古籍图像中的文字切分问题,需要人工切分,并且只完成了字符识别功能。为了解决藏文古籍字切分的问题,Hedayati 等人[2]首次将广义隐马尔可夫模型应用在藏文古籍识别流程中。西藏大学赵栋材等人[3]首次将反向传播网络应用在藏文古籍文字识别研究。为了增加识别效果,西藏大学高飞[4]进行藏文古籍图像二值化研究。随着深度学习技术的不断发展,藏文古籍文字识别有了更多的研究。2018 年,王筱娟[5]首次将深度神经网络应用于藏文古籍相似字的识别,该方法有效提高了在藏文乌金体古籍相似字符的识别准确率。2019 年,西北民族大学李振江[6]提出基于边缘对比的二值化方法,西北民族大学韩跃辉[7]进行基于色彩空间转换的二值化研究。同年李振江[8]提出利用基线信息进行字符识别方法,将藏字分为上下两部分进行识别,提高了藏文字符的识别准确率。2021 年,由于藏文古籍数据稀少且难以收集的问题,西藏大学仁青东主[9]进行了藏文古籍文字识别数据的合成方法研究,一定程度上解决了藏文古籍训练规模小的问题。在藏文古籍的系统应用中,韩跃辉[10]采用基于卷积神经网络(Convolutional Neural Network,CNN)模型的字丁识别算法,设计并完成了藏文古籍识别系统,提高了藏文古籍 7 240 类字丁的识别率。胡鹏飞[11]采用藏文文本行数据集合成的方法以及端到端的深度学习模型,实现了文本行图像的整行识别。仁青东主[12]使用残差网络和双向循环长短期记忆循环神经网络以及基于滑动窗的行识别技术,解决了行文字较长的问题。2021 年,西藏大学完成承担的国家重点研发项目,设计并完成了藏文古籍木刻本版面分析于文字识别系统,可以完成对整页藏文乌金体古籍的识别。

现有的藏文乌金体古籍文字识别中的问题包括:(1)藏文乌金体古籍文字识别数据集资源稀少;(2)藏文乌金体古籍文字粘连图像和背景复杂图像识别效果不佳;(3)缺少一个行之有效的藏文识别评测指标。针对这些问题,该文的主要贡献为:(1)提出以藏文字丁为基本单位的藏文字丁准确率评测标准,并应用在西藏大学国家重点研发项目中;(2)在文献[13]提出的 Encoder-Decoder 模型以及文献[14]提出的注意力机制的基础上设计了识别模型算法,该模型在只有 616 张藏文乌金体古籍图像作为数据集的情况下,以藏文字丁准确率为标准取得了 90.55% 的字丁识别效果。

1 相关工作

1.1 文字识别

近些年来,主流的文字识别方法主要分为两种:基于连接时域分类(Connectionist Temporal Classification,CTC)的识别方法(如文献[15])和基于注意力机制的识别方法。

基于 CTC 的识别方法的框架模型,首先使用卷积神经网络对图像进行视觉特征提取,再将视觉特征沿着宽度方向进行切片以形成特征序列,将特征序列输入至序列建模之中,如 RNN。再生成具有序列上下文的特征序列,最后使用 CTC 解码每个序列特征进行字符类别预测并基于动态规划对预测结果进行去重。该识别方法只依赖于视觉特征和视觉特征之间的序列关系,所以面对模糊文本和低质量图像等难识别样本时性能不好。

基于注意力机制的识别算法,同样是先使用卷积神经网络进行图像特征提取,然后使用编码器生成具有序列上下文信息的特征序列,使用注意力机制取所有特征序列为键和值,取解码器中前一个时间步的预测为查询进行注意力权重的计算,并对特征序列进行加权求和生成当前时间的解码特征,将其送入解码器中进行结果预测,持续过程直到输出终止符或超过预定时间步。该方法可以自动寻找需要预测的文本区域,并将注意力集中在图像中字符对应像素点位置,显著地提高了模型的准确率。

1.2 藏文特点

藏文是一种拼音型文字,是由 4 个元音字母和 30 个辅音字母组成,藏文的每一个音节都是由藏文字符通过纵向、横向或者是纵向横向组合而成的,每个藏文音节之间是通过“·”分隔开的,现代藏文字的音节结构一般是由七种空间位置构成的,分别是前加字,上加

字,基字,下加字,元音,后加字,再后加字,但是存在两种特殊的情况“ཀྱ”和“བཀྱ”,属于是基字,下加字,再下加字的组成,所以现代藏文字的音节结构可以总结为八种空间位置的组合结构,如图 1 所示。其中按照纵向划分,每一列称之为字丁,如图 2 所示。

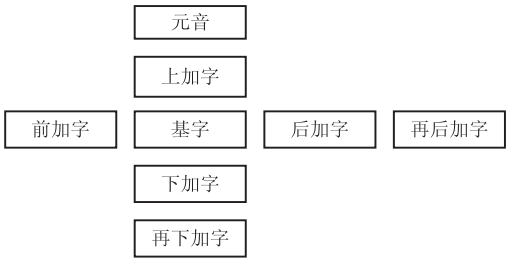


图 1　现代藏文音节结构

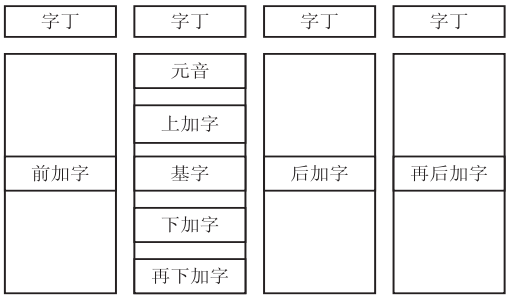


图 2　现代藏文字丁结构

由于受印度文化的影响,藏文中还存在特殊的梵文藏文转写形式,梵文藏文转写并不符合藏文文法规则,而是符合梵文的文法规则,在藏文古籍文献、藏文新闻等中时有出现,如图 3 所示。在识别中对藏文字进行字丁切分的主要目的有:

(1)保持藏文字的空间结构信息;

(2)简化识别任务。

图 3　梵文藏文转写

2　模型算法

2.1　基于注意力机制的卷积循环神经网络

模型使用编码器–解码器(Encoder–Decoder)的模型结构,如图 4 所示,其中 x 表示输入信息,c 表示通过 Encoder 层输出的语义编码,y 表示通过 Decoder 层获得的识别结果。该结构可以有效地将长度不同的图像特征与之对应的文本序列进行对齐,同时注意力机制会自动寻找需要预测的文本区域,将注意力集中在图像中字符对应的像素点位置从而显著提高模型的准确率。

该文使用的基于注意力机制的卷积循环神经网络(CRNN+ATTENTION)识别算法流程如图 5 所示。该算法可以支持的字丁长度是有限的,根据训练结果,该识别算法可识别的字丁个数为 25。网络对于输入图像的长宽并没有限制。通过对收集的藏文古籍乌金体数据的藏文字丁统计共获得了 1 353 个藏文字丁,并以此作为网络支持的类别数。

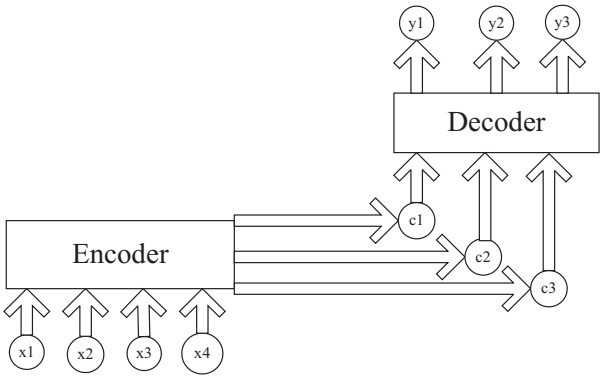


图 4　编码器–解码器结构

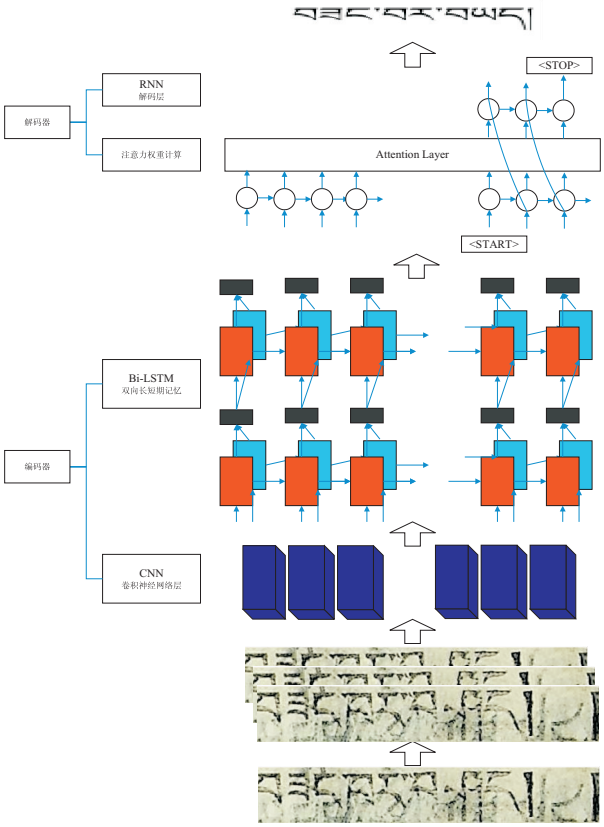


图 5　藏文古籍图像识别流程

在网络的前端,卷积神经网络自动从输入的图像中提取特征,将特征结果送入双向长短期记忆(Bidirectional Long Short Term Memory,Bi–LSTM)网络进行特征增强。接着注意力模型根据循环神经网络(Recurrent Neural Network,RNN)神经元的隐藏状态及上一时刻的输出计算出注意力权重,最后将卷积神经网络输出的特征图与注意力权重结合起来,输入循环神经网络进行编解码后,得到整个字符集的概率分

布,最后直接提取概率最高的编号所对应的字符作为最后的识别结果。

主要模型架构包括以下两个方面:

(1)编码器。

第一步,使用 CNN 网络提取输入图像的特征序列,输出为特征矩阵。在特征提取过程, imgH (图像高度)方向经过 4 个 pooling 和 1 个卷积(Valid 模式), imgW (图像宽度)方向经过 2 个 pooling 和 1 个卷积(Valid 模式),原图高度变为 imgH/32,原图宽度变为 imgW/4 + 1。获得图像的特征矩阵。

参数设置如表 1 所示。其中 K 、S 和 P 分别是卷积核(kernel size)、步长(stride)和填充大小(padding size)。BatchNorm2d 为参与特征的通道数。

表 1 卷积层参数

参数设置	配置
CONV2D	#OUT_CHANNELS:64 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
MAXPOOL2D	K:2 S:2 P:0
CONV2D	#OUT_CHANNELS:128 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
MAXPOOL2D	K:2 S:2 P:0
CONV2D	#OUT_CHANNELS: 256 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
BATCHNORM2D	256
CONV2D	#OUT_CHANNELS: 256 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
MAXPOOL2D	K:(2,2) S:(2,1) P:(0,1)
CONV2D	#OUT_CHANNELS: 512 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
BATCHNORM2D	512
CONV2D	#OUT_CHANNELS:512 K:(3,3) S:(1,1) P:(1,1)
RELU	TRUE
MAXPOOL2D	K:(2,2) S:(2,1) P:(0,1)
CONV2D	#OUT_CHANNES: 512 K:(2,2) S:(1,1) P:(1,1)
RELU	TRUE
BATCHNORM2D	512

第二步,使用 Bi-LSTM 的方法对卷积层结果进行前后序列特征的增强。BLSTM 在 LSTM 的基础上,进一步学习上下文特征,结合了输入序列在前向和后向两个方向上的信息。对于 t 时刻的输出,前向 LSTM 层具有输入序列中 t 时刻以及之前时刻的信息,而后向 LSTM 层中具有输入序列中 t 时刻以及之后时刻的信息。

循环参数设置如表 2 所示。其中 nIn 是输入特征数, nHidden 是 LSTM 中隐藏层的维度, Bidirectional 表示是否使用双向 LSTM, nOut 是输出特征数。

表 2 循环层参数

参数设置	配置
LSTM	NIN:512 NHIDDEN:256 BIDIRECTIONAL:TRUE
EMBEDDING	LINEAR(NHIDDEN * 2,NOUT:256)
LSTM	NIN:512 NHIDDEN:256 BIDIRECTIONAL:TRUE
EMBEDDING	LINEAR(NHIDDEN * 2,NOUT:256)

(2)解码器。

第一步,计算注意力权重之前先对前一次的输出进行词嵌入,并进行特征融合,然后计算注意力权重。

注意力权重的计算需要三个指定的输入 Q(query),K(key),V(value),分别表示查询,键值,值。然后通过计算得到注意力的权重结果。可以将其归纳为三个阶段:第一个阶段根据 Query 和 Key 计算两者的相似性或者相关性;第二阶段对第一阶段的原始分值进行归一化处理;第三个阶段根据权重系数对 Value 进行加权求和。第一阶段计算 Query 和 Key 某个的相似性,使用点向量积的方法进行计算。公式如下:

$$\mathrm{Sim}(\mathrm{Query},\mathrm{Key}_i) = \mathrm{Query} * \mathrm{Key}_i, \quad i \in I \tag{1}$$

第二阶段一方面可以进行归一化,将原始计算分值整理成所有元素权重之和为 1 的概率分布;另一方面也可以通过 SoftMax 的内在机制更加突出重要元素的权重。公式如下:

$$a_i = \mathrm{softmax}(\mathrm{Sim}_i) = \frac{e^{\mathrm{Sim}_i}}{\sum_{j=1}^{lx} e^{\mathrm{Sim}_j}} \tag{2}$$

式中, a_i 为 Value_i 对应的权重系数,第三阶段将每一个 a_i 进行加权求和即可获得注意力的权重,公式如下:

$$\mathrm{Att}(\mathrm{Query},\mathrm{Source}) = \sum_{i=1}^{lx} a_i * \mathrm{Value}_i \tag{3}$$

第二步,将卷积神经网络输出的特征图与注意力权重结合起来,根据 Attention 权重合并成 1 个最大概率的字符。

第三步,输入循环神经网络进行编解码后,得到整个字符集的概率分布,直接提取概率最高的编号所对应的字符作为最后的识别结果。

参数设置如表 3 所示。其中 out_size 表示字典的

维度, Dropout 表示每个神经元不被激活的可能性。

表 3　转录层参数

参数设置	配置
EMBEDDING	OUT_SIZE:1 353,NHIDDEN: 256
ATTN_COMBINE	LINEAR(NHIDDEN * 2,NOUT:256,TRUE)
DROPOUT	P:0.1
GRU	GRU(NHIDDEN: 256, NHIDDEN: 256)
LINEAR	NHIDDEN,OUT_SIZET:256,TRUE
LINEAR	NHIDDEN,NHIDDEN:256,TRUE

2.2　评测标准

对于藏文文字识别,目前并没有一个固定的评测标准。该文采取编辑距离作为藏文古籍乌金体文字识别的准确率计算标准。编辑距离可以充分反映出藏文古籍乌金体识别中出现的错识,漏识以及多识的情况。有利于对识别结果进行分析。藏文与中英文不同,每一个中英文都有对应的编码,而一个藏字是由多个藏文字符编码组成的,简单的理解就是一个藏字就是多个藏文字符组合在一起的字符串,不易于比较且计算量较大。考虑藏文文字的结构特点,该文以藏文字丁为基本单位进行准确率计算。

提出的藏文字丁准确率算法的计算公式如下所示:

$$Acc = rd/(rd + ld) \tag{4}$$

式中,Acc 是字丁准确率,rd 是字丁匹配中对应位置正确的字丁个数,ld 是字丁匹配中错误的字丁个数,包括识别中出现的多识,漏识,错识三种情况。rd + ld 是总共的比较次数,其计算结果并不一定等于标注文件的字丁个数。

3　实　验

实验运行环境:CPU 12th Gen Intel(R) Core(TM) i5-12400F 2.50 GHz;GPU NVIDIA GeForce RTX 3060;内存 12 G;程序为 Linux 系统 pytorch 框架编写运行。

表 4　训练参数

参数设置	配置
WORKERS	0
BATCHSIZE	128
THE_LSTM_HIDDEN_STATE	256
EPOCHS	500
LEARNING RATE	0.000 01
ADAM	TRUE
BETA1	0.5
CUDA	TRUE
ADADELTA	RMSPROP
RANDOM_SAMPLE	TRUE

以 500 张整页藏文乌金体古籍作为训练集,116 张藏文乌金体古籍作为测试集。实验训练参数如表 4 所示。图 6 为所使用的藏文乌金体古籍样本图。正常整页藏文乌金体古籍识别流程应该是先进行藏文古籍文本检测以及文本行切分处理,文本行切分处理结果送入文字识别模块最后将识别结果进行后处理。该文主要说明识别模型的识别效果,故文本检测,文本行切分处理和识别后处理这里不详细解释。

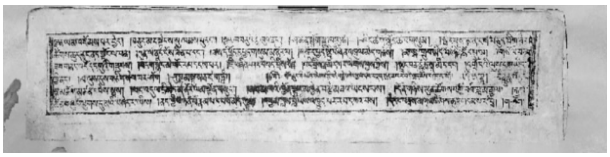


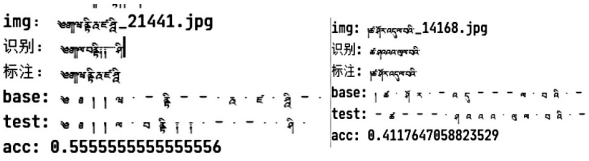
图 6　藏文古籍图像样本图

为了展示各模型的藏文乌金体古籍识别效果,特意截取两小块识别难度高的藏文乌金体古籍文本行图像,如图 7 所示,图 8 为各模型针对两小块的识别结果。图 9 为文中模型在 116 张整页藏文乌金体古籍中随机截取 300 个文本块的识别准确率曲线。

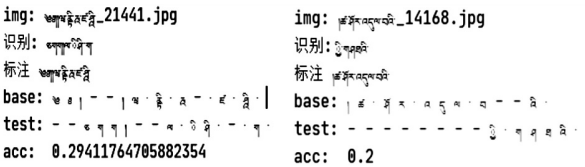
(a)文字粘连图像

(b)背景复杂图像

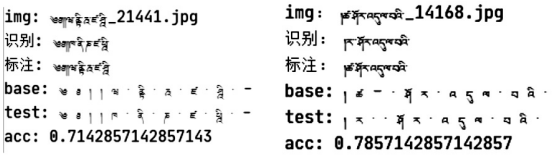
图 7　测试图像示例



(a)ABiNet 识别结果



(b)CRNN+CTC 识别结果



(c)CRNN+ATTENTION 识别结果

图 8　各模型识别结果

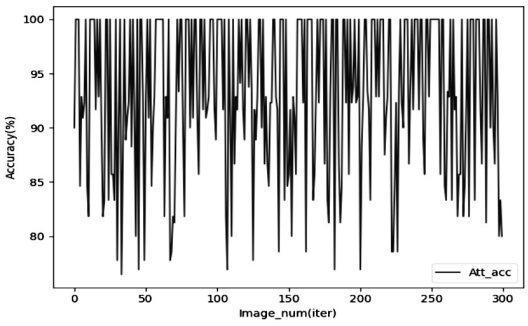


图 9 CRNN+ATTENTION 识别曲线

通过图 8 和图 9 的效果展示可以看出,文中模型在文字粘连和背景复杂严重的藏文乌金体古籍图像中的识别效果相较于其他模型有较为明显的提高,通过对 300 个文本块识别结果的分析,发现文中模型在识别结果中容易出现藏文音节点“་”和“|”的多识别以及在藏文音节中不该出现地方识别出来,此外对于藏文中相似字符的识别容易出现差错。针对这些问题,与学藏文同学交流发现,可以通过藏文的文法规则来进行约束和训练藏文古籍语言模型来做识别结果的后处理,以此提高藏文乌金体古籍的识别效果。

将文中识别模型与文献[16]提出的 CRNN+CTC 识别模型以及文献[17]提出的基于 ABINET 识别模型进行实验对比。同时为了进一步验证采用的注意力机制有效提高了藏文乌金体古籍识别效果,在文中算法基础上删去注意力机制进行实验,如表 5、表 6 所示,分别为文中模型与对比模型,文中模型与删去注意力机制的文中模型进行 500 epoch 训练之后使用 116 张样本测试获得的平均字丁准确率。

表 5 不同算法识别结果对比

模型	大小/M	字丁准确率/%
CRNN+ATTENTION	41.2	90.55
CRNN+CTC	77.2	81.94
ABINET	452	82.61

表 6 注意力机制的文中模型对比

模型	大小/M	字丁准确率/%
CRNN+ATTENTION	41.2	90.55
CRNN	41.2	69.95

由表 5 可以看出,在使用小样本的文字粘连和背景复杂的藏文乌金体古籍图像进行模型训练情况下,引入注意力机制能有效提高藏文乌金体古籍的识别准确率,使用 CTC 算法的模型其识别准确率明显低于基于注意力机制的识别模型。同时文中模型与去掉注意力机制的文中模型进行比较,充分说明注意力机制能有效提高对藏文乌金体古籍中文字粘连和背景复杂图像的识别效果。文中模型在少样本的情况下,能充分利用样本整体的上下文信息,并取得了较好的效果。同时,文中模型相比其他模型,在提升识别精度的同时,有效压缩了模型的大小,提升了算法的实用价值。

4 结束语

针对藏文乌金体古籍图像中的背景复杂和文字粘连的识别问题,采用卷积循环神经网络 CRNN 与 Attention 注意力机制相结合的模型解决行文字粘连问题;以动态规划的方法结合藏文字丁结构设计出来的藏文字丁识别准确率为评测指标;以统计藏文古籍中单独出现的藏文字丁为识别字典。通过与 CRNN+CTC 模型和 ABiNet 模型在相同条件下的实验结果进行对比,文中模型的识别效果最好,其字丁准确率为 90.55%,在只有 500 张藏文乌金体古籍进行模型训练的情况下取得了高效的识别结果。通过对文中模型测试的结果分析来看,后续计划训练藏文古籍语言模型以及添加藏文文法规则的方法来对识别结果进行后处理,以提高最终的识别效果。

参考文献:

[1] KOJIMA M,KAWAZOE Y,KIMURA M. Automatic recognition of wooden blocked tibetan image character by using object oriented design[J]. Ipsj Sig Notes,1998,98:39-44.

[2] HEDAYATI F,CHONG J,KEUTZER K. Recognition of Tibetan wood block prints with generalized hidden Markov and kernelized modified quadratic distance function[C]//Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data. New York:Association for Computing Machinery,2011:1-14.

[3] 赵栋材. 基于 BP 网络的木刻藏文经书文字识别研究[J]. 微处理机,2012,33(5):35-38.

[4] 高 飞,沈淑涛. 藏文古籍图像信息自适应补偿二值化算法研究[J]. 电子制作,2017(20):51-52.

[5] WANG Xiaojuan,HAO Zhanjun,HAN Yuehui,et al. A recognition method of the similarity character for uchen script Tibetan historical document based on DNN[C]//Chinese conference on pattern recognition and computer vision. Guangzhou:Springer,2018:52-62.

[6] LI Z,WANG W,CAI Z. Historical document image binarization based on edge contrast information[C]//Science and information conference.[s.l.]:Springer,2019:614-628.

[7] HAN Y,WANG W,LIU H,et al. A combined approach for the binarization of historical tibetan document images[J]. International Journal of Pattern Recognition and Artificial Intelligence,2019,33(14):1954038.

[8] LI Zhenjiang, WANG Weilan. Tibetan historical document recognition of uchen script using baseline information[C]// Tenth international conference on graphics and image pro-

(下转第 208 页)