ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 1,Jan. 2022,pp. 173－182
DOI:10. 16337/j. 1004-9037. 2022. 01. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax: ＋86-025-84892742

基于知识蒸馏的缅甸语光学字符识别方法

毛存礼[1,2], 谢旭阳[1,2], 余正涛[1,2], 高盛祥[1,2], 王振晗[1,2], 刘福浩[1,2]

(1. 昆明理工大学信息工程与自动化学院,昆明 650500;2. 昆明理工大学云南省人工智能重点实验室,昆明 650500)

摘　要: 与传统的图像文本识别任务不同,缅甸语光学字符识别(Optical character recognition, OCR)需要计算机在一个感受野内识别由多个字符嵌套组合的复杂字符,这给缅甸语OCR任务带来了巨大的挑战。为了解决该问题,提出了一种基于知识蒸馏的缅甸语OCR方法,构建了使用卷积神经网络(Convolutional neural networks, CNN)＋循环神经网络(Recurrent neural network, RNN)框架的教师网络和学生网络,以集成学习的方式进行训练的模型架构,在训练过程中通过教师集成的子网络与学生网络进行耦合,实现学生网络中单个感受野对应的局部字符图像特征与教师网络中整体字符图像特征的对齐,以此增强对长序列字符图像中局部特征的获取。实验结果表明,在没有背景噪声图像和有背景噪声图像作为训练数据集的情况下,本文模型的性能分别优于基线2.9%和2.7%。

关键词: 缅甸语;光学字符识别;卷积神经网络＋循环神经网络;知识蒸馏;图像特征对齐
中图分类号: TP391.1　　**文献标志码:** A

Burmese OCR Method Based on Knowledge Distillation

MAO Cunli[1,2], XIE Xuyang[1,2], YU Zhengtao[1,2], GAO Shengxiang[1,2], WANG Zhenhan[1,2], LIU Fuhao[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; 2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Different from traditional image text recognition tasks, the Burmese optical character recognition (OCR) requires computers to recognize complex characters nested and combined by multiple characters in a receptive field, which brings great challenges to Burmese OCR tasks. To solve this problem, a Burmese OCR method based on knowledge distillation is proposed. This paper constructs a model of teacher network and student network using the framework of convolutional neural networks (CNN)＋ recurrent neural networks (RNN) to train in an integrated learning way. In the training process, the teacher integrated sub-network is coupled with the student network to realize the alignment of the local character image features corresponding to a single receptive field in the student network and the overall character image features in the teacher network, so as to enhance the acquisition of local features in long sequence character images. The experimental results show that the performance of our model is better than the baseline by 2.9% and 2.7% respectively without and with background noise images as training data sets.
Key words: Burmese; OCR; CNN ＋ RNN; knowledge distillation; image feature alignment

基金项目: 国家自然科学基金重点项目(61732005);国家自然科学基金(62166023,61866019,61761026,61972186);云南省重大科技专项计划(202103AA080015);云南省应用基础研究计划重点项目(2019FA023);云南省中青年学术和技术带头人后备人才项目(2019HB006)。
收稿日期: 2020-08-01;**修订日期:** 2021-05-06

引　言

缅甸语文字有Zawgyi-One、Myanmar Three等多种字体编码，为避免网络中缅语文本内容显示乱码的问题，大多数缅语文本内容都是以图片形式呈现。这对于开展面向缅甸语的自然语言处理、机器翻译和信息检索等研究带来较大的困难。虽然结合深度学习的方法在中英文图像文本识别任务中已经取得了非常可观的效果，但由于缅甸语字符的特殊性，目前还没有关于缅甸语光学字符识别（Optical character recognition，OCR）研究方面的相关成果，因此开展缅甸语OCR研究具有非常重要的理论和实际应用价值。

光学字符识别通常用于识别图像中的自然语言。对于文本字符识别的早期工作，例如Anderson[1]主要将图像转换为结构化语言或标记，这些结构化语言或标记定义了文本本身及其现有语义。之后，在英语[2-3]、汉语[4-6]、德语[7]、阿拉伯语[8]、马拉雅拉姆语[9]和印地语[10]等OCR技术达到高识别率的相关报导陆续出现。利用卷积神经网络模型进行文本图像识别的相关工作有很多，例如文献[11]首次尝试对单个字符进行检测，然后利用深度卷积神经网络模型对这些检测到的特征进行识别，并用标记后的图像进行训练，但是该方法需要预先训练鲁棒的字符检测器，这样增加了文本图像识别任务的计算复杂度。而且缅甸语中的一个感受野内通常会出现由多个字符嵌套组合的复杂字符，很难切分成单个字符，因此该方法不适用于缅甸语图像文本识别任务。同时深度卷积神经网络[12-13]只能处理固定的输入和输出维度，但是缅甸语序列的长度变化相当大，例如，汉语“现在”的缅甸语翻译为“ခု”是由2个字符组成，而汉语“第二”的缅甸语翻译为“ဒုတိယအကြိမ်”是由11个字符组成，所以基于深度卷积神经网络的工作还不能直接用于基于缅甸语图像的序列识别任务。利用循环神经网络（Recurrent neural network，RNN）模型做图像文本识别任务也有一些相关的工作，然而在RNN处理序列之前，必须先将输入图像转换成图像特征序列。例如，Graves等[14]从手写文本中提取了一系列图像或几何特征，而Su等[15]将字符图像转换为一系列方向梯度直方图（Histogram of oriented gradient，HOG）特征。因此，目前基于循环神经网络的方法很难直接用于缅甸语光学字符识别。

缅甸语与中文或者英文不同，在一个感受野内英文字母或中文字由单个Unicode编码组成，然而缅甸语在1个感受野内可能由2个或者3个Unicode编码组成。例如，在图1(a)中，缅甸语“ဖြဲ”在感受野中由3个字符“ျ”(/ u107f)，“ဝ”(/ u1015)和“ဲ”(/ u102e)组成，但是在图1(b)中，感受野中的英语“n”由一个字符“n”(/ u006e)组成。缅甸语OCR任务不仅受到图像中的背景噪声、光照和图片质量等因素影响，还更难解决缅甸语多个字符嵌套组合的复杂字符的识别问题。在这种情况下，导致缅甸语OCR任务难度更大。

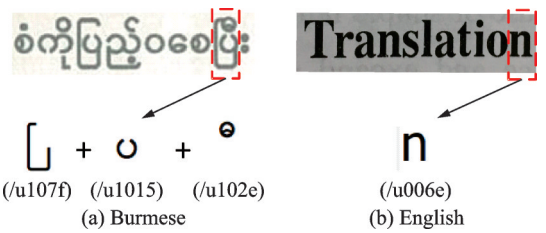


图1　1个感受野内不同语言的字符结构

Fig.1　Structure of characters in different languages in a receptive field

目前比较主流的方法是Shi等[16]提出的卷积循环神经网络（Convolutional recurrent neural network，CRNN）方法和Luo等[17-18]提出的Attention方法，它能端到端地有效解决英文序列识别问题，在英文识别方面达到了一定的效果。但是这些方法只能解决一个感受野内一个字符的识别问题，当处理缅甸文多个字符嵌套组合的复杂字符时识别准确率降低。因此本文提出了基于知识蒸馏的缅甸语OCR方法，构建教师网络和学生网络进行集成学习的网络框架，通过教师网络来指导学生网络，将来自教师的不同缅甸语组合字符和单字符特征知识提取到学生网络中，使学生网络能够学习到缅甸语组合字符的识别优点，从而解决复杂的缅甸文多字符组合词难以识别和提取的问题。

1 缅甸图像数据特征分析与预处理

缅甸语不同于一般的语言,具有非常复杂的字符空间组合结构,在计算机提取图像上的语言特征时非常困难。所以本文分析了缅甸语语言特征,利用基于知识蒸馏的缅甸语OCR方法,将教师网络提取到的单字符和多个字符嵌套组合的复杂字符特征对学生网络对应相同的位置字符信息进行特征增强,从而提高整句话的识别准确率。由于目前没有公开的缅甸图像文本识别数据,所以本文构造了缅甸语OCR模型训练测试的数据集。

1.1 特征分析

缅甸语音节字符构成结构与其他语言存在较大差异,具有基础字符、基础前字符、基础后字符、基础上字符和基础下字符,每个音节边界以基本辅音开头。缅甸语有33个辅音,辅音与元音结合,有时包含中音节,从而构成完整的缅甸语音节。此外,它在音节和单词之间没有分隔符,只有根据缅甸语的字符规则编码顺序,才能获得正确的缅甸语句子。这样就会引起相应的问题,当计算机提取图像特征时,1个感受野中可能包含多个字符,这增加了缅甸语OCR识别复杂度,而这种复杂字符在缅甸句子对中占大多数。

1.2 数据预处理

本文通过网站(www.nmdc.edu.mm)收集了120万个缅甸语句子。例如:“နဝမအကြိမ်မြောက် ပီကင်း ရုန်းဆန်း ဖိုရမ်ကို ၂၂ရက်နေ့က”“ပြောကြားပါသည်။”。然后,利用缅甸语片段切分工具将缅甸语音节和句子切成长序列缅甸语段文本数据。例如,汉语语义“论坛参会者”对应的缅语是“ယင်းဖိုရမ်ကို ကျင်းပရာတွင် တက်ရောက်သူများသည်”,分段后的缅语表示为“ယင်းဖိုရမ်ကို”“ကျင်းပရာတွင်”和“တက်ရောက်သူများသည်”。根据缅甸语的语言特点,对分段后的缅甸语文本数据进行人工分割成单字符和多个字符嵌套组合的复杂字符的缅甸语,并且保留其位置信息。

利用文本生成图像工具,将文本数据随机生成分辨率为10像素×5像素~500像素×300像素的含有背景噪音与不含有背景噪音的缅甸语图像,从而构造出训练任务所需的Zawgyi-One字体缅甸语图像,将其作为训练集、测试集和评估集数据。

若干个缅甸语音节构成一句缅甸语句子,一个缅甸语音节的Unicode编码可以分为5部分[19]:<辅音><元音><声调><韵母>和<中音>。这5个部分中只有辅音总是存在,在任何给定的音节中,一个或多个其他部分可能是空的。在实际中,元音可以显示在辅音之前,但是元音字符编码在辅音字符编码之后,例如“ကော”,但是它的编码为(/ u1000)(缅甸字母“က”)(/ u1031)(缅甸元音符号“ော”),所以需要对音节重新排序以进行归类,因为最后1个音节的优先级高于元音。因此,按照缅甸语Unicode编码算法顺序:<辅音><声调><元音><韵母>和<中音>对缅甸语图像进行规则性标注。

2 基于知识蒸馏的缅甸语OCR模型

本文提出模型架构如图2所示。图中的网络架构由教师网络和学生网络两部分组成,其中KD Loss表示知识蒸馏损失,其余的变量说明请见下文。利用单字符和嵌套组合字符的训练集来训练教师网络解决单个感受野内嵌套组合字符识别问题,利用长序列字符图像数据集来训练学生网络解决长序列字符识别问题。在训练过程中,学生网络与教师集成的子网络进行耦合,根据教师集成产生的组合字符特征和真实性标签对学生模型的参数进行优化,以此增强学生网络对缅甸语组合字符特征的提取,解决了缅甸语组合字符进入网络后容易被计算机误判,导致识别准确率低的问题。以下各节将详细地介绍学生网络、教师网络以及集成知识蒸馏的网络训练。

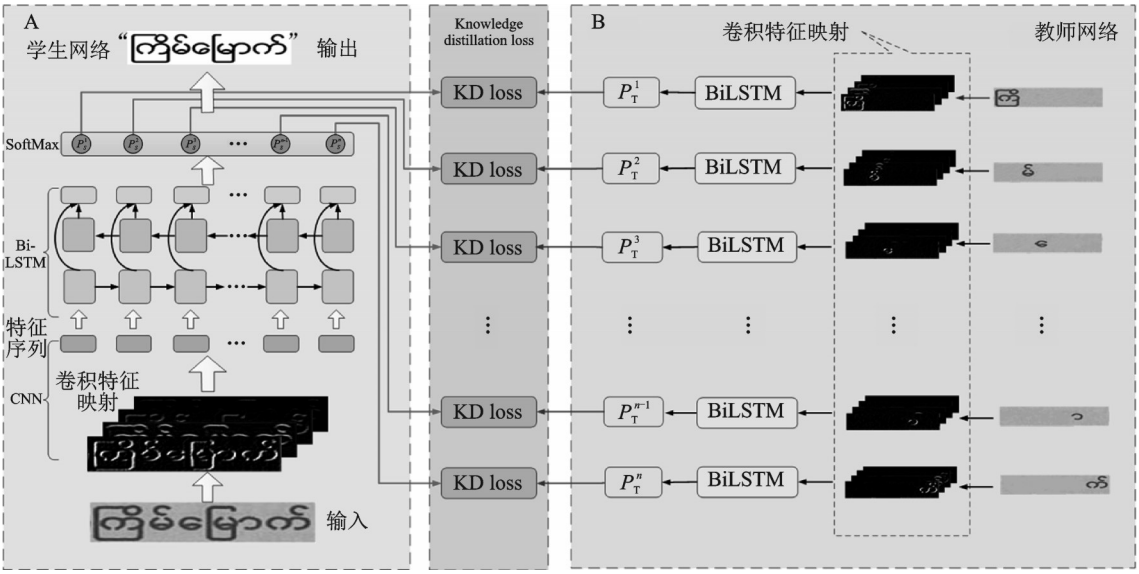


图2 缅甸语OCR模型网络框架图

Fig.2 Network framework diagram of Burmese OCR model

2.1 学生网络

2.1.1 缅甸语图像特征向量序列的提取与标注

本文采用了深度卷积神经网络模型中的卷积层、最大池化层和删除全连接层来构造适应缅甸语图像数据的卷积神经网络，所有的权重共享连接。同时在基于VGG-VeryDeep[20]架构的基础上构建了适应缅甸语OCR任务需求的深度卷积神经网络模型组件，在第3个和第4个最大池化层中采用1×2大小的矩形池化窗口，用以产生宽度较大的特征图，从而产生比较长的缅甸语的特征序列。本文设置输入的缅甸图像生成30帧的特征序列，特征序列的每个特征向量在特征图上从左到右逐列生成，使所有特征图的第x列映射到第x个特征向量上，从而保证图像上的信息全部转移到特征向量上。

本文选择双向长短期记忆网络（Bi-directional long short-term memory，BiLSTM）来处理深度卷积神经网络中获得的特征向量序列，从而获得特征的每个列的概率分布，即预测从前一层卷积提取的特征序列$X=x_1,\cdots,x_T$中每个帧x_t的标签分布y_t。使用长短期记忆网络（Long short-term memory，LSTM）用于解决传统的RNN单元梯度消失的问题。LSTM由输入、输出和遗忘门组成。存储单元的作用是存储过去的上下文，同时输入和输出门允许单元较长时间地保存到输入缅甸语图像中的上下文信息，并且单元里保存到的信息又可以被遗忘门删除。在提取的缅甸语图像特征序列中，不同方向的上下文信息具有互补作用，遇到一些模糊的字符在观察其上下文时更容易区分。例如：当遇到相似字符“ꩻ”时，不会识别成“ꩻꩻ”或者“ꩻꩻ”，这样可以使识别精度更加准确。因为LSTM通常是定向的，训练时只利用到过去的上下文信息，所以本文方法选用了BiLSTM，将向前向后的2个LSTM组合成1个BiLSTM，并且可以叠加多次，进而提升实验效果。同时BiLSTM能够从头到尾对任意长度的序列进行操作，这样就可以处理字符较多的缅甸语句子。

训练循环神经网络期间，当循环神经网络接收到特征序列中的帧x_t时，使用非线性函数来更新y_t，非线性函数同时接收当前输入x_t和过去状态y_{t-1}作为RNN的输入，即：$y_t=g(x_t,y_{t-1})$。在BiLSTM的底部，产生具有偏差的序列将会连接成映射，将缅甸语图像的特征映射转换为特征序列，然后再反转

并反馈到卷积层;也就是说,输出结果将会被送回到输入,之后在下一个时刻和下一个输入同时使用,这样就形成了卷积神经网络和循环神经网络之间联系桥梁。

2.1.2 转录

转录是将RNN所做的每帧预测转换成标签序列的过程。对于RNN得到的每帧预测的标签分布,本文依托Graves等[21]提出的联接时间分类(Connectionist temporal classification, CTC)计算出所有的标签序列概率。

定义符号“-”作为空白符加入到缅甸语字符集合中,从而用“-”填充RNN得到的每帧预测的标签分布中没有字符的位置。另外还定义了f变换,用以删除空白字符集合变换成真实输出。f变换的作用为:把连续的相同字符删减为1个字符并删去空白符,但是如果相同字符间有“-”符号的则把“-”符号删除,然后保留相同的字符。

例如:$f(\pi_1)=f($ ---အ--စ--ည်--းအ---ဝးတွ--ဒ်- $)=$ အစည်းအဝေးတွင် ,$f(\pi_2)=f($ --ဢဢအ--စ--ည်ည်--းအ---ဝးတွ--ဒ်- $)=$ အစည်းအဝေးတွင် ,$f(\pi_3)=f($ အအအ----စ--ည်ည်--းအ---ဝးတွ--ဒ် $)=$ အစည်းအဝေးတွင် 。

例子中$\pi\in Q'^{T}$,$Q'=Q\cup\{-\}$,这里Q包含了训练中的所有缅甸语标注好的正确标签。计算出所有的标签序列概率后,在输入$\boldsymbol{y}=y_1,\ldots,y_T$的情况下,输出为$\boldsymbol{q}$的标签序列概率为

$$p(\boldsymbol{q}|\boldsymbol{y})=\sum_{\pi:f(\pi)=q}p(\pi|\boldsymbol{y}) \tag{1}$$

式中:T为序列长度;输入$\boldsymbol{y}$为$T\times n$的后验概率矩阵,n为缅甸文字符数;$p(\pi|\boldsymbol{y})=\prod_{t=1}^{T}y_{\pi_t}^{t}$,$y_{\pi_t}^{t}$为时刻$t$时有标签$\pi_t$的概率,使用文献[21]中描述的前向后向算法可以计算存在指数级数量的求和项的方程$p(\pi|\boldsymbol{y})$。

如果词典较大时,对整个词典进行搜索非常费时,所以选用了基于最近邻候选目标的方式来选取最大的标签序列概率所对应的标签序列作为图像中每一帧的缅甸语最终的预测结果,即

$$q^{*}=\arg\max_{\boldsymbol{q}\in M_\delta(q')}p(\boldsymbol{q}|\boldsymbol{y}) \tag{2}$$

式中:$M_\delta(q')$为最近邻候选目标(可以使用BK树数据结构[22]快速有效地找到);δ为最大编辑距离;q'为在没有词典的情况下输入为$\boldsymbol{y}$转录的序列。

2.2 教师网络

本文的教师网络模块由多个CNN+RNN+CTC模型组成。教师子网使用与学生网络位置信息对应的缅甸语单字符和组合字符图像作为网络输入,以便在最后的卷积层产生不同的特征表示。教师网络最后产生与目标类有关的输入数据的概率分布,如图2中B部分所示。

由于缅甸语单字符和组合字符是有一定规则数量限制的,因此教师网络选用基于词典的转录方式更加准确便捷,词典是一组标签序列,当预测缅甸语图像识别结果时将会受到拼写检查字典的约束。在基于词典的情况下,选取最大的标签序列概率所对应的标签序列作为最终的预测结果,即

$$q^{*}=\arg\max_{\boldsymbol{q}\in D}p(\boldsymbol{q}|\boldsymbol{y}) \tag{3}$$

式中D为词典。

2.3 集成知识蒸馏的网络训练

采用知识蒸馏方法将教师学习到的对齐片段的单字符和组合字符特征对学生模型的参数进行优化指导,使学生网络能够强化学习到缅甸语组合字符的识别,从而提高整体缅甸语句子识别的准确率。将图像和标签$(x',y')\in(X',Y')$组成训练的数据集,其中每个样本属于K个类($Y'=1,2,\cdots,K$),教师网络输入的单字符或者组合字符图像位置与学生网络所输入的单字符或者组合字符图像具有相同的位

置信息。为了学习映射$f_s(x): X' \rightarrow Y'$,本文通过$f_s(x', \theta^*)$训练学生网络的参数,其中$\theta^*$是通过最小化训练目标函数$L_{train}$获得的学习参数,表达式为

$$\theta^* = \arg\min_{\theta} L_{train}(\boldsymbol{y}', f_s(x', \theta)) \tag{4}$$

本文的训练函数是3个损失项的加权组合。教师网络和学生网络的损失值分别用L_{CET}和L_{CES}表示,真实标签用$\boldsymbol{y}'$表示,知识蒸馏损失值用L_{KD}表示,知识蒸馏损失值与教师集成子网络的输出和学生模型的输出相匹配,即

$$L_{train} = \alpha L_{CET}(P_T, \boldsymbol{y}') + \beta L_{CES}(P_S, \boldsymbol{y}') + \gamma L_{KD} \tag{5}$$

式中$P_T = f_t(x)$和$P_S = f_s(x)$分别表示教师网络和学生模型中图像对应缅甸语单字符和组合字符字符所在相同感受野内输出y_t等时刻所对应的映射函数概率,通过计算损失值来优化学生模型的权重与参数,从而实现教师网络对学生网络的图像特征增强。$\alpha \in [0, 0.5, 1]$、$\beta \in [0, 0.5, 1]$和$\gamma \in [0, 0.5, 1]$是平衡单个损失项的超参数。从数学上讲,交叉熵损失值可以写成

$$L_{CET}(P_T, \boldsymbol{y}') = \sum_{k=1}^{K} \eta(k = \boldsymbol{y}') \log q_T^* \tag{6}$$

$$L_{CES}(P_S, \boldsymbol{y}') = \eta \log p_S' \tag{7}$$

式中:η为指示函数;q_T^*为教师网络对应单字符或者组合字符的输出概率;p_S'为学生网络与教师网络输出y_t等对应位置时刻单字符或者组合字符SoftMax输出概率;L_{KD}由散度损失值L_{KL}和均方误差损失值L_{MSE}组成,即有

$$L_{KD} = \sum_{i=1}^{n} (L_{KL}(p_{si}', q_{Ti}^*/W) + L_{MSE}(p_{si}', q_{Ti}^*)) \tag{8}$$

式中W是一个温度超参数,它控制教师子网络输出的软化。W值越大,目标类上的概率分布越软。L_{KL}公式为

$$L_{KL}(q_{T1}^*, p_{S1}') = \sum_{k=1}^{K} q_{T1}^* (\log(q_{T1}^*) - \log(p_{S1}')) \tag{9}$$

3 实验过程与分析

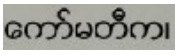
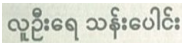
在缅甸语场景文本识别任务上,对所提出的基于知识蒸馏的缅甸语图像文本识别方法的有效性进行了评估。本文在构造的缅甸语图像数据集上进行了实验。

3.1 数据集

实验涉及以下6个可用的缅甸语图像数据集,所采用的实验数据来自网络采集的缅甸语文本数据随机生成分辨率为10像素×5像素~500像素×300像素的图像数据集。选用了80万张含有噪声的缅甸语场景文本图像作为评估数据集和80万张含有噪声的缅甸语场景文本图像作为测试数据集,数据集内的图像为".jpg"格式,对应的数据标签为缅甸语图像内对应的文本信息,如表1所示。神经网络训练前将数据保存为tfrecord格式以提升数据读取速率。训练数据集内包含以下6种缅甸语图像数据集。

表1 数据集格式及对应标签示例

Table 1 Example of data set format and corresponding label

数据集图像	标签文本信息
 1.jpg	ကော်မတီက၊
 2.jpg	လူဦးရေ သန်းပေါင်း
 3.jpg	ကျော်ရှိသော

数据集1 该数据集包含600万张无背景噪声的长序列的训练缅甸语图像数据集,例如“ကျင်းပရာတွင်”“အကြိမ်မြောက်”“ဆင်ဟွာသတင်းအရ”。

数据集2 该数据集图像为与数据集1中每张图像的位置特征信息一一对应的短序列的单字符缅甸语训练数据集。例如:数据集1中“ကျင်းပရာတွင်”第6个字符“ပ”对应的图像为“ ”;“ကျင်းပရာတွင်”中第7个字符“ရ”对应的图像为“ ”,将“ ”“ ”等图像作为数据集2。

数据集3 该数据集图像为与数据集1中每张图像的位置特征信息一一对应的短序列的组合字符缅甸语训练数据集。例如:数据集1中“ကျင်းပရာတွင်”第1个和第2个字符的组合字符“ကျ”对应的图像为“ ”;“အကြိမ်မြောက်”中第8个、第9个与第10个的字符的组合字符“မြာ”对应的图像为“ ”,将“ ”“ ”等图像作为数据集3。

数据集4 该数据集包含600万张具有背景噪声的长序列训练缅甸语图像数据集。例如“သိရပါသည်။”“ကော်မတီက၊”和“ရှေးဟောင်း”。

数据集5 该数据集图像为与数据集4中每张图像的位置特征信息一一对应的单字符缅甸语训练数据集。例如:数据集4中“ကော်မတီက၊”第8个字符“က”对应的图像为“ ”;“ရှေးဟောင်း”第1个字符“ဂ”对应的图像“ ”,将“ ”“ ”等图像作为数据集5。

数据集6 该数据集图像为与数据集4中每张图像的位置特征信息一一对应的短序列组合字符缅甸语训练数据集,例如:数据集4中“ကော်မတီက၊”第2个、第3个与第4个字符的组合字符“ကတ်”对应的图像为“ ”,第6个与第7个字符的组合字符“တီ”对应的图像为“ ”,将“ ”“ ”等图像作为数据集6。

3.2 实验结果及分析

本文的实验基于Tensorflow框架实现,服务器配置配置为Intel(R) Xeon(R) Gold 6132 CPU @ 2.60 GHz,NVIDIA Corporation GP100GL GPU。

实验中严格按照标准评价指标单字符(Per char, PC)和全序列(Full sequence,FS)精确率的公式为

$$PC=\frac{CS}{SN}\times 100\% \tag{10}$$

$$FS=\frac{SL}{LN}\times 100\% \tag{11}$$

式中:PC、CS和SN分别代表每个字符的准确率、正确的字符总数和所有字符的总数;FS、SL和LN分别代表全序列精确率、正确的序列数和序列总数。

在确保其他变量都一致的情况下,对比模型参数均基于原给出的超参数设置。在没有噪音的缅甸语图像情况下进行了实验1与实验2。

实验1 首先选用数据集1作为学生网络的训练数据,数据集3作为教师网络的训练数据进行了实验,对比实验的训练集为数据集1和数据集3的总和,识别结果如表2所示。

表2 训练集为数据集1和3时的识别结果

Table 2 Recognition results with the training set of datasets 1 and 3 %

方法	PC	FS
LSTM + CTC	75.2	70.8
CNN + LSTM + CTC	81.2	77.4
CNN + BLSTM + Attention	88.1	82.3
CNN + BLSTM + CTC	87.2	85.1
本文方法	**91.5**	**88.5**

从表2实验结果可以看出:采用“CNN + BLSTM + CTC”方法的单字符的准确率、全序列精确率分别为87.2%和85.1%,采用“CNN + BLSTM + Attention”方法单字符的准确率、全序列精确率分别为88.1%和82.3%,本文方法在单

字符的准确率、全序列精确率最好效果达到了91.5%和88.5%。实验中将教师网络学习到对齐片段的缅甸语组合字符特征对学生网络进行优化,从而对学生网络具有缅甸语组合字符的位置信息进行了特征增强,使多个字符嵌套组合的复杂字符识别准确率提高。对比实验中虽然在处理识别单字符方面比较擅长,但是在识别缅甸语组合字符时会产生误判或者输出字符顺序错乱等结果,所以导致识别准确率低于本文的值。

实验2 选用数据集1作为学生网络,数据集2数据集3作为教师网络的训练数据进行了实验,对比实验的训练集为数据集1、2、3的总和,识别结果如表3所示。

表3 训练集为数据集1、2和3时的识别结果

Table 3 Recognition results with training set of datasets 1, 2 and 3 %

方法	PC	FS
LSTM + CTC	75.3	72.1
CNN + LSTM + CTC	83.4	80.4
CNN + BLSTM + Attention	88.7	84.4
CNN + BLSTM + CTC	88.9	87.2
本文方法	**94.5**	**90.1**

从表3可见,增加了数据集2后,与表2相比模型识别结果均有所提升,本文方法在单字符的准确率、全序列精确率分别提升了3%和1.6%。因为数据集2包含了位置特征的短序列的单字符缅甸语,实现学生网络中单个感受野对应的局部字符图像特征与教师网络单字符图像特征的对齐,以此增强长序列字符图像中单字符特征的获取,从而提高了模型的准确性。

以上训练数据集是在不含有背景噪音的情况下进行模型训练,在处理实际生活中具有背景噪音的缅甸语图像时识别效果就会较差,为此本文在训练数据使用具有背景图像的情况下进行了实验3,以此来提高模型在应对不同场景下的缅甸语图像识别。

实验3 将数据集4作为学生网络的训练数据,数据集5、6作为教师网络的训练数据,在该情况下选用数据集4+5,数据集4+6和数据集4+5+6分别进行了1组实验。对比实验的训练集为所对应数据集的总和,识别结果如表4所示。

表4 具有背景噪声的情况下每个字符准确率和全序列准确率的实验结果

Table 4 Experimental results of accuracy of per character and accuracy of full sequence with background noise %

方法	数据集4+5		数据集4+6		数据集4+5+6	
	PC	FS	PC	FS	PC	FS
LSTM + CTC	76.4	73.5	76.2	75.8	78.2	76.1
CNN + LSTM + CTC	80.1	77.6	81.8	80.5	82.2	80.8
CNN + BLSTM + Attention	85.9	85.3	87.1	86.7	91.3	87.1
CNN + BLSTM + CTC	86.1	84.6	89.8	88.5	93.4	91.5
本文方法	**88.6**	**85.8**	**93.1**	**91.2**	**95.6**	**94.2**

从表4中可以观察到,在训练集使用具有背景噪声图像比使用无背景噪声图像时识别精度更准确。在该情况下,本文实验在采用数据集4+5+6时,即在同时考虑单字符和组合字符特征以及添加背景噪声因素后,模型达到了最好的效果。

实验训练数据集的大小也有可能影响模型识别图像的准确度,所以通过更改实验数据集的大小来比较测试结果,该数据集大小为学生网络训练集大小,教师网络训练集数量不计入其中,即与学生网络输入图像每张图像所对应的对齐片段特征的缅甸语单字符或者组合字符图像,但是教师网络训练集依然参与教师网络训练。单字符和全序列句子识别准确率结果如图3,4所示。

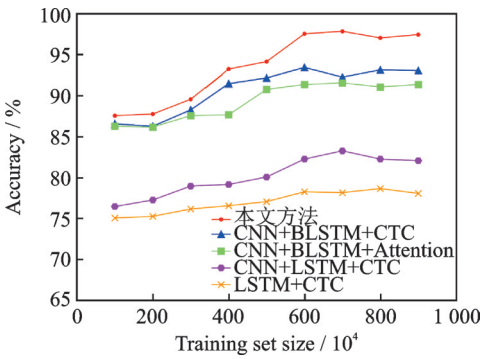


图3 不同数据集大小的单字符准确率

Fig.3 Accuracy of per character for different sizes of datasets

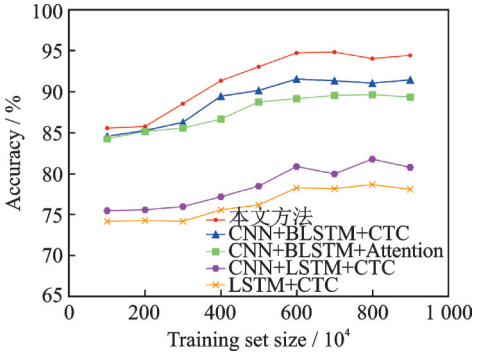


图4 不同数据集大小的全序列句子准确率

Fig.4 Accuracy of full sequence sentences with different sizes of datasets

通过实验结果可以得出结论,使用深度学习方法时训练模型数据集的大小会影响实验效果,并且通过实验比较分析结果可以看出,当训练数据集到600万张图片时,随着训练数据的提升,准确值提升不再明显,所以可以取600万训练数据来训练最优模型。

4 结束语

针对缅甸语图像中1个感受野内多个字符嵌套组合的复杂字符难以提取识别的问题,提出了一种基于知识蒸馏的缅甸语OCR方法,根据缅甸语文字特点,构建了适应缅甸语OCR任务需求的网络框架。首次将基于知识蒸馏的思想运用到缅甸语图像文本识别研究,构建了学生网络和教师网络对长序列中局部特征的增强,实现局部特征对齐,从而解决缅甸语嵌套组合字符识别的问题。本文构建了训练网络模型所需的数据集,并在该数据集的基础上进行了实验,在没有背景噪声图像与具有背景噪声图像作为训练数据的情况下,本文模型的性能分别优于基线2.9%和2.7%。在以后的工作中,本文将融合语言模型以优化结果,从而进一步提高识别的准确性。

参考文献:

[1] ANDERSON R H. Syntax-directed recognition of hand-printed two-dimensional mathematics[C]// Proceedings of Symposium on Interactive Systems for Experimental Applied Mathematics: Symposium. New York, United States: ACM, 1967: 436-459.

[2] BAI J, CHEN Z, FENG B, et al. Image character recognition using deep convolutional neural network learned from different languages[C]//Proceedings of 2014 IEEE International Conference on Image Processing (ICIP). Paris, France: IEEE, 2014: 2560-2564.

[3] YUAN A, BAI G, JIAO L, et al. Offline handwritten English character recognition based on convolutional neural network [C]// Proceedings of 2012 10th IAPR International Workshop on Document Analysis Systems. Gold Cost, QLD, Australia: IEEE, 2012: 125-129.

[4] YANG W, JIN L, XIE Z, et al. Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge[C]// Proceedings of 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Montreal, Quebec, Canada: IEEE, 2015: 551-555.

[5] HE M, ZHANG S, MAO H, et al. Recognition confidence analysis of handwritten Chinese character with CNN[C]// Proceedings of 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Montreal, Quebec, Canada: IEEE, 2015: 61-65.

[6] ZHONG Z, JIN L, XIE Z. High performance offline handwritten chinese character recognition using googlenet and directional feature maps[C]// Proceedings of 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Montreal, Quebec, Canada: IEEE, 2015: 846-850.

[7] BREUEL T M, UL-HASAN A, AL-AZAWI M A, et al. High-performance OCR for printed English and Fraktur using LSTM networks[C]// Proceedings of 2013 12th International Conference on Document Analysis and Recognition. Washington, DC: IEEE, 2013: 683-687.

[8] ELADEL A, EJBALI R, ZAIED M, et al. Dyadic multi-resolution analysis-based deep learning for Arabic handwritten character classification[C]// Proceedings of 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI). Vietri sul Mare, Italy: IEEE, 2015: 807-812.

[9] ANIL R, MANJUSHA K, KUMAR S S, et al. Convolutional neural networks for the recognition of Malayalam characters [C]//Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Odisha, India: Springer, 2015: 493-500.

[10] GHOSH D, DUBE T, SHIVAPRASAD A. Script recognition—A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(12): 2142-2161.

[11] ZHANG Y, WANG W, WANG L, et al. Scene text recognition with deeper convolutional neural networks[C]// Proceedings of 2015 IEEE International Conference on Image Processing (ICIP). Quebec City, Canada: IEEE, 2015: 2384-2388.

[12] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2016, 116(1): 1-20.

[13] WIGINGTON C, TENSMEYER C, DAVIS B, et al. Start, follow, read: End-to-end full-page handwriting recognition[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer Science, 2018: 367-383.

[14] GRAVES A, LIWICKI M, FERNÁNDEZ S, et al. A novel connectionist system for unconstrained handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 855-868.

[15] SU B, LU S. Accurate scene text recognition based on recurrent neural network[C]// Proceedings of Asian Conference on Computer Vision. Singapore: Springer, 2014: 35-48.

[16] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and machine Intelligence, 2016, 39(11): 2298-2304.

[17] LUO C, JIN L, SUN Z. MORAN: A multi-object rectified attention network for scene text recognition[J]. Pattern Recognition, 2019, 90: 109-118.

[18] XIE H, FANG S, ZHA Z J, et al. Convolutional attention networks for scene text recognition[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(1S): 1-17.

[19] HOSKEN M, TUNTUNLWIN M. Representing Myanmar in unicode[J]. Unicode Technical Note, 2012, 13: 1-67.

[20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-04-10) [2020-06-01]. https://arxiv.org/abs/1409.1556.

[21] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. New York, United States: [s.n.], 2006: 369-376.

[22] BURKHARD W A, KELLER R M. Some approaches to best-match file searching[J]. Communications of the ACM, 1973, 16 (4): 230-236.

作者简介:

毛存礼(1977-),男,博士,教授,研究方向:自然语言处理、信息检索、机器翻译,E-mail:maocunli@163.com。

谢旭阳(1995-),男,硕士研究生,研究方向:自然语言处理、图像处理和文本识别。

余正涛(1970-),**通信作者**,男,博士,教授,研究方向:自然语言处理、信息检索、机器翻译,E-mail:ztyu@hotmail.com。

高盛祥(1977-),女,博士,副教授,研究方向:自然语言处理。

王振晗(1993-),男,博士研究生,研究方向:自然语言处理、机器翻译。

刘福浩(1997-),男,硕士研究生,研究方向:自然语言处理、图像处理和文本识别。

(编辑:刘彦东)