融合多层语义特征图的缅甸语图像文本识别方法

刘福浩[1,2]，毛存礼[*1,2]，余正涛[1,2]，高盛祥[1,2]，王琳钦[1,2]，谢旭阳[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
1519195149@qq.com, maocunli@163.com, ztyu@hotmail.com
gaoshengxiang.yn@foxmail.com, 2424172505@qq.com, 2212166977@qq.com

摘要

由于缅甸语存在特殊的字符组合结构，在图像文本识别研究方面存在较大的困难，直接利用现有的图像文本识别方法识别缅甸语图片存在字符缺失和复杂背景下识别效果不佳的问题。因此，本文提出一种融合多层语义特征图的缅甸语图像文本识别方法，利用深度卷积网络获得多层图像特征并对其融合获取多层语义信息，缓解缅甸语图像中由于字符嵌套导致特征丢失的问题。另外，在训练阶段采用MIX UP的策略进行网络参数优化，提高模型的泛化能力，降低模型在测试阶段对训练样本产生的依赖。实验结果表明，提出方法相比基线模型准确率提升了2.2%。

关键词： 缅甸语；图像文本识别；语义信息；特征图融合；MIX UP

Burmese Image Text Recognition Method Fused with Multi-layer Semantic Feature Maps

Fuhao Liu[1,2], Cunli Mao[*1,2], Zhengtao Yu[1,2], Shengxiang Gao[1,2], Linqin Wang[1,2], Xuyang Xie[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
1519195149@qq.com, maocunli@163.com, ztyu@hotmail.com
gaoshengxiang.yn@foxmail.com, 2424172505@qq.com, 2212166977@qq.com

Abstract

Due to the special structure of character combinations in Burmese language, there are great difficulties in the research of image text recognition. Using directly the existing methods to recognize Burmese images has the problems of missing characters and poor recognition under complex background. Therefore, this paper proposes a Burmese image text recognition method based on fused with multi-layer semantic feature maps. It uses deep convolutional networks to obtain multi-layer image features and fuses them to obtain multi-layer semantic information, which alleviates the problem of feature loss in the case of Burmese combining characters . In addition, during the training phase, the MIX UP strategy is used to optimize the network parameters,which improves the generalization ability of the model and reduces the model's dependence on training samples in the testing phase. The experimental results show that the accuracy of the proposed method is improved by 2.2% compared with the baseline model.

Keywords: Burmese , Image text recognition , Semantic information , Feature map fusion , MIX UP

*毛存礼(通信作者):maocunli@163.com
国家自然科学基金重点项目(61732005)；国家自然科学基金(61866019，61761026，61972186)；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)

1 引言

互联网中存在大量含有缅甸语文本的图像，由于缅甸语拼写特殊性，人工方式进行图像文本转换存在较大困难。因此，研究利用图像文本识别技术（Optical Character Recognition，OCR）实现缅甸语文本图片自动识别具有重要的价值。

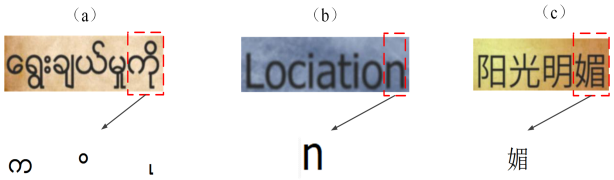


Figure 1: 样例分析图

缅甸语字符组合与英文、中文不同，缅甸语由基础字符、基础前字符、基础后字符、基础上字符和基础下字符构成，所以缅甸语在图像中的组合字符是由多个单字符组合而成，如图1所示，缅甸语图像中的组合字符“ကို”，实际上由基础字符“က”、基础上字符“ ◌ိ ”以及基础下字符“ ◌ု ”组成，但是在图1（b,c）中，感受野内的中文和英语都是由单个字符构成的，没有明显的边缘特征，现有针对中文、英文的图像文本识别模型 (Cheng et al., 2017; Wang et al., 2020; Xie et al., 2019; Shi et al., 2018)所利用特征序列信息主要来源于深度卷积神经网络中最后一层网络输出的特征图，取得了较好的效果，但是，这样的方式会造成部分语义信息丢失，尤其是针对缅甸语这种字符嵌套组合的语言，语义信息丢失更加明显，例如，一个感受野中的缅甸语经过卷积神经网络时，◌ု 、◌် 、◌း 等类型的微小特征在高层语义特征图存在丢失问题，所以，主流的图像文本识别方法直接应用于缅甸语上并不适用。

目前，图像文本识别方法在中英文等大规模训练集上展现出非常强大的性能 (Cheng et al., 2017; Wang et al., 2020; Xie et al., 2019; Shi et al., 2018; Yan and Huang, 2019)，但这依赖于大规模、高质量的训练数据，由于目前没有公开的高质量、大规模缅甸语文字识别数据集，通过合成方式构建的缅甸语图像数据集图像背景复杂、图片质量低下，导致特征缺失的问题更为严重，利用常规的网络训练策略得到的模型往往会使模型泛化能力较弱。

针对以上问题，提出一种融合多层语义特征图的缅甸语图像文本识别方法，本文方法是受到目标检测中特征金字塔网络(Fully Convolutional Networks，FCN) (Long et al., 2015)的启发，在基于编码-解码框架的基础上将深度卷积神经多层特征图进行上采样融合，该网络可以在特征提取阶段获取更强的特征表征能力，我们同时使用基于MIX UP (Zhang et al., 2017)的训练策略来提高缅甸语图像文本识别模型在不同背景下的识别能力。

本文工作主要有以下贡献：

（1）针对缅甸语图像特征提取方面，我们构建了基于多层语义特征图融合的神经网络，利用深度卷积网络各层语义特征图进行融合，来处理缅甸语图像文本识别过程中上下标字符特征丢失问题；

（2）在缅甸语图像识别网络训练方面，我们首次利用了MIX UP的数据增强策略进行网络训练，将缅甸语图像训练样本进行融合，用于提高缅甸语图像文本识别模型在不同背景下的泛化能力。

2 相关工作

现有图像文本识别方法大致分为基于字符分割的图像文本识别方法和基于序列到序列的图像文本识别方法，具体如下：

（1）**基于字符分割的图像文本识别方法**；首先进行字符级检测或分割，将其送入字符分类器，利用分类器实现英文或中文的识别 (Wang et al., 2011; Yao et al., 2014; Alsharif and Pineau, 2013)。例如，王凯等人 (Wang et al., 2011)利用基于滑动窗口分类的方式进行英文字符检测，再利用支持向量机进行分类以实现字符识别；白翔等人 (Yao et al., 2014)利用基于霍夫投票的检测算法得到每个字符的霍夫特征，结合笔画特征送入随机森林分类器进行分类以实

现识别；但是上述方法识别对象仅仅局限于单个字符，它们的性能很大程度上受字符分割好坏的影响。而缅甸语作为一种存在字符组合的语言，利用传统的字符分割方法无法准确分割出每一个缅甸语字符，而且独立识别每个字符丢弃单词的上下文信息，大大降低了模型的可靠性和鲁棒性。

（2）**基于序列到序列的图像文本识别方法**；为解决基于字符分割的图像文本识别方法忽略字符之间的上下文语义信息问题，研究者在卷积神经网络和循环神经网络的基础上设计了识别模型，用于捕捉字符之间的上下文语义信息以提高识别效果；例如，Jaderberg等人 (Jaderberg et al., 2014)利用卷积神经网络创新性地将单词识别任务转换为整个单词的分类任务，但是模型性能依赖于定义的单词词表，而且整个单词的分类任务不能应用于缅甸语识别任务上；为解决单词词表依赖问题，各种序列到序列的模型逐渐涌现出来，这类模型通常由编码模块和解码模块组成。编码模块通常基于卷积神经网络或循环神经网络的编码技术将输入图像编码成具有文字表征意义的特征向量；解码模块一般利用连接时序分类算法(Connectionist Temporal Classification，CTC) (Shi et al., 2016a; Wang and Hu, 2017; He et al., 2016; Shi et al., 2016b)、注意力机制 (Qiao et al., 2020; Yu et al., 2020; Bai et al., 2018; Lyu et al., 2019)、聚合交叉熵算法 (Xie et al., 2019)将编码特征向量解码成字符串。例如：白翔等人 (Shi et al., 2016a)创新性的将CTC算法应用于英语自然场景识别网络的解码模块，华南理工大学、复旦大学以及海康威视 (Cheng et al., 2017; Wang et al., 2020)受到注意力机制在机器翻译任务上成功应用的启发，将注意力机制应用于解码，提高了图像文本识别精确度。随着深度学习的图像文本识别算法的快速发展，一些特殊性问题也逐渐被更多研究人员所关注，例如，弯曲字符识别、低分辨率识别以及词汇依赖等问题。白翔等人 (Shi et al., 2018)提出了基于复杂变换的注意力场景文本识别模型(Attentional Scene Text Recognizer with Flexible Rectification，ASTER)，在基础的图像文本识别模型上设计了基于空间变换的矫正网络(Spatial Transformer Networks，STN)，提高了对不规则文字识别的精度。Yan R等人 (Yan and Huang, 2019)提出面向低质文本的超分辨率学习网络，用于提高在模糊、低质量图片下的识别准确率；Qiao Z等人 (Qiao et al., 2020)在ASTER模型的基础上提出了一个语义监督模块，增强语义信息，以改善部分遮挡、图像模糊下英文识别模型性能；Yu D等人 (Qiao et al., 2020)提出了全局语义推理模块(Global Semantic Reasoning Module，GSRM)捕捉全局语义信息，融合视觉特征以提高识别的精确度。

由于缅甸语字符组合的特点，上述针对中英文的图像文本识别方法并不适用，无法精确的提取到缅甸语特征导致缅甸语图像识别准确率大大降低，本文设计了一个特征增强网络，融合多层语义特征图，以提高模型对缅甸语中上下标字符特征的捕捉能力。中英文识别算法效果显著主要得益于高质量、大规模的图像数据集，合成得到的缅甸语图像数据集缺乏一定的多样性，导致缅甸语在复杂背景下识别效果极差，为此本文首次采用数据增强策略MIX UP来提高模型的泛化能力。

3 融合多层语义特征图的缅甸语图像文本识别方法

本文提出的方法包括多层语义特征图融合和MIX UP数据增强策略。

3.1 多层语义特征图融合网络

融合多层语义特征图的模型架构如图2所示，主要由特征提取网络、特征增强网络、识别网络三部分组成。下面各节将详细介绍特征提取网络、特征增强网络、识别网络。

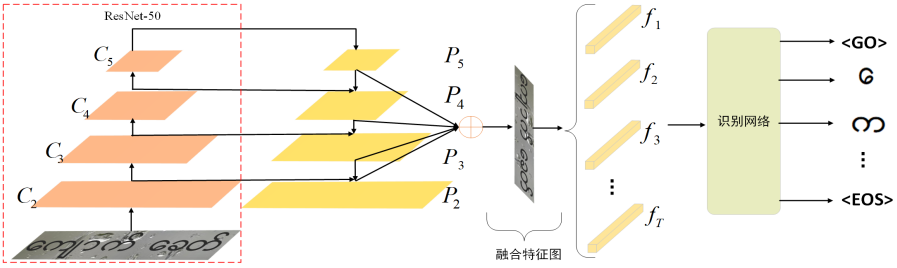


Figure 2: 多层语义特征图融合网络

3.1.1 缅甸语图像特征提取网络

我们在50层残差网络(Residual Network，ResNet-50)的基础上构建了适应缅甸语图像特征提取的主干网络。为了将深度卷积神经网络ResNet-50提取到的特征图使用于缅甸语图像文本识别中，本文网络架构中去除了ResNet-50中的全连接层；为了保证特征图更好的体现缅甸语文字图像特征，考虑在缅甸语图像中缅甸语字符的长宽比为2，所以我们在ResNet-50的第三、四阶段采用池化窗口为1*2的最大池化层得到长宽比为2的特征图，保证送入识别模块特征序列具有完整的缅甸语特征。

3.1.2 缅甸语图像特征增强

残差网络ResNet-50利用了残差块保证增加网络深度的同时提升训练效率，使用ResNet-50对图像进行特征提取时，一般经过五个阶段的残差融合，ResNet-50根据这五个阶段将50层的残差网络划分为5部分，本文根据网络位置的深度依次命名为Layer1、Layer2、Layer3、Layer4以及Layer5，缅甸语图像经过ResNet-50时，这五个阶段能够得到不同尺度的语义特征图，底层语义特征图具备的特征不够丰富，而高层语义特征图通常会忽略一些微小的特征信息。本文设计特征增强网络对五个阶段的语义特征图进行融合，本文将这五个阶段的语义特征图数学表示为$\{C_1, C_2, C_3, C_4, C_5\}$，特征增强模块中特征图融合对象分别为主干网络ResNet-50中得到的四层特征图$\{C_2, C_3, C_4, C_5\}$。为避免计算机内存溢出，融合之前我们首先减少C_2, C_3, C_4, C_5特征图的通道数以得到新的特征图$\{M_2, M_3, M_4, M_5\}$，如公式(1)所示。

$$M_i = Reduce(C_i) \tag{1}$$

减少特征图的策略是利用通道数为128、大小为1*1卷积核以1的步长进行计算，最终得到的特征图$\{M_2, M_3, M_4, M_5\}$的通道数都为128。由于特征图$\{M_2, M_3, M_4, M_5\}$大小不一，不能直接进行融合操作，本文采用的策略是通过上采样方法将$\{M_2, M_3, M_4, M_5\}$依次采样至统一大小，再进行特征图叠加以实现特征增强。具体步骤如下：首先利用基于双线性插值的上采样方法对高层语义特征图P_5进行上采样，将其放大至上一个阶段语义特征图的大小，利用采样之后的特征图与上一个阶段的特征图进行相加操作以实现特征增强，后续采用同样的思想得到P_3, P_2，融合P_2, P_3, P_4, P_5得到最终的增强特征图H_5。

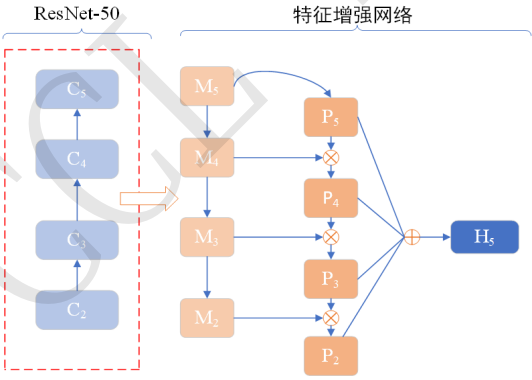


Figure 3: 特征增强网络图

3.1.3 缅甸语图像识别网络

为了将增强之后的特征图识别为缅甸语文字，我们首先利用卷积核大小为1的卷积对特征图H_5进行卷积计算，以解决展开特征图H_5得到的特征序列维度过高问题，最终得到8*25*128大小的结果。文字在图像中通常以一个文本行的形式展现，跟自然语言处理任务中的一句话类似，具有丰富的上下文语义信息，为了保证将对特征图进行解码时不丢失这类信息，本方法将第一维和第三维合并1024长度的特征向量，一共有25个，数学表示为$f_1, f_2, \cdots, f_T$，其中T最大为25。这25个特征向量用于后续的双向LSTM输入。

$$Enconder(M) = (f_1, f_2, \cdots, f_T) \tag{2}$$

其中，M表示输入图像，Enconder表示以残差网络ResNet-50为特征提取网络的编码网络。

BiLSTM可以有效获取上下文信息，采用BiLSTM从25个特征序列中获取缅甸图像中上下文信息。为解决训练时缅甸语训练标签的对齐问题，识别模块的解码策略采取基于注意力机制的解码方法。基于注意力机制的解码器利用t时刻BiLSTM编码向量$(h_1, h_2, \cdots, h_T)$的加权和g_t实现第t个序列的字符识别。由于识别图像中文本长度不一，缅甸语识别网络的解码器需要生成长度可变的序列，所以本文在对真实标签的编码中加入[GO]、[EOS]两个标签，其中[GO]表示文字序列的开始，[EOS]表示文字序列的结束。基于注意力机制的解码器具体识别过程如下，特征序列$(f_1, f_2, \cdots, f_T)$通过BiLSTM编码成具有上下文语义信息的向量$(h_1, h_2, \cdots, h_T)$，利用该向量和BiLSTM上一时刻隐层输出s_{t-1}生成注意力权重分布$[\alpha_{t,1}, \alpha_{t,2}, \cdots, \alpha_{t,T}]$；再利用注意力权重对每个BiLSTM编码之后的向量进行加权求和，得到加权和g_t；最后将g_t送入基于Softmax的分类器实现字符识别。公式表示如下。

$$g_t = \sum_{j=1}^{T} \alpha_{t,j} h_j \tag{3}$$

$$s_t = BiLSTM(f_t, s_{t-1}) \tag{4}$$

$$y_t = Softmax(g_t) \tag{5}$$

其中，y_t表示解码器t时刻的预测，s_t表示为BiLSTM网络中t时刻的隐层输出，α_t表示为注意力权重向量，用于对特征序列$(f_1, f_2, \cdots, f_T)$进行注意力分布计算，并对最终结果进行归一化计算，注意力分布计算公式如下。

$$e_{t,j} = v^T \tanh(W s_{t-1} + V h_j + b) \tag{6}$$

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{j=1}^{T} \exp(e_{t,j})} \tag{7}$$

在注意力分布计算公式中，v, W, V, b属于网络参数。

缅甸语图像文本识别的本质是对特征序列进行多分类，为保证网络训练时识别网络预测分布接近真实标签分布，所以本文采取交叉熵损失作为目标优化函数，注意力机制损失函数如下。

$$Loss_{Attention} = -\sum_{t} \ln P(\hat{y}_t | M, \theta) \tag{8}$$

其中，M表示为输入的缅甸语图像，θ表示为当前识别网络的模型参数，$\hat{y}_t$表示为缅甸语图像的第t个特征序列对应的真实标签。

3.2 网络训练

数据增强方法MIX UP在图像分类中已经取得了非常好的效果，受其启发本文将MIX UP应用于缅甸语自然场景识别的神经网络训练中，具体训练策略如图4所示。

假设一个批次的缅甸语图像样本为$[image_1, image_2, \cdots, image_n]$，$image_A$、$image_B$为该批次中的两个缅甸语图像样本，其对应的标签分别为$LabelA$,$LabelB$；首先根据超参数α,β的贝塔分布(Beta Distribution，Beta)计算得到融合系数λ，融合系数λ控制着两张缅甸语样本图片的融合程度，以融合系数λ为权重值对两个样本进行加权求和，得到混合结果mixed_result。混合结果mixed_result分别和原样本对应的标签进行交叉熵损失计算，将两损失加权和设为目标优化函数，利用反向传播算法求解目标函数的最优解从而实现网络参数更新，MIX UP策略公式如下。

$$\lambda = Beta(\alpha, \beta) \tag{9}$$

$$mixed\_result = \lambda * image_A + (1 - \lambda) * image_B \tag{10}$$

$$Loss_{all} = \lambda Loss_A(mixed\_result, Label\ \mathrm{A}) + (1 - \lambda) Loss_B(mixed\_result, Label\ \mathrm{B}) \tag{11}$$

其中,α, β为控制贝塔分布的超参数，λ为贝塔分布(Beta Distribution，Beta)计算所得的融合系数，$mixed_result$为融合之后的图像，$Loss$表示交叉熵损失的计算。

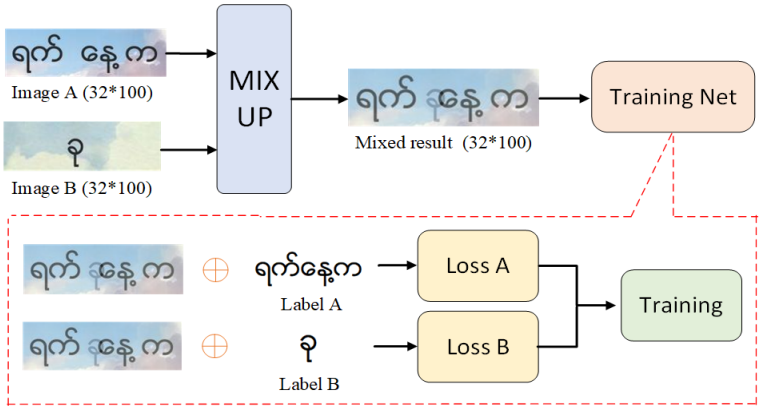


Figure 4: 网络训练策略图

4 实验与分析

为验证融合多层语义特征图的缅甸语图像文本识别方法的有效性，我们在缅甸语图像数据集上进行实验分析。

4.1 数据集及实验设置

本文方法所用缅甸语数据集总共包含了500万张缅甸语图像，数据集是由合成和人工标注的方法构建的，人工标注数据为3万条，剩余数据是由合成方法得到。其中，分别随机选取50万缅甸语图像作为评估数据集和测试数据集。为提升模型训练速度，数据预处理阶段采用“.mdb”文件存储方式来存储训练集、测试集、评估集以此提高模型读取速率，具体规模及样例如表1所示。

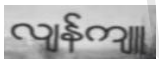
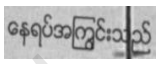
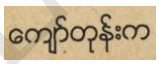
数据集	数量	样例	标签
训练集	400万		လျှောက်ကြေး
评估集	50万		နေရပ်အကြွင်းသည်
测试集	50万		ကျော်တုန်းက

Table 1: 数据集格式及对应标签示例

本文的神经网络架构是基于Pytorch框架开发设计实现，实验服务器的配置为Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz，NVIDIA Corporation GP100GL GPU。

实验采用缅甸语序列率精确率（Sequence Accuracy, SA）作为评价指标，如公式(12)所示：

$$SA = \frac{SL}{LN} \times 100\% \tag{12}$$

其中，SA、SL、LN分别代表缅甸语文本图像识别的序列精确率、正确的序列总数、序列的总数。

4.2 主要实验结果

为保证对比实验的公平性，本文将所有的缅甸语识别模型放置在同一实验条件下进行实验，实验所选优化器为Adam，初始学习率为1，训练时采用CosineAnnealing策略，基于余弦函数实现学习率动态变换，以保证网络的目标函数接近最优解时具备更小的学习率；模型训练的批处理大小设置为100，训练步长设为400000，训练epoch为10，实验结果选择评测中最高的准确率。

实验一：主要实验结果

本文选取CNN+ BiLSTM+Attention的方法作为基线模型，并与LSTM模型、卷积神经网络模型、基于LSTM的CRNN模型以及基于BiLSTM的CRNN模型进行比较，实验结果如表2所示。

方法	SA
LSTM+CTC	79.8
CNN+CTC	81.7
CNN+LSTM+CTC	84.5
CNN+ BiLSTM+Attention	92.2
CNN+BiLSTM+CTC	92.5
本文方法	**94.4**

Table 2: 基线模型和本文的方法在数据集上的结果

与基线模型（CNN+ BiLSTM+Attention）相比，本文的方法在识别缅甸语的过程中更好地提取到缅甸语上下标特征信息，以及识别不同背景下地缅甸语图像具有更好的泛化能力。在缅甸语数据集上准确率相比基线模型提升了2.2%。

与使用CTC解码器的图像识别模型(CNN+BiLSTM+CTC、CNN+LSTM+CTC、CNN+CTC、LSTM+CTC)相比，本文针对缅甸语的图像识别方法也展现出来明显的优势。尽管识别模型都是在最高层语义特征图的基础上进行文字提取，但不同的是，本文所融合得到的多层语义特征图不仅包含了高层语义信息，还融合了具有上下标特征信息的低层语义信息。与其相比，本文方法识别效果得到了明显的提升。

为保证验证实验的真实性以及有效性，本文用人工标注的方式额外标注了1000张真实场景图像，并将其作为测试集。本文在这1000张真实场景测试集上进行测试实验，实验结果如表3所示。

方法	SA
LSTM+CTC	78.9
CNN+CTC	80.5
CNN+LSTM+CTC	83.9
CNN+ BiLSTM+Attention	91.1
CNN+BiLSTM+CTC	91.6
本文方法	**92.9**

Table 3: 基线模型和本文的方法在真实场景测试集上的结果

本文的方法在对1000张真实场景测试集图像的识别中仍然保持着最优的效果，同比基线模型的准确率能够提升1.8个百分点，融合特征图的方式能够帮助后续的缅甸语识别解码器获取更多的特征，利用丰富的缅甸语图像特征，解码器能够很大程度上提升准确率；MIX UP数据增强策略能够在大量的合成数据集上起到数据扩充，保证识别模型面对真实场景图像时具有强大的鲁棒性。

实验二：消融实验结果对比

为验证缅甸语多层语义特征图融合策略和MIX UP网络增强策略各自的有效性，我们分别对其做了消融试验。我们分别对以VGG-16为主干网络和以ResNet-50为主干网络的基线模型进行消融实验，实验结果如表4所示，其中Mix Mut表示是否使用MIX UP数据增强策略，Feature Mut表示是否使用多层语义特征图融合。从实验结果可以看出，以VGG-16为主干网络的缅甸语图像识别模型在仅使用多层语义特征图融合策略时，识别准确率可以提高0.7百分点；在仅使用MIX UP数据增强策略时，识别准确率可以提高0.9个百分点。以ResNet-50为主干网络的缅甸语图像识别模型在上述两种情况下准确率分别可以提高0.9，0.4个百分点。使用不同的主干网络识别模型也展现出了性能的差异，在不使MIX UP数据增强策略以及特征图融合策略

时，利用残差网络ResNet-50进行缅甸语特征提取之后的识别准确率达到92.7%，与VGG-16作为特征提取网络的识别模型准确率高0.5个百分点，说明残差网络ResNet-50的特征提取能力优于VGG-16，为此本文方法的主干网络基于残差网络进行设计。

方法	MIX Mut	Feature Mut	SA
VGG-16+ BiLSTM + Attention	×	×	92.2
VGG-16+ BiLSTM + Attention	×	✓	92.9
VGG-16+ BiLSTM + Attention	✓	×	93.1
VGG-16+ BiLSTM + Attention	✓	✓	93.4
ResNet-50+ BiLSTM + Attention	×	×	92.7
ResNet-50+ BiLSTM + Attention	×	✓	93.6
ResNet-50+ BiLSTM + Attention	✓	×	93.1
ResNet-50+ BiLSTM + Attention	✓	✓	94.4

Table 4: 语义特征图融合和MIX UP对识别的影响

为了进一步验证消融实验的可靠性，我们对以ResNet-50为主干网络的缅甸语图像识别模型每次Epoch的准确率进行了统计分析，识别准确率结果如图5所示。

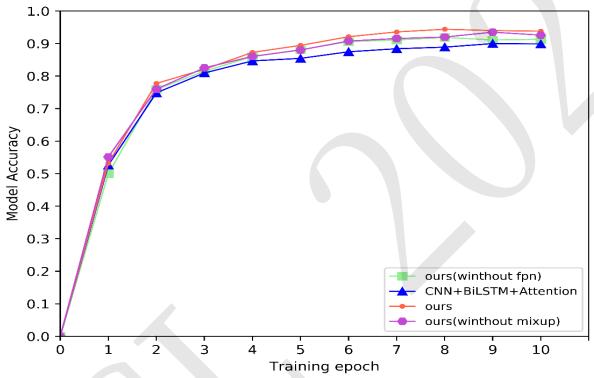


Figure 5: 不同Epoch下识别准确率

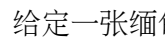
通过实验结果可以得出结论，在模型训练七个轮次之后，模型准确率区间趋于稳定，所提方法具有最优效果；由此可以得知，多层语义特征图融合策略具有更好的缅甸语图像特征提取能力，MIX UP数据增强策略可以提高模型的泛化能力。

4.3 样例分析

表5给出了缅甸语图像识别的实例。

样例	CNN+BiLSTM+Attention	ours
	အကြာကို	အကြံ့ကို
	မြုပန္နကော်မ	မြုပ်နှုမ္ဘကော်မ
	ဆိုရှယ်လစ်ဝါဒ	ဆိုရှယ်လစ်ဝါဒ
	မှပုစကိုပြည့်ဝစေပြီး	မှပုံစံကိုပြည့်ဝစေပြီး

Table 5: 缅甸语图像文本识别样例分析

给定一张缅甸语图像“  ”，正确的识别结果应该是“”，CNN+BiLSTM+Attn.模型直接利用最高层语义特征图进行缅甸语文字识别，错误地将其识别为“”，其中“”、“”以及“”等字符存在丢失，面对低质图像这类问题更为明显。而本文方法在融合特征图的基础上利用MIX UP数据增强策略训练之后，识别低质或者组合字符数量多的缅甸语图像时有着更好的性能，能够保证低质图像下的识别准确率,同时缓解字符丢失问题。

5 总结

针对字符组合导致上下标特征丢失问题，本文提出了一种基于多层语义特征融合的缅甸语图像文本识别方法，将卷积神经网络提取的具有缅甸语特征信息特征图进行融合操作，实现缅甸语特征提取能力的增强，缓解了缅甸语图像识别过程中上下标字符缺失问题；同时为缓解模型在复杂背景下缅甸语图像的识别不佳问题，首次将MIX UP数据增强策略用于缅甸语图像识别网络训练上，从而提升缅甸语复杂背景下模型的识别能力及鲁棒性。为验证方法的有效性，我们在缅甸语数据集上进行了实验，在同样训练条件下，所提模型相比基线模型准确率提升了2.2%。在下一步研究中，我们将考虑在特征图的基础上进行MIX UP数据增强策略来提升识别性能。

参考文献

Ouais Alsharif and Joelle Pineau. 2013. End-to-end text recognition with hybrid hmm maxout models. *arXiv preprint arXiv:1310.1811*.

Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. 2018. Edit probability for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1508–1516.

Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084.

Pan He, Weilin Huang, Yu Qiao, Chen Loy, and Xiaoou Tang. 2016. Reading scene text in deep convolutional sequences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep features for text spotting. In *European conference on computer vision*, pages 512–528. Springer.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Pengyuan Lyu, Zhicheng Yang, Xinhang Leng, Xiaojun Wu, Ruiyu Li, and Xiaoyong Shen. 2019. 2d attentional irregular scene text recognizer. *arXiv preprint arXiv:1906.05708*.

Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. 2020. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13528–13537.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016a. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016b. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048.

Jianfeng Wang and Xiaolin Hu. 2017. Gated recurrent convolution neural network for ocr. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 334–343.

Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE.

Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. 2020. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12216–12224.

Zecheng Xie, Yaoxiong Huang, Yuanzhi Zhu, Lianwen Jin, Yuliang Liu, and Lele Xie. 2019. Aggregation cross-entropy for sequence recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6538–6547.

Rui Yan and Yaohong Huang. 2019. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit.

Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. 2014. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4042–4049.

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.