融合双重注意力机制的缅甸语图像文本识别方法

王奉孝[1,2]，毛存礼[*1,2]，余正涛[1,2]，高盛祥[1,2]，黄于欣[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

1499539796@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, huangyuxin2004@163.com, 1519195149@qq.com

摘要

由于缅甸语字符具有独特的语言编码结构以及字符组合规则，现有图像文本识别方法在缅甸语图像识别任务中无法充分关注文字边缘的特征，会导致缅甸语字符上下标丢失的问题。因此，本文基于Transformer框架的图像文本识别方法做出改进，提出一种融合通道和空间注意力机制的视觉关注模块，旨在捕获像素级成对关系和通道依赖关系，降低缅甸语图像中噪声干扰从而获得语义更完整的特征图。此外，在解码过程中，将基于多头注意力的解码单元组合为解码器，用于将特征序列转化为缅甸语文字。实验结果表明，该方法在自构的缅甸语图像文本识别数据集上相比Transformer识别准确率提高0.5%，达到95.3%。

关键词： 缅甸语；文本识别；通道和空间注意力；特征增强；文字边缘特征

Burmese image text recognition method with dual attention mechanism

Fengxiao Wang[1,2], Cunli Mao[*1,2], Zhengtao Yu[1,2], Shengxiang Gao[1,2], Yuxin Huang[1,2], Fuhao Liu[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

1499539796@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, huangyuxin2004@163.com, 1519195149@qq.com

Abstract

Due to the unique language coding structure and character combination rules of Burmese characters, the existing image text recognition methods cannot fully pay attention to the features of text edges in the Burmese image recognition task, which will lead to the loss of subscripts and subscripts of Burmese characters. Therefore, this paper improves the image text recognition method based on the Transformer framework, and proposes a visual attention module that fuses channel and spatial attention mechanisms, aiming to capture pixel-level pairwise relationships and channel dependencies and reduce noise interference in Burmese images. Thereby a more semantically complete feature map is obtained. Furthermore, in the decoding process, multi-head attention-based decoding units are combined into a decoder for converting feature sequences into Burmese scripts. The experimental results show that the recognition accuracy of this method is 0.5% higher than that of Transformer on the self-constructed Burmese image text recognition dataset, reaching 95.3%.

国家自然科学基金重点项目（61732005,U21B2027）；国家自然科学基金（62166023，61866019）；云南省自然科学基金重点项目(2019FA023);云南省重大科技专项计划项目(202103AA080015，202002AD080001)

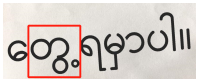
第二十一届中国计算语言学大会论文集，第355页-第365页，南昌，中国，2022年10月14日至16日。

Keywords: Burmese , Text recognition , Channels and Spatial Attention , Feature enhancement , Text edge features

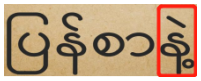
1 引言

由于缅甸语属于一种典型的低资源语言，互联网中存在大量的缅甸语文本图像，因此，快速精准地提取缅甸语文本图像中的文本信息对于开展面向缅甸语的自然语言处理、机器翻译、信息检索等研究具有重要的意义。

(a) 含上标图像



(b) 含下标图像



(c) 含上下标图像

Figure 1: 缅甸语字符边缘特征图像示例

现有方法在针对中英文的图像识别任务上已经取得很好的效果，但缅甸语字符的语言编码结构以及字符组合规则与中英文具有巨大的差异性，其字符主要由基础字符、基础前字符、基础后字符、基础上字符以及基础下字符构成，缅甸语中存在大量的由多个字符组成一个音节的情况，如Figure1中(c)的"န့်"是由"န"、"◌့"以及"◌်"等三个字符组成，这样的字符组成结构，在图像识别过程中会由于其上下标字符边缘特征不明显导致部分语义信息丢失，例如在识别Figure1（c）中"န့်"容易丢失"◌့"或"◌်"等上下标信息，从而极大地影响了缅甸语图像文本识别的准确率。

此外，研究人员针对缅甸语文本图像识别任务也尝试了很多有意义的工作，毛存礼et al. (2022)提出了利用知识蒸馏的方式将单字符特征的相关知识传递给学生模型，用于提升学生模型的特征提取能力，从而缓解缅甸语字符丢失及识别错误的问题，但是该方法忽略了深度卷积神经网络中的底层语义特征。Liu et al. (2021)提出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，实现主干网络对缅甸语特征提取能力的增强。然而，其特征提取网络对缅甸语文本边缘特征的提取并不充分，并且在效率上仍然有待提高。目前，随着深度神经网络的发展，Transformer在各大图像任务中展现出优异的性能(Dosovitskiy et al., 2020; Ali et al., 2021; Fan et al., 2021)，基于Transformer架构能够捕捉到整张图像的全局信息以及序列之间的依赖关系，这样的解码方式虽然有利于提升缅甸语文字的识别性能，但是针对缅甸语的字符上下标等边缘特征的识别仍然存在较大的挑战。

为了解决上述问题，本文主要针对缅甸语图像文本识别过程中字符上下标等边缘特征容易丢失的问题展开研究，受到卷积块注意模块(Woo et al., 2018)思想的启发，提出一种融合通道和空间注意力机制的缅甸语图像文本识别方法，我们考虑对经过图像特征提取网络得到的特征图同时构建空间注意力和通道注意力来获取缅甸语图像更细粒度的位置特征和通道映射特征，并将获取的两个特征进行融合，最后利用多头注意力机制对融合结果进行注意力计算，捕捉文本之间的全局信息。

本文的工作主要有以下贡献:

（1）我们提出一种融合双重注意力机制的缅甸语图像特征提取方法，使得模型可以更多地关注到缅甸语文本图像的上下标区域;

（2）我们基于Transformer模型设计一个适用于缅甸语图像文本识别的框架，使模型可以进行并行训练，极大提升了识别效率。

（3）在自构的缅甸语文本图像数据集上，实验结果表明所提方法的缅甸语识别准确率达到了95.3%，优于多个对比模型。

2 相关工作

现有图像文本识别方法大致分为联结主义时间分类的图像文本识别方法、基于序列到序列

的图像文本识别方法以及基于Transformer的图像文本识别方法，具体如下：

（1）基于联结主义时间分类的方法

基于（Connectionist Temporal Classification，CTC）的文本识别方法引入CTC 损失作为目标优化函数。该算法的本质是先定义预测结果到真实标签之间的转化方式，采用动态规划的策略从输出概率分布中获取多条状态转移路径，将所有路径概率之和的最大值作为目标优化函数。因此，CTC方法使其只需要文字级注释就可以进行端到端训练，而不需要字符级注释。Graves et al. (2007)提出首次将CTC应用在OCR领域的手写识别系统。随着深度神经网络的发展，(Shi et al., 2016)利用不受卷积神经网络(convolutional neural networks，CNN)输入空间大小限制的特性，提出了一个将卷积神经网络与循环神经网络(recurrent neural networks，RNN)一起识别场景文本图像的模型。采用全卷积方法对输入图像进行整体编码生成特征切片，引入了长短时记忆(Long Short-Term Memory，简称LSTM)用来增强上下文建模，最终将输出的特征序列输入到CTC模块，直接解码序列结果。(Gao et al., 2017)代替RNN采用堆叠的卷积层来有效地捕获输入序列的上下文依存关系，其主要优点在于较低的计算复杂度和较容易的并行计算。Yin et al. (2017)也避免在模型中使用RNN，他们通过使用字符模型滑动文本行图像来同时检测和识别字符，这些字符模型是在标记有文本记录的文本行图像上端对端学习的。

（2）基于序列到序列的图像文本识别方法

基于注意力机制的文本识别方法首先通过编码器将图像特征转化为中间语义特征，再利用基于注意力模型的解码器将中间语义特征转化为文本。这类方法通过训练可以学习到任意长度的序列之间的对齐关系，一定程度上缓解了序列对齐问题。受注意力机制在机器翻译任务中的成果应用，Lee and Osindero (2016)将注意力模型与循环神经网络进行融合，以提升文字预测效果。为了解决注意机制应用到文字识别中的注意力偏移问题，Cheng et al. (2017)设计聚焦注意力网络来解决该问题。为了让模型更加关注于文字识别相关的图像区域，Li et al. (2019)将1D attention拓展到2D attention上，用2D attention可以更精准的选取字符区域特征，忽略掉背景信息。具体来讲，相比于已有的1D attention，2D attention可以在纵向进行特征筛选与融合。Shi et al. (2018)将注意力序列-序列模型引入到场景文本识别问题中，设计的矫正网络采用（Spatial Transformer Networks，STN）Jaderberg et al. (2015)和薄板样条插值算法结合，将输入图像中不规则的文本区域变换成规则的文本区域图像，提高了不规则文本图像的识别准确率。

（3）基于Transformer的图像文本识别方法

随着Transformer 的快速发展，分类和检测领域都验证了Transformer 在视觉任务中的有效性。在针对规则文本识别的过程中，CNN在长依赖建模上会存在局限性，Transformer结构恰好解决了这一问题，它可以在特征提取器中关注全局信息。Yu et al. (2020)将Transformer的Encoder模块接在ResNet50后，增强了2D视觉特征。并提出了一个并行注意力模块，将读取顺序用作查询，使得计算与时间无关，最终并行输出所有时间步长的对齐视觉特征。Sheng et al. (2019)使用了完整的Transformer结构对输入图片进行编码和解码，只使用了简单的几个卷积层做高层特征提取，在文本识别上验证了Transformer结构的有效性。Yang et al. (2020)使用Transformer的解码器替换LSTM，再一次验证了并行训练的高效性和精度优势。

以上方法为本文解决缅甸语图像文本识别任务提供了较好的思路，本文方法与现有工作主要区别是提出一种融合通道注意力和空间注意力的视觉关注模块，对深度卷积神经网络提取的缅甸语图像特征图分别获取通道域和空间域的注意力图，融合后对原特征图重构，使缅甸语图像文字边缘特征能够获得更多的注意力关注，进而缓解缅甸语图像文本识别中上下标字符易丢失的问题。

3 融合双重注意力机制的缅甸语图像文本识别模型

本文提出的网络架构如图所示，模型架构由基于ResNet(He et al., 2016)的特征提取模块、融合通道注意力和空间注意力的视觉信息关注模块和解码模块三部分组成。特征提取模块主要对输入的缅甸语文本图像经过卷积神经网络提取到其文本信息特征，再通过视觉信息关注模块进行注意力计算，增强缅甸语图像的文本特征表征能力，最后通过解码器解码转录出对应的缅甸语文字。

3.1 缅甸语图像特征提取网络

我们在残差网络(Residual Network，ResNet)的基础上构建了适应缅甸语图像特征提取的主干网络，通过特征提取网络获得512维的缅甸语图像特征图。

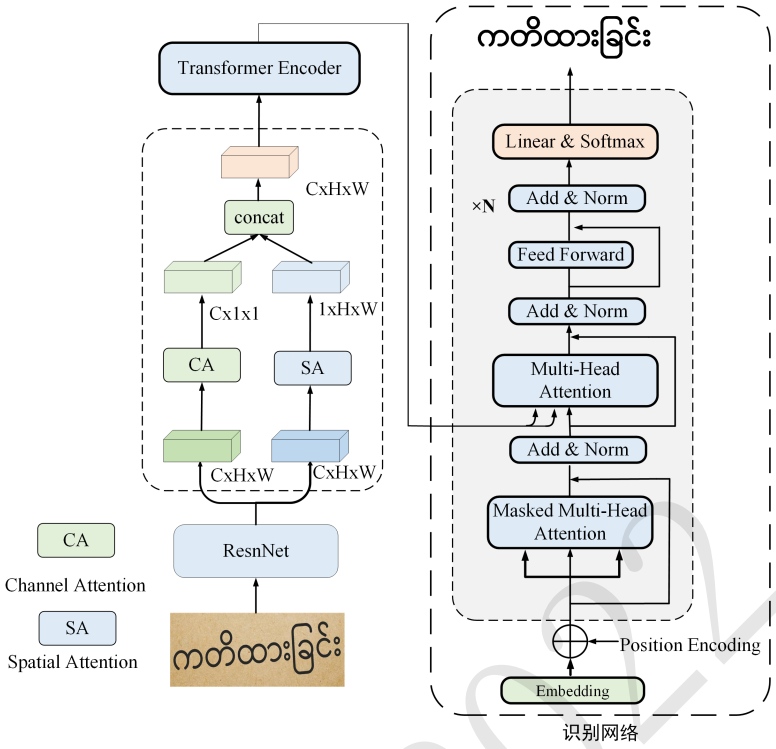


Figure 2: 融合双重注意力机制的缅甸语文本图像识别模型结构图

3.2 缅甸语图像语义特征增强

为了能够更好地捕捉缅甸语图像的像素级成对关系和通道依赖关系，排除图像中噪声干扰，从而获取到语义更精准的特征图，我们设计了通道注意力机制和空间注意力机制，以生成混合域的注意力向量，并对原特征进行重构，提高缅甸语图像的文字区域的表征能力。

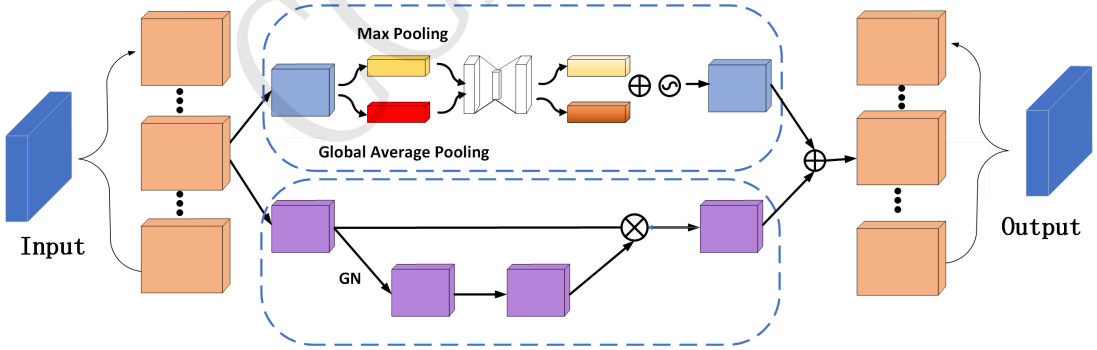


Figure 3: 双重注意力模块结构图

（1）特征图分组

输入的缅甸语图像通过特征提取网络得到通道数为512维的特征图，假设这些输入特征为$X \in R^{C\times H\times W}$，其中C，H，W分别表示通道数、空间高度和宽度，将特征X沿着通道维度拆分为K组：$X = [X_1, ...X_K]$，$X_i \in R^{C\times H\times W}$，其中每个子特征$X_i$在训练过程中逐渐捕获特定的语义响应对于每组特征，我们通过进行通道分组操作，在每个注意单元的开头，X_i的输入沿着通道维度被分成两个分支，即X_{i1}、$X_{i2} \in R^{C/2K\times H\times W}$。一个分支通过利用通道的相互关系

在[C] 维度上获取注意力权重来生成通道注意力图，而另一个分支则通过利用特征的空间关系在[H，W]维度上进行注意力权重计算来生成空间注意力图。

（2）通道注意力

通道注意力能够显式地建模特征通道之间的相互依赖关系。就是通过学习的方式来自动获取到每个特征通道的注意力权重，然后根据这个注意力权重去提升缅甸语图像中文本相关区域的特征并抑制背景及其他噪声信息的干扰。

我们首先通过使用平均池化和最大池化操作来聚合缅甸语图像的文本特征信息，生成两个不同的空间上下文特征描述：X'_{i1avg}和X'_{i1max}，分别表示平均池化特征图和最大池化特征图，其维度大小都为$C/2K \times 1 \times 1$，然后将这两个特征图分别送入两层的全连接神经网络，并且这个两层的全连接神经网络的参数是共享的，再将得到的两个特征图相加，通过Sigmoid函数得到0∼1之间的权重系数，得到最终输出通道注意力图为$M_c \in R^{C/2K \times 1 \times 1}$。其中，为了减少参数开销，共享网络的隐藏激活大小设置为$R^{C_t \times 1 \times 1}$，其中t为缩减率。简而言之，通道注意力权重计算如下:

$$\begin{aligned} M_c\left(X'_{i1}\right) &= \sigma\left(MLP\left(AvgPool\left(X'_{i1}\right)\right)\right) + MLP\left(MaxPool\left(X'_{i1}\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(X'_{i1avg}\right)\right)\right) + W_1\left(W_0\left(X'_{i1max}\right)\right) \end{aligned} \tag{1}$$

其中，σ表示sigmoid函数，$W_0 \in R^{C/t \times C}$，$W_1 \in R^{C \times C/t}$表示两个输入共享MLP的权重。

（3）空间注意力

不同的维度所代表的意义是不同的，它们本身所携带的信息也是不同的。相比较图像的通道信息而言，其空间所拥有的位置信息更为丰富。在实现方面，我们首先采用Group Norm(GN)对X_{i2}进行处理得到空间域层面的统计信息，然后采用$F_C(\cdot)$进行增强，得到空间注意力图为$M_s \in R^{C/2K \times H \times W}$，该过程可以描述如下：

$$M_s\left(X'_{i2}\right) = \sigma\left(W_2 \cdot GN\left(X_{i2}\right) + b_2\right) \cdot X_{i2} \tag{2}$$

其中，$W_2 \in R^{C/2K \times H \times W}$，$b_2 \in R^{C/2K \times H \times W}$。

（4）特征融合

在完成通道和空间注意力计算后，我们需要对其进行集成，首先通过简单的Concat进行融合得到：$M = \left[M_c\left(X'_{i1}\right), M_s\left(X'_{i2}\right)\right] \in R^{C/K \times H \times W}$，最后采用通道置换操作进行组间通信。

3.3 缅甸语图像特征表示

设缅甸语文本行的输入图像，图像的宽度可能具有任意长度，先用卷积神经网络ResNet对缅甸语图像进行处理，再利用通道和空间注意力网络对处理之后的结果进行特征增强，最后我们得到了一个大小为$H \times W \times C$的中间视觉特征表示F_C，这种视觉特征表示具有整个缅甸语输入图像的上下文化的全局表示，特征结构紧凑。缅甸语文本图像在本质上是连续的信号，缅甸语文的读取顺序是从左到右，为此我们视觉特征表示F_C转化为视觉特征向量$\{v_1, v_2..., v_w\}$，其中$v_i \in R^{C \times H}$。

我们采用Muti-Attention对视觉特征向量进行编码，由于输入视觉特征向量本身是缺乏位置信息，我们采用原始Transformer的位置编码方式对视觉特征向量进行位置编码。位置信息编码之前，维度大小为的视觉特征向量进行维度压缩，维度压缩方式为将其输入到一个全连接层实现维度转化，最终维度压缩之后视觉特征向量向量$\tilde{F}_C$的大小为(C, W)。为了有效地、明确地引导注意机制和让视觉向量$\tilde{F}_C$失去水平位移不变性，根据Vaswani et al. (2017)的研究，采用了基于正弦和余弦函数的位置编码。

$$TE\left(pos, 2i\right) = \sin\left(\frac{pos}{10000^{2i/f}}\right) \tag{3}$$

$$TE\left(pos, 2i+1\right) = \cos\left(\frac{pos}{10000^{2i/f}}\right) \tag{4}$$

其中，$pos \in \{0, 1, 2, ..., w-1\}$，$i \in \{0, 1, 2, ..., c-1\}$。

将$\tilde{F}_C$与位置编码进行融合得到向量$\hat{F}_c$，为了进一步提取视觉特征，在$\hat{F}_c$上应用了四次自注意模块。该注意模块输入为Q_c，K_c和V_c，其中$Q_c = K_c = V_c$。相关性信息计算方式如下：

$$\tilde{v}_c^i = Soft\max\left(\frac{q_c^i K_c}{\sqrt{c}}\right) V_c \tag{5}$$

其中，$q_c^i \in Q_c$，$i \in \{0,1,2,...,w-1\}$，$\tilde{F}_c = \left\{\tilde{v}_c^0, \tilde{v}_c^1, ..., \tilde{v}_c^{w-1}\right\}$，经过注意力计算得到增强之后的视觉特征$\tilde{F}_c$，用于后续的文字转录模块。

3.4 缅甸语文字转录

文字转录模块负责将视觉特征$\tilde{F}_c = \left\{\tilde{v}_c^0, \tilde{v}_c^1, ..., \tilde{v}_c^{w-1}\right\}$解码为字符，关注视觉特征以及从文本特征中学习到的语言特定知识。文字转录模块是由4个Tranformer解码器组成。选择Tranformer而不是基于RNN的体系结构的原因是，RNN结构在对当前时刻进行文字分类时依赖上一时刻不能实现并行计算。每个解码器层由三个子层组成：两个多头注意机制层和一个前馈神经网络组成。以前关于基于注意力机制的文字识别方法只在每个解码步骤的编码状态上使用一个注意力分布，相比之下，每个解码层我们采用多头注意力机制对编码器特征进行建模计算，并解决了解码时输出字符与编码特征之间的复杂对齐关系。

模型训练时采用交叉熵损失函数作为缅甸语识别模型的目标优化函数，计算方式如公式6所示：

$$Loss_{Att} = -\sum \ln P\left(\hat{y}_t | M, \theta\right) \tag{6}$$

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

4 实验结果及分析

为验证融合通道注意力和空间注意力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

由于缅甸语属于典型的资源稀缺性语言，目前没有公开的缅甸语文本图像数据集，因此本文将在自构的缅甸语数据集上来验证方法的有效性，该数据集总共包含了800万张缅甸语图像，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3万张图片，剩余数据是通过合成算法得到的包含不同背景颜色、不同倾斜角度的缅甸语文本图像，以此增加训练样本的多样性。其中，分别随机选取20万缅甸语图像作为测试数据集和验证数据集。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、验证集以此提高模型读取速率，具体规模如表1所示。

数据集	数量	样例	标签
训练集	800万		စိတ်ဆိုးသော
测试集	20万		ရယူလိုသော
验证集	20万		အရိပ်အမြွက်။

Table 1: 缅甸语图像数据集样例及对应标签实例

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7所示：

$$SA = \frac{SL}{LN} \times 100\% \tag{7}$$

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 实验结果及分析

为验证融合通道注意力和空间注意力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。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200，训练步长设为700000，实验结果选择评测中最高的准确率，实验结果如表2所示。

实验一：主要实验结果及分析

本文在缅甸语数据集上进行了实验，并与以下的模型的实验结果进行了对比：

CNN+BiLSTM+CTC: (Shi et al., 2016)首先使用标准的CNN网络提取文本图像的特征，再利用BiLSTM将特征向量进行融合以提取字符序列的上下文特征，然后得到每列特征的概率分布，最后通过CTC进行预测得到文本序列

CNN+BiLSTM+Attention(Baek et al., 2019)：解码部分采用注意力解码器对序列进行解码。

毛等人(毛存礼et al., 2022)：构建了基于卷积神经网络和循环神经网络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。

刘等人(Liu et al., 2021)：提出利用深度卷积网络获取并融合多层语义特征图，来缓解缅甸语图像文本识别过程中上下标字符特征丢失得问题，并采用MIX UP(Zhang et al., 2017)的训练策略。

方法类别	具体方法	SA(%)	Time(s)
联结主义时间分类的方法	CNN+LSTM+CTC	84.5	*
	CNN+BiLSTM+CTC	90.4	1250
序列到序列的方法	CNN+BiLSTM+Attention	90.6	16897
现有缅甸语图像文本识别的方法	谢等人	93.5	*
	刘等人	94.2	11560
基于Transformer的方法	Resnet+Transformer	94.8	1630
	Ours	95.3	1632

Table 2: 实验结果

如表2 所示，所提方法在缅甸语图像文本识别任务上准确率达到95.3%，达到了最高水平。相比联结主义时间分类的方法，提升了4.9%，说明本文方法能够获取更丰富的缅甸语图像文本特征信息，识别结果显示了明显的优势；相比序列到序列的方法，提升了4.7%，说明本文的方法在识别缅甸语的过程中提取到更为细粒度的缅甸语图像文本特征并进行特征图注意力计算，赋予了一些边缘特征更高的权重；相比已有缅甸语识别的方法，提升了1.1%，说明本文的方法缅甸语图像特征提取过程中更多地关注到缅甸语字符上下标等文字边缘特征，减少了缅甸语字符上下标丢失或错误识别的情况。

为了验证本文方法在缅甸语文本图像识别效率方面的提升效果，我们在相同的数据集和实验参数下对不同的方法进行了实验，并取平均每训练2000步长所需的时间作为对比结果。由表2的实验结果分析可知，本文方法大幅度缩短了训练时间，相比较刘等人的方法训练时间缩短将近7倍，与“CNN+BiLSTM+Attention”方法相比更是缩短到接近原来的十分之一，说明本文方法在能较好提高识别准确率的情况下，极大地提升了识别效率；同时与“Resnet+Transformer”相比训练时间相差无几，说明本文融合通道注意力和空间注意力模块的方法在几乎没有增加训练成本的前提下也能提升识别的准确率；此外，我们注意到“CNN+BiLSTM+CTC”的训练时间比本文方法更短，这是因为基于CTC的解码方式没有太多的针对图像上下文特征的注意力计算，考虑到本文方法的识别准确率相比“CNN+BiLSTM+CTC”有较大的提升，因此仍然能够说明方法的有效性与实用性。

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(%)
CNN+LSTM+CTC	82.5
CNN+BiLSTM+Attention	89.7
CNN+BiLSTM+CTC	89.5
Ours	94.1

Table 3: 真实测试集上的实验结果

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，同比基于注意力的识别模型的准确率能够提升4.4个百分点，融合通道注意力和空间注意力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。

实验二：通道和空间注意力融合消融实验结果对比

为验证缅甸语通道和空间注意力融合策略的有效性，我们分别对其做了消融试验。我们分别对以ResNet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（“✗”代表未融合，“✓”代表融合）

方法	Channel Attention	Spatial Attention	SA（%）
ResNet+Transformer	✗	✗	94.8
ResNet+Transformer	✓	✗	94.8
ResNet+Transformer	✗	✓	94.9
ResNet+Transformer	✓	✓	95.3

Table 4: 通道和空间注意力融合对识别的影响

如表4所示，其中Channel Attention表示通道注意力，Spatial Attention表示空间注意力，从实验结果可以看出，在只融合通道注意力或空间注意力中的情况下，以ResNet为主干网络的缅甸语图像识别模型性能提升非常小，但同时融合两种注意力时对模型的准确率可以提高0.5个百分点，说明同时对缅甸语图像的通道域和空间域做注意力计算并融合能够更充分关注到文本信息相关的特征。

实验三：针对注意力头数和的消融实验对比

为了验证多头注意力个数对识别模型的影响，我们对其进行了消融实验，实验结果如表5所示。其中，当注意力头数为6时，识别模型的性能最优。

注意力头数	SA(%)
2	94.5
4	94.8
6	95.3

Table 5: 注意力头数对识别的影响

实验四：针对视觉注意力单元和解码单元个数的消融实验对比

为了验证视觉注意力单元和解码单元个数对识别模型的影响，我们将注意力头数设为6，对单元个数进行了消融实验，实验结果如表6所示。其中，当单元个数为4时，识别模型的性能最优。

单元个数	SA(%)
2	94.6
4	95.3
6	95.0

Table 6: 单元个数对识别的影响

为了验证本文所提的注意力关注模块能够更加充分关注到缅甸语图像中文字所在区域，我们对其进行注意力可视化，注意力可视化结果如Figure4所示。我们的方法在识别缅甸语字符时能够有效关注到各个字符在图像中的位置，比如在Figure4的第一张图像中，模型在识别字符“ူ”和“ိ”时，对于图像中上标字符“ေ”以及下标字符“ျ”所在区域给予了较高的注意力权重，充分关注到缅甸语图像中的文字边缘特征，提高了模型对缅甸语字符序列的识别精度。

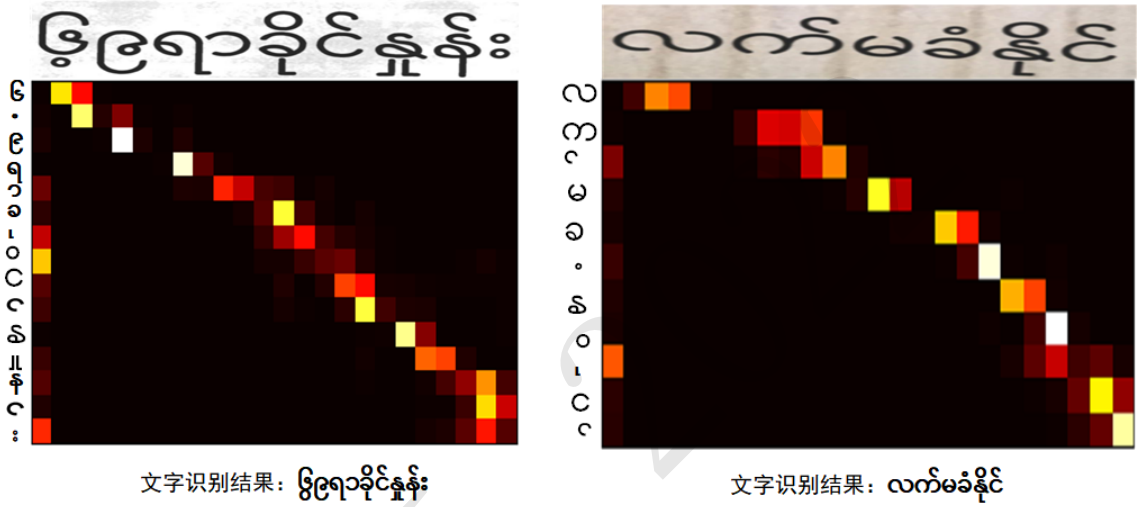


Figure 4: 字符识别注意力分布图

4.3 测试样例展示

表7 给出了缅甸语图像识别的实例。在针对有多个上下标的缅甸语嵌套字符图像的识别时，基于Resnet+transformer的识别模型会存在图像文字边缘特征的缺失，导致识别结果的上下标丢失，面对低质图像这类问题更为明显。而本文方法融合通道注意力和空间注意力的识别模型在面对低质或者组合字符数量多的缅甸语图像，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。

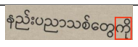
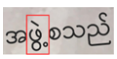
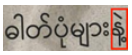
测试样例	Resnet+transformer	Ours
	နည်းပညာသစ်တွေကို	နည်းပညာသစ်တွေကို
	အဖွဲ့စသည်	အဖွဲ့စသည်
	ခါတ်ပုံများနှ	ခါတ်ပုံများနှ

Table 7: 测试样例及结果

5 结论

针对缅甸语图像文本识别中会存在上下标导致识别不佳的问题，提出了一种融合通道和空间注意力的缅甸语图像文本识别方法，将在通道域和空间域分别得到的注意力特征图融合后对原特征图进行重构，提高了模型对缅甸语图像文字边缘特征的提取能力。并在自构的缅甸语数据集的基础上进行了实验，相较于Resnet+Transformer的基线提升了0.5%，验证了所提方法的有效性。本文工作不仅缓解了缅甸语图像文本识别过程中字符上下标丢失的问题，还探索了类似缅甸语这类以音节为基本组成单位的低资源语言的语言特征在图像文本识别任务中所面临的问题和挑战，为其它类似的语言提供了较好的借鉴。在下一步工作中，在开展针对缅甸语这类具有复杂嵌套字符组合语言的图像文本识别的研究中，我们将进一步探索预训练模型以及Mask机制对其图像文本识别性能的影响。

参考文献

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. 2021. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34.

Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4715–4723.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835.

Yunze Gao, Yingying Chen, Jinqiao Wang, and Hanqing Lu. 2017. Reading scene text with attention convolutional sequence modeling. *arXiv preprint arXiv:1709.04303*.

Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, and Santiago Fernández. 2007. Unconstrained on-line handwriting recognition with recurrent neural networks. *Advances in neural information processing systems*, 20.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Chen-Yu Lee and Simon Osindero. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2231–2239.

Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8610–8617.

Fuhao Liu, Cunli Mao, Zhengtao Yu, Chengxiang Gao, Linqin Wang, and Xuyang Xie. 2021. 融合多层语义特征图的缅甸语图像文本识别方法(burmese image text recognition method fused with multi-layer semantic feature maps). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 176–185.

Fenfen Sheng, Zhineng Chen, and Bo Xu. 2019. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 781–786. IEEE.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Lu Yang, Peng Wang, Hui Li, Zhen Li, and Yanning Zhang. 2020. A holistic representation guided attention network for scene text recognition. *Neurocomputing*, 414:67–75.

Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. 2017. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*.

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*.

毛存礼, 谢旭阳, 余正涛, 高盛祥, 王振晗, 刘福浩. 2022. 基于知识蒸馏的缅甸语光学字符识别方法. 数据采集与处理, 37(1):10.