融合多粒度特征的缅甸语文本图像识别方法

何恩宇[1,2]，陈蕊[1,2]，毛存礼[*1,2]，黄于欣[1,2]，高盛祥[1,2]，余正涛[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
2329863182@qq.com, 1226211036@qq.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com

摘要

缅甸语属于东南亚低资源语言，缅甸语文本图像识别对开展缅甸语机器翻译等任务具有重要意义。由于缅甸语属于典型的字符组合型语言，一个感受野内存在多个字符嵌套，现有缅甸语识别方法主要是从字符粒度进行识别，在解码时会出现某些字符未能正确识别而导致局部乱码。考虑到缅甸语存在特殊的字符组合规则，本文提出了一种融合多粒度特征的缅甸语文本图像识别方法，将较细粒度的字符粒度和较粗粒度的字符簇粒度进行序列建模，然后将两种粒度特征序列进行融合后利用解码器进行解码。实验结果表明，该方法能够有效缓解识别结果乱码的现象，并且在人工构建的数据集上相比“VGG16+BiLSTM+Transformer”的基线模型识别准确率提高2.4%，达到97.35%。

关键词： 缅甸语文本图像识别；多粒度识别；字符簇

Burmese Language Recognition Method Fused with Multi-Granularity Features

Enyu He[1,2], Rui Chen[1,2], Cunli Mao[*1,2], Yuxin Huang[1,2], Shengxiang Gao[1,2], Zhengtao Yu[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
2329863182@qq.com, 1226211036@qq.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com

Abstract

Burmese is a low-resource language in Southeast Asia, and Burmese text image recognition is of great significance for carrying out tasks such as Burmese machine translation. Since Burmese language is a typical character-combined language, there are multiple character nests in one receptive field. The existing Burmese language recognition methods mainly recognize character granularity, and some characters may not be recognized correctly during decoding, resulting in partial Garbled characters. Considering that there are special character combination rules in Burmese, this paper proposes a Burmese text image recognition method that integrates multi-granularity features. The two granular feature sequences are fused and then decoded by a decoder. The experimental results show that this method can effectively alleviate the phenomenon of garbled recognition results, and the recognition accuracy rate of the baseline model of "VGG16+BiLSTM+Transformer" is increased by 2.4% to 97.35% on the artificially constructed dataset.

*毛存礼（通讯作者）：maocunli@163.com
国家自然科学基金（62166023，U21B2027，61972186），云南省科技重大专项（202103AA080015，202203AA080004），云南省基础研究计划项目（202201AT070858）

Keywords: Burmese text image recognition , Multi-granularity recognition , character clusters

1 引言

缅甸语是一种东南亚低资源语言，其文字具有独特的形态和结构。随着数字化技术的迅速发展，缅甸语文本图像识别逐渐成为了一个重要的研究领域。缅甸语文本图像识别可以帮助我们将印刷或手写的缅甸语文本转换为可编辑的数字形式，这对于数字化文献、信息检索和自然语言处理等领域都具有重要的应用价值。然而现有的缅甸语识别模型在识别缅甸语时，由于缅甸语中存在着大量组合字符，导致识别过程中易发生漏识、错识某些关键字符，容易出现乱码的现象。

(a) 含上下标图像

(b) 含左拼/右拼字符图像

Figure 1: 缅甸语字符组合规则示例

缅甸语字符编码顺序以及组合规则和中英文字符存在较大的差异。中英文字符编码遵循从左到右的顺序，同时中英文不存在组合字符的现象，然而由于缅甸语字符包括辅音字符、左拼元音字符、上拼元音字符、右拼元音字符、下拼音调字符/后拼音调字符、后拼其他字符。由于缅甸语中大量存在组合字符，如Figure1中的(a)的“ငှဲ့”是由辅音字符“င”、下拼音调字符“့”和上拼元音字符“ှ”组成的。我们将纵向堆叠的字符定义为一个字符簇，切分后的结果如表1中示例所示。同时缅甸语编码顺序和视觉上呈现的顺序一致。如Figure1中的(b)的字符簇“ဖတ့်”是由辅音字符“တ”、左拼元音字符“ဖ”、上拼元音字符“ိ”和下拼音调字符“့”组成的。虽然我们看到该字符簇的第一个字符为左拼元音字符“ဖ”，但是在文本编码端“ဖ”其实是该字符簇的第二个字符。在传统CRNN(Shi et al., 2016)网络中，一个感受野只会对应一个特征，但在缅甸语文本图像中一个感受野往往包含多个缅甸语字符，按照单个字符解码的方式会导致一些关键字符的丢失，从而出现局部乱码的现象。

缅甸语文本	字符级	字符簇级
တထေရာတည်း	တ ေ ထ ရ ာ တ ည ် း	တ ေ ထ ရာ တ ည် း
နေဝင်မှနေထွက်ချိန်။	ေ န ဝ င ် မ ့ ေ န ထ ွ က ် ခ ျ ိ န ် ။	ေ န ဝ င် မ့ ေ န ထွ က် ချ ိ န် ။
အောက်ခြေ	ေ အ ာ က ် ေ ခ ြ	ေ အာ က် ေ ခြ

Table 1: 缅甸语文本切分实例

此外，研究人员针对缅甸语文本图像的识别任务做了一些尝试，毛存礼et al. (2022)提出了利用知识蒸馏的方式，将教师网络学习到的单个字符的特征传递给学生网络，以提升学生网络对于缅甸语字符的特征提取能力，从而在一定程度上解决了缅甸语文本图像识别过程中的某些字符丢失的情况，但是该方法忽略了深度卷积神经网络的底层语义信息及相关特征。Liu et al. (2021)提出一种基于融合多层语义特征的缅甸语文本图像识别方法，将特征提取层提取到的多层缅甸语特征进行融合，达到提高主干网络对缅甸语文本图像的特征提取的能力的目的。但是该主干网络在提取缅甸语文本图像的边缘特征上表现的并不是很好，同时模型的编解码效率较低。Wang et al. (2022)提出一种融合通道注意力和空间注意力的缅甸语文本图像识别方法，在提取图像特征的同时构建空间注意力和通道注意力，最后利用多头注意力机制对融合结果进行注意力计算，但是该方法忽略了缅甸字符的组合规则，在真实场景应用中依然会出现由于识别结果中某些缅甸语字符丢失导致的乱码现象。

为了解决上述问题，本文针对某些关键字符识别丢失或者出现识别错误导致的局部乱码现象，受到缅甸语分词林颂凯et al. (2018)的启发，本文将纵向堆叠字符作为一个字符簇，提出一

种融合多粒度特征的缅甸语文本图像识别方法，在编码端获得字符粒度和字符簇粒度的两种特征序列，然后将两种粒度的特征序列进行融合，然后将利用字符组合规则切分得到的字符簇作为解码字典，最后通过解码器得到识别结果。

本文工作主要有以下贡献：

(1)我们提出了一种融合多粒度特征的缅甸语文本图像识别方法，使用了字符簇粒度的编解码字典，解决了现有缅甸语识别模型未能正确识别某些字符导致的乱码现象。

(2)我们在图像特征提取层做了改进，分别提取到字符粒度和字符簇粒度的图像特征并进行序列建模，将两种粒度特征序列进行融合后利用解码器进行解码，提高了缅甸语识别模型的精度。

(3)在人工构建的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文本图像识别的精度达到97.35%，优于多个对比模型。

2 相关工作

(1)基于联结主义时间分类的文本图像识别方法

基于联结主义时间分类（Connectionist Temporal Classification, CTC）的文本识别方法，使用CTCLoss作为目标优化函数。该算法的核心思想是定义如何将预测结果转化为真实标签，并使用动态规划算法从输出概率分布中获取多条状态转移路径，将路径概率之和或最大值作为目标优化函数。因此，CTC算法可以进行端到端的训练，只需要输入文字级标签，而不需要字符级标签。这种方法使得文本识别的训练更加高效，同时减少了标注数据的成本。Shi et al. (2016)利用卷积神经网络（Convolutional Neural Networks, CNN）提取文本图像中的文本特征，利用循环神经网络（Recurrent Neural Networks, RNN）对特征进行编码，提出了一种将CNN和RNN相结合的识别模型（CRNN）。通过CRNN将文本图像转化为特征序列，然后通过长短时记忆（Long Short-Term Memory, LSTM）增强上下文的语义建模，最后将输出的特征序列输入到CTC模块进行解码得到最后的识别结果。Chandio et al. (2022)使用视觉几何组网络（Visual Geometry Group Network, VGGNet）来提取图像特征，使用基于RNN 的结构将特征序列解码为概率分布，最后，将CTC 函数应用于RNN序列之上，以将每帧预测转换为标签的目标序列。Bhatt et al. (2023)提出了一种混合模型，将新颖的属性签名表示（表征单词中基本视觉形状和字符的出现和位置）与CTC 框架中的LSTM 结合在一起。然而CTC算法假设每个时间片都是相互独立的，但在OCR中，相邻几个时间片中往往包含着高度相关的语义信息，它们并非相互独立的，这种特性使得CTC算法存在一定的缺陷。

(2)基于注意力机制的文本图像识别方法

基于注意力机制的文本识别方法首先使用编码器将文本图片转化为中间语义特征。接着，基于注意力模型的解码器可以将这些中间语义特征转化为识别结果。这种方法可以学习任意长度序列之间的对齐关系，从而减轻了序列对齐的问题。Wojna et al. (2017)使用CNN特征提取器处理图像，然后通过注意力机制进行加权，然后将加权后的数据传递给RNN进行解码。Liao et al. (2019)将文本图片编码为二维特征，用一个结合注意力机制的全卷积网络做像素级的分类，再用一个后处理模块输出字符序列以实现任意形状文本的识别。Zhong et al. (2022)首先利用语义生成对抗网络（Generative Adversarial Network, GAN）生成简单的语义特征，然后利用平衡注意力模块对场景文本进行识别。但是此类模型的计算量较大、对图像噪声和畸变的鲁棒性不高、对长文本的识别效果不佳以及对于语义的理解能力有限。

(3)基于Transformer的文本图像识别方法

如今Vision Transformer(Dosovitskiy et al., 2020)在计算机视觉取得了广泛的应用。在文本图像识别方面，CNN在长依赖建模上存在局限性。而Transformer因为可以在提取特征的同时关注到全局的信息，解决了这一问题。Zhao and Gao (2022)等人使用DenseNet作为图像的文本特征提取网络，将提取到的特征通过Transformer进行编解码，结合注意力精炼模块（Attention Refinement Module, ARM）得到最终输出。Xie et al. (2022)等人分别利用角点检测器和传统图像特征提取网络得到角点特征和图像特征，然后将图像的特征将通过多头自注意力机制进一步建模全局特征，同时角点图的特征将通过多头交叉注意力机制与图像全局特征融合，编码器的输出和字符序列Embedding输入Transformer解码器获得特征序列，得到最终输出。Li et al. (2021)首先将输入文本图像调整为384×384，然后将图像分割成16x16patch的序列，送入Transformer中进行编解码并最终得到输出。

以上方法均为本文解决缅甸语文本图像识别任务提供了较好的思路，本文的主要方法与现有方法的主要区别是提出了一种融合多粒度的编解码方法，并在识别的过程中融合了缅甸语的字符组合规则，进而缓解了在识别缅甸语文本图像时由于某些关键字符丢失导致的乱码现象。

3 模型架构

本文提出的识别模型架构如Figure2所示，整个模型分为基于视觉几何组网络VGGNet(Sengupta et al., 2019)的图像特征提取模块、多粒度图像特征编码模块、多粒度特征融合模块、缅甸语文本解码模块。图像特征提取模块将文本图像进行特征提取，多粒度图像特征编码模块将多粒度图像特征进行序列建模，多粒度特征融合模块将字符粒度和字符簇粒度的序列进行融合、缅甸语文本解码模块将融合特征序列解码得到缅甸语文本输出。

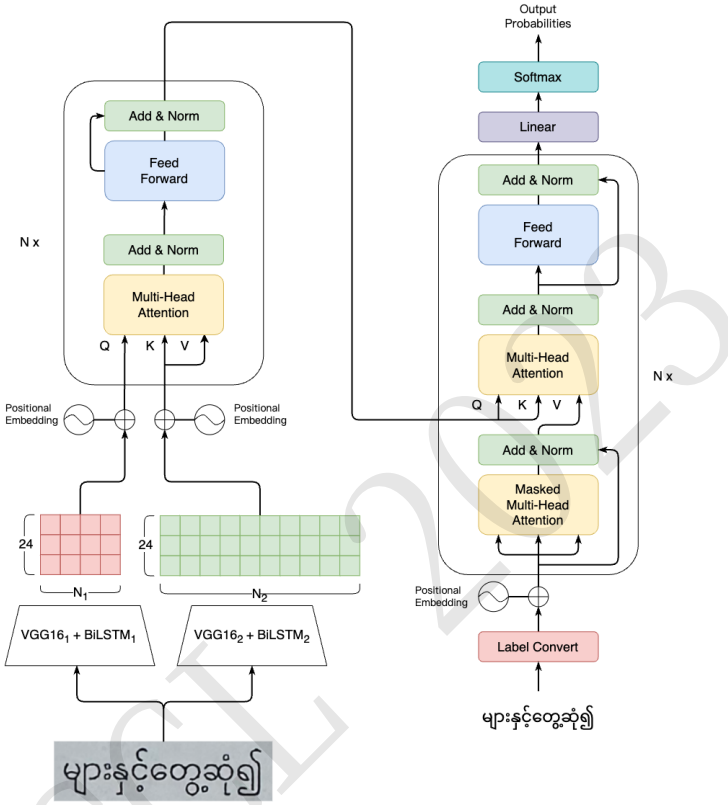


Figure 2: 融合多粒度特征的缅甸语识别模型框架

3.1 缅甸语字符簇字典构建

我们基于缅甸语拼写规则构建了字符簇单独编码字典。以字符簇为单位的编解码字典的构建：我们研究缅甸语拼写规则后发现，缅甸语文本组成的基本单位是缅甸语音节。其中每个缅甸语音节是由辅音字符为主题，然后加上左拼、上拼、右拼元音字符或其他上拼、右拼字符组成。参考现有中英文OCR识别方法不难发现，其编码顺序均是将文字序列从左往右依次编码，然而缅甸语中的编码规则却和中英文的编码方式有着较大的差异，一般来说，缅甸语的编码规则为：辅音字符、左拼元音字符、上拼元音字符、上拼其他字符、右拼元音字符/右拼其他字符。同时我们使用一些算法规则并以字符簇为单位，将缅甸语文文本从左到右切分为纵向字符簇。切分字符簇伪算法如下所示。

3.2 缅甸语文本图像特征提取模块

为了提取到字符粒度的图像特征和字符簇粒度的图像特征，我们在VGGNet的基础上分别构建了适应于提取缅甸语字符粒度和字符簇粒度的特征提取网络。通过特征提取网络分别得到512维的字符粒度特征$X_1 \in R^{C\times H\times W}$和字符簇粒度特征$X_2 \in R^{C\times H\times W}$,其中$C$，$H$，$W$ 分别为通道数、高度和宽度。

Algorithm 1 缅甸语切分字符簇伪算法

Input: တထေရာတည်း

Output: တ ေ ထ ရ တ ည် ်း

```
lst = list( တထေရာတည်း )
clusterList = [" "]
for index in range(len(lst)) do
    if (clusterList[-1] + lst[index]) is not a cluster then
        clusterList.add(lst[idx])
    else
        clusterList[-1] += lst[index]
    end if
end for
return တ ေ ထ ရ တ ည် ်း
```

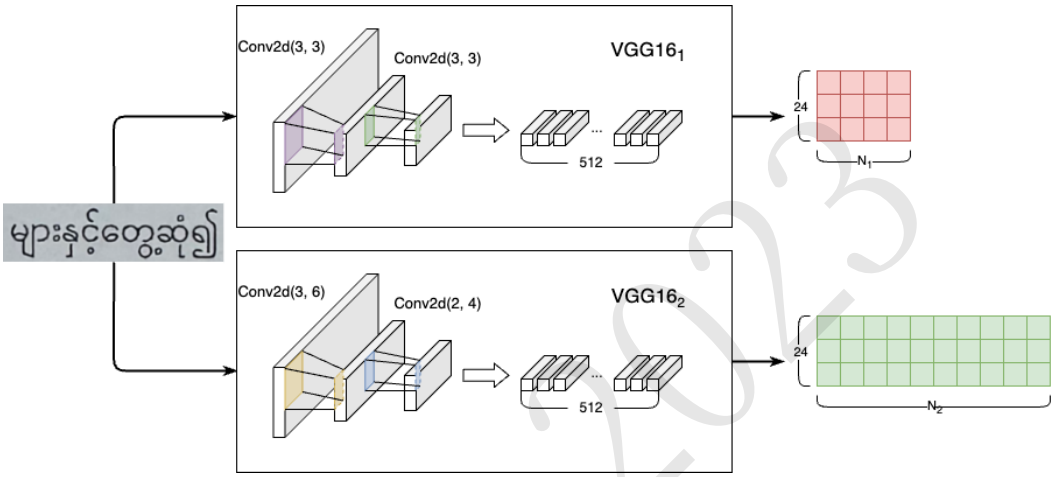


Figure 3: 缅甸语文本图像特征提取模块

考虑到字符簇粒度在纵向上的分布相比于横向分布的占比更大，考虑使用横纵比为1：2的卷积核来提取字符簇粒度的图像特征。

如Figure3中所示，$VGG16_1$为提取字符粒度特征的特征提取网络，$VGG16_2$为提取字符簇粒度特征的特征提取网络。$VGG16_1$将输入图片分别经过(3×3)、(2×2)的卷积核以提取缅甸语文本图像的字符粒度的图像特征。$VGG16_2$将输入图片分别经过(3×6)、(2×4)、(3×3)、(2×2)的卷积核以提取缅甸语文本图像的字符簇粒度的图像特征。

3.3 缅甸语文本图像特征编码模块

为了更好地对文本图像的文本特征进行建模，排除图像中噪声、畸变等条件的干扰，从而获取质量更好的图像特征，我们使用BiLSTM(Bi-directional Long Short-Term Memory, BiLSTM)对通过特征提取网络获取到512维的缅甸语文本图像特征图进行建模，以提高模型对图像噪声和畸变的鲁棒性以及对缅甸语文本图像的表征能力。

$$F_1 = BiLSTM_1(X_1) \tag{1}$$

$$F_2 = BiLSTM_2(X_2) \tag{2}$$

$F_1 \in R^{C \times B \times N_1}$，$F_2 \in R^{C \times B \times N_2}$。其中$C$，$B$，$N_1$，$N_2$分别为通道数、最大字符预测长度、字符级编解码字典长度和字符簇级编解码字典长度。

3.4 多粒度特征融合模块

为了使用字符粒度的特征来优化字符簇粒度的特征，我们使用基于Transformer的多粒度特征融合模块来将$F_1 = \{v_1, v_2, ..., v_{N_1}\}$、$F_2 = \{V_1, V_2, ..., V_{N_2}\}$进行融合，其中$v_i \in R^{B \times N_1}$，$V_i \in R^{B \times N_2}$。

我们的模型采用了Multi-Attention来对视觉特征向量进行编码。由于视觉特征向量缺乏位置信息，因此我们使用Transformer中的位置编码方法来对其进行编码。在进行位置编码之前，我们先将视觉特征向量按维度大小进行压缩，得到两个视觉特征向量F_1和F_2，它们的大小分别为(C, W_1)和(C, W_2)。为了让注意力机制更加有效，同时让F_1和F_2在水平方向上具有位移不变性，我们采用了一种基于正弦和余弦函数的位置编码方式，该方式已在Vaswani et al. (2017)的研究中得到了证明。

$$TE_1(pos_1, 2i) = sin\left(\frac{pos_1}{10000^{2i/f}}\right) \tag{3}$$

$$TE_1(pos_1, 2i+1) = cos\left(\frac{pos_1}{10000^{2i/f}}\right) \tag{4}$$

$$TE_2(pos_2, 2i) = sin\left(\frac{pos_2}{10000^{2i/f}}\right) \tag{5}$$

$$TE_2(pos_2, 2i+1) = cos\left(\frac{pos_2}{10000^{2i/f}}\right) \tag{6}$$

其中，$pos_1 \in \{0, 1, 2, ..., N_1-1\}$，$pos_2 \in \{0, 1, 2, ..., N_2-1\}$，$i \in \{0, 1, 2, ..., c-1\}$。将$F_1$、$F_2$和位置编码融合得到$F_1'$、$F_2'$，为了使用$F_1'$优化$F_2'$，我们使用交叉注意力对$F_1'$和$F_2'$进行融合。该注意力模块输入为$Q$，$K$，$V$。这里我们将$F_1'$ 作为Q，F_2'作为K，V。

$$F_{fusion} = Softmax\left(\frac{Q^i \times K^T}{\sqrt{c}}\right)V \tag{7}$$

3.5 缅甸语文本解码模块

文本解码模块将视觉特征F_{fusion}转化为字符，关注视觉特征、从文本特征中学到特定的语言知识以及隐式的学习到相关的字符组合规则。文本解码模块由4个Transformer解码器组成。使用Transformer解码器而不使用RNN作为解码器的原因是：RNN在解码的时候是串行解码的，并且RNN在某一时刻的输出依赖上一时刻的输出。我们通过将视觉特征F_{fusion}通过Transformer模块得到最终的预测序列F_{pred}。最后将F_{pred}输入解码器Convert得到对应的缅甸语文本text。

$$F_{pred} = Transformer\left(F_{fusion}\right) \tag{8}$$

$$text = Convert(F_{pred}) \tag{9}$$

模型训练时我们使用交叉熵损失作为整个模型的损失函数。

$$Loss_{attn} = -\sum \ln P(y_t|M, \theta) \tag{10}$$

其中，M为输入的缅甸语文本图像，θ为当前识别网络的模型参数，$y_t|M$为缅甸语文本图像的第t个特征序列对应的标签。

4 实验结果与分析

为了验证融合多粒度的缅甸语文本图像识别方法的有效性，我们在缅甸语文本图像数据集上进行实验分析。

4.1 数据集

本文中使用了自构的缅甸语文本图像数据集进行实验，因为缅甸语属于南亚东南亚低资源语言，目前没有公开的缅甸语数据集。该数据集包括800万张缅甸语文本图像，其中大约3万张是人工标注的，其余数据是通过算法合成的。合成数据考虑了不同的背景、噪声和倾斜度等因素，以最大程度模拟真实场景中的数据情况。为了验证模型的有效性，我们随机抽取了20万张图像作为测试数据集和验证数据集。为了提高模型的训练速度，我们使用“.mdb”文件来存储数据集。具体规模如表2所示。

实验采用评价指标为序列率精确率（Sequence Accuracy, SA），如公式11所示：

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

其中，SA、SL、LN分别表示缅甸语文本图像识别的字符串正确率、正确识别文本图像中的字符长度、正确文本长度。

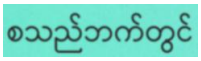
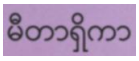
数据集	数量	样例	标签
训练集	800万		အနားယူပါ။
测试集	20万		စသည်ဘက်တွင်
验证集	20万		မီတာရှိကာ

Table 2: 缅甸语文本图像数据集样例及实例

4.2 实验结果分析

为了评估融合多粒度特征的缅甸语文本图像识别方法的有效性，我们在自构的缅甸语文本图像数据集上进行了实验。所有模型在相同的实验环境下进行训练和测试。我们使用Adam优化器，初始学习率设置为1，并使用CosineAnnealing策略逐渐减小学习率。批处理大小设置为128，训练步长为1,000,000。实验结果以测试集的字符错误率（CER）和词错误率（WER）表示，同时选择训练过程中精度最高的模型作为最终结果。实验结果如表3所示。

实验一：主要实验结果及分析

本文在自构的缅甸语文本图像数据集上进行了实验，同时与当前主流识别模型的实验结果做了对比。

CNN+BiLSTM+CTC(Shi et al., 2016)：该文本图像识别方法首先通过卷积神经网络提取文本图像的特征，接着采用BiLSTM网络融合特征向量来捕获字符序列中的语义信息。然后，对每一列特征进行分类，得到其概率分布。最后，采用CTC方法对概率分布进行预测，以得到最终的文本序列。

CNN+BiLSTM+Attention(Baek et al., 2019)：解码部分采用基于注意力机制的解码器对序列进行解码。

CNN+Transformer：使用CNN从文本图像中提取出高层次的特征表示，利用Transformer进行序列建模，学习特征之间的依赖关系，得到更加准确的文本序列预测结果。

毛等人(毛存礼et al., 2022)：构建了教师网络和学生网络，并利用卷积神经网络和循环神经网络的框架进行特征提取和序列建模，教师网络和学生网络共同进行训练，以提高模型的泛化能力和减少过拟合风险。

刘等人(Liu et al., 2021)：利用深度卷积网络获取多层语义特征图并进行融合，以缓解上下标字符特征丢失的问题；使用MIX UP的训练策略进行模型的训练，以提高模型的鲁棒性和泛化能力。

王等人(Wang et al., 2022)：提出了融合通道注意力和空间注意力的缅甸语文本图像识别方法，在提取图像特征的同时构建空间注意力和通道注意力，最后利用多头注意力机制对融合结果进行注意力计算，得到文本图像的预测。

具体方法	SA(%)
VGG16+LSTM+CTC	84.5
VGG16+BiLSTM+CTC	90.4
VGG16+BiLSTM+Attention	90.6
VGG16+Transformer	93.3
毛等人	93.5
刘等人	94.2
王等人	95.3
Resnet50+LSTM+CTC	85.3
Resnet50+BiLSTM+CTC	91.5
Resnet50+BiLSTM+Attention	92.1
Resnet50+Transformer	94.8
Ours	**97.3**

Table 3: 主要对比实验结果

如表3所示，本文所提的识别方法在缅甸语文本图像识别任务上准确率达到97.3%，到达了当前最高水平。相比基于“VGG16+BiLSTM+CTC”和“ResNet50+BiLSTM+CTC”的识别方法，分别提升了6.9%、5.8%，说明本文方法使用新的编解码字典，极大程度上避免了识别结果中的乱码现象；相比基于“VGG16+BiLSTM+Attention”和“ResNet50+BiLSTM+Attention”识别方法，分别提升了6.7%、5.2%，说明本文方法在识别含有大量噪声的文本图像时，也能有较好的表现；相比“VGG16+Transformer”和“ResNet50+Transformer”的识别方法，提升了4.0%、2.5%，说明本文方法使用字符粒度的特征优化了字符簇粒度的特征，提高了字符簇识别的准确率；相比于现有的缅甸语识别方法，提升了2.0%，说明本文方法在关注到上下标的同时也关注到了整个字符簇整体，减少了缅甸语文本图像识别过程中某些关键辅音字符丢失导致的乱码现象。

实验二：多粒度特征和字符簇解码字典消融实验结果对比

为了验证多粒度特征以及字符簇解码字典的有效性，我们分别在基线模型上对其做了消融实验。实验结果如表4所示（“✓”表示融合，“×”表示未融合）。

方法类别	字符粒度	字符簇粒度	SA(%)
VGG16+BiLSTM+Transformer	✓	×	94.9
	×	✓	96.5
	✓	✓	**97.3**

Table 4: 关于多粒度特征的消融实验结果

如表4所示，其中字符粒度、字符簇粒度分别表示模型使用字符粒度和字符簇的特征进行推理。从实验结果可以看出，模型使用字符簇粒度的特征可以更好的提取缅甸语中纵向堆叠字符的特征。当模型使用字符粒度和字符簇粒度的融合特征后，模型既可以关注到缅甸语文本图像的边缘特征，又可以关注到图像中的纵向堆叠字符的特征，提高了基线模型的精度，证明了所提方法的有效性。

实验三：融合多粒度的编解码端对不同识别模型识别效果的影响

为了验证多粒度特征融合策略的有效性，我们分别在多个主流识别模型上做了实验。实验结果如表5所示（“✓”表示融合，“×”表示未融合）。

方法类别	字符粒度	字符簇粒度	SA(%)
VGG16+BiLSTM+CTC	✓	×	90.4
	×	✓	91.2
	✓	✓	**92.1**
VGG16+BiLSTM+Attention	✓	×	90.6
	×	✓	91.4
	✓	✓	**92.6**

Table 5: 融合多粒度特征编解码端对其他模型的影响

如表5所示，其中字符粒度表示模型提取字符粒度的特征，字符簇粒度表示模型提取字符簇粒度的特征。从实验结果可以看出，只使用字符簇粒度特征的情况下，当前主流识别网络均有一定的提升，证明字符簇粒度特征确实使模型关注到了整个纵向堆叠的字符簇；当融合字符粒度和字符簇粒度特征后，这些模型均有明显的提升，证明了融合两种粒度的缅甸语识别方法，既可以使模型关注到字符的边缘特征，又可以关注到整个纵向字符簇的特征。

实验四：真实场景缅甸语文本图片数据测试

为了保证模型在真实场景中的应用，本文使用人工标注的1000张真实场景缅甸语文本图像作为测试集。本文在该真实场景缅甸语文本图像测试集上进行实验，实验结果如表6所示。

本文的方法在对1000张真实场景中的缅甸语文本图像的识别中仍然保持着最高的精度，相比基于CTC的识别模型有着5.6个百分点的提高，证明了本文使用的字符簇粒度的编解码字典确实可以规避大量的乱码现象；比基于注意力机制的模型有着5.8个百分点的提高，证明了本文方

第二十二届中国计算语言学大会论文集，第124页-第134页，哈尔滨，中国，2023年8月3日至5日。

方法类别	SA(%)
VGG16+LSTM+CTC	82.5
VGG16+BiLSTM+CTC	89.7
VGG16+BiLSTM+Attention	89.5
Ours	**95.3**

Table 6: 真实场景缅甸语文本图片测试结果

法可以对图像特征进行更好的建模以及在面对图像中的噪声以及文本图像畸变时拥有着更高的鲁棒性。

实验五：多种模型训练速度与推理速度对比

为了验证我们所提模型的训练速度与推理性能，本文对多个主流识别模型进行了训练速度与推理速度的对比实验。

具体方法	训练时间(s)	推理时间(s)
VGG16+LSTM+CTC	*	5.7
VGG16+BiLSTM+CTC	1250	7.3
VGG16+BiLSTM+Attention	16897	12.4
VGG16+Transformer	1590	6.6
Resnet50+LSTM+CTC	*	6.1
Resnet50+BiLSTM+CTC	1750	7.6
Resnet50+BiLSTM+Attention	23631	13.2
Resnet50+Transformer	2206	6.9
毛等人	*	*
刘等人	11560	11.2
王等人	1632	7.4
Ours	**1664**	7.8

Table 7: 模型训练速度与推理速度对比

如表7所示，我们统计了多个主流模型在缅甸语文本图像识别任务上的训练时间以及推理时间。为了验证我们的模型在训练以及推理速度上的性能，我们在相同的数据集上做了多种主流模型的性能测试，测试过程中我们取模型训练2000步的时长作为训练速度，并使用训练的模型对同样的缅甸语文本图像进行推理预测得到模型的推理速度。从实验结果可以看出，我们的模型相比于“VGG16+Transformer”和“ResNet50+Transformer”模型的训练速度和推理速度变化不大，说明本文所提方法在几乎没有增加多余的时间开销的同时提高了模型识别的精度。此外，虽然我们的模型在推理速度上不及“VGG16+BiLSTM+CTC”等模型，但是我们的模型在识别精度上相较于这些模型均有提升，依然可以说明本文所提方法的有效性。

4.3 测试样例

表8给出了缅甸语文本图像识别的实例。在针对组合字符的识别上，基于“VGG16+BiLSTM+CTC”的识别模型会存在某些关键辅音字符错识、漏识，进而导致识别结果出现乱码。而本文方法使用字符簇粒度的编解码字典，从而避免了识别结果因为错识、漏识导致的乱码。在针低质图片或有畸变存在的文本图像上的识别中，基于“CTC+BiLSTM+Attention”的识别模型由于模型鲁棒性较差，导致识别结果较差。而本文方法使用鲁棒性更好的Transformer框架对图像特征序列进行建模，从而在一定程度上解决了低质图像识别结果查的问题。在针对复杂场景中或含有大量组合字符的文本图像的识别中，基于“VGG16+BiLSTM+Transformer”的识别模型由于未能较好的关注到字符的边缘特征和纵向字符簇的整体特征导致未能准确的识别出图像中的文本。而本文方法将字符粒度特征和字符簇粒度的特征进行融合，使模型可以同时关注到文本的边缘特征和整个纵向字符簇的特征，进而提升识别模型的精度。

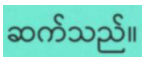
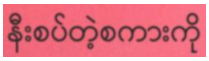
测试样例	CTC	Attention	Transformer	ours
	ကျ ်းကျင်မှ	ကျငြးကျင်ျမှု	ကျမ်းကျင်မွ	ကျမ်းကျင်မှ
	ဆက်သ ်။	ဆတ်သည်။	ဆက်သညီ။	ဆက်သည်။
	ိးစပ်တဲ့စကားကို	နီးစပ်ီတဲ့စကားကို	နီးစပ်တဲ့စကားကဲ့	နီးစပ်တဲ့စကားကို

Table 8: 测试样例

5 结束语

针对现有缅甸语识别模型识别过程中容易出现局部乱码的问题，提出了一种融合多粒度特征的缅甸语文本图像识别方法，使用特征提取网络提取到的字符粒度和字符簇粒度图像特征并进行序列建模，将两种粒度特征序列进行融合后利用解码器进行解码，大幅度减少缅甸语文本图像识别中的乱码情况，提高了缅甸语识别的精度。并在自构的数据集上进行了实验，准确率达到97.35%，验证了所提方法的可行性。本文工作不仅解决了缅甸语识别中字符丢失导致的乱码现象，还为类似缅甸语的南亚东南亚小语种基于字符簇的识别方法提供了解决思路。在下一步的工作中，针对南亚东南亚小语种等具有组合字符语言的文本图像识别的研究中，我们将进一步探索将两种粒度的识别结果进行更好的融合。

参考文献

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.

Ravi Bhatt, Anuj Rai, Sukalpa Chanda, and Narayanan C Krishnan. 2023. Pho (sc)-ctc—a hybrid approach towards zero-shot word image recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(1):51–63.

Asghar Ali Chandio, MD Asikuzzaman, Mark R Pickering, and Mehwish Leghari. 2022. Cursive text recognition in natural scene images using deep convolutional recurrent neural network. *IEEE Access*, 10:10062–10078.

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*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2019. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8714–8721.

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.

Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. 2019. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95.

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.

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.

Fengxiao Wang, Cunli Mao, Zhengtao Yu, Shengxiang Gao, Huang Yuxin, and Fuhao Liu. 2022. 融合双重注意力机制的缅甸语图像文本识别方法(burmese image text recognition method with dual attention mechanism). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 355–365.

Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. 2017. Attention-based extraction of structured information from street view imagery. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 844–850. IEEE.

Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. 2022. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 303–321. Springer.

Wenqi Zhao and Liangcai Gao. 2022. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 392–408. Springer.

Dajian Zhong, Shujing Lyu, Palaiahnakote Shivakumara, Bing Yin, Jiajia Wu, Umapada Pal, and Yue Lu. 2022. Sgbanet: Semantic gan and balanced attention network for arbitrarily oriented scene text recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 464–480. Springer.

林颂凯, 毛存礼, 余正涛, 郭剑毅, 王红斌, and 张家富. 2018. 基于卷积神经网络的缅甸语分词方法. 中文信息学报, 32(6):62–70.

毛存礼, 谢旭阳, 余正涛, 高盛祥, 王振晗, and 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.