ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 5,Sep. 2023,pp. 1058－1068
DOI:10. 16337/j. 1004-9037. 2023. 05. 005

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

基于改进DAN的自然场景下越南文字的识别

王利兵[1]，俸亚特[2]，文益民[1,2]

(1. 广西图像图形与智能处理重点实验室(桂林电子科技大学)，桂林 541004；2. 广西文化和旅游智慧技术重点实验室(桂林旅游学院)，桂林 541006)

摘　要：越南语字符由拉丁字符结合变音符号组成，由于变音符号的存在易导致注意力漂移，并且越南语文字字符类别较多，字符间差异性较小，部分字符仅为变音符号的差异，使得越南文字的识别具有挑战性。本文在解耦注意力网络(Decoupled attention network，DAN)的基础上，设计了视觉特征与序列特征融合模块(Visual feature and sequence feature fusion module，VSFM)，分别利用双向门控循环单元(Bidirectional gated recurrent unit，Bi-GRU)在水平方向和竖直方向进行序列建模，进一步缓解注意力漂移，增强变音符号与拉丁字符间的关联性。然后设计了增强型解耦文本解码器模块(Enhanced decoupled text decoder module，ETDM)，在解码器中分类时结合了更多的特征信息，可以更加有效地识别相似字符。一系列的实验验证了本文提出方法的有效性。

关键词：声调语言文字；越南语文字；变音符号；注意力漂移；场景文本识别

中图分类号：TP391　　**文献标志码**：A

Recognition of Vietnamese Text in Natural Scene Based on Modified DAN

WANG Libing[1]，FENG Yate[2]，WEN Yimin[1,2]

(1. Guangxi Key Laboratory of Image and Graphic Intelligent Processing(Guilin University of Electronic Technology)，Guilin 541004，China；2. Guangxi Key Laboratory of Culture and Tourism Smart Technology(Guilin Tourism University)，Guilin 541006 China)

Abstract：Vietnamese characters which are composed of Latin characters and diacritic symbols make recognition more challenging. On the one hand，diacritic symbols are more likely to lead to attention drift. On the other hand，Vietnamese characters include many categories，and the differences between characters are small，for example some characters only differ from diacritical symbols，which further increases difficulty of recognition. Based on the decoupled attention network (DAN) algorithm，this paper designs a visual feature and sequence feature fusion module (VSFM)，which utilizes bidirectional gated recurrent unit (Bi-GRU) to model sequences in the horizontal and vertical directions，further alleviating attention drift and enhancing correlation between diacritics and Latin characters. And an enhanced decoupled text decoder module (ETDM) is designed，which employs more feature information to identify similar characters more effectively. A series of experiments validate the effectiveness of the proposed method.

Key words: tonal language; Vietnamese text; diacritic sign; attention drift; scene text recognition

---

基金项目：广西重点研发计划项目(桂科AB21220023)；国家自然科学基金(62366011)；广西图像图形与智能处理重点实验室项目(GIIP2306)。

收稿日期：2022-03-13；**修订日期**：2023-04-19

引　言

在过去几十年中,自然场景文本识别吸引了很多研究者的关注,并被广泛应用,如:自动驾驶、视觉辅助、招牌识别等。随着深度学习的发展,自然场景文本识别已经取得巨大进步[1-6]。然而,现有自然场景文本的研究以针对英文等非声调语言文字为主,鲜有针对越南文字等声调语言文字的研究。越南语是一种声调语言,可根据声调区别词义,与泰语、汉语有诸多相似之处。越南文字使用拉丁字母书写,共有29个字母。其中借用拉丁字母22个,即"a,e,i,o,u,y,b,c,d,g,h,k,l,m,n,p,q,r,s,t,v,x",越南文字专用字母7个,即"ă,â,đ,ê,ô,o',u'"。越南文字使用变音符号区分读音,变音符号位于拉丁字母的上部和下部。其中,3个变音符号表示元音("˘,ˆ,'"),比如越南语字母"ă,â,o'",5个变音符号表示声调("~,‘,',?,."),比如字母"a"添加不同声调符号后的字符为"ā,á,à,ả,ạ"。这5个声调符号的变换可以改变越南文字的读音进而改变语义。比如"thiên"为"天空"的意思,而"thiên"为"德善"的意思。为了识别越南文字,本文结合越南语词典以及越南场景中出现的字符,生成了如图1所示包含196个字符的字符集。越南语字符独特的构成,使得自然场景中越南文字的识别面临更大的困难和挑战。从图1中可以观察到:一方面,越南文字字符类别多,比如字符"A"的相似字符就有多种;另一方面,字符间差异性较小,部分字符仅为变音符号的细微差别。变音符号相较于对应的主体字符形状较小,识别会更加困难,变音符号识别错误将导致整个识别结果出错。除此之外,变音符号的存在更易导致注意力漂移问题[7]的发生,从而导致错误的识别结果。

图1　越南文字字符集

Fig.1　Vietnamese characters

基于上述困难与挑战,本文提出了改进解耦注意力网络(Decoupled attention network, DAN)[8]算法,以识别自然场景中的越南文字。并且本文构建了自然场景中越南文字识别数据集,该数据集包含利用场景文字图像生成算法生成的数据以及从真实拍摄的越南自然场景图片中截取的数据;设计了视觉特征与序列特征融合模块(Visual feature and sequence feature fusion module, VSFM),有效缓解注意力漂移问题,增强变音符号与拉丁字符的关联性,同时对字符的笔画和大小写有更好的敏感性;还设计了增强型解耦文本解码器模块(Enhanced decoupled text decoder module, ETDM),可以更精确地解码差异性较小的越南文字字符。

1　相关工作

早期的场景文字识别主要是利用低级的特征,比如方向梯度直方图(Histogram of oriented gradient, HOG)[9]、尺度不变特征变换(Scale-invariant feature transform, SIFT)[10]、连通域[11]等。随着深度学习的发展,场景文字识别进入新时代,场景文字识别使用卷积神经网络(Convolutional neural network, CNN)对图像进行特征编码。其中一类是基于时序连接分类(Connectionist temporal classification, CTC)的方法[3, 12-14]。该方法通过计算条件概率,将深度卷积神经网络或循环神经网络提取的特征序列直接解码为目标字符串序列;另一类是基于Encoder-Decoder的方法[7, 8, 15-19]。该方法包含两个模块:一个是编码器模块,利用神经网络得到场景文字图像的中间特征编码向量;一个是解码器模块,对中间特征编码向量进行时序性解码,从而得到识别的结果。这种方法主要是结合注意力机制,通过连续地学习将输入序列与输出序列对齐。然而,基于Encoder-Decoder的方法会面临严重的注意力漂移问题,即注意力机制不能准确地定位到与当前解码位置相对应的文本图像特征序列[20]。文献[7]首次提出注意力漂移的概念,并且采用Focusing Attention Mechanism改善注意力漂移。文献[19]提出位置

增强分支，并将其输出与注意力模块的输出动态融合，有效缓解注意力漂移问题。

DAN是一种基于Encoder-Decoder的方法[8]，其结构图如图2所示。图中H是图像的垂直方向的像素数；W是图像水平方向的像素数；r_h和r_w分别为图像高度和宽度的抽样比率，以确定输出特征图的高度和宽度。该方法提出卷积对齐模块(Convolutional alignment module, CAM)，将注意力对齐操作从Decoder中解耦，可有效缓解注意力漂移问题。CAM模块是由特征金字塔网络(Feature Pyramid network, FPN)[21]和全卷积网络(Fully convolutional network, FCN)[22]组成，输出为不同时刻的注意力图序列，参数max$T$代表解码的最大时间步长，设为25。在DAN中，Encoder使用ResNet[23]对图像进行特征编码，在Decoder中，将特征图和注意力图相乘得到的上下文向量$\boldsymbol{C}$，再利用$\boldsymbol{C}$与前一时刻的解码向量$\boldsymbol{E}$拼接输入门控循环单元(Gated recurrent unit, GRU)，接着再将GRU的输出结果送入分类器，得到识别结果。需要说明的是，Decoder进行解码时是一个时刻接一个时刻按序进行。

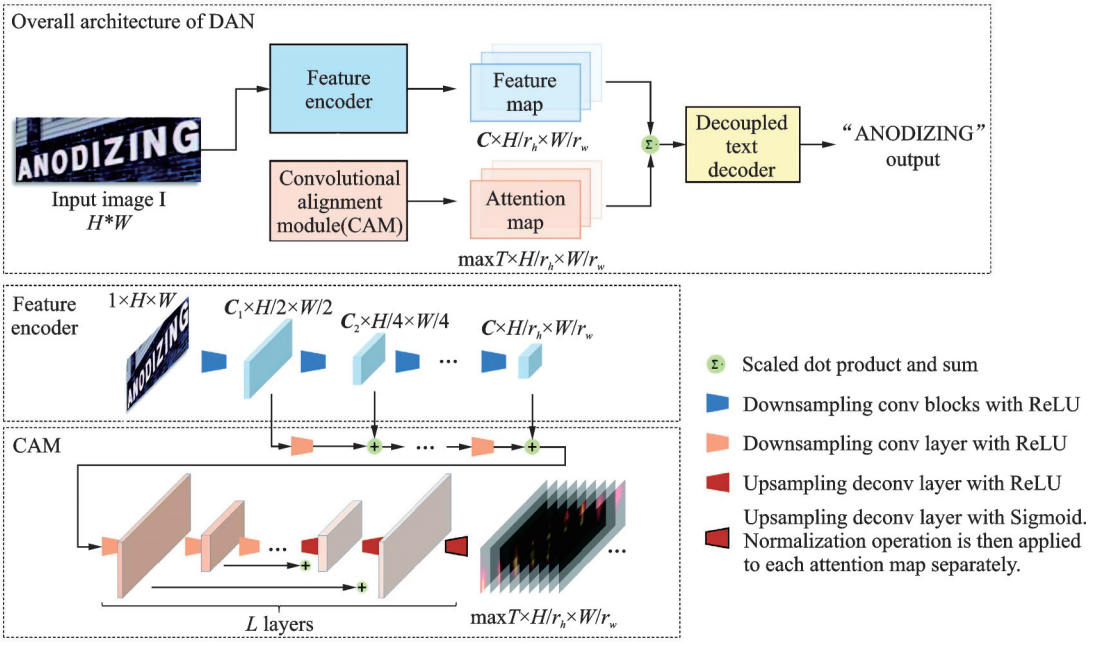

图2 DAN的网络结构图[8]

Fig.2 Overall architecture of DAN [8]

DAN算法的不足之处在于CAM模块在得到注意力图时仅仅使用视觉特征信息，并不能完全保证准确定位到与当前解码位置相对应的文本图像特征序列，即其在形成注意力时缺乏序列特征信息的引导。其次，DAN在Decoder中使用GRU进行序列性解码，而GRU采用一些门控机制，不可避免会丢失一些信息。若应用于越南语这种字符数量多、字符间差异性较小的声调语言，识别准确率将大打折扣。

上述工作主要是聚焦于一些英文、中文等文字，而对于越南文字这种声调语言的场景文字识别鲜有相关的研究。本文从越南语手写文字识别的工作[24-26]获得一些灵感。但是，手写文字和场景文字差异性很大，比如字体风格、字体的连续性和字体背景等，因此这些工作也并不适用于场景越南文字的识别。

2 本文方法

本节主要介绍本文提出的算法框架，并详细介绍了VSFM和ETDM模块。

2.1 改进DAN算法

图3为改进DAN算法的网络结构图。本文在DAN算法的基础上,在CAM模块中加入本文设计的VSFM模块,使得CAM模块得到的注意力图序列性更强,进一步缓解注意力漂移问题,同时也使得算法对变音符号和拉丁字符有同样的关注度,并且增强两者间的关联性。同时,本文提出ETDM模块,替换掉DAN算法中的解耦文本解码器。该增强型解码器在解码每个时刻的字符时,不仅将GRU输出的结果送入分类器,同时还将在VSFM模块中利用Bi-GRU在竖直方向序列建模得到的特征以及原始的特征融合后输入分类器。为了适应输入解耦文本解码器的特征信息的形状,本文将max$T$调整为32。

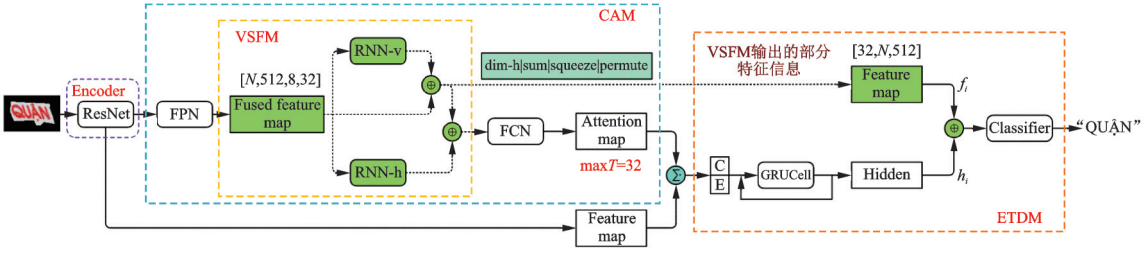

图3 改进DAN的网络结构图

Fig.3 Network structure diagram of the modified DAN

2.2 VSFM模块

VSFM的结构如图4所示。本文将VSFM模块直接嵌入至CAM模块中。在VSFM模块中,本文将FPN网络层得到的特征信息(视觉特征Fused feature map)分别输入两个不同方向的Bi-GRU(即图4中的RNN),其中RNN-v代表使用双向门控循环单元(Bidirectional gated recurrent unit,Bi-GRU)对视觉特征进行竖直方向的序列建模(以特征信息的h维度为时间步长),RNN-h代表使用Bi-GRU对视觉特征进行水平方向的序列建模(以特征信息的w维度为时间步长)。最终,本文将原始的视觉特征、竖直方向的序列建模特征以及水平方向的序列建模特征融合(Element-wise add)后,输入FCN,进而得到注意力图。

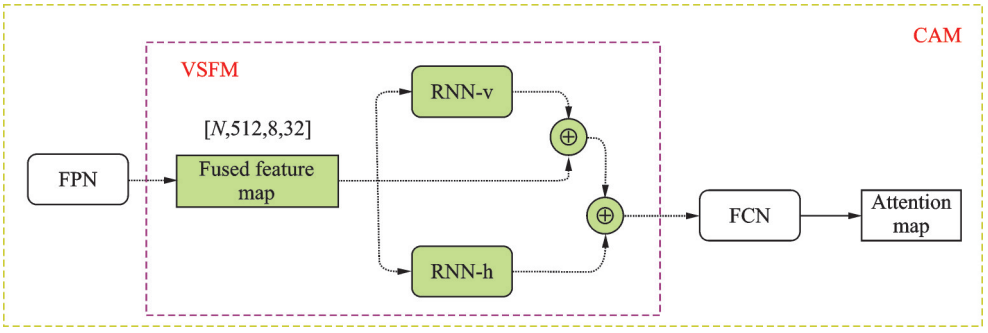

图4 VSFM结构图

Fig.4 Architecture of VSFM

本文使用两个不同方向的Bi-GRU对视觉特征进行不同方向的序列建模基于以下的考虑:在前述相关工作部分提到DAN算法的不足之处——在生成注意力图时缺乏序列信息的引导。由于字符是连续排列,因此本文使用GRU在视觉特征的水平方向进行序列建模;考虑到越南语字符的变音符号是位

于拉丁字符的上部和下部，本质上也是一种隐含的序列关系，因此本文使用GRU在视觉特征的竖直方向进行序列建模。

VSFM模块使用Bi-GRU在两个不同的方向进行序列建模，增强了特征信息间的时序性以及变音字符和拉丁字符间的关联性，同时也使得本文方法对字符的大小写更敏感，进一步缓解注意力漂移问题。

2.3 ETDM模块

由于DAN算法中解码器使用GRU并结合分类器进行解码，在对越南文字图像解码时，由于GRU含有门控机制，使用GRU后不可避免会丢失一些信息，而越南文字字符间的差异性较小（相似性高），这部分丢失的信息必然会对识别过程产生影响。因此，本文提出一种ETDM模块，其结构如图3中右边蓝色虚线框所示。在解码时，分类器的输入不仅为将上下文向量$\boldsymbol{C}$与前一时刻的解码向量$\boldsymbol{E}$输入GRU后得到的输出h_t，同时还包含VSFM中原始视觉特征和RNN-v输出的特征f_t，最终分类器的输入为h_t+f_t。该设计可以有效区分差异性较小的字符，提高识别准确率。考虑到h_t和f_t相加需有相同的维度，本文将max$T$的大小由原来的25调整为32。

3 实验验证

本节主要介绍实验所用的数据集，实验实施的相关细节，对本文所提出的两个模块的有效性验证及和其他算法的比较实验。

3.1 数据集

本文使用场景图像文字生成算法，结合在网络收集的越南语语料生成24 757张越南文字图片，并对其进行标注（生成数据集），具体示例如图5所示。同时，收集了200张在越南真实拍摄的图片，并将越南文字从图片中截取出来，共计3 261个图片。为了更好地验证本文算法的优势，本文制作了两种真实数据集：一种数据集（Data_no_bg）如图6所示，是将背景信息掩盖掉（用Labelme标注）；另一种数据集（Data_have_bg）如图7所示，保留了背景信息。字符类别（分类器类别数）共计198类，其中包括图1中的196类，以及未知字符类和结束符类（Unknown + End_token）类别。

图5 算法生成的数据集

Fig.5 Algorithm-generated dataset

图6 无背景的数据

Fig.6 Data without background

图7 有背景的数据

Fig.7 Data with background

3.2 实施细节

本文使用24 757张算法生成的越南文字数据集做预训练,然后将真实数据集进行随机划分,训练集共2 608个样本,测试集共653个样本。在所有模型训练中,本文采用Adadelta优化器,epochs设定为200,学习率为1.0,batch-size设定为256,解码器Decoder中分类器的类别数为198(196待识别文字字符+Unknown+End_token),评估指标为测试集的单词识别准确率(Word recognition accuracy,WRA),max$T$的设定稍有变化,在下文中详述。需进一步说明的是:本文在有背景的数据和无背景的数据上都做了相关的实验。一方面是因为在DAN算法中设计了类似的鲁棒性实验,另一方面是因为越南语字符含有变音符号,此时场景文字图片中会存在着更多的干扰信息,为了简单探究这种干扰信息的影响,本文对这两种数据进行了实验。

除此之外，由于本文方法是专门针对越南场景文字识别这一任务设计，所设计的模块也是以越南语字符的构造为出发点，因此本文并未在一些公开数据集，比如ICDAR 2013[27]、ICDAR 2015[28]或SVT[29]等数据集上验证本文方法的性能。

3.3 VSFM有效性验证

为了验证VSFM的有效性，本文做了两组实验：一组为在Baseline算法（DAN）的基础上，将max$T$设置为25，分别验证融合VSFM模块中不同序列特征的有效性；另一组实验为探究加入VSFM模块后，将max$T$分别设置为25和32的影响。需说明的是，最终本文方法是将max$T$参数调整为32，这是因为在ETDM中两种不同的特征信息相加，需要相同的维度形状，而在进行VSFM模块相关实验时并未加入本文设计的ETDM模块，因此将max$T$设置为25和32都可以进行。

在第1组实验中，主要是验证以下方面：利用Bi-GRU在水平和竖直方向的序列建模特征的有效性，即在DAN中的CAM模块中加入利用Bi-GRU在水平和竖直方向的序列建模的特征（CAM+RNN-h/RNN-v），同时结合两者的有效性（CAM+RNN-hv），融合原始视觉特征的有效性（即表1中RNN-h-cat、RNN-v-cat）以及VSFM（既融合水平和竖直方向的序列特征又融合视觉特征，CAM+VSFM）的有效性。

第1组实验的WRA结果如表1所示。从表1中可以看出，无论是有背景的数据，还是无背景的数据，在分别结合不同方向的序列建模特征后（CAM+RNN-h/RNN-v），单词识别准确率提升近3%，若同时结合两种方向的序列特征（CAM+RNN-hv），识别准确率同样提升3%，说明融合不同方向的序列特征去形成注意力图有助于识别。除此之外，可以看到在融入原始的视觉特征后（CAM+RNN-h-cat/RNN-v-cat），单词识别准确率再次提升，而最终结合VSFM后（CAM+VSFM），单词识别准确率相较于Baseline提升近4%。

第2组实验是在加入VSFM模块后，将max$T$调整为32，WRA结果会稍微提升一些，如表1所示。从表1中可以发现：无背景数据集的测试准确率总是相较于有背景数据集的测试准确率高一些，这说明背景对文字识别有一定影响，尤其是越南语这种声调语言，会引入更多的多余背景信息。

表1 VSFM有效性验证的实验结果

Table 1 Experimental results of VSFM effectiveness verification %

Dataset	DAN	CAM+RNN-h	CAM+RNN-v	CAM+RNN-hv	CAM+RNN-h-cat	CAM+RNN-v-cat	CAM+VSFM (max$T$=25)	CAM+VSFM (max$T$=32)
Data_no_bg	70.4	73.2	73.4	73.4	74.0	73.8	74.4	74.9
Data_have_bg	69.5	72.6	73.2	73.3	73.4	73.5	73.8	74.5

为了验证加入VSFM后可以进一步缓解注意力漂移问题的发生，本文对注意力图进行了可视化，如图8所示。示例①中，DAN算法在t=3和t=4时，发生注意力漂移，因此多识别出一个字符“T”。而加入VSFM后，注意力图也是正确的，并识别出正确结果，除此之外，t=2时，DAN算法并未关注到变音符号，导致其识别结果为“E”，而在t=3时错误地关注到变音符号，但是在类别中并未有带变音符号的字符“T”，因此其识别结果变为“T”，而加入VSFM后，可以明显看到在t=2时正确关注到变音符号，所以识别为“É”；示例②中可以看到，加入VSFM，识别出的结果为“w”，而DAN算法的结果为“W”，这说明VSFM对字符的笔画、大小写较敏感。同时在t=3时，DAN算法发生注意力漂移，而加入VSFM后，生成了正确的注意力图序列；在示例③中，t=2时刻，从注意力图中可以看到对变音符号的关注度相

图8 注意力图可视化

Fig.8 Visualization of attention maps

较于拉丁字符是弱的(颜色稍淡一些),因此其识别结果为“Y”,当加入VSFM后,从注意力图中看到变音符号和字符“Y”有同样的关注度(颜色相近),因此其识别出正确结果“Ý”。示例④和示例⑤同样如此,并且示例④也显示出变音符号对注意力的影响。因此可以得到:结合VSFM后,算法可以有效缓解注意力漂移问题的发生,同时对变音符号同样有较大的关注度,并加强了变音符号和拉丁字符间的关联性,因此在识别时正确识别出了变音符号。

图9是DAN算法和加入VSFM后的识别结果示例,本文可以进一步发现VSFM对字符笔画以及大小写的敏感性,对相似的字符有更准确的识别度,以及对注意力漂移的鲁棒性,同时对变音符号的识别也更准确。

3.4 ETDM有效性验证

ETDM的“亮点”在于其在分类时引入丰富尺度的视觉特征和竖直方向的序列建模特征。为了验证ETDM有效性,本文将DAN算法中的解码器替换为ETDM模块后进行相关的比较实验。在做该实验时,本文并未添加完整的VSFM模块,只是在CAM模块中加入竖直方向的序列建模(RNN-v),并将

图9 识别结果示例

Fig.9 Examples of recognition results

其输出特征与FPN输出的特征图融合后输入本文设计的解码器。实验结果如表2所示，使用ETDM模块相较于Baseline算法在无背景数据集上单词识别准确率有4%的提升，而在有背景的数据集上准确率有4.4%的提升。实验结果可以说明在分类时结合视觉特征和利用Bi-GRU进行竖直方向序列建模的特征信息可以有效提升模型对越南语字符的识别能力，增加更多丰富的特征信息，使得分类器的分类能力更强；同时也利用竖直方向序列建模的优势——增强变音符号和拉丁字符的关联性以及对字符笔画和大小写的敏感性，更有效地识别相似字符。

表2 ETDM有效性验证实验结果(max*T*=32)

Table 2 Experimental results of ETDM effectiveness verification (max*T*=32) %

Dataset	DAN	ETDM
Data_no_bg	71.1	75.5
Data_have_bg	70.8	75.2

3.5 算法比较

表3为本文方法和其他算法的比较结果。从表3中可以看出无论是在有背景的真实数据集还是在无背景的数据集，本文方法都表现出良好的性能，其优势正是在于其可以有效适应越南语字符的构成，准确识别变音符号，对相似的字符有较强的分类性能，同时可以缓解注意力漂移问题的发生。除此之外，从表3可见，DAN与其他几个方法相比，在越南场景文字数据集所表现的单词识别准确率相差比较大，这与在通用数据集上的表现似乎不太一致。本文认为原因有以下两个方面：一方面为CRNN、ASTER、SEED这几个方法在对输入图像“编码”时，将卷积网络得到的视觉特征输入双向长短时记忆网络(Bidirectional long short-term memorynet work，Bi-LSTM)，使得特征信息含有丰富的上下文信息，然而在面对含有变音符号的越南语字符时更容易产生错误信息的积累和传播，导致识别性能退化，而DAN算法在“编码”时并未利用Bi-LSTM，减少错误信息的积累的风险，并且其在CAM模块中利用FPN融合丰富的上下文信息。另一方面，这几个算法对避免注意力漂移或者说在每个时刻将待识别的区域与所对应的区域特征对齐的能力不同，在面对由于变音符号更易发生注意力漂移的越南场景文字时，这种能力的差异被放大，所导致性能的差异更加明显。

表3 与其他算法的性能比较

Table 3 Performance comparison with other algorithms %

Dataset	CRNN	ASTER[17]	SEED[18]	DAN	Ours
Data_no_bg	51.7	62.5	41.5	71.1	78.3
Data_have_bg	50.7	60.8	42.7	70.8	76.9

4 结束语

本文提出一种自然场景中越南文字的识别方法，在DAN算法的基础上，设计了VSFM和ETDM模块。VSFM模块通过在视觉特征的水平方向和竖直方向进行序列建模以进一步缓解注意力漂移问

题的发生,同时增强变音符号和拉丁字符的关联性,ETDM模块通过在分类时引入更多的特征信息,有效识别相似字符,对繁多的越南语字符有较强的分类能力。实验结果也验证了本文方法的有效性。

参考文献:

[1] JADERBERG M, KAREN S, ANDREA V, et al. Deep structured output learning for unconstrained text recognition[C]//Proceedings of International on Learning Representations. Ithaca, NY: [s.n.], 2015: 1-10.

[2] 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.

[3] SHI Baoguang, BAI Xiang, YAO Cong. 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.

[4] SHI Baoguang, WANG Xinggang, LYU Pengyuan, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4168-4176.

[5] LI Hui, WANG Peng, SHEN Chunhua, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 8610-8617.

[6] HUANG Yunlong, SUN Zenghui, JIN Lianwen, et al. EPAN: Effective parts attention network for scene text recognition[J]. Neurocomputing, 2020, 376: 202-213.

[7] CHENG Zhanzhan, BAI Fan, XU Yunlu, et al. Focusing attention: Towards accurate text recognition in natural images[C]//Proceedings of the IEEE International Conf. on Computer Vision. Piscataway, NJ: IEEE, 2017: 5076-5084.

[8] WANG Tianwei, ZHU Yuanzhi, JIN Lianwen, et al. Decoupled attention network for text recognition[C]//Proceedings of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2020: 12216-12224.

[9] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 886-893.

[10] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 1999: 1150-1157.

[11] THILOU C, FERREIRA S, GOSSELIN B. An embedded application for degraded text recognition[J]. EURASIP Journal on Advances in Signal Processing, 2005, 2005(13): 2127-2135.

[12] LIU Hu, JIN Sheng, ZHANG Changshui. Connectionist temporal classification with maximum entropy regularization[C]//Annual Conference on Neural Information Processing Systems (NIPS). Montréal, Canada: Curran Associates Inc., 2018: 831-841.

[13] FENG Xinjie, YAO Hongxun, ZHANG Shengping. Focal CTC loss for Chinese optical character recognition on unbalanced datasets[J]. Complexity, 2019, 2019(1): 1-11.

[14] HU Wenyang, CAI Xiaocong, HOU Jun, et al. GTC: Guided training of CTC towards efficient and accurate scene text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Menlo Park, CA: AAAI, 2020: 11005-11012.

[15] LEE Chenyu, OSINDERO S. Recursive recurrent nets with attention modeling for OCR in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2231-2239.

[16] LIU Zichuan, LI Yixing, REN Fengbo, et al. Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 7194-7201.

[17] SHI Baoguang, YANG Mingkun, WANG Xinggang, et al. ASTER: An attentional scene text recognizer with flexible rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2035-2048.

[18] QIAO Zhi, ZHOU Yu, YANG Dongbao, et al. SEED: Semantics enhanced encoder-decoder framework for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:

IEEE, 2020: 13528-13537.

[19] YUE Xiaoyu, KUANG Zhanghui, LIN Chenhao, et al. RobustScanner: Dynamically enhancing positional clues for robust text recognition[C]//Proceedings of European Conference on Computer Vision. Glasgow, UK: Springer International Publishing, 2020: 135-151.

[20] 刘崇宇, 陈晓雪, 罗灿杰, 等. 自然场景文本检测与识别的深度学习方法[J]. 中国图象图形学报, 2021, 26(6): 1330-1367.
LIU Chongyu, CHEN Xiaoxue, LUO Canjie, et al. Deep learning methods for scene text detection and recognition[J]. Journal of Image and Graphics, 2021, 26(6): 1330-1367.

[21] LIN T, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[22] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3431-3440.

[23] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

[24] Nguyen D K, Bui T D. Recognizing Vietnamese online handwritten separated characters[C]//Proceedings of 2008 International Conference on Advanced Language Processing and Web Information Technology. Piscataway, NJ: IEEE, 2008: 279-284.

[25] LE A D, NGUYEN H T, NAKAGAWA M. Recognizing unconstrained Vietnamese handwriting by attention-based encoder decoder model[C]//Proceedings of 2018 International Conference on Advanced Computing and Applications (ACOMP). Piscataway, NJ: IEEE, 2018: 83-87.

[26] LY V L, TUAN D, NGOC Q L. Transformer-based model for Vietnamese handwritten word image recognition[C]//Proceedings of 2020 7th NAFOSTED Conference Information and Computer Science (NICS). Piscataway, NJ: IEEE, 2020: 163-168.

[27] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]//Proceedings of 2013 12th International Conference on Document Analysis and Recognition. Piscataway, NJ: IEEE, 2013: 1484-1493.

[28] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]//Proceedings of 2015 13th International Conference on Document Analysis and Recognition. Piscataway, NJ: IEEE, 2015: 1156-1160.

[29] WANG K, BABENKO B, BELONGIE S. End-to-end scene text recognition[C]//Proceedings of 2011 International Conference on Computer Vision. Piscataway, NJ: IEEE, 2011: 1457-1464.

作者简介:

王利兵(1996-),男,硕士研究生,研究方向:人工智能、计算机视觉。

俸亚特(1996-),男,硕士研究生,研究方向:人工智能、计算机视觉。

文益民(1969-),**通信作者**,男,博士,教授,研究方向:机器学习与数据挖掘、推荐系统、计算机视觉,E-mail: ymwen2004@aliyun.com。

(编辑:刘彦东)