ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 665－675
DOI:10. 16337/j. 1004-9037. 2023. 03. 015

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

视觉注意与语义感知联合推理实现场景文本识别

佟国香，董田荣，胡珩彰

（上海理工大学光电信息与计算机工程学院，上海 200093）

摘　要：场景中的不规则文本识别仍然是一个具有挑战性的问题。针对场景中的任意形状以及低质量文本，本文提出了融合视觉注意模块与语义感知模块的多模态网络模型。视觉注意模块采用基于并行注意的方式，与位置感知编码结合提取图像的视觉特征。基于弱监督学习的语义感知模块用于学习语言信息以弥补视觉特征的缺陷，采用基于Transformer的变体，通过随机遮罩单词中的一个字符进行训练提高模型的上下文语义推理能力。视觉语义融合模块通过选通机制将不同模态的信息进行交互以产生用于字符预测的鲁棒特征。通过大量的实验证明，所提出的方法可以有效地对任意形状和低质量的场景文本进行识别，并且在多个基准数据集上获得了具有竞争力的结果。特别地，对于包含低质量文本的数据集SVT和SVTP，识别准确率分别达到了93.6%和86.2%。与只使用视觉模块的模型相比，准确率分别提升了3.5%和3.9%，充分表明了语义信息对于文本识别的重要性。

关键词：场景文本识别；不规则文本；视觉注意模块；语义感知模块；多模态

中图分类号：TP391　　**文献标志码**：A

Joint Inference of Visual Attention and Semantic Perception for Scene Text Recognition

TONG Guoxiang, DONG Tianrong, HU Hengzhang

(College of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Irregular text recognition in scenes is still a challenging problem. For arbitrary shapes and low-quality text in scenes, this paper proposes a multimodal network that combines a visual attention module and a semantic perception module. The visual attention module uses a parallel attention-based approach to extract visual features of images combined with positional encoding. The semantic perception module based on weak supervised learning is used to learn linguistic information to compensate for the deficiencies of visual features. The module uses a Transformer-based variant that improves the model's contextual semantic inference by randomly masking a character in a word for training. The visual semantic fusion module interacts information from different modalities through a gating mechanism to generate robust features for character prediction. The proposed approach is demonstrated through extensive experiments to be effective in recognizing arbitrarily shaped and low-quality scene text, and competitive results are obtained on several benchmark datasets. In particular, accuracy rates of 93.6% and 86.2% are achieved for the datasets SVT and SVTP, which contain low-quality text, respectively. Compared with

基金项目：国家重点研发计划(2018YFB1700902)。
收稿日期：2022-11-23；**修订日期**：2023-03-21

the method containing only the visual module, the accuracy is improved by 3.5% and 3.9%, respectively, which fully demonstrates the importance of semantic information for text recognition.

Key words: scene text recognition; irregular text; visual attention module; semantic perception module; multimodal

引　言

作为计算机视觉领域的研究热点,场景文本识别在自动驾驶、路标识别、产品检索等场景下具有广泛的应用。场景文本识别方法[1-11]在深度学习领域的推动下取得了比较好的进展。然而,由于背景复杂、字体多变、形状不规则、分辨率低、光照不可控以及遮挡等因素的影响,自然场景下的文本识别仍然是一项具有挑战性的任务。

近年来,大多数方法[1-3]将规则文本识别建模为序列学习问题并取得了显著进展。与规则文本相比,不规则文本如任意形状文本或者低质量文本的识别更具有挑战性。现有的不规则文本识别方法大致可以概括为基于校正的方法[4-6]、基于多方向编码的方法[7]、基于字符级监督的方法[8]和基于2D注意的方法[10-11]。基于校正的方法使用空间变换网络对文本行进行校正或者对单个字符进行校正[12]。基于多方向编码的方法中,Cheng等[7]采用基于序列的模型从4个方向提取并融合图像特征,但这种方法会造成冗余的表示。基于字符级监督的方法[8]使用字符级注释对模型进行训练。基于2D注意的方法[9]在特征地图上应用2D注意机制处理不规则文本。上述对不规则文本进行识别的方法需要将输入图像转换为中间序列表示,但将分布在二维空间的任意形状文本转换为一维特征序列进行处理会导致大量信息丢失。

不规则文本识别任务的难点,主要来源于文本形状的不规则以及诸如模糊、遮挡、低对比度等问题。现有的不规则文本识别方法更倾向于任意形状的处理,忽略了低质量文本的处理问题。采用升级主干网络、添加修正模块、改进注意力机制等方法可以从提取更有效的视觉特征的角度提高文本识别的推理能力。但所有这些尝试只丰富了视觉特征,很大程度上缺乏语义推理能力。对于低质量文本而言,仅使用视觉特征提高识别准确率的效果受到一定限制。引入语言模型可以通过学习语言信息帮助字符预测。然而,大多数文本识别方法[5,7]将视觉模型和语言模型耦合,导致不能有效地学习固定模态的信息。Qiao等[13]提出了使用单独的语言模块进行处理。SRN[14]使用全局语义模块用于语言文本建模。SEED[13]使用预训练的语义模块提供语义特征,该语义模块的不可微具有很大的局限性。JVSR[15]提出的联合语义模块中当前阶段的解码器依赖于前一阶段的输出导致效率低下。

针对上述问题,本文基于Transformer分别设计了视觉注意模块与语义感知模块,多模态信息的有效融合增强了文本识别能力。本文的贡献主要有:

(1)使用视觉注意模块和语义感知模块分别学习特定的模态知识,语义感知模块提取的语义信息可以辅助视觉注意模块进行更准确的识别。

(2)视觉注意模块采用基于2D注意的方式,直接对不规则文本进行识别,无需额外的校正方法,此外,加入位置编码解决注意力漂移的问题。

(3)在语义感知模块中提出了一种基于弱监督训练的方式,利用文本语义信息来学习有效的视觉文本表示。通过随机遮挡字符训练模型增强了对于模糊文本识别的鲁棒性。

(4)为了评估模型对于低质量文本的识别效果,构建了一个低质量文本数据集STD。与先进的SRN[14]在STD上进行实验对比,本文提出的视觉语义联合推理的文本模型在准确率上提升了6.1%。

1 本文方法

1.1 网络结构

本文所提出的方法由视觉注意模块、语义感知模块以及视觉语义融合3部分组成,整体结构如图1所示。假设模型的输入图像是对自然场景中图片的文本区域进行裁剪得到的,输出是对裁剪图像识别得到的文本字符串。对于给定输入图像,首先将其高度和宽度分别调整为32和128。然后,通过带有残差连接的卷积神经网络ResNet45提取高级视觉特征,将Transformer单元作为序列建模网络堆叠在ResNet45的顶部,以捕获像素的长距离相关性。视觉注意模块将提取的二维特征与位置编码结合,通过字符感知模块捕获对齐的视觉信息。语义感知模块将视觉模块的输出与随机索引作为输入,通过对字符进行遮罩模拟视觉模糊的情况,以弱监督的方式学习语言信息。随机索引用来指示被遮挡字符的索引,其值在$[0,N_L]$之间随机取得,N_L表示输入图像中单词的长度。最后将视觉注意模块和语义感知模块通过门控机制融合以预测N个字符。对长度小于N的字符进行填充。

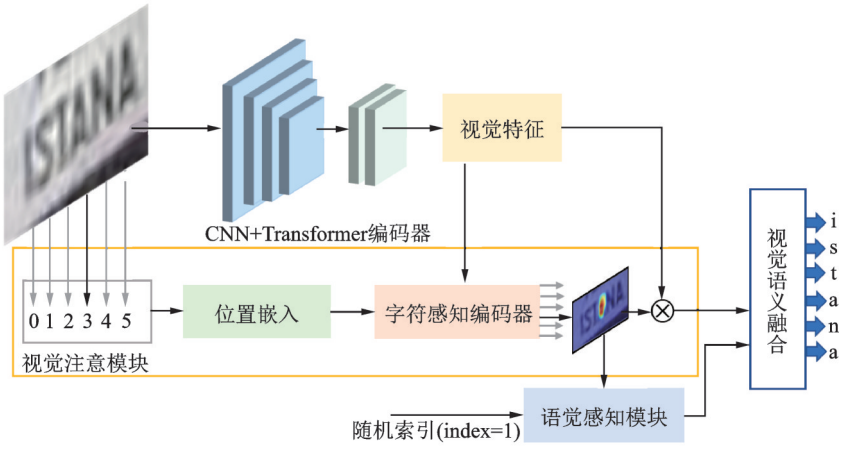

图1 网络整体结构

Fig.1 Overall network structure

1.2 视觉注意模块

注意力机制广泛应用于序列识别,对于文本序列$T=\{c_0,c_1,c_2,\cdots,c_{N-1}\}$,可以生成$N$个特征,每个特征对应序列中的一个字符。现有的基于注意力的方法在解码过程中严重依赖之前步骤的预测,不仅效率低下,而且对于无上下文语义的文本实例会出现注意力漂移[16]的问题。

受Lyu等[10-11,14]的启发,本文的视觉注意模块由位置嵌入模块和字符感知模块两部分组成。对于给定的图像X,采用多级特征融合策略[17]提取视觉特征$\boldsymbol{V}_h\in\mathbf{R}^{H\times W\times C}$,同时利用低级特征的局部细节以及高层特征信息。$H$和$W$分别表示图像的长和宽,$C$表示通道数。为了使字符感知模块可以适用于任意形状的输入,将视觉特征向量$\boldsymbol{V}_h$映射为形状为$l\times c$的特征序列I。l表示特征序列的长度,c表示特征序列I中每个特征向量的维度。对于每个特征向量$\boldsymbol{I}_i$,通过位置嵌入模块将字符的位置索引$N=\{0,1,2,\cdots,N-1\}$编码为位置嵌入向量$\boldsymbol{PE}=\{PE_0,PE_1,PE_2,\cdots,PE_{N-1}\}$,该位置嵌入向量$\boldsymbol{PE}$与特征向量$\boldsymbol{I}_i$的维度保持一致。之后,将特征序列$I$与位置嵌入向量$\boldsymbol{PE}$拼接形成位置敏感的视觉特征$V_f$。字符感知模块将$V_f$作为键值对,字符位置索引作为查询进行并行处理,如式(1)所示。

$$F_v = \mathrm{Soft\,max}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{C}}\right)V \tag{1}$$

式中：$\boldsymbol{K}=\boldsymbol{G}\left(V_f\right)\in\mathbf{R}^{\frac{HW}{16}\times C}$，由 U-Net 实现；$\boldsymbol{V}=\boldsymbol{R}\left(V_f\right)\in\mathbf{R}^{\frac{HW}{16}\times C}$ 表示恒等映射；Q 表示字符的顺序编码；T 表示文本实例的长度。

由于自然场景中的一些文本只是字符的简单组合，并没有明确的语义信息，因此对这些文本进行上下文语义推理会导致错误。视觉注意模块可以捕获文本实例的视觉注意力，通过字符感知编码器对文本进行字符级建模，有效地编码了每个字符的文本信息，并且帮助更好地学习视觉文本表示。图 2 展示了注意力图的可视化结果，可以看出字符感知编码器更好地关注到每个字符，这证明了所提出的编码器在学习视觉文本特征方面的优越性。它可以提取每个字符更精确的特征，从而更好地约束特征序列，有利于语义推理模块的上下文建模。

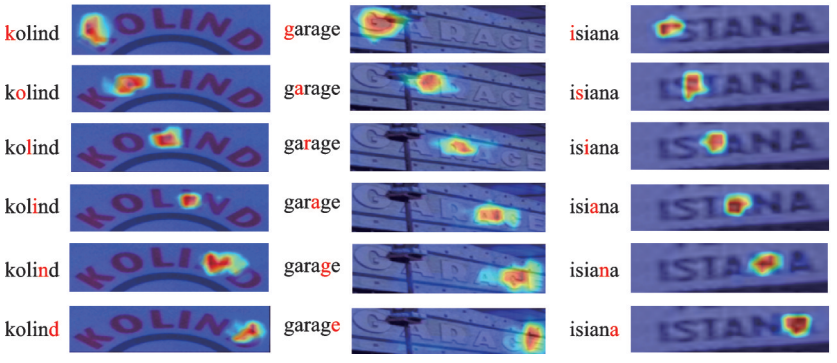

图 2 注意力图

Fig.2 Attention map

1.3 语义感知模块

大部分场景文本识别方法通过注意力机制将视觉模块和语言模块耦合，阻碍了特定的模态信息学习。由于注意力解码器的自回归性质：只有之前步骤预测的字符才能为下一步推理提供语义上下文，从而使语义信息在推理过程中单向流动。不难看出，一个错误预测将对之后的步骤产生累积的负面影响。因此，基于单阶段的注意力解码器无法有效地进行语义推理。受 BERT 中的掩码语言模型[18]和 simMIM[19]的启发，本文提出了语义感知模块以学习双向语义推理能力。如图 3 所示，将视觉注意模块输出的视觉特征 V_h 通过 Transformer 单元获得增强的视觉特征，与随机字符索引 index 集成，之后连接 Sigmoid 层生成遮挡字符的掩码图 mask。其中，字符索引 index 用来表示被遮挡字符的索引，该索引是随机的。视觉特征 V_h 与掩码图 mask 逐元素相乘得到文本掩码嵌入向量，随后，将视觉特征作为键和值，文本掩码嵌入向量作为查询，输入到 Transformer 预测文本实例中的遮罩字符。

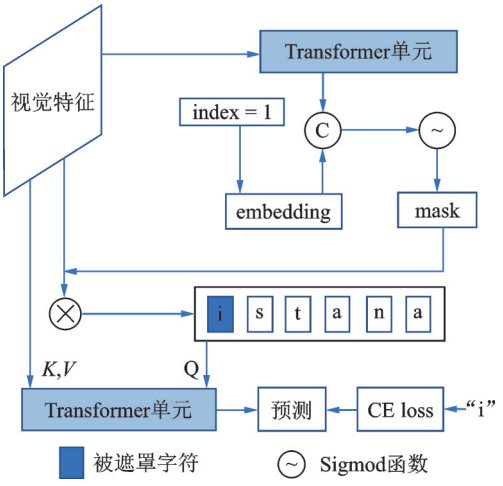

图 3 语义感知模块

Fig.3 Semantic perception module

对于第 i 个字符的推理过程需要完全依赖于其他字符的信息，而不包含当前字符信息。语义感知模块是

Transformer解码器的变体,每一层都包含一系列的多头注意、前馈神经网络、残差连接和层归一化,去掉了Transformer中的自注意力层以避免跨时间步的信息泄露。单个字符的推理过程如式(2)和式(3)所示。

$$\text{attn}=\text{Softmax}\left(\boldsymbol{W}_k^{\mathrm{T}}\tanh\left(\boldsymbol{W}_p P_t+\boldsymbol{W}_v V_h\right)\right) \tag{2}$$

$$F_{\mathrm{s}}=\text{attn}_t V_{\text{fea}} \tag{3}$$

式中:$\boldsymbol{V}_h$表示视觉特征图;attn_t为注意力图;P_t表示位置索引编码;$\boldsymbol{W}_k$、$\boldsymbol{W}_p$、$\boldsymbol{W}_v$均代表权重矩阵;F_{s}为输出的语言特征。通过借鉴类似于完型填空的方式对遮罩字符进行预测,语义感知模块可以学习到更多的特征信息,由于采用了类似于Transformer的体系结构,所提出的语言模块只需要一半的计算和参数,因此效率会更高。

大多数文本行中包含丰富的语义信息,因此采用语义感知模块对单词文本上下文进行建模具有重要的意义。对于一个单词中存在遮挡或者模糊的单词字符,语义感知模块可以借助单词中没有被遮挡或者不模糊的字符进行上下文语义推理,从而推理出遮挡或模糊字符的正确结果。

1.4 视觉语义融合

基于图像的视觉注意模块和基于文本的语义感知模块属于不同的模态,针对不同的情况应该分配不同的权重。因此,通过选通机制引入一些可训练的权重来平衡不同模态特征的贡献。具体可以表示为

$$M=\sigma\left(\left[F_{\mathrm{v}},F_{\mathrm{s}}\right]\boldsymbol{W}_m\right) \tag{4}$$

$$F_t=F_{\mathrm{v}}\odot M+F_{\mathrm{s}}\odot(1-M) \tag{5}$$

式中:$\boldsymbol{W}_m$为权重矩阵;σ为Sigmoid函数;F_{v}为视觉特征;F_t表示第t个文本向量的视觉语义融合特征。

1.5 方法优化

采用式(6)对所提出方法进行优化。

$$L=\alpha_1 L_v+\alpha_2 L_m+\alpha_3 L_f \tag{6}$$

式中:L_v、L_m、L_f分别代表视觉注意模块、语义感知模块以及视觉语义融合模块的损失函数,均采用交叉熵损失$L^{*}=-\frac{1}{N}\sum_{t=1}^{N}\lg\left(p_t\left|g_t\right.\right)$进行优化。$N$代表文本实例输出的最大长度,实验中设置为25。$\alpha_1$、$\alpha_2$、$\alpha_3$表示平衡因子,在本文的实验中分别设置为1.0、0.5和1.0。

2 本文实验

2.1 实验设计

本文模型在单个RTX 3090GPU上进行训练和测试。使用ResNet45作为主干网络进行特征提取,在第2、3、4阶段将步长设置为2。输入图像的宽度为32,高度为128。数据增强包括随机旋转、仿射变换、颜色抖动和透视失真。采用Adam优化器对模型进行训练,批次大小设置为128,初始学习率设置为$1e^{-4}$。分为37个类,分别是a~z、0~9和序列结束符号'EOS'。输出序列的最大长度N设置为25。

本文使用了两个合成数据集:MJSynth和SynthText对网络进行训练。在6个基准数据集上进行评估,6个基准数据集分别为ICADR2013、ICADR2015、IIIT5K、SVT、SVTP、CUTE。为了验证所提出模型对于低质量场景文本的识别能力,从6个基准数据集中收集了2 000张图像,并对这些图像进行遮挡处理以形成一个新的数据集STD。具体地,STD由1 000张低分辨率(低于16像素×64像素)图像和1 000张正常图像组成。该数据集中的图像采用手动遮挡的方式进行处理,对于每张图像,随机选择一个角度用一条线对其中的一个字符进行遮盖,图4展示了STD的部分图片。

2.2 视觉注意模块的参数影响

为了验证视觉注意模块中Transformer单元数量对于实验结果的影响，设计了不同数量的Transformer单元进行比较，计算出具有几层堆叠的Transformer单元可以最大化视觉模块的潜力。如图5所示，可以看出，Transformer单元的数量设置为3时在各个数据集上的准确率达到最高。当继续增加Transformer单元的数量时，模型的性能下降或者保持不变，并且推理时间会增加。具体地，当数量为4时，在IC13上的准确率下降了0.8%，在CUTE上的准确率下降了0.7%。

2.3 损失函数的超参数分析

使用了3个损失函数对所提出模型进行优化。超参数α_1、α_2、α_3用于平衡视觉注意损失、语义感知损失和视觉语义融合损失所占的权重。根据经验，通常情况下超参数α_1、α_2、α_3均简单地设置为1。但是，考虑到自然场景中部分文本行仅仅是字符的简单组合，并不包含语义信息，若使用语义信息推理会导致识别错误，因此通过降低语义感知损失的权重来进行平衡，以降低语义感知对于视觉感知的影响能力。将α_2分别设置为0.3、0.5、0.8和1.0进行实验对比，见表1。

图4 STD数据集的一些示例

Fig.4 Some examples of STD dataset

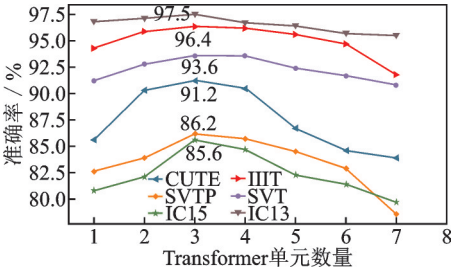

图5 Transformer单元数量的影响

Fig.5 Effect of the number of Transformer units

表1 语义感知损失在不同权重下的实验结果

Table 1 Experimental results of semantic perceptual loss under different weights %

α_2	IC13	SVT	IIIT	IC15	SVTP	CUTE
0.3	96.7	92.1	95.9	82.7	84.4	89.8
0.5	97.5	93.6	96.4	84.6	86.2	91.2
0.8	97.3	93.3	96.3	84.1	85.5	90.8
1.0	97.1	92.8	96.0	83.9	85.3	90.9

2.4 消融实验分析

2.4.1 不同模块的消融实验

为了验证所提出的视觉注意模块和语义感知模块的有效性。将视觉注意模块和语义感知模块分开进行训练，使两个模块各自拥有稳定的学习过程。接下来的实验部分，本文所提出的模型采用VSNet表示，视觉注意模块和语义感知模块分别使用VM和SM表示。表2展示了不同模块下的文本识别准确率。可以看出，单独使用视觉注意模块或者语义感知模块得到的效果明显弱于视觉语义联合推理的性能。特别是对于包含低质量文本的数据集SVT和SVTP，加入语义感知模块可以明显提高识别准确率，与仅使用视觉注意模块相比，联合推理在数据集SVT和SVTP上的准确率分别提高了2.0%和3.8%。

表2 消融实验

Table 2 Ablation experiments %

方法	IC13	SVT	IIIT	IC15	SVTP	CUTE
VM	96.3	91.6	95.3	81.4	82.4	88.3
SM	96.1	90.2	94.4	81.6	82.6	89.7
VSNet	97.5	93.6	96.4	84.6	86.2	91.2

2.4.2 语义感知模块的有效性

在本节中，将本文所用的语义感知模块与SRN[14]提出的语言模块进行了比较，为了得到更公正的结果，二者的视觉模块均采用本文的视觉注意模块。与SRN[14]提出的语言模块相比，本文的语义感知模块SM引入了具有遮罩的掩码图进行训练，这种基于弱监督的方式有利于语义建模，可以更好地学习语义信息推理。从表3可以看出，语义感知模块SM在准确率上优于SRN[14]提出的语言模块，表明了该

模块的有效性与设计的合理性。

此外,为了评估语义感知模块SM的鲁棒性,将SM与经典方法ASTER[5]、2D[10]以及HRGA[11]融合之后进行评估,具体结果见表4。可以看出,各个模型在加入语义感知模块后的准确率明显得到了提升。这是因为语义感知模块采用模拟完形填空的方式进行推理学习,建立了全局语义推理能力,不仅加强了视觉感知模块的识别能力,而且在视觉感知模块产生错误预测时可以通过上下文语义推理进行纠正。

表3 语义感知模块与SRN语言模块的对比

Table 3 Comparison of semantic-aware module and language module of SRN %

方法	IC13	SVT	IIIT	IC15	SVTP	CUTE
SRN[14]	95.9	92.6	95.3	83.3	85.4	86.9
SM	96.6	92.3	95.9	83.6	86.1	89.2

表4 语义感知模块的鲁棒性

Table 4 Robustness of semantic perception module %

方法	IC13	SVT	IIIT	IC15	SVTP	CUTE
ASTER	91.8	89.5	93.4	76.1	78.5	79.5
ASTER +SM	93.4	89.8	93.8	76.9	85.4	83.6
2D	92.7	90.1	94.0	76.3	82.3	86.8
2D +SM	95.4	93.7	96.4	79.3	85.1	89.9
HRGA	93.2	88.9	94.7	79.5	80.9	85.4
HRGA+SM	95.6	90.2	95.7	76.1	84.2	87.8

2.5 与先进方法进行比较

在6个数据集上将VSNet与当前先进的方法进行比较,结果见表5。实验结果表明,VSNet在6个基准数据集上的表现均优于其他方法。对于常规的文本数据集ICADR2013、SVT和IIIT5k,准确率分别达到了97.5%、93.6%和96.4%。在识别包含低质量文本的数据集SVT和SVTP时,所提出的VSNet表现出来很好的鲁棒性,说明语义感知模块对于低质量文本识别准确率的提升作出了贡献。

表5 所提出的VSNet与之前的方法进行比较的结果

Table 5 Results of the proposed method VSNet compared with the previous methods %

方法	IC13	SVT	IIIT	IC15	SVTP	CUTE
CRNN[4]	98.4	80.1	81.8	60.4	65.9	61.5
2D (并行)[10]	92.7	90.1	94.0	76.3	82.3	86.8
ViTSTR[20]	92.4	87.7	88.4	72.6	81.8	81.3
SRCAN[21]	91.3	88.1	93.3	74.0	80.2	85.1
ASTER[5]	91.8	89.5	93.4	76.1	78.5	79.5
HRGA[11]	93.2	88.9	94.7	79.5	80.9	85.4
TextScanner[22]	92.9	90.1	93.9	79.4	84.3	83.3
SEED[13]	92.8	89.6	93.8	80.0	81.4	83.6
PlugNet[23]	95.0	92.3	94.4	82.2	84.3	85.0
SRN[14]	95.5	91.5	94.8	82.7	85.1	87.8
JVSR[15]	95.5	92.2	95.2	84.0	85.7	89.7
SCATTER[24]	93.9	92.7	93.7	82.2	84.3	85.0
VSNet	97.5	93.6	96.4	84.6	86.2	91.2

本文的VSNet在没有使用校正模块等额外措施对不规则文本实例进行处理的情况下，在模糊、失真等低质量数据集上实现了更好的性能。与包含超分辨率单元的PlugNet[23]相比，VSNet的性能仍然是最优的。如1.3节所述，在对低质量文本单词进行识别时，语言感知模块对于单词中被遮挡字符的推理完全借助于其他清晰的字符信息，模型通过上下文语言语义进行了稳健的字符预测。因此，可以认为语义信息对视觉信息进行了重要补充，并且在面对较为困难的场景时，它会发挥出更有效的作用。

2.6 时间消耗分析

为了验证所提方法的效率，将VSNet与其他场景文本识别模型进行了评估。为了公平地进行比较，在相同的硬件平台上评估所有的方法，并将测试的批次大小设置为1。单张图片的时间消耗如表6所示。ASTER[5]先对文本进行校正，然后基于串行注意模型进行识别，效率较低。JVSR模型[15]以阶段性的方式进行细化，在进行下一个阶段的处理之前，之前的每个阶段都需要完全展开，因此效率也较低。尽管SRN[14]与本文所提出模型均使用了并行注意力的方式，但是所提出方法的语言模型采用基于Transformer改进的模型实现，SRN的语言模型采用了类似递归神经网络(RNN)的方式实现，虽然使用与时间无关的近似字符嵌入向量作出了改进，但是由于RNN本身固有的缺陷，与所提出模型相比SRN的效率相对要低。

表6 时间消耗实验对比

Table 6 Experimental comparison of time consumption

方法	注意力	单张图片时间消耗/ms
ASTER[5]	串行	37.4
JVSR[15]		26.1
SRN[14]	并行	24.8
VSNet	并行	24.4

2.7 校正模块的影响

所提出的模型直接对输入图像进行处理，不需要进行额外的校正。常规的文本识别方法需要将文本图像转化为一维特征序列，对于不规则文本，直接转化为一维特征序列会包含大量的无关文本的信息，因此需要将校正后的图像(图6)转化为一维特征序列，然后使用RNN对一维特征序列进行处理。由于不规则文本分布在二维特征空间，转换为一维特征序列会导致大量信息丢失。另外，对于严重弯曲或者失真的图像很难进行校正。因此在视觉注意模块考虑直接对二维特征图进行处理。考虑到RNN难以并行优化，而且存在梯度消失或者爆炸的问题，采用RNN直接对二维特征图进行处理并不是好的方法。因此在视觉注意模块中采用了2D注意力机制进行处理，该模块以编码器编码的二维特征图作为输入，然后通过注意力机制对图片的单个字符进行聚焦。从图7可以看出注意力机制可以准确

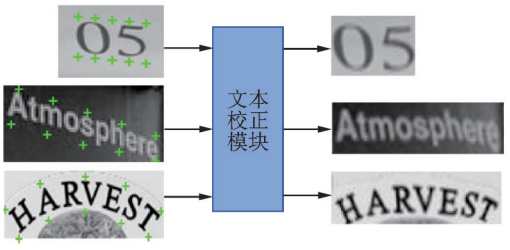

图6 文本校正示例

Fig.6 Examples of text rectification

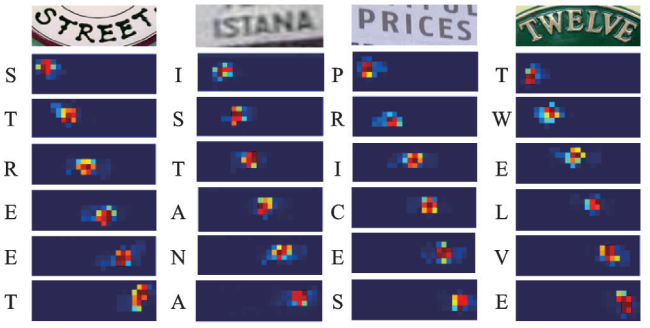

图7 二维空间注意地图

Fig.7 Attention map in two dimensions

地关注到分布在二维空间中的每个字符的所在区域,因此不需要对图像进行校正。另外,将校正模块与所提出的模型进行整合后进行预测,如表7所示,加入校正模块的模型并没有带来准确率的提升,反而在部分数据集上识别性能下降了,分析原因可能是校正模块将二维空间的文本转化为一维,导致文本扭曲以及大量信息的丢失。

表7 加入校正模块的影响

Table 7 Impact of the text rectification module %

方法	IC13	SVT	IIIT	IC15	SVTP	CUTE
VSNet	97.5	93.6	96.4	84.6	86.2	91.2
+校正	97.5	93.4	96.4	84.2	86.1	90.7

2.8 实验结果可视化分析

图8显示了仅使用视觉模块进行预测以及视觉语义模块联合进行预测的结果。可以看出,加入语义模块可以得到更精确的字符检测结果。特别地,视觉语义联合预测可以处理容易混淆的字符。例如,图8中的"istana"图像,由于小写的"l"和大写的"I"具有相似的视觉线索,因此仅仅使用视觉模块预测会给出错误的检测结果"l",而带有语言模块的模型可以通过上下文推理得出正确结果"I"。此外,针对图8中的被遮挡的以及模糊字符的正确识别也证实了语言模块的有效性。

图8 文本识别结果展示

Fig.8 Text recognition result display

为了更好地理解不同文本与图像的匹配效果,图9中计算了模型预测的单词与其他相似单词之间的余弦相似性,值越大表明相似性越高。可以看到,VSNet模型成功地学习了与当前图像相匹配的文本特征。例如,"pizza"和"przza"的拼写相似,编辑距离为1,但是"przza"不具有实际意义。借助语义感知模块,可以对被遮挡的"i"进行双向推理得到正确的结果。同样对于"yong"和"yomg"也是如此。

图9 预测的单词与其他单词的余弦相似性

Fig.9 Cosine similarity of the predicted word to other words

2.9 在STD数据集上的评估

为了更深入了解本文的VSNet语言能力，将其与包含语言模型的先进方法在构建的低质量数据集STD上进行比较。从表8可以看出，VSNet的准确率较高，比SRN[14]和PlugNet[23]分别高出了6.1%和2.8%，这表明了所提出的基于弱监督学习的语义感知模块的有效性。

表8 在STD数据集上的实验结果

Table 8 Results on STD dataset

方法	准确率/%
SRN[14]	54.1
PlugNet[24]	57.4
SEED[13]	54.3
VSNet	60.2

3 结束语

本文提出了融合视觉注意模块与语义感知模块的文本识别器，在视觉注意模块识别的基础上加强了模型的语言感知能力，使得模型不仅可以处理任意形状文本，还可以提升对低质量文本的识别能力。通过在6个基准数据集以及构建的低质量文本数据集上进行大量的实验证明了所提出方法的有效性，也表明语义推理能力对于文本识别的重要性。在未来的工作中，希望可以提升对于艺术字、手写文本的识别能力。

参考文献：

[1] BAI F, CHENG Z, NIU Y, et al. Edit probability for scene text recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 1508-1516.

[2] GAO Y, CHEN Y, WANG J, et al. Reading scene text with fully convolutional sequence modeling[J]. Neurocomputing, 2019, 339: 161-170.

[3] 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.

[4] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 4168-4176.

[5] SHI B, YANG M, WANG X, 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.

[6] LUO C, JIN L, SUN Z. Moran: A multi-object rectified attention network for scene text recognition[J]. Pattern Recognition, 2019, 90: 109-118.
[7] CHENG Z, XU Y, BAI F, et al. Aon: Towards arbitrarily-oriented text recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 5571-5579.
[8] CHENG Z, BAI F, XU Y, et al. Focusing attention: Towards accurate text recognition in natural images[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 5076-5084.
[9] LI H, WANG P, SHEN C, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019: 8610-8617.
[10] LYU P, YANG Z, LENG X, et al. 2D attentional irregular scene text recognizer[EB/OL].(2019-06-13)[2022-11-10]. http://arxiv.org/abs/1906.05708.
[11] YANG L, WANG P, LI H, et al. A holistic representation guided attention network for scene text recognition[J]. Neurocomputing, 2020, 414: 67-75.
[12] LIU W, CHEN C, WONG K Y. Char-net: A character-aware neural network for distorted scene text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018: 7154-7161.
[13] QIAO Z, ZHOU Y, YANG D, 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. [S.l.]: IEEE, 2020: 13528-13537.
[14] YU D, LI X, ZHANG C, et al. Towards accurate scene text recognition with semantic reasoning network-s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 12113-12122.
[15] BHUNIA A K, SAIN A, KUMAR A, et al. Joint visual semantic reasoning: Multi-stage decoder for text recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2021: 14940-14949.
[16] YUE X, KUANG Z, LIN C, et al. Robustscanner: Dynamically enhancing positional clues for robust text recognition[C]//European Conference on Computer Vision. Cham: Springer, 2020: 135-151.
[17] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2117-2125.
[18] 谢润忠, 李烨. 基于BERT和双通道注意力的文本情感分类模型[J]. 数据采集与处理, 2020, 35(4): 642-652.
XIE Runzhong, LI Ye. Text sentiment classification model based on BERT and dual channel attention[J]. Journal of Data Acquisition and Processing, 2020, 35(4): 642-652.
[19] XIE Z, ZHANG Z, CAO Y, et al. Simmim: A simple framework for masked image modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 9653-9663.
[20] ATIENZA R. Vision transformer for fast and efficient scene text recognition[C]//Proceedings of International Conference on Document Analysis and Recognition. Cham: Springer, 2021: 319-334.
[21] WANG P, YANG L, LI H, et al. A simple and robust convolutional-attention network for irregular text recognition[EB/OL](2019-04-20)[2022-11-10]. http://arxiv.org/abs/1904.01375.
[22] WAN Z, HE M, CHEN H, et al. Textscanner: Reading characters in order for robust scene text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2020: 12120-12127.
[23] MOU Y, TAN L, YANG H, et al. Plugnet: Degradation aware scene text recognition supervised by a plug-able super-resolution unit[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 158-174.
[24] LITMAN R, ANSCHEL O, TSIPER S, et al. Scatter: Selective context attentional scene text recognizer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 11962-11972.

作者简介:

佟国香(1968-),**通信作者**,女,博士,副教授,研究方向:嵌入式系统设计与开发、人工智能及数据科学,E-mail:tonggx@usst.edu.cn。

董田荣(1997-),女,硕士研究生,研究方向:深度学习、场景文本检测与识别,E-mail:15735185267@163.com。

胡桁彰(2001-),男,本科,研究方向:图像处理,E-mail:huhengzhang1@163.com。

(编辑:夏道家)