

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2024)05-1381-11

论文引用格式: Chen W D, Wang L F and Tao D P. 2024. SAME-net: scene text recognition method based on soft attention mask embedding. Journal of Image and Graphics, 29(05):1381-1391(陈威达, 王林飞, 陶大鹏. 2024. 融合软注意力掩码嵌入的场景文本识别方法. 中国图象图形学报, 29(05):1381-1391)[DOI:10.11834/jig.230081]

融合软注意力掩码嵌入的场景文本识别方法

陈威达, 王林飞, 陶大鹏*

云南大学信息学院, 昆明 650500

摘要: 目的 基于深度学习的端到端场景文本识别任务已经取得了很大的进展。然而受限于多尺度、任意形状以及背景干扰等问题, 大多数端到端文本识别器依然会面临掩码提议不完整的问题, 进而影响模型的文本识别结果。为了提高掩码预测的准确率, 提出了一种基于软注意力的掩码嵌入模块(soft attention mask embedding, SAME), 方法 利用Transformer更好的全局感受野, 将高层特征进行编码并计算软注意力, 然后将编码特征与预测掩码层级嵌入, 生成更贴近文本边界的掩码来抑制背景噪声。基于SAME强大的文本掩码优化及细粒度文本特征提取能力, 进一步提出了一个健壮的文本识别框架 SAME-Net, 开展无需字符级注释的端到端精准文本识别。具体来说, 由于软注意力是可微的, 所提出的 SAME-Net 可以将识别损失传播回检测分支, 以通过学习注意力的权重来指导文本检测, 使检测分支可以由检测和识别目标联合优化。结果 在多个文本识别公开数据集上的实验表明了所提方法的有效性。其中, SAME-Net 在任意形状文本数据集 Total-Text 上实现了 84.02% 的 H-mean, 相比于 2022 年的 GLASS (global to local attention for scene-text spotting), 在不增加额外训练数据的情况下, 全词典的识别准确率提升 1.02%。所提方法在多向数据集 ICDAR 2015 (International Conference on Document Analysis and Recognition) 也获得了与同期工作相当的性能, 取得 83.4% 的强词典识别结果。结论 提出了一种基于 SAME 的端到端文本识别方法。该方法利用Transformer的全局感受野生成靠近文本边界的掩码来抑制背景噪声, 提出的 SAME 模块可以将识别损失反向传输到检测模块, 并且不需要额外的文本校正模块。通过检测和识别模块的联合优化, 可以在没有字符级标注的情况下实现出色的文本定位性能。

关键词: 自然场景文本检测; 自然场景文本识别; 软注意力嵌入; 深度学习; 端到端自然场景文本检测与识别

SAME-net: scene text recognition method based on soft attention mask embedding

Chen Weida, Wang Linfei, Tao Dapeng*

School of Information, Yunnan University, Kunming 650500, China

Abstract: **Objective** Text detection and recognition of natural scenes is a long-standing and challenging problem. Hence, this study aims to detect and recognize text information in natural scene images. Owing to its wide applications (e. g., traffic sign recognition and content-based image retrieval), text detection and recognition has attracted much attention in the

收稿日期: 2023-03-12; 修回日期: 2023-09-14; 预印本日期: 2023-09-21

* 通信作者: 陶大鹏 dptao@ynu.edu.cn

基金项目: 国家科技创新 2030-“新一代人工智能”重大项目(2021ZD0111700); 国家自然科学基金项目(62172354); 云南省重大科技专项资助项目(202202AD080003)

Supported by: National Science and Technology Innovation 2030-“New Generation Artificial Intelligence” Major Project (2021ZD00111700); National Natural Science Foundation of China(62172354); Yunnan Provincial Major Science and Technology Special Plan Projects(202202AD080003)

field of computer vision. The traditional scene text detection and recognition method regards detection and recognition as two independent tasks. This method first locates and then clips to predict the text area of the input image and to clip the relevant area and then sends the clipped area into the recognizer for recognition. However, this process has some limitations, such as: 1) inaccurate detection results may seriously affect the performance of image text recognition owing to the accumulation of errors between the two tasks, and 2) the separate optimization of the two tasks may not improve the results of text recognition. In recent years, the end-to-end scene text recognition task based on deep learning has made great progress. Many studies have found that detection and recognition are closely related. End-to-end recognition, which integrates detection and recognition tasks, can promote each other and gradually become an important research direction. In the end-to-end recognition task, the natural scene image contains disturbing factors, such as light, deformation, and stain. In addition, scene text can be represented by different colors, fonts, sizes, directions, and shapes, making text detection very difficult. Limited by multi-scale, arbitrary shapes, background interference, and other issues, most end-to-end text recognizers still face the problem of incomplete mask proposals, which will affect the text recognition results of the model. Hence, we propose a mask embedding module (SAME) based on soft attention to improve the accuracy of mask prediction. This module effectively improves the robustness and accuracy of the model. **Method** High-level features are coded, and soft attention is calculated using the global receptive field of Transformer. Then, the coding features and prediction mask are embedded to generate a mask close to the text boundary to suppress background noise. Based on these designs, we propose a simple and robust end-to-end text recognition framework, SAME-Net, because soft attention is differentiable. The proposed SAME module can propagate the recognition loss back to the detection branch to guide the text detection by learning the weight of attention so that the detection and recognition targets can jointly optimize the detection branch. SAME-Net does not need additional recognition modules, nor does it need to annotate the text at the character level. **Result** This method can effectively detect multi-scale and arbitrarily shaped text. The recall rate, accuracy rate, and H-mean value on the public arbitrarily shaped data set Total-Text are 0.884 8 and 0.879 6. Compared with the best results in the comparison method, without adding further training data, the recognition accuracy rate without dictionary guidance is increased by 2.36%, and the recognition accuracy rate of the full dictionary is increased by 5.62%. In terms of detection, the recall rate and H-mean value of this method increased from 0.868 to 0.884 8 and from 0.861 to 0.879 6, respectively, which greatly exceeded the previous method in terms of end-to-end recognition. Both obtained 83.4% strong dictionary recognition results in the multi-directional dataset ICDAR 2015 (International Conference on Document Analysis and Recognition). In short, our method is superior to others. **Conclusion** The performance of SAME-Net proposed in this study has significantly improved on the two scene text data machines of ICDAR 2015 and Total-Text. The best results in this task were obtained. This study proposes an end-to-end text recognition method based on SAME. The proposed method has two advantages. First, the method uses the global receptive field of Transformer to embed high-level coding features and prediction mask levels to generate a mask close to the text boundary to suppress background noise. Second, the proposed SAME module can reverse transmit the recognition loss to the detection module, and no additional text correction module is needed. Great text positioning performance can be achieved without character-level comments through the joint optimization of the detection and recognition modules.

Key words: natural scene text detection; natural scene text recognition; soft attention embedding; deep learning; end-to-end natural scene text detection and recognition

0 引言

自然场景文本识别旨在检测出自然场景中文本的位置,并识别出文本区域的单词或句子。由于其在拍照翻译、自动驾驶、文档分析和表格识别(高良才等,2022)等领域具有丰富的现实应用,近年来已经引起了学术界和工业界的广泛关注,对文本检测和

识别的精确度提出了更高的要求。

自然场景文本识别目前有两种主流技术路线,第1种是采用级联检测和识别的方法,首先通过文本检测器检测出相应的文本区域,并将文本区域裁剪,然后通过特征对齐等操作后送入识别器进行识别。在这种级联的方法中检测和识别通常单独优化,识别器会受到检测器性能的限制;第2种是端到端检测识别架构,该方法将检测和识别放到了同一

个模型中进行优化,以特征采样代替裁剪操作,并将识别器直接作用于检测特征,允许检测和识别联合优化。与级联模型相比,端到端的文本识别方法中检测和识别两项任务关系更加密切,更容易协同优化并迁移到新的领域,因此通过端到端的联合架构来建模文本识别变得更加流行。然而,现阶段大多数端到端场景文本识别方法面临着以下挑战:

1)检测器在提取掩码过程中,通常只使用图像的低层特征来获得更准确的位置,却忽略了高层的全局特征,因为局部感受野有限,模型无法很好地区分相邻或不同尺度的文本实例是否属于同一个分段类别,进而影响识别结果。

2)在任意形状文本表示方面,目前方法通常使用文本实例的最小外接矩形框来表示文本区域,然而这引入了大量的背景噪声。同时,如果掩码分支没有预测到本应包含的文本信息,使用掩码提取文本感兴趣区域(region of interest, RoI)特征送入识别的方法,虽然有效地屏蔽了部分噪声,但依然不能识别到完整的文本实例。

为了共同改进检测和识别,ARTS (auto-rectification text spotter)(Zhong等,2021)通过使用可微分空间变换网络(spatial Transformer network, STN)将损失从识别分支反向传播到检测分支,提高端到端性能,ABCNet v2(adaptive bezier-curve network)(Liu等,2022)通过自适应贝塞尔曲线提取特征,解决文本检测和识别的优化不一致性,但这类通过引入文本矫正模块的方法尽管可以轻微地提升识别精度,却依然没有很好地解决端到端文本识别的优化问题。

为此,提出了一个端到端的任意形状文本识别框架,称为 SAME-Net,通过引入软注意力掩码嵌入

模块来更好地区分不同尺度的文本实例。

具体来说,首先使用掩码分支对低层特征进行文本区域预测,进而使用 Transformer 对特征金字塔网络(feature pyramid network, FPN)输出的高层级特征进行特征序列编码以获得全局感受野,并更好地学习到全局语义信息,提高模型对缩放文本的鲁棒性。

其次,提出软注意力掩码嵌入模块(soft attention mask embedding, SAME),通过层级嵌入带注意力权重的高层特征来补充出掩码分支没有检测到的文本区域,同时将细化后的掩码和文本注意力权重图来辅助引导识别,并抑制用于识别的融合特征图中的背景噪声,如图 1 所示。基于上述工作,进一步提出了 SAME-Net 文本识别框架,利用 Transformer 软注意力编码模块开展检测识别联合优化,该方法不需要使用字符级注释和矫正模块,训练完成的模型可以在掩码分支检测不精确的情况下,仍然通过嵌入高层语义特征的掩码正确地预测文本序列。

最后,通过大量的实验来验证 SAME-Net 在两个具有挑战性的基准数据集上的有效性,包括多向数据集 ICDAR 2015(International Conference on Document Analysis and Recognition)(Karatzas等,2015),任意形状文本数据集 Total-Text(Ch'ng 和 Chan, 2017),本文方法在 Total-Text 上实现了 84.02% 端到端定位 H-mean,大大超过了 ARTS v2(Zhong等, 2021)的端到端识别网络。

本文方法的主要贡献总结如下:1)设计了一个名为 SAME 的软注意力掩码嵌入模块,通过特征嵌入模块可以将掩码不断细化并通过高层语义补充出掩码检测分支未检测到的文本区域,通过该方法可以有效地提高检测的提议置信度,使得掩码更加精

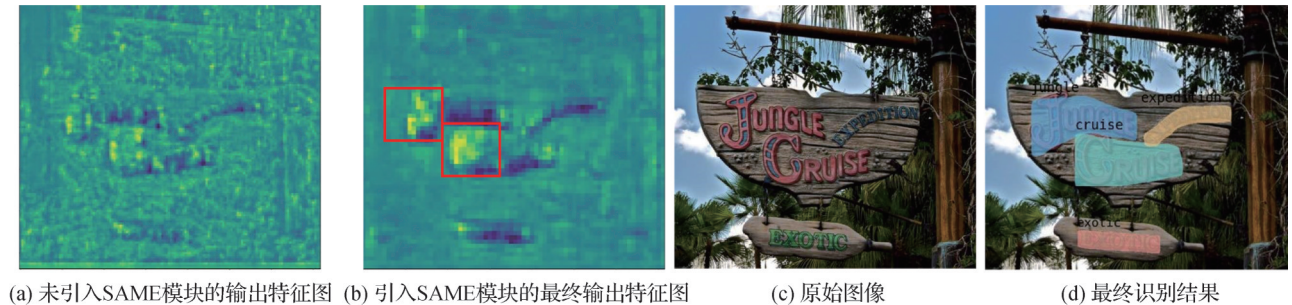


图1 SAME-Net网络效果

Fig. 1 Effect result of SAME-Net network ((a) output feature map without introducing the SAME module; (b) final output feature map with introducing the SAME module; (c) original image; (d) final recognition result)

确,并抑制背景噪声。2)提出了一个新的任意形状文本识别框架 SAME-Net,引入掩码检测分支和软注意力特征增强分支,通过识别损失反向传播来优化软注意力权重图,进而获得更适合文本识别的检测特征。3)大量实验表明,该方法在多方向和任意形状文本基准上都取得了有竞争力的甚至是最先进的结果。

1 相关工作

在过去的几十年里,深度学习方法的出现极大地促进了自然场景文本识别发展,为解决场景文本识别存在的问题,研究者通常会设计级联或端到端的文本检测识别方法。

1.1 级联的文本检测识别

采用级联的文本检测识别方法,一般通过独立的检测器,将单词区域裁剪并统一地缩放和旋转到文本的水平阅读方向,识别器在固定的输入尺寸上进行操作,这些方法大多应用于水平文本或多方向文本识别上,在任意形状文本检测和识别中效果很差。

Wang 等人(2011)使用基于滑动窗口的检测器来检测文本字符,然后对每个字符进行分类,Bissacco 等人(2013)集合深度神经网络(deep neural network, DNN)和尺度不变特征变换(histogram of oriented gradient, HOG)特征,利用字符分类建立文本提取系统,Liao 等人(2017)提出 TextBoxes,采用单次检测和文本识别以两阶段的形式解决问题。前述方法将检测和识别作为单独的任务,且在识别期间,预测的单词之间不共享上下文信息,这些方法会由于单独工作之间的误差累计,形成错误识别。

1.2 端到端文本检测识别

相比之下,在端到端方法中,识别器利用包含文本的局部特征,通过对检测的文本区域解码,来提升识别困难文本序列的能力,此外,受刘崇宇等人(2021)工作的启发,端到端联合训练有利于更好地结合检测和识别这两项任务,为场景文本识别带来了巨大的收益。

Li 等人(2017)将检测和识别统一到一个端到端可训练场景文本识别框架中,为了更好地桥接文本检测和识别,许多方法引入了文本矫正模块,TextDragon (Feng 等, 2019) 利用 RoI Slide (return on investment slide)将文本的预测段转换成水平姿态文

本送入识别,以更好地将检测和识别任务结合起来,ARTS(Zhong 等, 2021)使用空间变换网络(STN)将识别损失的梯度反向传递给检测器,更好地结合了文本检测与识别之间的协同作用,ABCNet(Liu 等, 2020)及其改进版本 ABCNet V2(Liu 等, 2022)使用 BezierAlign 将任意形状的文本转换为常规文本。上述几类方法通过引入额外的文本矫正模块来优化文本检测识别,但是造成了大量的额外开销。

为了解决任意形状文本的定位问题,FOTS(fast oriented text spotting)(Liu 等, 2018)使用一阶段检测器生成旋转框,并采用 RoI Rotate 将有方向的文本特征采样到水平网络,Mask TextSpotter v1(Lyu 等, 2018)与 Mask TextSpotter v3(Liao 等, 2020)通过提出一种基于分割的分段建议网络,其中检测目标以分割图的形式表示,进而通过 Hard RoI Masking 对特征进行采样,以提高识别器的性能。Qin 等人(2019)使用 RoI Masking 来提取任意形状文本的检测特征,送入识别器进行文本识别,He 等人(2018)提出了一个使用基于注意力的识别器框架,来解决任意形状场景文本识别的任务。上述方法仅通过掩码来提取特征,忽略了未在掩码区域中的全局语义信息,且识别损失也没有反向传播到检测器来优化检测,检测与识别仍然彼此相对独立。

2 模型方法

本文方法架构如图 2(a)所示,框架包含 4 个组件:1)由 Swin-Transformer(Liu 等, 2021)组成的主干;2)基于 Transformer 的高层特征编码及掩码预测分支;3)软注意力掩码嵌入模块(SAME);4)文本识别分支。

2.1 SAME-Net 模型概述

为更好地对比方法的优越性,模型使用了 Mask TextSpotter v3(Liao 等, 2020)中的掩码检测分支及识别分支作为基线,同图 2(b),受到 Zhang 等人(2023)方法的启发,模型将 FPN 输出的 P3, P4, P5 这 3 个阶段特征作为 Transformer 编码器的输入,进而对特征进行分层编码计算,以获得软注意力权值。在 SAME 模块,分段预测掩码和软注意力权重矩阵进行嵌入计算,获得最终的融合注意力权重图及掩码图像,将融合后的掩码图像与 RoI 特征进行点乘来获得掩码特征图,并将掩码特征图和软注意力权重

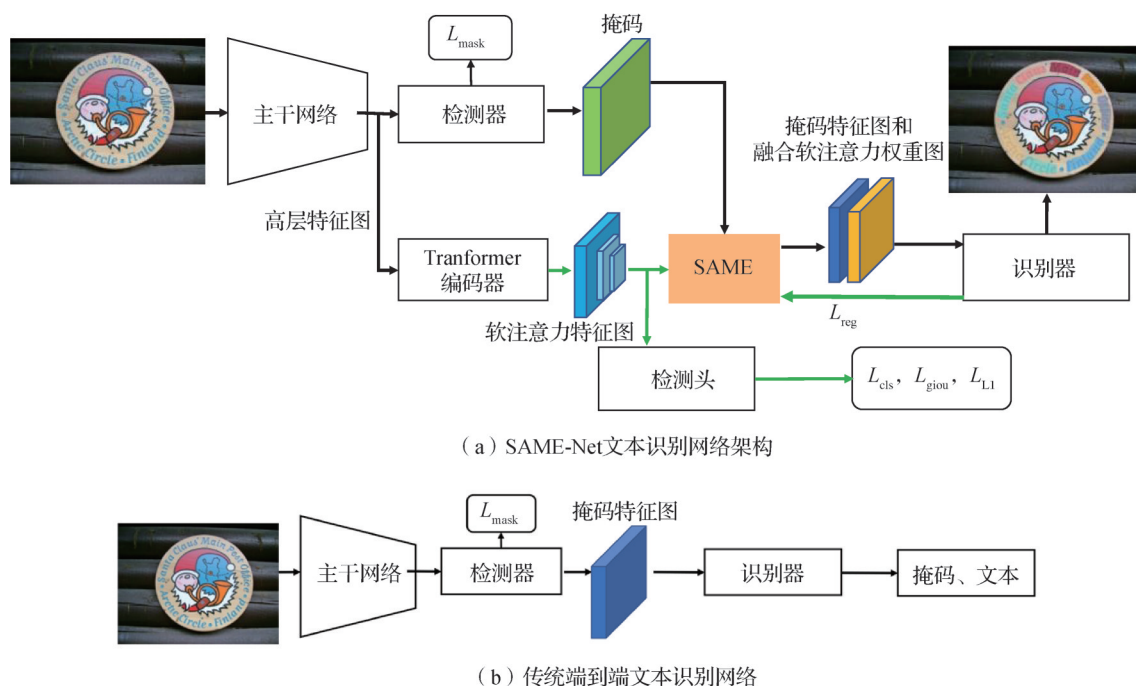


图2 端到端自然场景文本识别方法 SAME-Net 的框架

Fig. 2 The framework of SAME-Net for end-to-end natural scene text recognition

((a) SAME-Net text recognition network architecture; (b) traditional end-to-end text recognition network)

图传递给结合空间注意力的识别器(Liao等,2020)进行序列解码。

对于文本定位方法,选择当下效果优秀的Swin-Transformer结合特征金字塔网络(FPN)来构建模型主干。在前向阶段,首先将输入图像送到主干网络,并输出RoI特征图,之后将特征图送入掩码预测分支预测掩码,为了保持嵌入特征的分辨率与掩码分辨率对应,将FPN输出的P3, P4, P5这3个阶段特征,送入Transformer编码器进行计算,得到 $f_1^{\text{soft_at}}, f_2^{\text{soft_at}}, f_3^{\text{soft_at}}$ 这3个阶段的注意力特征,之后将掩码提议和注意力特征送入软注意力掩码嵌入模块(SAME)进行掩码细化和特征嵌入,最后将细化掩码提取的RoI特征及软注意力权重图,由于识别阶段需要比检测更高的分辨率,使用最终阶段输出的掩码 $Mask_3$ 来提取RoI特征,该特征的分辨率是检测阶段分辨率的4倍,识别器采用结合空间注意模块(space attention module, SAM)的解码器对特征进行解码,得到最终的文本序列。

2.2 软注意力掩码嵌入模块

首先将提议掩码图像送入Transformer编码器进行编码,将高层特征的软注意力权重嵌入掩码,进而补充由于局部感受野受限而没有分割出的掩码区

域。为了增强对多尺度文本序列的识别效果,将掩码特征和不同尺度的文本特征进行层级嵌入,进而获得软注意力权重图来辅助文本识别,识别损失可以通过反向传播到Transformer多头编码器,以通过优化注意力权重图的形式指导检测。具体来说,软注意力掩码嵌入模块包括1个Transformer编码器和4个上采样结构,SAME的输入是掩码图 $Mask$ 和3个层级的软注意力特征 $f_1^{\text{soft_at}}, f_2^{\text{soft_at}}, f_3^{\text{soft_at}}$, SAME模块结构图如图3所示。

掩码 $Mask$ 被送到Transformer编码器 $TrE()$ 进行编码,通过上采样操作 $U_{up}()$ 和注意力特征层级嵌入的方式得到特征嵌入图 $\{d_1, d_2, d_3\}$,经过sigmoid函数 $Sig()$ 生成文本区域掩码 $\{M_1, M_2, M_3\}$,具体为

$$d_1 = TrE(Mask) \quad (1)$$

$$d_2 = U_{up}(d_1) + f_2^{\text{soft_at}} \quad (2)$$

$$d_3 = U_{up}(d_2) + f_3^{\text{soft_at}} \quad (3)$$

$$M_i = Sig(d_i), i = 1, 2, 3 \quad (4)$$

为了获得全局文本特征,利用 $\{M_1, M_2, M_3\}$ 和不同层级的软注意力特征来进一步集成,具体为

$$SM_1 = M_1 d_1 + f_1^{\text{soft_at}} \quad (5)$$

$$SM_2 = M_2 (SM_1 + f_2^{\text{soft_at}}) \quad (6)$$

$$SM_3 = M_3 (SM_2 + f_3^{\text{soft_at}}) \quad (7)$$

为 2.5×10^{-7} , 然后在相应的 ICDAR 2015 和 Total-Text 数据集上微调模型。

在 ICDAR 2015 和 Total-Text 数据上对模型进行调优时, 对于 Total-Text 及 ICDAR 2015, 将输入图像的短边均调整为 1 000, 同时使用数据扩充策略。

3.3 与最先进方法的比较

3.3.1 任意形状文本识别

SAME-Net 主要关注任意形状的文本识别, 为

了验证其有效性, 在具有挑战性的 Total-Text 上进行了实验, 同时遵循标准文本识别评估指标来进行公平的比较。Total-Text 的评估结果如表 1 所示, 其中, None 和 Full 分别表示没有词典和组合了所有图像完整词典的结果。SAME-Net 方法在准确性方面优于其他对比方法, 具体来说, SAME-Net 实现了 84.02% 其他对比的端到端 H-mean 指标。

表 1 端到端文本识别算法在 Total-Text 上的性能对比

Table 1 Performance comparison of end-to-end text recognition algorithms on Total-Text

方法	特征提取主干网络	检测			端到端识别	
		精确率	召回率	H-mean	None	Full
ABCNet(Liu 等, 2020)	ResNet50	—	—	—	64.2	75.7
Mask v3(Liao 等, 2020)	ResNet50	—	—	—	71.2	78.4
ABCNet V2(Liu 等, 2022)	ResNet50	—	—	—	70.4	78.1
Boundary(Wang 等, 2020)	ResNet50	85.0	88.9	87.0	68.6	78.6
PGNet(Wang 等, 2021)	FPN	86.8	85.5	86.1	63.1	—
MANGO(Qiao 等, 2021)	ResNet50	—	—	—	72.9	82.6
TextDragon(Feng 等, 2019)	VGG16-FPN	75.7	85.6	80.3	48.8	—
GLASS(Ronen 等, 2022)	ResNet50	—	—	—	76.6	83.0
SAME-Net(本文)	Swin-Transformer	88.48	87.44	87.96	73.56	84.02

注: 加粗字体表示各列最优结果, “—”表示无实验数据。

3.3.2 多向文本识别

端到端文本识别算法在 ICDAR 2015 上的性能对比如表 2 所示, 其中, Strong、Weak 和 Generic 分别表示

具有强、弱和通用词汇的结果。可以看出, 本文方法在多方向数据集 ICDAR 2015 上依然具有十分优秀的性能, 在端到端强词典中达到了 83.4% 的识别结果。

表 2 端到端文本识别算法在 ICDAR 2015 上的性能对比

Table 2 Performance comparison of end-to-end text recognition algorithms on ICDAR 2015

方法	特征提取主干网络	端到端		
		Strong	Weak	Generic
ABCNet v2(Liu 等, 2022)	ResNet50	82.7	77.3	70.5
PAN++(Wang 等, 2022)	ResNet18	82.7	78.2	69.2
PGNet(Wang 等, 2021)	FPN	83.3	78.3	63.5
MaskTextSpotter v3(Liao 等, 2020)	ResNet50	83.3	78.1	74.2
MANGO(Qiao 等, 2021)	ResNet50	81.8	78.9	67.3
SAME-Net(本文)	Swin-Transformer	83.4	77.1	70.2

注: 加粗字体表示各列最优结果。

3.4 消融研究

为了验证框架中各个部分功能的有效性,在 Total-Text 数据集上进行了消融研究,每个模型版本都经过了预训练和微调,基线架构由 Mask TextSpotter v3(Liao 等,2020)构成,表 3 展示了在 Total-Text 上的测试结果,其中,SW 表示 Swin-Transformer(Liu 等,2021),TR 为 Transformer 语义编码组件。通过充分的消融实验,以证明反向传播对提升端到端网络性能的有效性。

1)SAME 对检测和识别的有效性。从图 4 可以看出,图 4(a)为没有使用 SAME 模块的效果图,“R”

被误识别为“P”,图 4(b)为使用了 SAME 模块的效果图,“R”被正确识别,在引入 SAME 和特征增强后,掩码更加完整,使得文本能被正确识别。

2)识别损失反向传播对检测和识别的有效性。为了验证反向传播识别损失的有效性,遵循 ARTS (Zhong 等,2021)的方法,对于第 3 组实验,使用真值标注来训练识别分支,从而阻断识别损失的反向传播。对于第 4 组,通过使用细化的 Mask 来提取嵌入特征并启用识别损失反向传播,以便检测器由检测和识别目标进行联合优化。结果可以在表 3 中看到。

表 3 消融实验
Table 3 Ablation experiment

方法	融合方式	梯度阻断	H-mean		
			检测	端到端(无字典)	端到端(有字典)
基线	×	√	—	71.20	78.40
基线+SW+TR	×	√	81.25	71.61	80.42
基线+SW+TR	基线融合方法	√	84.65	70.64	81.35
基线+SW+TR	SAME	√	85.94	72.13	82.51
本文	SAME	×	87.96	73.56	84.02

注:加粗字体表示各列最优结果,“—”表示无实验数据,“√”表示使用该方法,“×”表示未使用该方法。



图 4 消融实验效果图

Fig. 4 Effect of ablation experiment ((a) not using the SAME module; (b) using the SAME module)

3.5 定性研究

选取了具有挑战性的文本图像来体现 SAME-Net 对于识别困难样本的健壮性,如图 5 所示,其中,3 幅子图从左到右依次为包含了遮挡、密集分段、变形和多尺度的文本实例图像,图 5(a)为标

准的 E2E(end to end)文本识别框架 Mask TextSpotter v3 的预测结果,图 5(b)为本文提出的 SAME-Net 方法的预测结果。通过实验表明,使用软注意力嵌入模块(SAME)后能够有效识别任意形状文本。



图5 在Total-Text上的定性结果

Fig. 5 Qualitative results on Total-Text ((a) Mask TextSpotter v3; (b) SAME-Net (ours))

4 结 论

本文提出了SAME-Net,一个融合软注意力掩码嵌入的场景文本识别方法,利用Transformer更好的全局感受野,将高层编码特征与预测掩码层级嵌入,生成更贴近文本边界的掩码来抑制背景噪声,所提出的SAME模块能够通过将识别损失反向传播到检测模块,并且没有额外文本矫正模块,通过检测和识别模块的联合优化在没有字符级注释的情况下实现更好的文本定位性能。

相比现有方法,本文提出的SAME-Net虽然带来了显著的端到端性能提升,但模型的整体性能依然受限于识别器的解码能力的限制。同时该模型在密集小文本场景中(比如ICDAR 2023 DStext数据集)表现效果不佳,经分析得知这是由于在下采样图像的过程中将相邻文本字段的特征进行了叠加,使得在密集场景中的效果较差。在后续工作中,将在ICDAR 2023 DStext这个难度更高的数据集进行优化分析,来解决在密集小文本场景中该模型不足的问题。此外,如何突破人工标注的真值框来训练检测,并利用天然的弱标注或无标注实现模型的端到

端识别也是值得探索的重要方向。

参考文献(References)

- Bissacco A, Cummins M, Netzer Y and Neven H. 2013. PhotoOCR: reading text in uncontrolled conditions//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 785-792 [DOI: 10.1109/ICCV.2013.102]
- Ch'ng C K and Chan C S. 2017. Total-Text: a comprehensive dataset for scene text detection and recognition//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 935-942 [DOI: 10.1109/ICDAR.2017.157]
- Feng W, He W H, Yin F, Zhang X Y and Liu C L. 2019. TextDragon: an end-to-end framework for arbitrary shaped text spotting//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9075-9084 [DOI: 10.1109/ICCV.2019.00917]
- Gao L C, Li Y B, Du L, Zhang X P, Zhu Z Y, Lu N, Jin L W, Huang Y S and Tang Z. 2022. A survey on table recognition technology. Journal of Image and Graphics, 27(6): 1898-1917 (高良才, 李一博, 都林, 张新鹏, 朱子仪, 卢宁, 金连文, 黄永帅, 汤帆. 2022. 表格识别技术研究进展. 中国图象图形学报, 27(6): 1898-1917) [DOI: 10.11834/jig.220152]
- Gupta A, Vedaldi A and Zisserman A. 2016. Synthetic data for text localization in natural image [EB/OL]. [2023-02-25].

- <https://arxiv.org/pdf/1604.06646.pdf>
- He T, Tian Z, Huang W L, Shen C H, Qiao Y and Sun C M. 2018. An end-to-end textspotter with explicit alignment and attention//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5020-5029 [DOI: 10.1109/CVPR.2018.00527]
- Hu J, Cao L J, Lu Y, Zhang S C, Wang Y, Li K, Huang F Y, Shao L and Ji R R. 2021. ISTR: end-to-end instance segmentation with Transformers [EB/OL]. [2023-03-12]. <https://arxiv.org/pdf/2105.00637.pdf>
- Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S J, Shafait F, Uchida S and Valveny E. 2015. ICDAR 2015 competition on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE: 1156-1160 [DOI: 10.1109/ICDAR.2015.7333942]
- Li H, Wang P and Shen C H. 2017. Towards end-to-end text spotting with convolutional recurrent neural networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5248-5256 [DOI: 10.1109/ICCV.2017.560]
- Liao M H, Pang G, Huang J, Hassner T and Bai X. 2020. Mask TextSpotter v3: segmentation proposal network for robust scene text spotting//Proceeding of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 706-722 [DOI: 10.1007/978-3-030-58621-8_41]
- Liao M H, Shi B G, Bai X, Wang X G and Liu W Y. 2017. TextBoxes: a fast text detector with a single deep neural network//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 4161-4167 [DOI: 10.1609/aaai.v31i1.11196]
- Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]
- Liu C Y, Chen X X, Luo C J, Jin L W, Xue Y and Liu Y L. 2021. Deep learning methods for scene text detection and recognition. Journal of Image and Graphics, 26(6): 1330-1367 (刘崇宇, 陈晓雪, 罗灿杰, 金连文, 薛洋, 刘禹良. 2021. 自然场景文本检测与识别的深度学习方法. 中国图象图形学报, 26(6): 1330-1367) [DOI: 10.11834/jig.210044]
- Liu X B, Liang D, Yan S, Chen D G, Qiao Y and Yan J J. 2018. FOTS: fast oriented text spotting with a unified network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5676-5685 [DOI: 10.1109/CVPR.2018.00595]
- Liu Y L, Chen H, Shen C H, He T, Jin L W and Wang L W. 2020. ABCNet: real-time scene text spotting with adaptive bezier-curve network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9806-9815 [DOI: 10.1109/CVPR42600.2020.00983]
- Liu Y L, Shen C H, Jin L W, He T, Chen P, Liu C Y and Chen H. 2022. ABCNet v2: adaptive bezier-curve network for real-time end-to-end text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (11): 8048-8064 [DOI: 10.1109/TPAMI.2021.3107437]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Lyu P, Liao M H, Yao C, Wu W H and Bai X. 2018. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 71-88 [DOI: 10.1007/978-3-030-01264-9_5]
- Qiao L, Chen Y, Cheng Z Z, Xu Y L, Niu Y, Pu S L and Wu F. 2021. MANGO: a mask attention guided one-stage scene text spotter [EB/OL]. [2023-02-25]. <https://arxiv.org/pdf/2012.04350v1.pdf>
- Qin S Y, Bissaco A, Raptis M, Fujii Y and Xiao Y. 2019. Towards unconstrained end-to-end text spotting//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4703-4713 [DOI: 10.1109/ICCV.2019.00480]
- Rezatofghi H, Tsoi N, Gwak J, Sadeghian A, Reid I and Savarese S. 2019. Generalized intersection over union: a metric and a loss for bounding box regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 658-666 [DOI: 10.1109/CVPR.2019.00075]
- Ronen R, Tsiper S, Anshel O, Lavi I, Markovitz A and Manmatha R. 2022. GLASS: global to local attention for scene-text spotting [EB/OL]. [2023-02-25]. <https://arxiv.org/pdf/2208.03364.pdf>
- Wang H, Lu P, Zhang H, Yang M K, Bai X, Xu Y C, He M C, Wang Y P and Liu W Y. 2020. All you need is boundary: toward arbitrary-shaped text spotting//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 12160-12167 [DOI: 10.48550/AAAI.2020.v34i07.6896]
- Wang K, Babenko B and Belongie S. 2011. End-to-end scene text recognition//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 1457-1464 [DOI: 10.1109/ICCV.2011.6126402]
- Wang P F, Zhang C Q, Qi F, Liu S S, Zhang X Q, Lyu P Y, Han J Y, Liu J T, Ding E R and Shi G M. 2021. PGNet: real-time arbitrarily-shaped text spotting with point gathering network [EB/OL]. [2023-03-12]. <https://arxiv.org/pdf/2104.05458.pdf>
- Wang W H, Xie E Z, Li X, Liu X B, Liang D, Yang Z B, Lu T and Shen C H. 2022. PAN++: towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Transactions on Pat-

tern Analysis and Machine Intelligence, 44(9): 5349-5367 [DOI: 10.1109/TPAMI.2021.3077555]

Zhang Q M, Xu Y F, Zhang J and Tao D C. 2023. ViTAEv2: vision Transformer advanced by exploring inductive bias for image recognition and beyond. International Journal of Computer Vision, 131(5): 1141-1162 [DOI: 10.1007/s11263-022-01739-w]

Zhong H M, Tang J, Wang W H, Yang Z B, Yao C and Lu T. 2021. ARTS: eliminating inconsistency between text detection and recognition with auto-rectification text spotter [EB/OL]. [2023-02-25].

<https://arxiv.org/pdf/2110.10405.pdf>

作者简介

陈威达,男,硕士研究生,主要研究方向为自然场景文本检测和识别。E-mail:chenweida96@163.com

陶大鹏,通信作者,男,教授,博士生导师,主要研究方向为人工智能。E-mail:dptao@ynu.edu.cn

王林飞,男,博士研究生,主要研究方向为目标检测。E-mail:linfeiwang@mail.ynu.edu.cn