Computer Science and Application 计算机科学与应用, 2021, 11(1), 167-175
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111018

Hans汉斯

基于改进EAST的文本检测算法

王　俊[1]，苗　军[1*]，卿来云[2]，乔元华[3]

[1]北京信息科技大学，网络文化与数字传播北京市重点实验室，北京
[2]中国科学院大学，计算机科学与技术学院，北京
[3]北京工业大学，数理学院，北京

Email: 1410888563@qq.com, *jmiao@bistu.edu.cn

收稿日期：2020年12月26日；录用日期：2021年1月20日；发布日期：2021年1月28日

摘　要

自然场景文本定位检测是文本识别的研究热点之一。EAST算法是目前自然场景文本定位检测算法较为出色的算法之一，在ICDAR2015数据集上，有着较高的准确率和召回率。但EAST算法仍存在着感受野不够大、长文本检测效果不佳的问题。因此本实验对EAST算法进行改进，通过改进EAST算法的结构，加入了ASPP网络，扩大感受野，加入了BLSTM神经网络，增强了文本之间的关联，提高文本定位效果。实验结果表明，该算法在ICDAR2015文本定位任务上的召回率为77.84%，精确率为86.24%，F-score为81.82%，优于经典EAST算法。

关键词

文本识别，EAST，ASPP网络，BLSTM神经网络

Text Detection Algorithm Based on Improved EAST

Jun Wang[1], Jun Miao[1*], Laiyun Qin[2], Yuanhua Qiao[3]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing
[3]College of Applied Sciences, Beijing University of Technology, Beijing

Email: 1410888563@qq.com, *jmiao@bistu.edu.cn

Received: Dec. 26th, 2020; accepted: Jan. 20th, 2021; published: Jan. 28th, 2021

*通讯作者。

文章引用: 王俊, 苗军, 卿来云, 乔元华. 基于改进 EAST 的文本检测算法[J]. 计算机科学与应用, 2021, 11(1): 167-175.
DOI: 10.12677/csa.2021.111018

Abstract

Text location and detection in natural scenes is one of the research hotspots of text recognition. East algorithm is one of the most excellent algorithms for text location and detection in natural scenes. It has high accuracy and recall rate in ICDAR2015 Dataset. However, the sensitivity field of EAST is not large enough and the effect of long text detection is not good. Therefore, this experiment improves the EAST algorithm by improving the structure of the EAST algorithm, adding the ASPP network, expanding the receptive field, adding the BLSTM neural network, enhancing the relevance between texts, and improving the text location effect. Experimental results show that the recall rate, precision rate and F-score of ICDAR2015 are 77.84%, 86.24% and 81.82% respectively, which are better than the classical EAST algorithm.

Keywords

Text Recognition, EAST, ASPP Network, BLSTM Neural Network

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着互联网的发展和计算机技术的提高，光学字符识别(Optical Character Recognition, OCR)技术[1]也得到了有效的提高。通过 OCR 识别技术可以有效地从图片上面提取到所需要的文字信息。但是目前通用的 OCR 算法只针对于简单的运用场景，一旦场景掺杂过大的因素，识别效率和召回率都会急剧下降。OCR 识别技术主要分文本检测和文本识别两部分[2] [3]。文本检测作为文本识别的前提，在整个文本信息提取和理解过程中起着重要的作用。只有正确的定位到文本区域才能进行正确的文本识别。正确的文本区域检测对提高文本识别准确有着重要的作用，因此，如何提高文本检测是一个重要的课题。

国内外学者利用不同的方法解决了文本检测问题。Tian [4]提出的一种新型的连接主义文本提议网络(CTPN)，使用一种垂直锚定机制，共同预测每个固定宽度候选位置和文本/非文本的分数。通过最大分数确定文本行位置，这大大提高文字定位的准确度。但是 CTPN 只针对于水平文字检测有很高的效率。为此，Shi 等人[5]提出一种定向文本检测方法 SegLink，在 CTPN 的基础上进行改进。主要的思想是将文本分为两个本地可以检测的元素，通过端对端训练的完全卷积神经网络在多个尺度上密集检测这两个元素。最终检测时通过连接段的组合。CTPN 检测法、SegLink 检测法是通过先预测 proposals (预选框)、segment (切片)，然后再回归、合并等方式实现对文本的检测。由于 CTPN 模型过于冗余复杂，Xinyu Zhou [6]等人提出 EAST 检测法，将中间过程缩减为只有 FCN (全卷积网络)、NMS (非极大值抑制) [7]两个阶段，而且输出结果支持文本行、单词的多个角度检测，既高效准确，又能适应多种自然应用场景。但是 EAST 算法仍然存在着感受野不够大，长文本检测效果不佳的问题。

因此，为了解决 EAST 算法的存在的问题，本文在 EAST 算法上进行改进，通过改进 EAST 算法的结构，利用 ASPP 网络替代 EAST 算法中的部分结构，引入 BLSTM 神经网络[8]，增加输出特征图之间的关联性，从而改善了 EAST 算法的文本检测效果，提高算法的性能。

2. 改进 EAST 算法

2.1. EAST 算法介绍

大部分传统的文本检测算法都是由多个阶段组成，在准确性和效率上面表现不是很好。EAST 算法提出端到端的文本定位方法，消除多个中间的 stage，直接预测文本行。它只有两个阶段，第一个阶段基于全卷积网络(FCN)模型，直接产生文本框预测；第二个阶段对生成的文本框进行非极大值抑制(NMS)以产生最终结果。该模型放弃了不必要的中间步骤，进行端到端的训练和优化。

如图 1 所示，EAST 算法网络结构分为三个部分：特征提取层，特征合并和输出层。

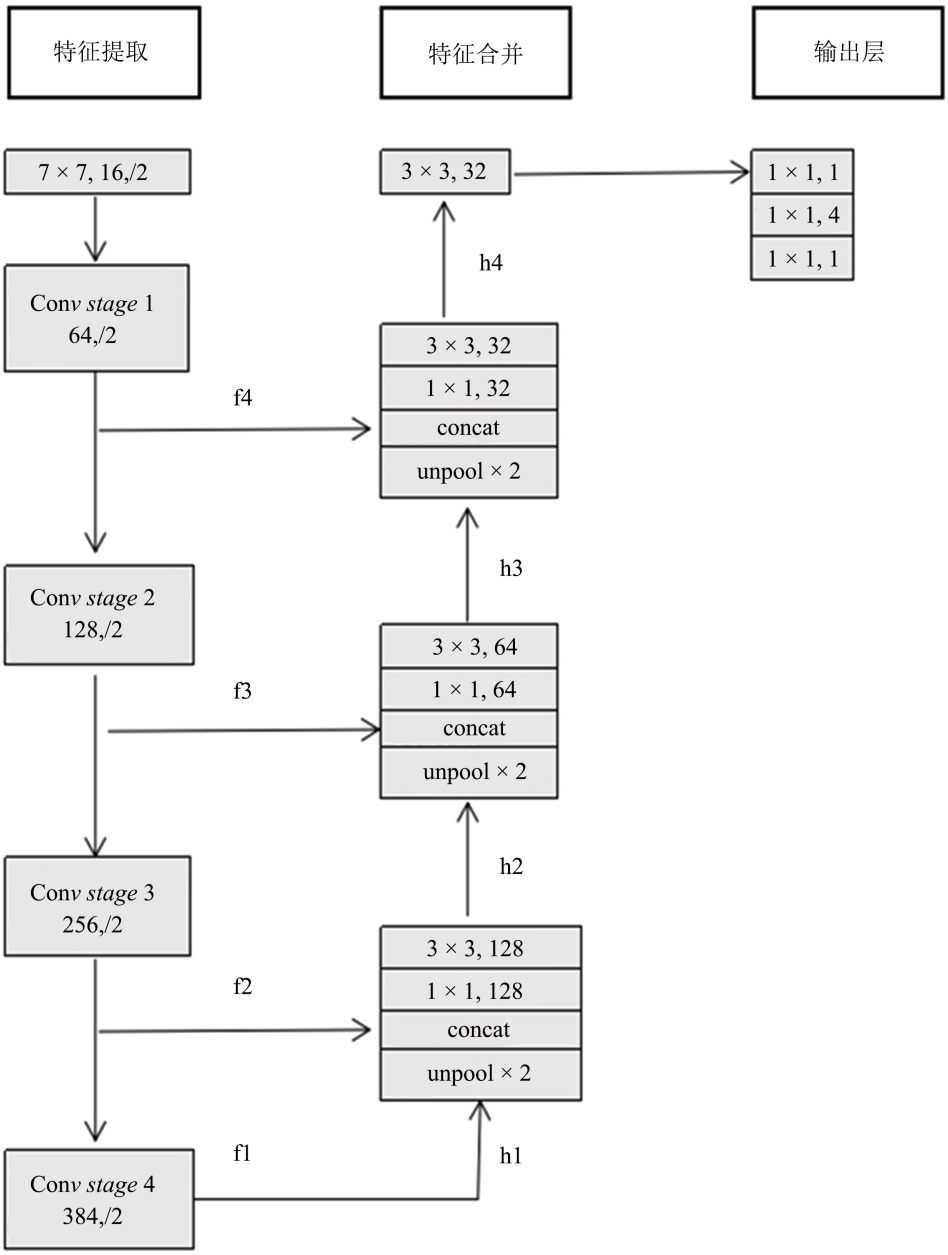

Figure 1. Architecture of EAST algorithm network
图 1. EAST 算法网络结构

特征提取层：利用在 ImageNet 数据集上预训练的卷积网络参数初始化。基于 VGG16 模型[9] (或者 ResNet-50 模型[10])作为主干神经网络提取文本的浅层和深层的纹理特征。提取模型的最后四个级别的特征图，其大小分别为输入图像的 1/4，1/8，1/16，1/32。通过提取出不同尺度的特征图，实现对不同尺度文本行的检测(大的特征图擅长检测小物体，小的特征图擅长检测大物体)。

特征合并：特征合并主要采取 U-net 思想，通过上采样，将上一级别的特征图和上采用的特征图进行合并，再通过 1 × 1 的卷积进行减少通道数量和计算量。接着通过 3 × 3 的卷积核运算得到新的特征图。

输出层：输出的特征图进行不同的卷积操作，最后得到分数特征图和多通道几何图形特征图。

EAST 算法的损失函数主要由两个部分组成，分类损失函数 L_s 和几何损失函数 L_g：

$$L = L_s + \lambda_g L_g \tag{1}$$

上式中，L_s 代表该像素是否存在文字的损失，L_g 代表 IOU 和角度的损失，λ_g 代表两个损失之间的重要性。原文的实验中将 λ_g 设置为 1 [11]。

$$L_s = balanced - xent\left(\hat{Y}, Y^*\right) = -\beta Y^* \ln \hat{Y} - \left(1-\beta\right)\left(1-Y^*\right)\ln\left(1-\hat{Y}\right) \tag{2}$$

上式中，$\hat{Y}$ 代表分数图的预测，Y^* 代表标注值。β 代表正负样本之间的平衡因子。

L_g 几何损失分为 IOU 损失和旋转角度的损失。公式如下：

$$L_{AABB} = -\ln \mathrm{IOU}\left(\hat{R}, R^*\right) = -\ln \frac{\left|\hat{R} \cap R^*\right|}{\left|\hat{R} \cup R^*\right|} \tag{3}$$

$$L_\theta\left(\hat{\theta}, \theta^*\right) = 1 - \cos\left(\hat{\theta} - \theta^*\right) \tag{4}$$

上式中，$\hat{\theta}$ 是对旋转角度的预测，θ^* 表示标注值。总体的损失为：

$$L_g = L_{AABB} + \lambda_\theta L_\theta \tag{5}$$

2.2. 加入 ASPP 网络

传统的卷积神经网络模型中，下采样过程是为了扩大感受野，使得每个卷积输出都包含较大范围的信息，对于提取抽象化信息有很大帮助，但在这个过程中，图像的分辨率不断下降，包含的信息越来越抽象，而图像的局部信息与细节信息会逐渐丢失，虽然现在也有通过线性插值上采样来恢复分辨率的手段存在，但在这个过程，还是不可避免的会造成信息的损失。而 EAST 算法采用的主干网络无论是 VGG16 还是 resnet50 都存在利用下采样用来增大感受野，但都不可避免的导致分辨率下降。而空洞卷积的出现解决了下采用带来的分辨率下降的问题。利用空洞卷积可以实现网络不进行下采样，同样能启到扩大感受野的目的。

本文利用空洞空间卷积池化金字塔 ASPP (Atrous Spatial Pyramid Pooling)网络来替代 EAST 算法的主干网络 VGG16 (或者 resnet50)的 stage 4 模块。ASPP 对所给定的输入以不同采样率的空洞卷积并行采样，相当于以多个比例捕捉图像的上下文。

如图 2 所示，本文 EAST 算法的主干网络采用的是 resnet50，利用 ASPP 网络替代 EAST 算法的主干网络的 stage 4 模块，ASPP 网络包括 5 个模块，通过将 stage 3 的输出特征图进行 5 种不同的操作，第一个模块是进行平均池化，1 × 1 的卷积层进行通道数变换，最后通过双线性插值恢复分辨率。第二个到第 5 个模型都是空洞卷积，但每个的卷积核的扩展率不同，分别取了 1，6，12，18；之后将这五个模块的输出拼接到一起，通过一个 1 × 1 的卷积层，降低通道数到需要的数值，作为下一步操作的输入。

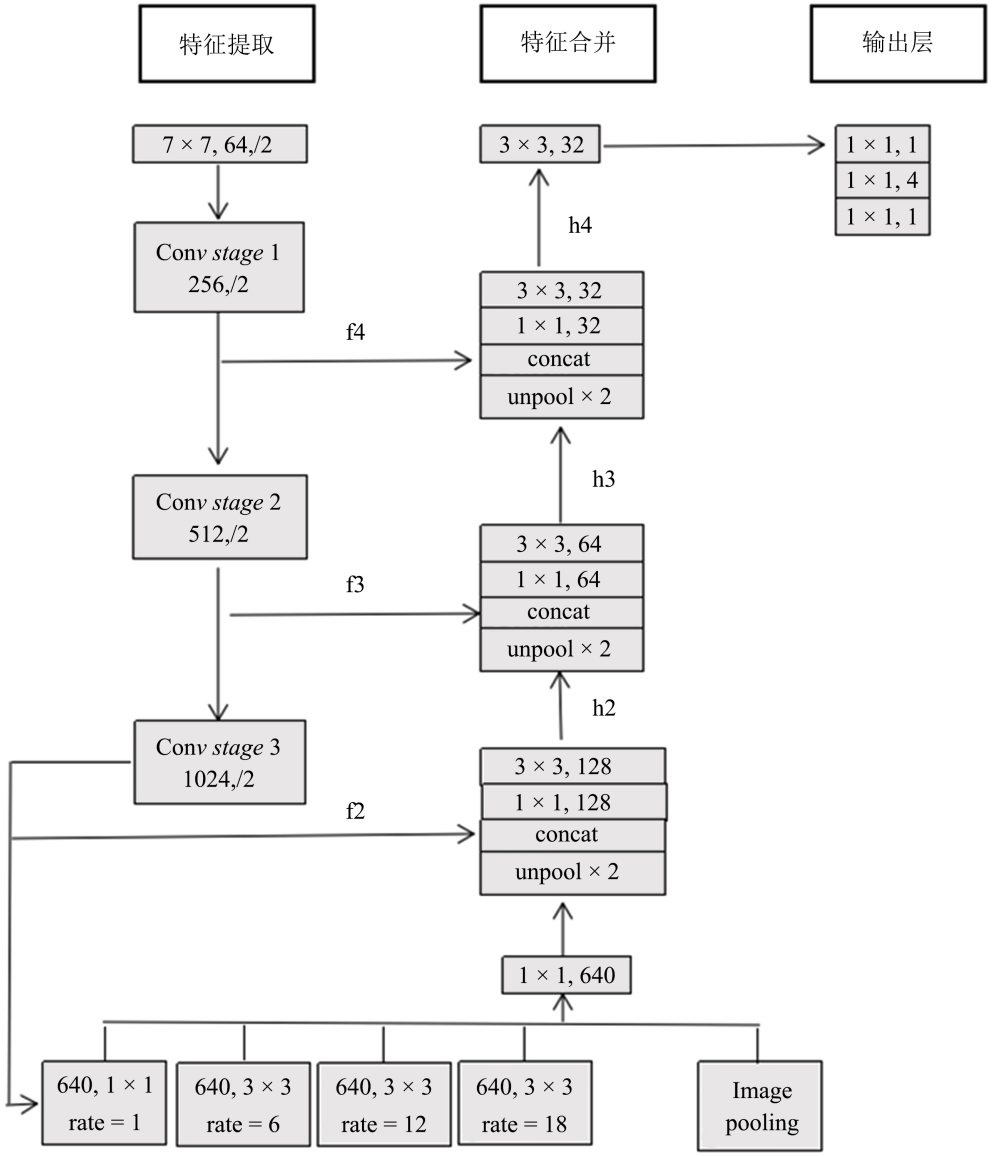

Figure 2. Add ASPP network structure
图 2. 加入 ASPP 网络结构

2.3. 增加 BLSTM 网络

文本检测不只是检测某个位置是否是单个文字，同样检测文字之间是否存在连续性。文字存在特征信息，同时邻近的文字存在关联关系。CNN 学习的是感受野内的空间信息，LSTM 学习的是序列特征。对于文本序列检测，显然既需要 CNN 抽象空间特征，也需要序列特征。

本文在 EAST 网络的特征合并和输出层之间加入 BLSTM 网络。BLSTM 是一种特殊的循环长短期记忆神经网络，由双向 LSTM 神经网络组成。

如图 3 所示，本文在 EAST 算法特征合并层和输出层之间插入 BLSTM 网络，BLSTM 网络结构如图 4 所示。BLSTM 网络能将每个特征的前后序列呈现为两个单独的隐藏状态，以分别捕获序列过去和未来的信息，然后再将两个隐藏的特征序列连接起来形成一个新的特征样本进行最终输出[12]。本文通过 EAST 算法特征合并后输出的特征图进行序列化关联，使得得到的序列样本更加合理均匀具有连续性。

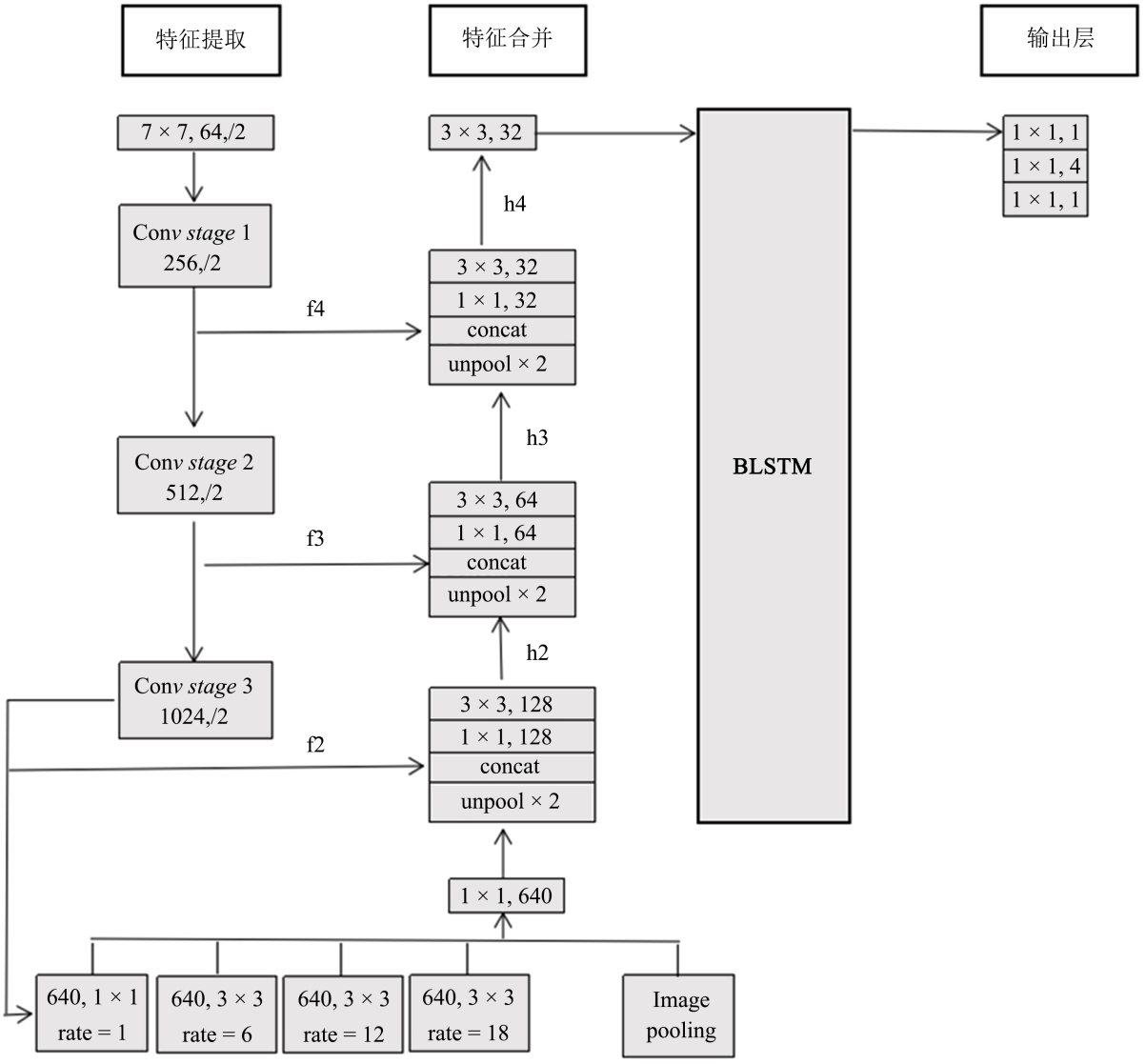

Figure 3. Adding the network structure of BLSTM
图 3. 加入 BLSTM 网络结构

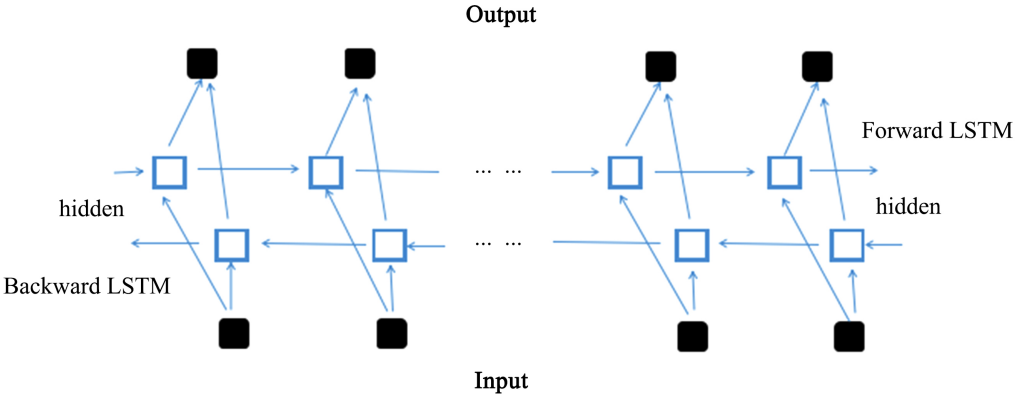

Figure 4. Network structure of BLSTM
图 4. BLSTM 网络结构

3. 实验结果

本实验在 tensorflow 深度学习框架上进行。采用 GTX2080Ti 的显卡进行改进的 EAST 算法实验。实验使用 resnet-50 网络模型作为 EAST 算法的主干网络进行预训练模型。使用的数据集是 ICDAR2013 和 ICDAR2015 的训练数据集。ICDAR2013 训练集具有 229 张标注训练集，ICDAR2015 训练集具有 1000 张标注训练集。以 ICDAR2015 测试集作为测试，ICDAR2015 测试集有 500 张图片。这些数据集是由谷歌公司制作，数据集均是在自然场景下采集的，图中的文本是任意方向和位置的。

实验过程主要将图片送入 resnet-50 网络模型，经过 4 次下采用，将获取到特征图送入 ASPP 网络，ASPP 网络在不损失图片分辨率的情况下，提高感受野。再通过 3 次上采用并且每次和原有对应的下采用特征图进行融合，最终获取特征图。特征图经过不同的卷积获取出来的数据和标注的数据集进行对比，通过 ADAM 优化器训练网络模型，获取较优的模型参数。本实验的优化器采用 ADAM 优化器，每个 batch size 等于 8，ADAM 的学习率从 1E−3 开始，每 10,000 批次衰减十分之一，训练到 34 万次获取到最优解。

将文中算法与其他算法[13] [14] [15]在 ICDAR2015 数据集上进行比较，结果如表 1 所示。

Table 1. Comparison of different text detection algorithms
表 1. 不同文本检测算法比较

序号	模型	Precision (%)	Recall (%)	F-score (%)
1	CTPN + VGG16	74.20	51.60	60.90
2	Seglink + VGG16	76.80	73.10	75.00
3	EAST + VGG16	80.50	72.80	76.40
4	EAST + ResNet-50	81.66	77.32	79.43
5	EAST + PVANET2x	83.60	73.50	78.20
6	EAST + PVANET2x MS	84.64	77.23	80.77
7	本文算法	86.24	77.84	81.82

其中本算法在 ICDAR2015 文本定位任务上的召回率(针对原样本而言的，它的含义是在实际为正的样本中被预测为正样本的概率)为 77.84%，精准率(针对预测结果而言的，它的含义是在所有被预测为正的样本中实际为正的样本的概率)为 86.24%，F-score 为 81.82%，优于经典 EAST 算法。

通过对比原有 EAST 算法以及改进 EAST 算法的检测效果，如图所示。

图 5 中包含两组自然场景文本图片，每组图片中的左侧为原始 EAST 算法的检测效果，右侧为改进 EAST 算法的检测效果。从图中可以看出原始 EAST 算法在检测长文本会遗漏部分文本，以及文本检测的边界过长导致部分没有联系的文本本框选，而本算法通过扩大感受野和增强文本之间的连续性，可以检测出跟多的文本，以及正确的框选文本的合理边界，更好地检测出自然场景文本区域，提高检测的准确率。

实验通过利用 ASPP 网络替代原来 EAST 算法的主干网络 resnet-50 的 stage 4，再通过添加 BLSTM 网络增强特征图序列化关联。实验结果比经典的 EAST 算法具有更高的精确率和召回率，F-score 到达 81.82%。总体相对于经典的 EAST 算法有着一定的提高。

Figure 5. The detection effect of two EAST algorithms
图 5. 两种 EAST 算法检测效果

4. 结论

本文在经典的 EAST 算法的基础上进行改进来实现文本检测。由于经典的 EAST 算法存在感受野不够大问题，通过利用 ASPP 网络来替代 EAST 算法的主干网络的 stage 4 模块，在不损失分辨率的情况下提高 EAST 算法的感受野。同时 EAST 算法也存在文本检测边框过长和过短的问题，通过添加 BLSTM 网络，增加文本特征图序列之间的关联，提高了文本检测分界线的效果。相比于经典的 EAST 算法，本文实现的算法在精确率和召回率上都提高了。同时本文仍存在不足，在后续的实验中可以通过调整参数和利用 Dense ASPP 网络来替代 ASPP 网络改进算法，进一步提高文本检测的效果。

基金项目

北京市自然科学基金项目(4202025)，国家自然科学基金项目(61872333)，北京教委科技计划项目(KM201911232003)，北京未来芯片技术高精尖创新中心科研基金(KYJJ2018004)。

参考文献

[1] Mori, S. (1992) Historical Review of OCR Research and Development. *Proceedings of the IEEE*, **80**, 1029-1058. https://doi.org/10.1109/5.156468

[2] Goodfellow, I.J., Bulatov, Y., Ibarz, J., *et al*. (2013) Multi-Digit Number Recognition from Street View Imagery Using Deep Convolutional Neural Networks. Computer Science.

[3] Graves, A., Fern'andez, S., Gomez, F., *et al*. (2006) Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, 25-29 June 2006, 369-376.

[4] Tian, Z., Huang, W., He, T., *et al*. (2016) Detecting Text in Natural Image with Connectionist Text Proposal Network. *European Conference on Computer Vision*, Amsterdam, 11-14 October 2016, 56-72. https://doi.org/10.1007/978-3-319-46484-8_4

[5] Shi, B., Bai, X. and Belongie, S. (2017) Detecting Oriented Text in Natural Images by Linking Segments. 2017 *IEEE*

Conference on Computer Vision and Pattern Recognition, Honolulu, 21-26 July 2017, 3482-3490. https://doi.org/10.1109/CVPR.2017.371

[6] Zhou, X.-Y., Yao, C., Wen, H., *et al*. (2017) EAST: An Efficient and Accurate Scene Text Detector. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2642-2651. https://doi.org/10.1109/CVPR.2017.283

[7] He, Y., Zhu, C., Wang, J., *et al*. (2019) Bounding Box Regression with Uncertainty for Accurate Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 2888-2897. https://doi.org/10.1109/CVPR.2019.00300

[8] Hu, H., Zhang, C., Luo, Y., *et al*. (2017) WordSup: Exploiting Word Annotations for Character Based Text Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4950-4959. https://doi.org/10.1109/ICCV.2017.529

[9] Redmon, J., Divwvala, S., Girshick, R., *et al*. (2016) You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[10] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[11] 杨飚, 杜晓宇. 基于改进 EAST 的自然场景文本定位算法[J]. 计算机工程与应用, 2019, 55(18): 161-165.

[12] 池凯, 赵逢禹. 改进 EAST 算法的游戏场景文本检测[J]. 小型微型计算机系统, 2020, 41(10): 2189-2193.

[13] Yang, P., Rong, G., Peng, G., *et al*. (2011) Research on Lip Detection Based on Opencv. 2011 *International Conference on Transportation*, *Mechanical and Electrical Engineering* (*TMEE*), Changchun, 16-18 December 2011, 1465-1468.

[14] Zhi, T., Huang, W., Tong, H., *et al*. (2016) Detecting Text in Natural Image with Connectionist Text Proposal Network. In: *European Conference on Computer Vision* (*ECCV*), Springer, Cham, 56-72. https://doi.org/10.1007/978-3-319-46484-8_4

[15] Deng, D., Liu, H., Li, X., *et al*. (2018) PixelLink: Detecting Scene Text via Instance Segmentation. 2018 *the Association for the Advance of Artificial Intelligence* (*AAAI*), New Orleans, 2-7 February 2018, 6773-6780.