计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):171−175 [doi: 10.15888/j.cnki.csa.007779]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

自然场景下的密集文本检测方法①

牟　森, 陈洪刚, 卿粼波, 何小海, 王思怡

(四川大学 电子信息学院, 成都 610065)
通讯作者: 陈洪刚, E-mail: honggangchen.scu@gmail.com

摘　要: 自然场景下的文本检测任务是图像处理领域中的难点之一. EAST (Efficient and Accurate Scene Text detector) 算法是近年来比较出色的文本检测算法, 但是增加后置处理之后的 AdvancedEAST 算法仍存在由于激活像素的头尾边界丢失导致的漏检情况, 对密集文本的检测效果也不是很理想. 因此提出了 Dilated-Corner Attention EAST (DCA_EAST) 改进算法, 对网络结构加入空洞卷积模块以及角点注意力模块, 改善了漏检情况. 针对损失函数, 加入类别权重因子和样本难度权重因子, 有效提升了密集文本的检测效果. 实验结果表明, 该算法在 ICDAR2019 的 ReCTS 数据集上准确率为 93.02%, 召回率为 76.69%, F-measured 值为 84.07%, 优于 AdvancedEAST 算法.

关键词: 密集文本检测; AdvancedEAST 算法; 空洞卷积; 角点注意力; 样本难度权重

引用格式: 牟森,陈洪刚,卿粼波,何小海,王思怡.自然场景下的密集文本检测方法.计算机系统应用,2021,30(2):171–175. http://www.c-s-a.org.cn/1003-3254/7779.html

Dense Text Detection Method in Natural Scene

MOU Sen, CHEN Hong-Gang, QING Lin-Bo, HE Xiao-Hai, WANG Si-Yi

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Text detection in natural scenes is one of the difficulties in the field of image processing. An efficient and accurate scene text detector (EAST) algorithm is an excellent text detection algorithm in recent years, but the AdvancedEAST algorithm after the addition of post processing still has the problem of missed detection caused by the loss of the head and tail boundaries of the activated pixels. Thus, the detection effect of dense texts is not ideal. For this reason, an improved algorithm of dilated-corner attention EAST (DCA_EAST) is proposed, and a dilated convolution module and a corner attention module are added to the network structure to improve the missed detection. For the loss function, weight factors of category and sample difficulty are introduced to effectively improve the detection effect of dense texts. The experimental results show that the proposed algorithm has an accuracy of 93.02%, a recall rate of 76.69%, and an F-measured value of 84.07% on the ReCTS dataset of ICDAR2019, thus being superior to the AdvancedEAST algorithm.

Key words: dense text detection; AdvancedEAST algorithm; dilated convolution; corner attention; sample difficulty weight

自然场景下的文本检测与识别被认为是目标检测领域中最具有挑战性的难点之一, 它在图像处理、无人驾驶、文档分析、自然语言处理等诸多机器视觉领域都存在大量的应用. 相较于通用物体的目标检测, 复杂场景下的文本检测存在诸多难点: (1) 场景中的文本行颜色、字体、尺度多样化并且相关性较小. (2) 背景

① 收稿时间: 2020-06-18; 修改时间: 2020-07-14; 采用时间: 2020-07-23; csa 在线出版时间: 2021-01-27

多样化. 在自然场景下, 文本行的背景是任意的, 还可能会受到结构相近的背景的影响 (如栅栏). (3) 文本行的形状和方向多样化. 如水平、垂直、倾斜、弯曲等. (4) 存在诸多艺术字、手写字、多种语言混合以及不同程度的扭曲. (5) 恶劣的光照条件和不同程度的遮挡.

近年来, 文本检测领域的深度学习策略主要有: (1) 基于字符的文本检测. Baek 等[1] 提出先检测单个字符 (character region score) 及字符间的连接关系 (affinity score), 然后根据这些连接关系确定最后的文本行, 再采用高斯热度图来生成区域分数和连接分数两个特征图, 最后借助文本行的长度进行弱监督训练. (2) 基于文本框的坐标回归的文本检测. Tian 等[2] 使用一连串小尺度文本框来实现文本检测的任务, 并且引入 RNN 模型提高文本的检测效果, 用边界优化使文本框的边界预测更加精准; Liao 等[3] 提出的端到端的神经网络模型, 修改了锚点 (anchors) 尺寸和卷积核尺寸, 采用多个尺度的预测, 来提高对 anchors 没有覆盖到的长文本的检测效果. Liao 等后来又针对该模型进行了改进[4], 实现了预测旋转的文本框; Shi 等[5] 提出文本行检测的两个基本组成元素: 分割 (segment) 和连接 (link), 并且提出了两种 link 类型: 层内连接 (within-layer link) 和跨层连接 (cross-layer link); Zhou 等[6] 提出一个快速、准确的两阶段文本检测方法. (3) 基于语义分割后进行实例分割的方法. Deng 等[7] 提出通过实例分割结果提取文本的位置, 并且将像素点进行连接得到文本框. 使用像素分类实现语义分割, 使用链接实现实例分割. Wang 等[8] 提出了一种渐进性的扩展网络, 它可以实现对任意形状文本实例的检测. 该方法使用了最小内核的思想完成实例分割, 在此基础上渐进式地使用不同内核来补充实例分割的区域. (4) 文本框回归和语义分割的组合方法. Zhang 等[9] 提出了一个新型端到端文本检测器, 它由 3 部分组成: 直接回归模块 (DR)、迭代修正模块 (IRM)、形状表征模块 (SEM). 首先由直接回归模块产生粗略的四边形候选文本框; 然后通过迭代修正得到完整的文本行的特征块; 最后根据文本行的区域、中心线及边界偏移得到最终的文本行.

Zhou 等[6] 提出的 EAST 算法在准确性和总体效率方面明显优于同领域内之前提出的其他方法, 后有人对其增加了后置处理 (AdvancedEAST[10]). 本文提出 Dilated-Corner Attention EAST (DCA_EAST) 改进算法, 在 AdvancedEAST 网络结构加入空洞卷积模块以及角点注意力模块, 改善了漏检情况. 对损失函数改进, 加入类别权重因子和样本难度权重因子, 有效提升了密集文本的检测效果.

1 AdvancedEAST 算法分析

AdvancedEAST 包括全卷积网络 (Fully Convolutional Networks, FCN) 阶段和非极大值抑制 (Non-Maximum Suppression, NMS) 合并阶段. FCN 可以直接生成文本区域, 消除冗余过程及复杂的中间步骤. 该方法既可以检测单词, 又可以检测文本行, 检测的形状可以为任意形状的四边形. 针对文本行的特点, 使用了位置感知 NMS (Locality-Aware NMS) 来对生成的文本区域进行过滤, 降低了 NMS 的复杂度. AdvancedEAST 网络结构图 (如图 1), 分为特征提取主网络 (4 个级别的特征图, 表示为f_i)、特征合并分支 (依次将主网络中 1/32, 1/16, 1/8, 1/4 特征图进行合并) 以及输出层: 是否在文本框内 (score map), 是否属于文本框边界像素以及是头还是尾 (vertex code), 预测的 2 个对角线顶点坐标 (vertex coord).

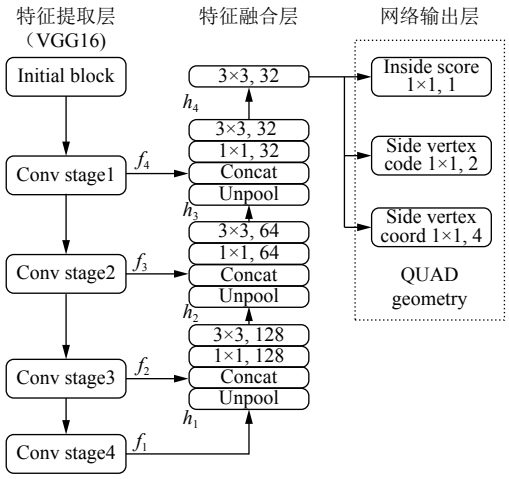


图 1 AdvancedEAST 网络结构图

对于密集文本的检测, AdvancedEAST 算法存在感受野受限的问题; 并且在预测生成激活像素的过程中, 存在头或尾边界像素丢失的情况, 导致文本框漏检, 如图 2 所示.

2 Dilated-Corner Attention EAST 结构

2.1 网络结构优化

针对上述问题, 本文在 AdvancedEAST 算法的基

础上引入了空洞卷积模块 (dilated conv module) 以及角点注意力机制 (corner attention module), 改进算法的网络结构如图 3 所示.

(a) 原图 (b) 激活像素

图 2 AdvancedEAST 算法的图象激活像素

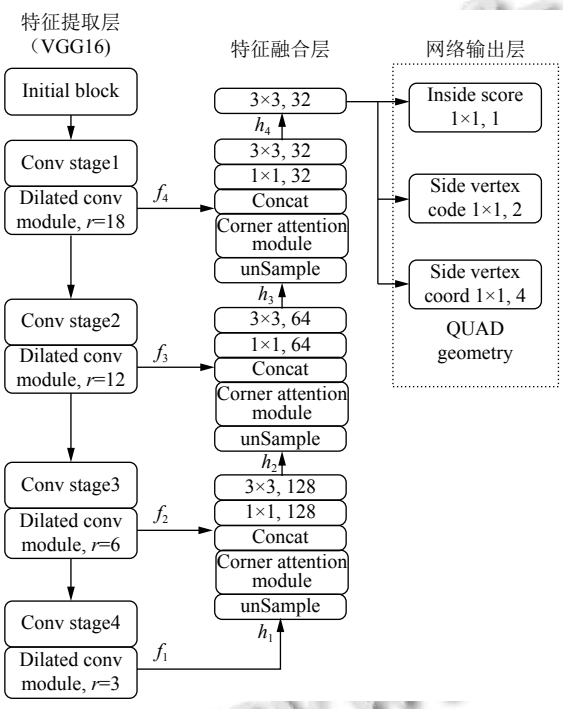


图 3 本文算法网络结构图

在 4 个特征图输出之前, 分别采用扩张率为 18、12、6 和 3 的 3×3 空洞卷积来增加网络的感受野.

为了减少激活过程中头或尾边界像素丢失的情况, 本文在对不同尺度的图片特征提取后, 由上至下对每一个层次的特征进行融合, 将特征融合阶段的上采样 (unpool) 改为双线性上采样, 并利用当前层次融合的特征对目标位置进行预测. 这样相较于标准方法来说, 可以生成更均匀的特征金字塔, 包含更多的上下文信息. 并且本文加入角点注意力模块, 目的是融入后置处理中的边界像素特征.

假设注意力模块需要处理的特征序列为$s=\{s_1, s_2, s_3, \cdots, s_n\}$, 其中 n 表示特征向量的个数. 最基础的形态注意力机制的公式如下:

$$c_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t \tag{1}$$

$$\alpha_{t't} = Softmax(\sigma(s_{t'-1}, h_t)) \tag{2}$$

其中, t 表示当前时间, $c_{t'}$表示输出变量, h_t表示隐藏层, $\alpha_{t't}$表示一个权重的概率分布, σ是一个单层的感知机.

常用的通道注意力机制和空间注意力机制对于特征图边界像素的关键信息提取效果并不理想, 故本文采用了角点注意力机制, 具体地是将特征图的输入边像素界特征与输出的边界像素特征通过一个标准的一维全连接层 (dense layer) 连接起来, 公式如下:

$$attention_i = Softmax(Dense(x_i, y_{i-1})) \tag{3}$$

$$c_i = \sum_{i=1}^{m} attention_i \times x_i \tag{4}$$

其中, i 表示当前时步, x_i为输入边界像素特征, y_{i-1}为输出的边界像素特征, $attention_i$表示 i 处的注意力权重, c_i表示输出的带有注意力的上下文信息.

2.2 针对密集文本的损失函数设计

在一般的数据集中, 负样本数量太大, 导致损失函数输入参数的大部分都是负样本, 并且很多是容易分类的, 因此会使得对密集文本的检测效果并不是很好. 之前也有一些算法来处理这种类别不均衡的问题, 比如 OHEM (Online Hard Example Mining), OHEM 算法虽然增加了错分类样本的权重, 但是 OHEM 算法忽略了容易分类的样本.

故本文在标准交叉熵损失函数[11] 的基础上引入了类别权重因子α和样本难度权重因子$(1-\tilde{Y})^{\gamma}$, 来缓解上述问题, 提升模型精确. α可以平衡正负样本, γ可以调节简单样本权重降低的速率, γ>0 可以减少易分类样本的损失, 使得模型更关注于困难的、错分的样本. 在产生区域文本框的阶段, 通过得分和 NMS 筛选可以过滤大量的负样本, 然后在分类和回归阶段又可以固定正负样本的比例. 对于不同的γ值, 模型的平均精确度 (Average Precision, AP) 具有不同的表现, 经测试, γ=2时表现最好, 结果如表 1 所示.

表 1 本文不同γ值对应的 AP 表现

γ	0	0.1	0.25	0.5	0.75	1	2	5
AP	30.2	30.5	31.2	32.3	33.2	34.0	**34.7**	32.6

Score map 和 vertex code 的损失函数公式如下:

$$L_s=-(\alpha Y^*(1-\tilde{Y})^\gamma \log\tilde{Y}+(1-\alpha)(1-Y^*)\tilde{Y}^\gamma\log(1-\tilde{Y})) \quad (5)$$

$$L_v=\frac{-\sum_i^N(\alpha y_i^*(1-\tilde{y}_i)^\gamma\log\tilde{y}_i+(1-\alpha)(1-y_i^*)\tilde{y}_i^\gamma\log(1-\tilde{y}_i))w_i}{\sum_i^N w_i} \quad (6)$$

其中, Y^*表示正确标注, $\tilde{Y}$表示预测值, N 表示样本数量. α表示所有训练图像中为 1 的像素点数量占总像素点数量的比例, 这是个先验值, 在标签生成中就可得到, 具体地定义为:

$$\alpha=1-\frac{\sum_{y^*\in Y^*}y^*}{|Y^*|} \quad (7)$$

其中, w为归属权重:

$$w_i=\begin{cases}1, & y_i^*=1\\ 0, & y_i^*=0\end{cases} \quad (8)$$

其中, $y_i^*\in Y^*$.

对于 vertex coord 的损失函数, 本文采用加权的 *Smooth* L_1 函数. 相比于 L_1 损失函数, *Smooth* L_1 可以收敛得更快, 相较于 L_2 损失函数来说, *Smooth* L_1 对异常值、离群点不敏感, 梯度的变化相对更小, 训练更稳定. 损失函数的定义如下:

$$L_g=\frac{\sum_i^N S_i w_i}{\sum_i^N w_i} \quad (9)$$

其中, w为式 (8) 中的权重, *Smooth* L_1 函数定义如下:

$$S=smooth_{L_1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{otherwise}\end{cases} \quad (10)$$

综上, 得到总的损失函数为:

$$L=\lambda_s L_s+\lambda_v L_v+\lambda_g L_g \quad (11)$$

其中, λ_s、λ_v和λ_g分别为 score map、vertex code 和 vertex coord 权重.

3 实验与分析

3.1 实验环境与设置

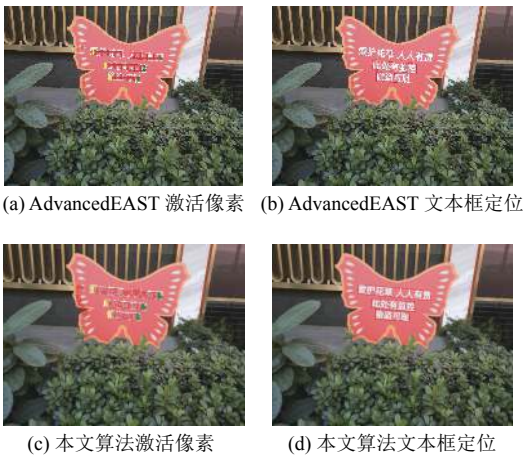
本次实验在 Ubuntu 18.04.3 LTS 上进行, 开发语言为 Python 3.6.9. GPU 版本为 NVIDIA GTX 1080Ti, 显存 11 GB.

3.2 模型训练

采用 Adam[12] 优化器对本文提出的模型进行端到端训练. 损失函数参数γ=2, λ_s=4, λ_v=1, λ_g=1. 数据集采用的是 ICDAR2019 挑战赛所用的 ReCTS, 该数据集主要是中英文招牌, 包括 20 000 张训练图片和 5000 张测试图片. 由于图片尺寸跨度较大, 故本次实验采用多尺度训练的方式对原始图像进行训练, 以改善模型对不同尺度的图片文本检测的鲁棒性. Batch size 设为 8, Adam 学习率从1e^{-3}开始, 5 个 epoch 后无改善则下降到1e^{-5}, 进行网络训练.

3.3 实验结果

本文算法与 AdvancedEAST 算法在自然场景下的文本检测结果对比如图 4、图 5 所示.



(a) AdvancedEAST 激活像素 (b) AdvancedEAST 文本框定位

(c) 本文算法激活像素 (d) 本文算法文本框定位

图 4 图象激活像素和文本框定位

对比可以发现, 图 4(a) 中存在头或尾边界像素丢失而导致的文本框漏检情况, 图 5(a) 中存在对于密集文本检测不到的情况. 通过本文算法处理后, 激活像素连通性更好, 头尾像素也更加丰富, 密集文本的检测效果明显改善, 如图 4(c)、图 5(b) 所示. 同时, 本文使用准确率 (Precision)、召回率 (Recall) 和加权调和平均值 F-measure 三个指标来评价本文算法的性能, 并与 AdvancedEAST 算法进行对比, 实验结果如表 2 所示. 可以看出, 本文算法相比于 AdvancedEAST 算法在文

本检测的各项指标上均有提升. 其中召回率提升比较明显, 这是因为本文算法增大了困难正样本的检测能力.

(a) AdvancedEAST 算法的密集文本检测效果图

(b) 本文算法的密集文本检测效果图

图 5 密集文本检测效果图

表 2 本文算法与 AdvancedEAST 文本检测算法实验结果对比

算法	Precision	Recall	F-measure
AdvancedEAST[10]	89.46	61.07	72.59
本文算法	**93.02**	**76.69**	**84.07**

4 结论

本文算法在 AdvancedEAST 算法的基础上, 引入了 Dilated-Corner Attention EAST, 增大网络特征提取的感受野, 可捕获更多激活过程中边界的上下文信息, 改善了文本定位中出现的文本框漏检情况; 同时, 对损失函数的改进, 平衡了样本的类别权重以及样本难度权重, 最终有效提升了密集文本的检测效果. 与 AdvancedEAST 相比, 准确率、召回率和 F-值均有提高.

参考文献

1 Baek Y, Lee B, Han D, *et al*. Character region awareness for text detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 9365–9374.

2 Tian Z, Huang WL, He T, *et al*. Detecting text in natural image with connectionist text proposal network. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 56–72.

3 Liao MH, Shi BG, Bai X, *et al*. TextBoxes: A fast text detector with a single deep neural network. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4161–4167.

4 Liao MH, Shi BG, Bai X. TextBoxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing, 2018, 27(8): 3676 –3690. [doi: 10.1109/TIP.2018.2825107]

5 Shi BG, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. Proceedings of 2017 IEEE Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3482–3490.

6 Zhou XY, Yao C, Wen H, *et al*. EAST: An efficient and accurate scene text detector. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2642–2651.

7 Deng D, Liu HF, Li XL, *et al*. PixelLink: Detecting scene text via instance segmentation. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, LA, USA. 2018. 6773–6780.

8 Wang WH, Xie EZ, Li X, *et al*. Shape robust text detection with progressive scale expansion network. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 9336–9345.

9 Zhang CQ, Liang BR, Huang ZM, *et al*. Look more than once: An accurate detector for text of arbitrary shapes. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 10552–10561.

10 AdvancedEAST.https://github.com/huoyijie/AdvancedEAST. [2020-05-12]

11 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2999–3007.

12 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2015.