

引用格式:贾颖,程艳云. 面向自然场景的中文文本检测[J]. 微电子学与计算机, 2022, 39(1): 25-33. [JIA Y, CHENG Y Y. Chinese text detection for natural scenes[J]. Microelectronics & Computer, 2022, 39(1): 25-33.]DOI: 10. 19304/J. ISSN1000-7180. 2021. 0897

面向自然场景的中文文本检测

贾 颖,程艳云

(南京邮电大学 自动化学院、人工智能学院,江苏 南京 210000)

摘 要: 随着深度学习的发展,自然场景的文本检测取得了进步,但多方向和弯曲中文文本检测效果仍不理想. 针对多方向和弯曲中文文本的检测问题,提出一种融合注意力机制的多尺度文本检测方法. 为了平衡模型准确性和降低计算复杂度,采用轻量级 Resnet18 为主干网络. 针对特征金字塔(FPN)提取的特征分布不确定性的问题,嵌入平衡注意力机制(BAM)提取有效文本特征并抑制低效特征通道,进而提升检测方法的鲁棒性. 针对空洞空间金字塔池化网络(ASPP)下采样时图像局部信息和细节信息丢失的问题,改进 ASPP 以降低特征图分辨率的损失. 针对 FPN 提取特征不足以及感受野小的问题,将嵌入注意力机制的 FPN 和改进的 ASPP 并行增强特征提取融合. 针对正负样本的不平衡性的问题,基于可微二值化模块在二值图损失中引入对数化的 AC Loss,从而增强检测模型的泛化能力. 在公开数据集 MSRA-TD500 上的实验结果表明,该算法与目前快速高效的 DBnet 相比,准确率、召回率和 F 值分别提升 0.1%、1.4%和 0.6%,并且该算法的检测速率也有较好表现.

关键词: 文本检测;特征金字塔;BAM 注意力机制;可微二值化;AC Loss

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-7180(2022)03-0025-09

Chinese text detection for natural scenes

JIA Ying, CHENG Yanyun

(School of Automation, School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: With the development of deep learning, the text detection of natural scenes has made progress, but the detection effect of multi-directional and curved Chinese text is still not ideal. A multi-scale text detection method is proposed, which integrates attention mechanism for the detection of multi-directional and curved Chinese text. In order to balance model accuracy and reduce computational complexity, a lightweight Resnet18 backbone network is adopted. Aiming at the problem of the uncertainty of the feature distribution extracted by the feature pyramid (FPN), the embedded balanced attention mechanism (BAM) extracts effective text features and suppresses inefficient feature channels, thereby improving the robustness of the detection method. Aiming at the problem of the loss of image local information and detail information during downsampling of the Hollow Space Pyramid Pooling Network (ASPP), ASPP is improved to reduce the loss of feature map resolution. Aiming at the problem of insufficient FPN feature extraction and small perception field, the FPN embedded in the attention mechanism and the improved ASPP parallel enhanced feature extraction are fused. Aiming at the problem of the imbalance of positive and negative samples, the logarithmic AC Loss is introduced into the binary graph loss based on the differentiable binarization module, thereby enhancing the generalization ability of the detection model. The experimental results on the public data set MSRA-TD500 show that compared with the current fast and efficient DBnet, the accuracy, recall and F value of this algorithm are increased by 0.1%, 1.4% and 0.6% respectively, and the detection rate of this

algorithm is also has a good performance.

Key words: text detection; feature pyramid; BAM attention mechanism; differentiable binarization; AC Loss

1 引言

文本在日常生活中无处不在,其携带着丰富且精确的信息. 自然场景文本检测技术可以广泛地应用到自动驾驶、网络安全、地理定位、智能交通等领域. 中文是承载文化的重要工具,因此,研究自然场景中文文本检测有着重要意义. 但自然场景图像背景复杂,且中文文本具有不同尺度和形状,包括水平、多方向和弯曲的文本,因此检测文本实例区域具有一定的挑战性. 传统文本检测方法主要基于滑窗和强连通分支,其步骤繁多且鲁棒性和可靠性差. 目前基于深度学习的场景文本检测方法大致分为基于回归,基于组件和基于分割三类.

基于回归的方法分为一阶段法和二阶段法. 一阶段法直接回归文本框的坐标,二阶段法包括生成候选文本框和细化文本框. 两阶段法检测精度高,但计算复杂度高. 廖等人^[1]提出的 Textboxes 算法,其修改卷积核锚定比例和形状以适应文本的各种纵横比. 大多数基于回归的算法利用四边形边界框定位文本,其检测任意形状文本效果差.

基于组件的方法首先定位单个部分或字符,而后处理将其组合成一个字符或单词. 黄等人^[2]提出了 CTPN 算法准确定位水平文本,但无法检测不规则文本. Shi 等人^[3]提出了 SegLink 算法,用多尺度 SSD^[4]作为主干网络,先检测局部片段,再按照其合并规则将相应片段进行连接得到最终的文本行,但其检测弯曲文本效果差.

基于分割的方法是在像素级别推断分割出图像中的文本. 2018 年, Lyu 等人^[5]将 Mask R-CNN^[6]应用于文本行检测并提出 Mask TextSpotter 算法,在字符级别上进行文本分割,其可以检测任意形状的文本,但需要字符级标注来进行训练. 邓等人^[7]提出 PixelLink 算法,通过预测不同文本实例之间的像素连接来分离彼此靠近的文本. 但其需要针对不同数据集调整 pixel 和 link 两个阈值并设计不同的后处理方法,其次是模型检测速度很慢,且无法有效处理背景复杂的数据. 2019 年,王等人^[8]提出 PSEnet 算法,以渐进比例扩展方式分割距离相近的文本实例,其检测弯曲文本的效果优于前者算法,但

它需要合理的选择和优化超参数. 2020 年,廖等人^[9]提出可微分模块 DB 解决了训练带来的梯度不可微问题. 但是,它只注重预测正确标签的准确性,而忽略了其他非正确标签的差异,从而导致学习到的特征比较分散. 实际自然场景中文本不规则且背景复杂,大多数基于分割的算法后处理步骤复杂,且无法解决彼此相邻文本的覆盖问题. 为了解决以上这些问题,本文提出一种融合注意力机制的多尺度自然场景中文文本检测方法 (Multi-scale natural scene Chinese text detection method fused with attention mechanism, MSAM).

2 融合注意力机制的多尺度网络

2.1 总体网络架构

本文提出的 MSAM 网络架构如图 1 所示,具体分为三个模块,分别是特征提取模块,特征融合模块和改进的可微二值化模块 (Enhanced Differentiable Binarization, EDB). 本文主要内容如下:为了显著减少模型的体积和推理时间,采用 Resnet18 作为轻量级主干网络. 与其他大型网络相比, Resnet18 的浅层网络导致 FPN 提取特征不足,此外,其感受野小会导致漏检长中文文本. 因此,将 FPN 与改进的空间空洞金字塔分支 (Improved Atrous Space Pyramid Pooling, IASPP) 并行提取足量的文本信息特征,同时增大感受野捕捉长文本特征信息. 针对上方模块 ASPP 下采样时特征图的局部信息与边缘信息的缺失问题,增加下方模块与上方模块 ASPP 联合降低分辨率的损失. 针对特征图的分布不确定性,在 FPN 分支中嵌入平衡注意力机制 (Balanced attention mechanism, BAM) 对特征进行处理,提取有效文本特征. 针对正负样本之间的不平衡性和模型检测速率慢的问题,采用 DB 模块并加权概率图与阈值图损失函数,同时在二值图中引入对数化的 AC Loss,将分割问题转变为求解能量泛函最小值,当能量达到最小时,活动轮廓收敛至检测的文本边缘,进而增强模型的泛化能力.

2.2 特征提取与注意力模块

FPN 如图 2 所示,自底向上分别提取 $1/2, 1/4, 1/8, 1/16, 1/32$ 不同尺度特征图,分别记为 $C_1, C_2,$

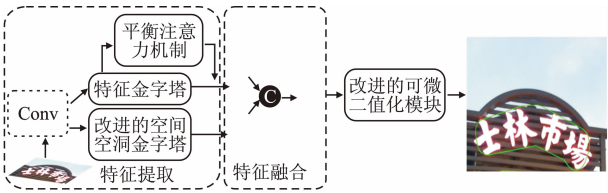


图 1 MSAM 网络架构

Fig. 1 MSAM Network Architecture

C_3, C_4, C_5 . 自顶向下将 C_2, C_3, C_4, C_5 分别上采样 1 倍、2 倍、4 倍和 8 倍至 $1/4$ 尺度大小并横向连接至前一层特征. 横向连接如图 2 右下角放大区域所示, 先将高层特征图进行 2 倍上采样再将其与经过 1×1 卷积层的前一层特征融合. 再者使用 3×3 卷积核分别处理已融合的特征图来生成具有相同的尺度大小的 P_2, P_3, P_4, P_5 特征映射. 最后, 将 P_2, P_3, P_4, P_5 级联得到特征 F_0 .

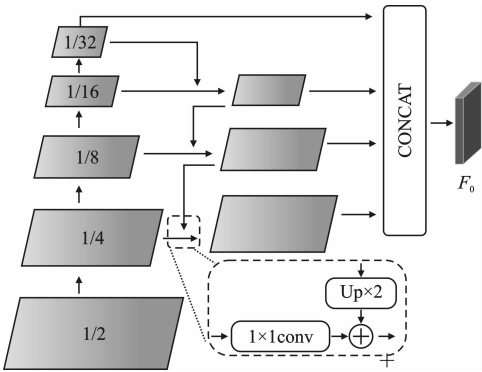


图 2 特征金字塔网络

Fig. 2 Feature Pyramid Networks

IASPP 分支如图 3 所示, 上方模块 ASPP 对主干网络输出的特征 F 分别以 1、6、12 和 18 不同采样率的空洞卷积并行采样. 其中 C 表示级联.

上方模块 ASPP 在下采样得到高语义信息的同时, 由于分辨率的降低会丢失输入图像的细节特征信息. 为了精确地从自然场景图像中分割出中文文本, 需要低层特征图提供重要的边缘特征信息. 所以设计将上方模块 ASPP 的输出特征通过一个 1×1 大小的卷积层来降低通道数至需要的数值, 然后进行 4 倍双线性插值上采样得到特征图 F_2 . 下方模块从主干网络中输出一张与特征图 F 分辨率一致的特征图 F_1 , F_1 经过 1×1 大小的卷积层扩大通道数得到特征图 F_3 , 最后将 F_2 与 F_3 融合得到 F^* . 上方模块使用 ASPP 结构提取不同尺

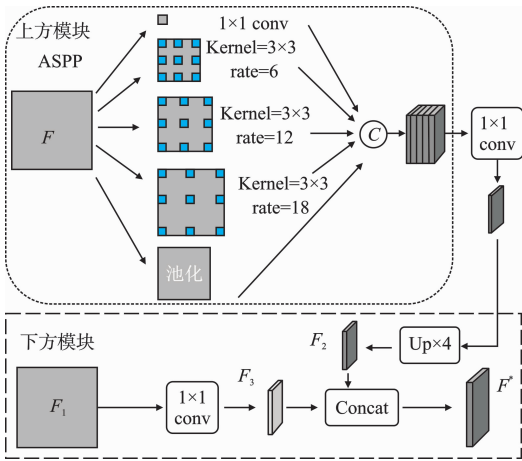


图 3 改进的空洞空间金字塔池化

Fig. 3 Improvedatrous spatial pyramid pooling

度的空间信息, 得到一个包含抽象化特征信息的输出, 而下方模块输出一个包含丰富局部信息和边缘信息的低层特征以补充细节信息再与上方模块联合预测特征图.

FPN 将多层级特征融合达到预测效果, IASPP 采用不同膨胀速率的空洞卷积层捕获多尺度特征信息. 两个分支互相补充, 提升了整个网络的性能.

针对特征图分布的不确定性问题, 在 FPN 中嵌入轻量级平衡注意力机制 (BAM), 如图 4 所示. BAM 既注重不同通道像素的重要性, 又注重同一通道不同位置像素的重要性. 在训练时不断调整各通道的权值, 提取有效文本特征信息并抑制低效特征通道.

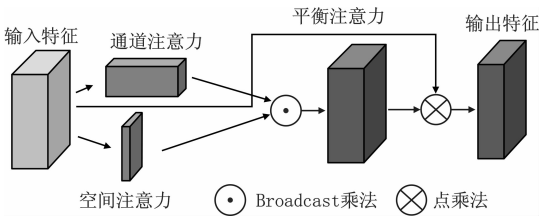


图 4 平衡注意力机制

Fig. 4 Balanced Attention mechanism

BAM 由平均池化通道注意力模块 (ACAM) 和最大值池化空间注意力模块 (MSAM) 组成. BAM 与 CBAM^[10] 的连续串联结构不同, 其并行结构可以避免误差累积以及通道注意力与空间注意力之间的串扰. 其次 BAM 能在控制参数数量的同时进一步提高文本检测模型的性能.

输入特征图通过 ACAM 和 MSAM 分别生成

垂直和水平注意力权重,其中 ACAM 输出大小为 $N \times C \times 1 \times 1$,MSAM 输出大小为 $N \times 1 \times H \times W$,使用 broadcast 乘法将它们融合成一个 $N \times C \times H \times W$ 张量再与输入特征图对应元素相乘. N 是批量大小, C 是特征图的通道数, H 和 W 分别是特征图的高和宽.

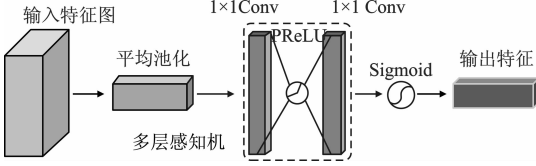


图 5 ACAM 模块

Fig. 5 ACAM module

ACAM 如图 5 所示. 通道注意力从输入特征图中找到重要信息的通道并赋值其更大的权重. 因为最大值池化会选择 $H \times W$ 维通道里的异常极值点并得到错误的注意力权重,所以只使用平均池化去提取通道信息,再使其通过两点式卷积层组成的多层感知机 (MLP). 为了增加 MLP 的非线性, PReLU 用于激活第一个卷积层的输出. 此外,为了减少 ACAM 的参数数量和计算复杂度, MLP 采用瓶颈架构. 第一层卷积层的输入通道数是输出通道数的 r 倍. 在 PReLU 激活后,通道数量由第二层卷积层恢复. 最后通道权重由 sigmoid 激活函数生成. ACAM 的公式为:

$$A(x) = S[f_{n/r \times n}^{1 \times 1}(\text{PReLU}(f_{n/r \times n}^{1 \times 1}(\text{AvgPool}(x))))] \quad (1)$$

式中, S 为 sigmoid 函数, $f_{n/r \times n}^{1 \times 1}$ 表示卷积核大小为 1×1 的卷积层, n 为输入通道数,输出通道数为 n/r . PReLU 和 Sigmoid 定义为:

$$\text{PReLU} = \begin{cases} x, & x > 0 \\ a \cdot x, & x \leq 0 \end{cases} \quad (2)$$

$$S(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

MSAM 如图 6 所示. 与通道注意力不同,空间注意力为输入特征图的水平部分生成权重. 其目标是找到对最后恢复高分辨率贡献最大的横向区域并赋予它们更大权重. 这些区域通常在通道中以极值的形式包含高频细节. 因此,空间注意力只使用最大值池化. 最大值池化的输出通过一个具有 7×7 大小的感受野卷积层后再被 sigmoid 函数激活后得到空间注意力权重. MSAM 的公式为:

$$M(x) = \text{Sigmoid}[f_{1 \times 7}^{7 \times 7}(\text{Maxpool}(x))] \quad (4)$$

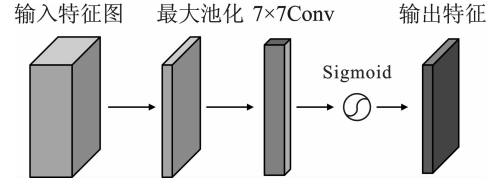


图 6 MSAM 模块

Fig. 6 MSAM module

2.3 改进的可微分二值化模块

2.3.1 标注与文本包围盒的生成

文本的标注以红实线显示. 收缩和扩张的多边形分别以蓝虚线和绿点线显示,如图 7 所示. 其中原始多边形 G 为:

$$G = \{S_k\}_{k=1}^n \quad (5)$$

式中, n 是文本多边形的顶点数量. 使用 Vatti 裁剪算法将 G 收缩为 G_s . 收缩偏移量 D 为:

$$D = \frac{A^2(1-r^2)}{L^2} \quad (6)$$

式中, A 和 L 分别是原始多边形面积和周长, r 是收缩率,根据经验设为 0.4. 同理阈值图使用相似步骤生成标注. G 以偏移量 D 扩张到 G_d . 前景与背景间的边界即为 G_s 和 G_d 之间的间隙.

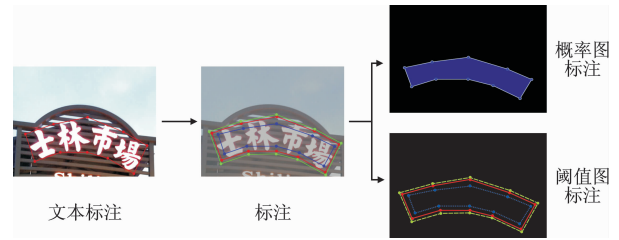


图 7 标注生成

Fig. 7 Differential binarized module

可微分二值化模块如图 8 所示,首先利用融合的特征图预测概率图 P 和阈值图 T ,将概率图或近似二值图使用固定的阈值进行二值化处理得到二值图. 使用近似阶跃函数进行二值化,即:

$$B_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \quad (7)$$

式中, B 是二值图. (i, j) 为像素点. 根据经验设置放大系数 $k = 50$.

其次,通过二值图获取文本的连通区域. 最后,利用偏移量 D' 将连通区域放大得到文本区域. 其中

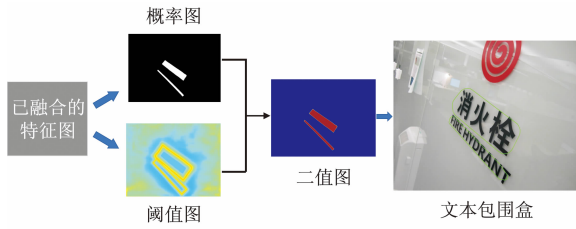


图8 可微分二值化模块

Fig. 8 Label generation

偏移量为:

$$D'^2 = \frac{A'^2 \times r'^2}{L^2} \quad (8)$$

式中, A' 是收缩多边形的面积, L' 是收缩多变形的周长, 根据经验设置收缩率 r' 为 1.5.

2.3.2 损失函数

改进的损失函数 L 为概率图损失 L_p 、对数化的二值图损失 L_b 和阈值图损失 L_t 的加权和:

$$L = \alpha \times L_p + \beta \times L_t + \lg(L_b) \quad (9)$$

根据损失数值, α 和 β 分别设置为 10 和 10. L_p 采用二元交叉 (BCE), 如公式 (10) 所示. 此外, 采用 hard negative mining 的方法解决正负样本之间不平衡的问题.

$$L_p = \sum_{i \in S_l} y_i \log x_i + (1 - y_i) \log (1 - x_i) \quad (10)$$

式中, S_l 是正负比为 1 : 3 的采样集.

L_b 采用 AC loss, 其基于曲线演化, 能够精确文本边缘的定位. AC loss 来源于 ACWE^[11] 模型. 其能量泛函方程为:

$$F_1(C) + F_2(C) = \int_{\text{inside}(C)} |c_1 - I|^2 dx dy + \int_{\text{outside}(C)} |c_2 - I|^2 dx dy \quad (11)$$

式中, C 表示待更新的曲线即包围文本的边界, $F_1(C)$ 为文本包围盒内的能量值, $F_2(C)$ 为文本包围盒外的能量值, c_1 为文本包围盒内像素的平均值, c_2 为文本包围盒外像素的平均值, I 为输入图像.

设计 AC loss 是将分割转变为求解能量泛函最小值, 当能量达到最小时, 活动轮廓收敛到所要检测文本的边缘, 改善正负样本之间的不平衡性, 提升分割文本精度. AC loss 定义如下:

$$\text{Loss}_{AC} = L_1 + \lambda \times \text{Area}$$

$$= \int_C |\nabla u| ds + \lambda \times \int_{\Omega} ((c_1 - v)^2 - (c_2 - v)^2) u dx dy \quad (12)$$

在公式 (12) 中, L_1 表示边界曲线 C 的长度, Area 表示包围文本曲线 C 内部面积, v 和 u 分别表示预测结果的掩码和 ground truth. 将公式 (12) 离散化为:

$$\begin{aligned} \text{Loss}_{AC} &= L_1 + \lambda \times \text{Area} \\ &= \sum_{i=1, j=1}^{\Omega} \sqrt{(\nabla u_{x^{i,j}})^2 + (\nabla u_{y^{i,j}})^2} + \epsilon \\ &\quad + \lambda \times \left(\left| \sum_{i=1, j=1}^{\Omega} u_{i,j} (c_1 - v_{i,j})^2 \right| + \left| \sum_{i=1, j=1}^{\Omega} (1 - u_{i,j}) (c_2 - v_{i,j})^2 \right| \right) \end{aligned} \quad (13)$$

式中, (i, j) 为像素点, ϵ 趋于 0. 因为掩模被用作标签, 因此 $c_1 = 1, c_2 = 0$. 显然, 原始可微二值化模块的输出没有因引入 AC loss 而改变.

L_t 是扩张多边形 G_d 内预测结果与标签之间的 L_1 距离之和:

$$L_t = \sum_{i \in R_d} |y_i^* - x_i^*| \quad (14)$$

式中, R_d 是扩张多边形 G_d 内像素的一组索引, y^* 是阈值图的标签.

3 实验与分析

3.1 数据集

SynthText 是由 800k 张图像组成的数据集. 这些图像由 8k 张背景图像合成.

MSRA-TD500 是多方向和弯曲文本数据集, 包含办公室、商场和街道等场景图片, 总共 900 张, 其中训练集 700 张, 测试集 200 张. 标注以行为单位, 每张图片都完全标注, 难以识别的有 difficult 标注.

MLT-2017 中文文本是一个任意方向文本的场景文本数据集, 包括训练集 800 张, 测试集 200 张. 它采用四边形框对出现的文本进行标注, 即给出每个顶点的坐标.

ShopSign 是河南大学张重生教授团队为了促进中文场景图像的识别研究, 纯手工构建的大规模中文标牌图像数据集. 其中 ShopSign1265 是部分公开的数据集, 将此数据集分为训练集 1 000 张, 测试集 115 张.

3.2 实验设计

首先使用 SynthText 数据集对所有数据集进行 100k 次迭代预训练,再使用预训练模型微调其他数据集 1 000 次,每批样本量设为 8. 本文采用动态调整学习率的 Poly 策略即指数变换的策略,设置迭代的学习率 $L_r = I_r \times (1 - \text{iter} / \text{maxiter})^p$ 其中设置初始学习速率 I_r 为 0. 001 以及 p 为 0. 009,权重衰减为 0. 000 4,动量为 0. 9,maxiter 表示最大迭代次数.

将图像进行几何变换增强数据,例如随机旋转、翻转和裁剪,其中旋转范围为 $(-10^\circ - 10^\circ)$. 将每张经过处理后的图像调整为 640×640 优化训练. 在训练过程中,概率图的训练标注与二值图的训练标注相同.

对 MLT-2017、ShopSign1265 和 MSRA-TD500 数据集进行消融实验来验证本文提出方法的有效性. 消融实验对比显著,其结果如图 9 中(a)~(f)所示,以下样本都是没有标注的测试验证集. 其中(a)是 FPN 加 DB 方法的实验结果. (b)和(c)是只增加 IASPP 分支的实验结果. (d)和(e)是增加 IASPP 分支和嵌入 BAM 注意力机制的实验结果. (f)是在增加 IASPP 分支和嵌入注意力机制的基础上改进 DB 模块的实验结果.

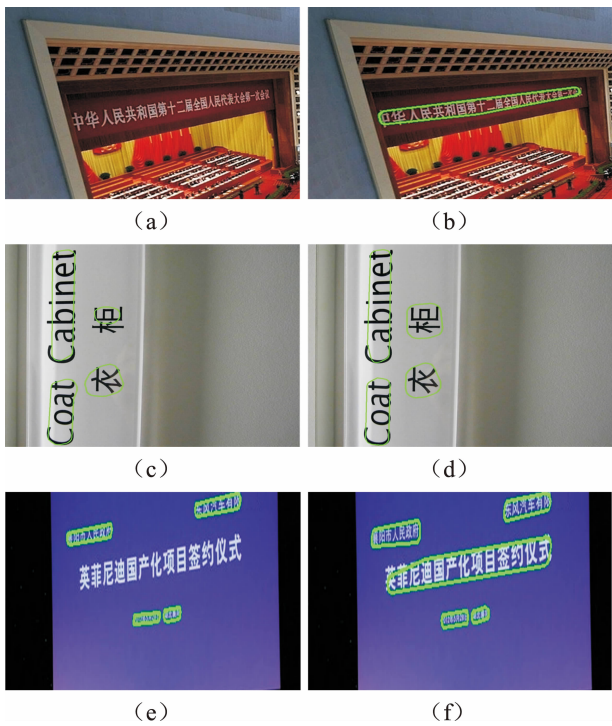


图 9 消融实验对比图

Fig. 9 Ablation experiment comparison
本文检测性能评测参数为准确率 Precision、召

回率 Recall 和综合评价指标 F-measure,其公式为:

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$F = \frac{2PR}{P + R} \tag{17}$$

若一个实例是正类,其被预测成为正类,即为真正类(True Postive, TP). 若一个实例是负类,其被预测成为负类,即为真负类(True Negative, TN). 若一个实例是负类,但是被预测成为正类,即为假正类(False Postive, FP). 若一个实例是正类,但是被预测成为负类,即为假负类(False Negative, FN). 其中 1 代表正类,0 代表负类.

3.3 实验环境

本文实验使用 Python3. 6 作为编程语言,pytorch1. 4. 0,所有的实验都在 Linux18. 04 操作系统进行,显卡配置为两张 NVIDIA RTX2080ti.

3.4 实验结果与分析

在 MLT-2017 中文文本数据集上的消融实验如表 1 所示,设 FPN+DB 为方法一,FPN+IASPP+DB 为方法二,FPN+IASPP+BAM+DB 为方法三,FPN+IASPP+BAM+EDB 为方法四. 方法二相比方法一,准确率上升了 2. 55%,召回率和 F 值分别增加 2. 88%和 1. 74%. 但是检测速率下降了 2. 34 fps. 方法三相比方法二,准确率,召回率和 F 值分别提升了 0. 43%,0. 74%和 0. 66%,检测速率降为 18. 63fps. 方法四相比方法三,准确率,召回率以及 F 值分别提升 1. 32%,0. 26%和0. 68%,检测速率为 18. 97fps,与前者相比变化不大.

表 1 MLT-2017 消融实验结果
Tab. 1 MLT-2017 ablation experiment

Backbone	Method	P	R	F	FPS
ResNet18	FPN+DB	83. 18	73. 89	78. 26	22. 02
ResNet18	FPN+IASPP+DB	85. 73	76. 77	81. 00	19. 68
ResNet18	FPN+IASPP+BAM+DB	86. 16	77. 51	81. 66	18. 63
ResNet18	FPN+IASPP+BAM+EDB	87. 48	77. 77	82. 34	18. 97

ShopSign1265 数据集上的消融实验如表 2 所示,方法二相比方法一,准确率,召回率和 F 值分别提升 2. 46%,0. 51%和 1. 08%,检测速率降至 21. 82 fps. 方法三相比方法二,准确率提升为 80. 02%,召回率和 F 值分别增加了 2. 08%和

2.53%,检测速率又下降了2.34 fps.方法四相比方法三,准确率上升了1.73%,召回率与F值分别提升0.51%和0.87%,检测速率为19.68 fps,与前者19.63 fps相比上升了0.05 fps.

表2 ShopSign1265 消融实验结果

Tab. 2 ShopSign1265 ablation experiment					
Backbone	Method	P	R	F	FPS
ResNet18	FPN+DB	74.35	44.06	55.33	23.20
ResNet18	FPN+IASPP+DB	76.81	44.57	56.41	21.82
ResNet18	FPN+IASPP+BAM+DB	80.02	46.65	58.94	19.63
ResNet18	FPN+IASPP+BAM+EDB	81.75	47.16	59.81	19.68

MSRA-TD500数据集上的消融实验结果如表3所示,方法二相比方法一,准确率增加了0.87%,召回率和F值分别上升了1.48%和1.22%,检测速率下降了4.84fps.方法三相比方法二,准确率,召回率和F值相比于前者都分别增加了1.25%,0.22%和0.67%,检测速率下降至19.75 fps.方法四相比方法三,模型的准确率,召回率以及F值分别上升了1.37%,0.53%和0.7%,检测速率上升至19.93 fps.

表3 MSRA-TD500 消融实验结果

Tab. 3 MSRA-TD500 ablation experiment					
Backbone	Method	P	R	F	FPS
ResNet18	FPN+DB	88.13	78.35	82.95	26.13
ResNet18	FPN+IASPP+DB	89.00	79.83	84.17	21.29
ResNet18	FPN+IASPP+BAM+DB	90.25	80.05	84.84	19.75
ResNet18	FPN+IASPP+BAM+EDB	91.62	80.58	85.75	19.93

将以上实验结果进行复杂度分析可得,在FPN和DB组成的原始算法框架中加入IASPP分支后,模型的检测准确率、召回率和F值都有明显的提升,这是因为IASPP弥补了FPN提取特征时感受野小的缺点,不仅增强了模型检测长文本的能力,而且降低了特征图分辨率损失.但是算法运行速度明显降低,例如MLT-2017检测速率从22.02fps降为19.68fps,MSRA-TD500检测速率从26.13fps降为21.29fps,这是因为加入IASPP后增加了整体模型的规模大小.在FPN中嵌入BAM注意力机制后准确率、召回率和F值都有提升,这是因为BAM增强了有效文本特征信息的提取.但算法的运行速度有所降低,例如ShopSign1265的检测速率从21.82fps降为19.63fps,这是因为BAM增加了检测网络的参数量.将DB模块换成优化损失函数后的EDB模

块后,模型检测效果都有明显提升,算法的运行速度也有一定提升,这是因为在二值图上采用对数化的AC Loss后,精确了文本边缘的定位,不仅改善了正负样本的不平衡性,而且增强了模型的泛化能力.

MSRA-TD500数据集的对比实验结果如表4所示.

表4 MSRA-TD500 对比实验

Tab. 4 MSRA-TD500 comparative experiment				
Method	P	R	F	FPS
EAST ^[12]	87.3	67.4	76.1	—
PixelLink ^[8]	83	73.2	77.8	3.00
Corner ^[13]	87.6	76.2	81.5	5.70
TextSnake ^[14]	83.2	73.9	78.3	1.10
CRAFT ^[15]	88.2	78.2	82.9	8.60
SBD ^[16]	89.6	80.5	84.8	3.20
DBnet ^[9]	91.5	79.2	84.9	32.00
文中方法	91.6	80.6	85.7	19.93

将实验结果分析可得,文中方法相比EAST,准确率,召回率和F值分别提升了3.8%,13.2%和9.4%.因为EAST的样本权重设置不合理且感受野较小.和PixelLink相比,准确率,召回率和F值分别提升了8.1%,7.4%和7.7%,本文算法检测速率提升了6倍多.这是由于PixelLink对于不同数据集需要不断调整pixel和link两个阈值,无法有效处理背景复杂的数据.文中方法相比Corner,准确率,召回率和F值分别提升了3.5%,4.4%和4%,检测速率约为其3.5倍.由于Corner任意两个角点都组合,所以导致很多无关的框被当做候选框.相比于TextSnake,准确率,召回率和F值分别提升了7.9%,6.1%和6.9%,本文算法检测速率提升18倍之多.TextSnake需要逐像素点预测,计算复杂度高.文中的方法与CRAFT网络相比,准确率,召回率和F值分别提高了3.9%,2.4%和2.6%,检测速率比其快2.3倍之多.CRAFT标注字符级数据集代价高,需要先进行弱监督训练得到字符级标注框再训练网络.文中方法与SBD相比,准确率和F值分别提高了1.5%和0.7%,但召回率只提升了0.1%,检测速率提升6倍之多.SBD利用与标签序列无关的不变表示形式来反推边界框坐标,简化参数且摆脱了训练目标的模糊性,检测任意形状的能力差.与DB网络相比,准确度提升了0.1%,召

回率和 F 值分别提高了 1.4% 和 0.6%, 但检测速率只有其 0.6 倍, 差距明显. DB 网络有轻量级的特征提取网络, 但其二值图和概率图都采用二元交叉熵损失函数, 只注重对于正确标签预测概率的准确性, 而忽略了其他非正确标签的差异, 从而导致学习到的特征比较分散.

4 结束语

本文研究具有一定的应用前景, 例如: 检测路牌文本, 辅助无人驾驶的导航; 检测商品包装文字, 追踪无人超市的商品; 检测门牌, 实现机器人送货上门等等. 但是, 目前自然场景中文文本的检测存在问题, 因此本文提出了一种融合注意力机制的多尺度自然场景中文文本检测方法 (MSAM). 首先改进 ASPP 网络以降低特征图的分辨率损失, 其次将其与 FPN 分支并行进行多尺度提取特征, 同时在 FPN 中嵌入平衡注意力机制对特征进行处理, 提取重要文本特征信息. 再者优化损失函数, 加权概率图与阈值图损失函数, 并将对数化的 AC Loss 引入二值图损失, 提升检测模型的泛化能力. 实验结果表明, 本文方法在多方向和弯曲中文文本的检测上有着优异的表现. 但是, 本文算法在检测速度与 DB-net 有着明显差距. 后续的研究工作将减小检测网络规模以提升模型检测速率, 并且改进特征提取融合方式, 进一步提高文本检测效果.

参考文献:

- [1] LIAO M H, SHI B G, BAI X, et al. Textboxes: a fast text detector with a single deep neural network [C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017: 4161-4167.
- [2] TIAN Z, HUANG W L, HE T, et al. Detecting text in natural image with connectionist text proposal network [C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 56-72. DOI: 10.1007/978-3-319-46484-8_4.
- [3] SHI B G, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 3482-3490. DOI: 10.1109/CVPR.2017.371.
- [4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiBox detector [C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 21-37. DOI: 10.1007/978-3-319-46448-0_2.
- [5] LYU P, LIAO M H, YAO C H, et al. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes [C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 71-88. DOI: 10.1007/978-3-030-01264-9_5.
- [6] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42 (2): 386-397. DOI: 10.1109/TPAMI.2018.2844175.
- [7] DENG D, LIU H F, LI X L, et al. PixelLink: detecting scene text via instance segmentation [C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018: 6773-6780.
- [8] WANG W H, XIE E Z, LI X, et al. Shape robust text detection with progressive scale expansion network [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 9328-9337. DOI: 10.1109/CVPR.2019.00956.
- [9] LIAO M H, WAN Z Y, YAO C, et al. Real-time scene text detection with differentiable Binarization [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020: 11474-11481. DOI: 10.1609/aaai.v34i07.6812.
- [10] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 3-19. DOI: 10.1007/978-3-030-01234-2_1.
- [11] CHAN T F, VESE L A. Active contours without edges [J]. IEEE Transactions on Image Processing, 2001, 10(2): 266-277. DOI: 10.1109/83.902291.
- [12] ZHOU X Y, YAO C, WEN H, et al. EAST: an efficient and accurate scene text detector [C]//Proceeding of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2642-2651. DOI: 10.1109/CVPR.2017.283.
- [13] LYU P, YAO C, WU W H, et al. Multi-oriented

- scene text detection via corner localization and region segmentation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7553-7563. DOI: 10.1109/CVPR.2018.00788.
- [14] LONG S B, RUAN J Q, ZHANG W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 19-35. DOI: 10.1007/978-3-030-01216-8_2.
- [15] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 9357-9366. DOI: 10.1109/CVPR.2019.00959.
- [16] LIU Y L, ZHANG S, JIN L W, et al. Omnidirectional scene text detection with sequential-free boxdiscretization[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2019: 3052-3058. DOI: 10.24963/ijcai.2019/423.

作者简介:

贾 颖 男,(1996-),硕士研究生.研究方向为图像处理与文本检测识别. E-mail:704664258@qq.com

程艳云 女,(1974-),硕士,副教授.研究方向为模式识别、机器学习、文字识别.