控制与决策
Control and Decision

基于双分支特征融合的场景文本检测方法

赵鹏， 徐本朋， 闫石， 刘政怡

引用本文:
赵鹏, 徐本朋, 闫石, 等. 基于双分支特征融合的场景文本检测方法[J]. 控制与决策, 2021, 36(9): 2179−2186.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0002

基于双分支特征融合的场景文本检测方法

赵 鹏†, 徐本朋, 闫 石, 刘政怡

(1. 安徽大学 计算智能与信号处理教育部重点实验室，合肥 230601；
2. 安徽大学 计算机科学与技术学院，合肥 230601)

摘 要: 现有的基于深度学习的自然场景文本检测方法一般采用大型深度神经网络作为主干网络进行特征提取,虽然效果显著但检测模型十分庞大,检测效率较低,若直接将主干网络换成轻量型网络则不能提取出足够的特征信息,直接导致检测效果大幅降低. 为了降低文本检测模型的规模以及更为高效地检测文本,提出基于双分支特征融合的场景文本检测方法,在采用相对轻量级的主干网络EfficientNet-b3的基础上,使用双路分支进行特征融合进而检测场景文本. 一路分支使用特征金字塔网络,融合不同层级的特征;另一路分支使用空洞卷积空间金字塔池化结构,扩大感受野,然后融合两个分支的特征,在小幅增加计算量的同时获取更多的特征,弥补小型网络提取特征不足的问题. 在3个公开数据集上的实验结果显示,所提出方法在保持较高检测水平的情况下,可以大幅度降低模型的参数量,大幅度提升检测速度.

关键词: 场景文本检测；深度学习；特征金字塔；特征融合；轻量级网络；注意力机制

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0002 **开放科学(资源服务)标识码(OSID):**

引用格式: 赵鹏,徐本朋,闫石,等. 基于双分支特征融合的场景文本检测方法[J]. 控制与决策, 2021, 36(9): 2179-2186.

A scene text detection based on dual-path feature fusion

ZHAO Peng, XU Ben-peng, YAN Shi, LIU Zheng-yi

(1. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education，Anhui University，Hefei 230601，China；2. School of Computer Science and Technology, Anhui University，Hefei 230601，China)

Abstract: The existing scene text detection methods based on deep learning generally use a deep neural network as the backbone network for feature extraction. Although it can achieve a striking detection effect,the entire detection model is very large which results in poor detection efficiency. If the large backbone network is replaced by a small backbone network directly, it will often fail to extract enough semantic features and can't achieve an ideal detection result. To reduce the size of the scene text detection model and promote the detection efficiency, a dual-path feature fusion based scene text detection (DPFFSTD) is proposed. Based on a relatively lightweight basic network EfficientNet-b3, the DPFF uses two branches for feature fusion to detect scene text. One branch uses a feature pyramid network to fuse the features with different levels. The other branch uses an atrous spatial pyramid pooling to enlarge receptive field and obtains the features of different scales. And then the features from the above two branches are fused to form more features only with a very small increasing computation, which makes up for the shortage of features caused by the small backbone network. The experimental results on three benchmark datasets show that the proposed method significantly reduces the number of the model parameters and greatly improves the detection efficiency while maintaining a high detection effect.

Keywords: scene text detection；deep learning；feature pyramid network；feature fusion；lightweight neural；attention mechanism

0 引 言

自然场景下图像中包含的文本信息通常能够对图像的理解提供很大帮助,如无人驾驶中对路况信息的判断、增强现实中对各种物体的识别等,因此场景文本检测吸引了众多研究者的关注. 但是,自然场景下文本的出现有着很大的随机性和多样性,常规的有

收稿日期: 2020-01-01；**修回日期:** 2020-03-23.

基金项目: 国家自然科学基金项目(61602004)；安徽省高校自然科学研究重点项目(KJ2018A0013, KJ2017A011)；安徽省自然科学基金项目(1908085MF188, 1908085MF182)；安徽省重点研究与开发计划项目(1804d08020309).

责任编委: 王凌.

†通讯作者. E-mail: zhaopeng_ad@163.com.

横向和竖向文本,复杂的有斜向文本,更为复杂的有曲形文本甚至不规则文本. 文字的大小、位置、形状等特征无法人为确定,加上背景或噪声的干扰等,自然场景下的文本检测依然是一项非常具有挑战性的任务. 目前行之有效的场景文本检测模型大都基于深度学习技术,这些模型为了追求检测的高精度往往参数量巨大,模型也很大. 而场景文本检测作为一项面向实际应用场景的技术,在诸如便携式设备、分布式设备上的应用很强调模型的小型化和高效性. 同时兼顾文本检测的高精度和模型的小型化目前是场景文本检测技术亟待解决的问题.

近年来,自然场景下文本检测的研究取得了长足的进步. 随着深度学习技术的发展,很多基于深度学习的场景文本检测方法被陆续提出,效果显著,已经成为场景文本检测的主流方法. 基于深度学习的场景文本检测算法主要分为两类:

1) 基于目标检测的方法. Ren等[1]提出了Faster-RCNN目标检测算法,Liu等[2]提出了SSD目标检测算法. 基于这些目标检测算法,很多研究者对其进行改进,用于场景文本的检测. Tian等[3]提出了CTPN算法,结合卷积神经网络和长短期记忆网络,在Faster-RCNN的基础上预测多个小矩形框,然后合并得到对文本框的预测. Ma等[4]提出了RRPN算法,在Faster-RCNN的基础上对检测框进行旋转检测倾斜的文本. Shi等[5]提出了SegLink算法,在SSD的基础上,将文本的检测划分为多个局部的检测,然后将多个局部片段进行连接. He等[6]提出了SSTD算法,在SSD中加入注意力机制,强化了文本区域的特征,从而对文本进行更好的识别. Hu等[7]提出了WordSup算法,引入弱监督调整字符的坐标,对字符进行合并得到所检测的文本行. Liu等[8]提出了CTD+TLOC算法,在Faster-RCNN和CTPN的基础上回归14个坐标,对曲形文本进行检测. 虽然这些基于目标检测的文本检测算法取得了不错的检测效果,但是由于四边形框的限制,往往对多方向文本和弯曲形文本的检测不够精准.

2) 基于分割的方法. 该方法能够不限于文本的方向和形状,对倾斜文本和弯曲形文本也能进行较好的检测. 场景文本检测任务中,还有很多文本位置十分靠近甚至连接在一起的情况,基于分割的方法往往不能很好地将其分离. Long等[9]在分割出文本区域的同时预测出文本中轴线的位置,然后使用大小不一的圆盘表示文本区域,从而进行文本的预测. Xu等[10]抽取VGG16不同阶段的特征图融合后进行分割,得到大概的文本区域,然后对分割结果中的像素点进行形态学的后处理,得到最终精确的文本分割图,并且该算法能够对较接近的文本进行很好地区分. PSENet[11]首先采用特征金字塔融合不同尺度的特征进行分割,同一个文本框会得到多个不同尺度的文本分割图,然后对不同尺度的分割图使用渐近扩展算法进行融合,最终得到完整的检测文本. 渐近尺度扩展算法不仅能够对场景文本进行较为准确的检测,而且能显著地区分开靠近或者粘连在一起的文本.

基于深度学习的场景文本检测方法往往都使用深度神经网络作为基础网络提取特征,常用的深度神经网络模型有VGG[12]、ResNet[13]等. 这些经典的深度神经网络有着较好的泛化能力,可以迁移到如场景文本检测之类的任务中. 然而,网络参数量巨大,基于网络的模型往往也很大,直接换用过于小型的神经网络往往泛化能力不够,使得检测效果不佳. 最近,谷歌的研究人员提出了新型的高效神经网络EfficientNet,在小规模的参数量下能够达到与传统神经网络相同甚至更好的效果[14].

为了降低自然场景文本检测模型的规模,同时又不降低检测的效果,本文提出一种基于双分支特征融合的自然场景文本检测方法(dual-path feature fusion based scene text detection, DPFF). 主要贡献如下:

1)针对传统的基于深度学习的场景文本检测方法参数过多、模型过大的问题,提出一种新的高效的自然场景文本检测方法,采用轻量型的基础网络,同时利用语义分割方法和渐近尺度扩展算法进行后处理;

2)通常自然场景文本检测模型直接换用小网络会导致性能变差,针对该问题提出双分支特征融合方法,一路分支为特征金字塔网络融合高层和低层的特征,另一路分支为使用并行空洞卷积扩大感受野,最后融合两路的特征,获取更丰富的特征以弥补小网络性能不足的问题.

1 相关工作

1.1 深度学习

自2012年AlexNet在ImageNet图像识别大赛斩获冠军以来,深度学习的研究和应用取得了巨大的进展. 深度学习也带动了目标检测、图像分割等任务的快速发展,研究者们在深度学习的基础上提出各种改进方法,提升各自任务的效果. 例如特征金字塔网络在深度神经网络的基础上融合不同层次的特征,有效提升了目标检测效果. 在图像分割中,Yu等[15]使用空洞卷积扩大感受野,证明了空洞卷积能够显著

提升图像分割效果. Deeplab系列对空洞卷积进行了改进[16-17],其中Deeplab-v2提出了空洞卷积空间金字塔池化,使用多个并行的空洞卷积获取不同尺度层次的特征. Deeplab-v3进一步改进了空洞卷积空间金字塔池化,加入了批正则化提升分割效果[17]. 深度神经网络在诸多任务中的应用,当网络增大到一定规模后,单纯加深网络并不能带来明显的效果提升,一些研究人员通过添加注意力机制来提升神经网络的性能. SENet对于不同通道的特征学习不同的权重,更准确和高效地利用特征. DANet同时使用空间注意力和通道注意力两种类型的注意力机制,有效提升了语义分割的效果[18].

1.2　EfficientNet

当前,很多计算机视觉任务中都会采用VGG[12]、ResNet[13]等深度神经网络模型作为主干网络进行特征提取. 这些模型迁移到其他任务中效果较好,但是规模较大[19]. 2019年,谷歌研究人员提出了一种新的神经网络模型EfficientNet,利用神经网络结构搜索技术和模型缩放方法,相比于其他神经网络模型,EfficientNet参数更少,推理速度更快,准确率更高. EfficientNet有8个不同版本,本文所使用的EfficientNet-b3的网络结构简洁,如表1所示. 整个网络由多个阶段组成,每个阶段由若干个重复卷积块组成. 表1中输入大小表示在该部分输入的特征图的尺寸大小,层数表示该操作模块在对应阶段重复的次数.

表1　EfficientNet-b3的网络结构

阶段	操作模块	输入大小	通道数	层数
1	Conv3×3	224×224	40	1
2	MBConv1, k3×3	112×112	24	2
3	MBConv6, k3×3	112×112	32	3
4	MBConv6, k5×5	56×56	48	3
6	MBConv6, k5×5	14×14	136	5
7	MBConv6, k5×5	14×14	232	6
8	MBConv6, k3×3	7×7	384	2
9	Conv1×1&Pooling&FC	7×7	1 536	1

1.3　PSENet

PSENet[11]是一种基于分割的场景文本检测方法,提出了一种渐近尺度扩展算法(progressive scale expansion algorithm, PSE). 实验表明,渐近扩展算法能够有效区分开粘连在一起的文本,因此本文模型在后处理中采用渐近扩展算法. PSENet采用特征金字塔网络作为网络结构,对于一个输入图片,首先提取不同层次的特征,然后逐层上采样至相同的尺寸,最后级联到一起得到融合特征F. 融合特征F经过处理得到n个不同的分割图$S_1, S_2, \ldots, S_n$,这些分割图是同一个图片内同一文本不同比例的分割图,小比例的分割图能够较容易地区分开文本的边缘,大比例的分割图会更准确. 最后使用渐近尺度扩展算法对n个分割图从小到大依次进行扩展,得到最终的预测结果.

2　本文方法

2.1　总体网络架构

本文提出的DPFF方法总体框架如图1所示,主要分为3个模块:特征提取模块、双分支特征融合模块和渐近扩展模块. 第1个模块采用EfficientNet-b3作为基础网络,首先在ImageNet数据集上进行预训练,然后删去全连接层,对输入图像提取特征. 第2个模块为双分支特征融合模块, 该模块分为两路分支:

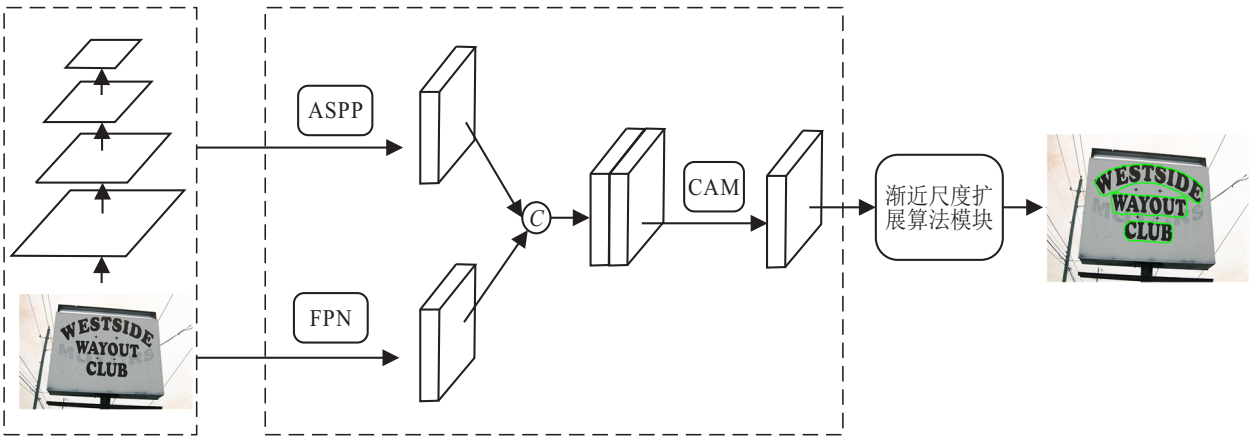

图1　DPFF的总体框架

1) 空洞卷积空间金字塔池化(atrous spatial pyramid pooling, ASPP)分支,该路分支为了扩大感受野,使用多分支并行的不同空洞率的空洞卷积获取不同尺度的特征; 2) 特征金字塔(feature pyramid networks, FPN)分支,该路分支融合基础网络中提取的不同层次的特征图,得到不同尺度特征融合特征,两路分支产生的不同特征采用通道注意力机制(channel attention module, CAM)进行融合,然后产生n个不同比例尺度的分割结果. 第3个模块采用渐近扩展算法作为后处理算法的渐近扩展模块,对前一部分产生的多个不同尺度的分割结果进行逐步扩展融合,得到最终检测结果.

2.2 特征融合

本文提出的特征融合方式如图1特征融合模块所示,该特征融合模块包括ASSP和FPN两个并行的分支.

ASPP分支如图2所示,其中C表示级联. 该分支使用4个并行的空洞卷积层和1个池化层同时对主干网络中的1/16特征图进行并行处理,然后级联5路24通道的特征图,得到120通道的空洞卷积特征. 4路空洞卷积的空洞率分别为1、6、12、18,池化层为最大化全局池化.

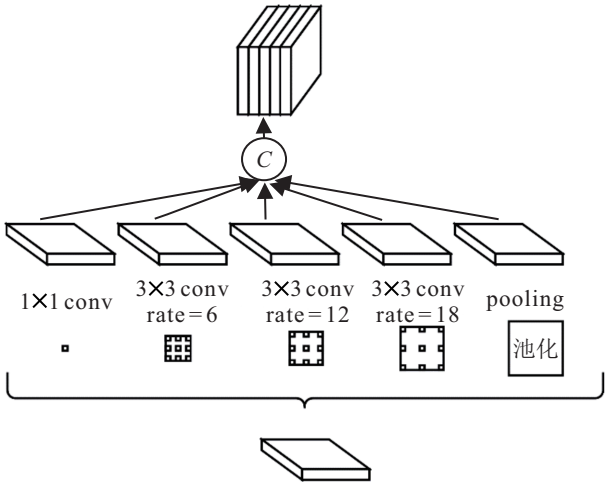

图 2 空洞卷积空间金字塔模块

FPN分支如图3所示,其中C表示级联. FPN分支对基础网络中提取的1/4、1/8、1/16、1/32特征图分别记为F_2、F_3、F_4、F_5,自上而下逐层上采样至相同的尺寸,并与下一层级联,分别得到P_5、P_4、P_3、P_2,其中$P_5 = F_5, P_4 = F_4||Up_{\times 2}(P_5), P_3 = F_2||Up_{\times 2}(P_4)$, $P_2 = F_2||Up_{\times 2}(P_3)$. 最后级联到一起得到融合特征$P$,即

$$P = C(P_2, P_3, P_4, P_5) = P_2||Up_{\times 2}(P_3)||Up_{\times 4}(P_4)||Up_{\times 8}(P5).$$

其中: "||"表示级联, $Up_{\times 2}(\cdot)$、$Up_{\times 4}(\cdot)$、$Up_{\times 8}(\cdot)$分别表示上采样2倍、4倍和8倍. 每层特征图有24个通道,最后级联得到96通道的特征金字塔特征图,融合高层和低层语义.

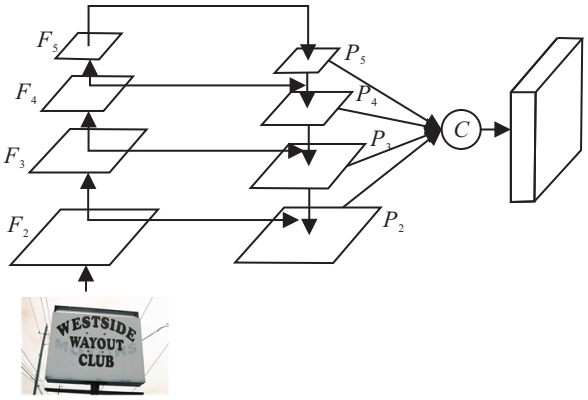

图 3 特征金字塔模块

FPN分支在使用小型基础网络的情况下,通道数较少,融合的特征不够丰富. ASPP分支使用不同空洞率的卷积扩大感受野,提供不同层级的特征. 两个分支互相补充,提升整个网络的性能. 由于特征图分布的不确定性,本文利用通道注意力机制对级联后的特征进行处理,在训练的过程中不断调整各通道的权值,增加重要特征通道的权值. 本文通道注意力机制实现方式如下: 1) 将输入的特征图$F \in R^{C\times H\times M}$转换为$F' \in R^{C\times N}$,即将每个通道中的特征图直接拉伸成一维的特征表示; 2) 将转换后的特征图F'与其转置$(F')^{\mathrm{T}}$进行相乘,即$F' \cdot (F')^{\mathrm{T}}$,然后使用softmax函数对其进行激活,得到通道注意力图$X \in R^{C\times C}$, X中每个元素x_{ji}计算为

$$x_{ji} = \frac{\exp(F'_i \cdot F'_j)}{\sum_{i=1}^{C} \exp(F'_i \cdot F'_j)}. \tag{1}$$

其中: x_{ji}为第i个通道对第j个通道的影响, F'_i为F'的第i列向量. 将得到的通道注意力图X的转置X^{T}与F'相乘,得到的结果转换为原始特征图的尺寸,即$E \in R^{C\times H\times W}$. 特征$E$中的每个通道的特征图$E_j$为

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} F'_i) + F'_j, \tag{2}$$

其中超参数β在训练过程中进行学习. 经过以上处理,得到的特征在通道数和尺寸上没有变化,但是每个通道的特征图是其他通道特征图的加权和,并且权值可以在训练的过程中学习到.

2.3 损失函数

对于同一个文本框,渐近尺度扩展算法需要对多个不同尺度的分割图进行融合,进而得到准确的预

测结果. 在训练过程中,这些不同尺度的分割图需要对应的标签进行指导训练,但是通常公开数据集只给出唯一完整的文本框标签. 为了得到不同尺度的标签,需要对数据集给定的标签进行多次缩放处理,每次对原始标签中的多边形p_n缩减d_i个像素得到缩放后的标签p_i,最终得到一系列不同尺度的文本标签$p_1, p_2, \ldots, p_{n-1}$. d_i计算为

$$d_i = \frac{\text{area}(p_n) \times (1 - r_i^2)}{\text{perimeter}(p_n)}. \tag{3}$$

其中: $\text{area}(p_n)$为多边形p_n的面积; $\text{perimeter}(p_n)$为多边形p_n的周长; r_i为缩放比,计算为

$$r_i = 1 - \frac{(1-m) \times (n-i)}{n-1}, \tag{4}$$

$m \in (0, 1]$为缩放的最小尺度,本文m设置为0.5, n为不同尺度的分割图的个数.

本文的损失函数在训练过程中由两部分组成: 完整的文本损失L_c和缩放后的文本损失L_s. 损失函数L计算如下:

$$L = \lambda L_c + (1 - \lambda) L_s, \tag{5}$$

其中超参数$\lambda \in (0, 1)$平衡L_c和L_s的权重. L_c和L_s均使用Dice系数进行损失函数的计算,Dice系数计算如下:

$$D(S_i, G_i) = \frac{2 \sum\limits_{x,y} (S_{i,x,y} \times G_{i,x,y})}{\sum\limits_{x,y} S_{i,x,y}^2 + \sum\limits_{x,y} G_{i,x,y}^2}. \tag{6}$$

其中: $S_{i,x,y}$为分割结果S_i上像素(x, y)的值, $G_{i,x,y}$为标签上像素(x, y)的值. 在训练过程中,使用在线难例挖掘算法(online hard example mining, OHEM)[20],则L_c计算为

$$L_c = 1 - D(S_n \cdot M, G_n \cdot M), \tag{7}$$

其中M为训练过程中OHEM预测的文本区域. L_s计算为

$$L_s = 1 - \frac{\sum\limits_{i=1}^{n-1} D(S_i \cdot W, G \cdot W)}{n-1};$$

$$W_{x,y} = \begin{cases} 1, & S_{n,x,y} \geqslant 0.5; \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

其中W为S_n中忽略非文本区域像素的掩模.

3　实验分析

3.1　数据集和实验设置

选择3个通用数据集ICDAR2015[21]、Total-Text[22]和SCUT-CTW1500[8]作为实验数据集. ICDAR2015是一个任意方向文本的场景文本检测数据集,包括训练集1 000张图片,测试集500张图片,共计1 500张图片,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. Total-Text是一个包含任意方向文本和弯曲文本的自然场景文本数据集,包含1 225张训练集图片,300张测试集图片. SCUT-CTW1500是一个包含水平方向文本、任意方向文本和弯曲文本的自然场景文本检测数据集,包含1 000张训练集图片,500张测试集图片,其标注采用14个顶点组成多边形对文本进行包围标注.

实验使用Python3.6作为编程语言,使用的深度学习框架为pytorch1.1,所有实验均在NVIDIA RTX2080显卡上进行. 实验中所使用的EfficientNet-b3在ImageNet数据集上进行预训练,网络在训练过程中使用Adma优化算法进行优化.

将渐近扩展算法中所需要的多尺度特征图的个数设置为原始PSENet中的最优值7, OHEM算法的正负比设为3,训练数据的批大小设置为4,初始学习率为1e-4,每隔50 000步学习率降低一个数量级,每个数据集共训练250 000步.

采用场景文本检测任务中常用的评价指标: 准确率(P), 召回率(R), F值(F)和帧率(FPS). 准确率、召回率和F值计算为

$$P = \frac{\text{match}(D, G)}{|D|}, \tag{9}$$

$$R = \frac{\text{match}(D, G)}{|G|}, \tag{10}$$

$$F = 2 \times \frac{P \times R}{P + R}. \tag{11}$$

其中: D为模型检测出的文本框, G为真实的文本框, $\text{match}(D, G)$为检测的正确的文本框数量, $|D|$为检测的文本框数量, $|G|$为真实的文本框数量. 帧率用来衡量模型的检测速度,通常用每秒处理图片的数量(frames per second, FPS)作为单位衡量模型的计算速度.

3.2　实验结果和分析

为了验证所提出DPFF的有效性,分别与多种主流方法(CTPN[3]、Seglink[5]、SSTD[6]、WordSup[7]、RRPN[4]、PSENet-1s[11]、CTD+TLOC[8]、TextSnake[9]、TextField[10])进行对比. 在公共数据集ICDAR201、CTW-1500和TotalText上进行实验,结果如表1～表4所示. 表中对比方法的数据均来自其对应的论文. 由表2可见, DPFF在ICDAR2015数据集上的准确率、召回率和F值分别为81.67 %、80.26 %和80.96 %. DPFF相比于CTPN,准确率、召回率和F值分别提升了7.67 %、28.26 %和19.96 %. DPFF相比于Seglink,

准确率、召回率和F值分别提升了8.57%、3.46%和5.96%. DPFF相比于SSTD,准确率、召回率和F值分别提升了8.67%、0.26%和3.96%. DPFF相比于WordSup,准确率、召回率和F值分别提升了4.64%、0.93%和2.8%. DPFF相比于RRPN,准确率略有下降,召回率提升了7.03%,F值提升了3.52%. ICDAR2015数据集以横向、纵向和倾斜文本为主,TextSnake和TextField方法针对自然场景文本本身的特点各自有自己的代表性检测方法,DPFF与之相比也达到了基本同等级别的检测效果. DPFF与PSENet-1s相比,准确率、召回率和F值都差别不大,表明DPFF检测效果达到了与之相同的水平. 与以上方法在ICDAR2015数据集上的对比表明,DPFF对于常规文本和倾斜文本的检测处在较高的水平.

表2 ICDAR2015数据集上的对比实验结果 %

方法	准确率	召回率	F值
CTPN	74.00	52.00	61.00
SegLink	73.10	76.80	75.00
SSTD	73.00	80.00	77.00
WordSup	77.03	79.33	78.16
RRPN	82.17	73.23	77.44
TextSnake	84.90	80.40	82.59
TextField	80.50	84.30	82.40
PSENet-1s	81.49	79.68	80.57
DPFF	81.67	80.26	80.96

在CTW1500数据集上,DPFF相比CTPN,准确率、召回率和F值分别提升24.28%、16.4%和19.86%. 相比Seglink,准确率、召回率、F值分别提升42.38%、30.2%和35.96%. CTW1500数据集上以弯曲型文本居多,CTPN和Seglink方法使用矩形框往往不能准确地进行检测和标注,DPFF能够不受矩形框的限制,对于任意形状的文本都能进行很好的检测. 与CTW1500数据集的基准方法CTD+TLOC相比,DPFF的准确率提升了7.28%,召回率提升了0.4%,F值提升了3.36%. 与TextSnake相比,DPFF的准确率较高但是召回率偏低,F值提升了1.16%. 与TextField相比,DPFF准确率较高,召回率偏低,F值稍有差距. 在CTW1500数据集上,DDPFF相比于PSENEt-1s,准确率有较大提升,但是召回率也有较大下降,体现综合检测能力的F值略有下降. 这是由于在CTW1500数据集中很多文本过于接近,甚至粘连重叠在一起,分开难度较大,即使渐近尺度扩展算法也不能很好地进行区分,当主干网络采用小网络时,该问题被进一步放大. DPFF在CTW1500数据集上的F值也达到了76.7%,表明DPFF能够很好地检测出弯曲型的文本.

表3 CTW-1500数据集上的对比实验结果 %

方法	准确率	召回率	F值
CTPN	60.40	53.80	56.90
SegLink	42.30	40.00	40.80
CTD+TLOC	77.40	69.80	73.40
TextSnake	67.90	85.30	75.61
TextField	79.80	83.00	81.37
PSENet-1s	80.60	75.60	78.10
DPFF	84.68	70.20	76.76

在Total-Text数据集上的实验结果如表4所示. DPFF相比基线方法DeconvNet,准确率、召回率和F值分别提升了48.53%、34.67%和41.95%. TextField、TextSnake和PSENet-1s方法略微优于本文方法,这是由于TextField、TextSnake和PSENet-1s在不考虑模型参数量规模的情况下采用VGG16或者ResNet50作为主干网络进行大量特征的提取带来的优势. DPFF在考虑降低模型规模的前提下选取小型网络进行特征提取,在特征提取不够丰富的不利前提下,能够取得与TextField、TextSnake和PSENet-1s相当的水平,表明本文方法在弯曲形文本的检测方面有着足够优秀的效果.

表4 Total-Text数据集上的对比实验结果 %

方法	准确率	召回率	F值
DeconvNet	33.00	40.00	36.00
TextField	79.00	81.20	80.54
TextSnake	82.70	74.50	78.39
PSENet-1s	81.77	75.11	78.30
DPFF	81.53	74.67	77.95

3.3 网络性能和复杂度分析

为了验证所提出方法的性能,在ICDAR2015数据集上进行对比实验,实验结果如表5所示. DPFF的参数量只有12 M,检测速度达到7.45 FPS. PSENet提出了渐近扩展算法并给出了基准方法PSENet-1s,基准方法使用ResNet50和ResNet152作为主干网络. DPFF方法与PSENe-1s(ResNet50)相比,参数量降低了50%,检测速度提升了3倍,F值没有降低. 与

PSENet-1s(ResNet152)相比,参数量降低80 %,检测速度提升4倍,F值只降低了1.38 %.

表5　模型参数和速度在数据集ICDAR2015上的对比结果

方法	主干网络	F值/%	参数量/M	速度/FPS
PSENet-1s	ResNet50	80.57	24	2.24
PSENet-1s	ResNet152	82.34	58	1.82
DPFF	EfficientNet-b3	80.96	12	7.45

为验证所提出的特征融合方法的有效性,在数据集ICDAR2015上进行对比实验. 对比实验分别为仅将PSENet-1s主干网络换为EfficientNet-b3(如表6中的EfficientNet-b3+FPN)和在EfficientNet-b3的基础上增加DPFF,实验结果如表6所示. 由表6可见,DPFF能够有效提升网络的性能.

表6　不同模块在数据集ICDAR2015上的对比结果 %

方法	准确率	召回率	F值
EfficientNet-b3+FPN	79.50	77.80	78.60
DPFF	81.67	80.26	80.96

由以上实验和分析可知,本文提出的场景文本检测算法能够有效地检测出图像中出现的文本,对于过于接近的文本不仅能准确地检测出来,还能有效地区分开来. 对于各种不同形状的文本,也能准确地检测出来,显示了本文方法有较好的准确性和鲁棒性.

DPFF方法参数量只有12 M,在训练和测试的过程中占用的内存空间较小,在空间复杂度方面占有较大的优势. 与此同时,DPFF网络的每秒浮点运算次数(floating-point operations per second, FLOPS)约为28×10^9 FLOPS,而PSENet-1s(ResNet50)约为$1\,170 \times 10^9$ FLOPS, PSENet-1s (ResNet152)约为$1\,778 \times 10^9$ FLOPS. 相比于PSENet的基准方法,DPFF将时间复杂度降低了两个数量级,极大地加快了推理速度. 无论是在空间复杂度还是在时间复杂度方面,所提出的方法均占有较大优势,能够极大地节省计算成本.

综合以上实验表明,所提出的DPFF能够快速且准确地检测出文本,其按照处理流程可分为特征提取和后处理两大部分,使用渐近扩展算法的后处理部分能够保障文本检测的准确性. 但是实验表明,在特征提取部分单纯地换用轻量型网络,由于特征提取不足会导致文本的检测效果下降. DPFF能够弥补检测效果下降的轻量型网络的不足,因此DPFF的特征提取部分和后处理部分在分别承担着降低模型规模和检测文本任务的同时,两者又能互相弥补,最终达到准确而又快速地检测文本的目的.

4　结　论

本文提出了一种新的基于特征融合的场景文本检测方法DPFF,该方法以EfficientNet-b3为基础网络,融合特征金字塔和空洞卷积空间金字塔池化两个分支的特征,通过特征融合弥补小网络提取特征不足的缺陷. 所提出的融合策略即使在轻量级网络的情况下,文本检测模型依然能够达到采用大型神经网络作为主干网络模型的同等检测水平. 最后使用渐近扩展算法进行后处理,使得整个模型快速而准确地检测出文本. 所提出方法在不降低检测效果的前提下,极大降低了网络的参数量,提升了检测速度,对于以实际应用场景为导向自然场景文本检测技术研究而言有着重要的意义.

下一步的工作将考虑在现有的工作基础上,保持网络较少参数量和较快检测速度的同时,尝试增加其他特征融合方式进一步提升场景文本检测的效果. 以及采用更小的基础网络,在保证检测效果不下降的同时,进一步减少网络的参数,降低网络的规模,使文本检测方法更为轻量化.

参考文献(References)

[1] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6): 1137-1149.

[2] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. Proceedings of European Conference on Computer. Cham: Springer, 2016: 21-37.

[3] Tian Z, Huang W L, He T, et al. Detecting text in natural image with connectionist text proposal network[C]. Proceedings of Europan Conference on Computer vision. Cham: Springer, 2016: 56-72.

[4] Ma J, Shao W Y, Ye H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[5] Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017: 2550-2558.

[6] He P, Huang W L, He T, et al. Single shot text detector with regional attention[C]. Proceedings of the 2017 IEEE International Conference on Computer Vision. Hawaii: IEEE, 2017: 3047-3055.

[7] Hu H, Zhang C Q, Luo Y X, et al. Wordsup: Exploiting

word annotations for character based text detection[C]. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 4940-4949.

[8] Liu Y L, Jin L W, Zhang S T, et al. Detecting curve text in the wild: New dataset and new solution[J]. 2017, arXiv: 1712.0217.

[9] 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 2018 European Conference on Computer Vision. Munich: Springer, 2018: 19-35.

[10] Xu Y C, Wang Y K, Zhou W, et al. Textfield: Learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5566-5579.

[11] Wan W, Xie E, Li X, et al. Shape robust text detection with progressive scale expansion network[C]. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 9328-9337.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. Proceedings of International Conference on Learning Representations. San Diego: IEEE, 2015: 1-14.

[13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[14] Tan M X, Quoc V L. EfficientNet: Rethinking model scaling for convolutional neural networks[C]. Proceedings of the 36th International Conference on Machine Learning. Piscataway: IEEE, 2019: 1-10.

[15] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]. Proceedings of 4th International Conference on Learning Representations. San Juan: IEEE, 2016: 1-10.

[16] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 40(4): 834-848.

[17] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. 2017, arXiv: 1706.05587.

[18] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]. Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3146-3154.

[19] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.

[20] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 761-769.

[21] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading[C]. Proceedings of 13th International Conference on Document Analysis and Recognition. Nancy: IEEE, 2015: 1156-1160.

[22] Ch'Ng C K, Chan C S. Total-Text: A comprehensive dataset for scene text detection and recognition[C]. Proceedings of 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017: 935-942.

作者简介

赵鹏 (1976−), 女, 副教授, 博士, 从事机器学习、模式识别等研究, E-mail: zhaopeng_ad@163.com;

徐本朋 (1993−), 男, 硕士生, 从事深度学习、场景文本检测等研究, E-mail: xubenpeng2019@163.com;

闫石 (1997−), 男, 硕士生, 从事场景文本检测与识别的研究, E-mail: bzyanshi1011@126.com;

刘政怡 (1978−), 女, 副教授, 博士, 从事计算机视觉、机器学习等研究, E-mail: 22927463@qq.com.

(责任编辑：郑晓蕾)