

doi: 10.12052/gdutxb.230011

嵌入拓扑特征的自然场景文本检测方法

郑侠聪, 程良伦, 黄国恒, 王敬超

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 传统的基于锚点框(anchor box)实现的自然场景文本检测方法中, 锚点框容易受到其他文本实例的干扰产生误判或精度降低, 且文本实例包含强烈的拓扑特征但并未得到重视, 导致在弯曲环形文本检测任务中表现不佳。针对这个问题提出了一种新颖的神经网络结构, 引入图卷积神经网络的概念, 充分考虑邻近锚点框之间的联系, 并融入锚点框的拓扑特征辅助图神经网络的学习, 提高整体网络的有效性。在两个公开的自然场景文本检测数据集上进行了消融实验, 在公开数据集CTW1500中, 本文提出的方法使模型在召回率、精确率、F分数这3个指标上分别提高了3.0%、1.9%以及2.5%, 在公开数据集Total-Text中这3个指标分别是2.2%、1.8%以及2.0%。此外, 本文方法还与近年提出的其他文本检测算法进行了比较, 实验结果证明本文提出的方法在复杂自然场景下文本检测效果优秀, 所提出的模块有利于文本检测性能的提高。

关键词: 文本检测; 自然场景; 图神经网络; 拓扑特征

中图分类号: TP391

文献标志码: A

文章编号: 1007-7162(2024)03-0102-08

Text Detection in Natural Scenes Embedded Topological Feature

Zheng Xia-cong, Cheng Liang-lun, Huang Guo-heng, Wang Jing-chao

(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In traditional anchor box-based text detection methods for natural scenes, anchor boxes are prone to interference from other text instances, resulting in erroneous judgments or affecting accuracy. Moreover, text instances contain strong topological features, which are usually be ignored, resulting in poor performance in curved circular text detection tasks. To solve this problem, a novel neural network structure is proposed, which introduces the concept of graph convolutional networks by fully considering the relationship between adjacent anchor frames, and incorporating the topological characteristics of anchor frames to assist the learning of graph neural networks, improving the effectiveness of the overall network. The ablation experiments were conducted on two publicly available natural scene text detection datasets. In the CTW1500 dataset, the proposed method improved the model by approximately 3.0%, 1.9%, and 2.5% in terms of recall, accuracy, and F-score, respectively, and in the Total-Text dataset, the three values were improved by approximately 2.2%, 1.8%, and 2.0%, respectively. In addition, the proposed method has also been compared with other text detection algorithms proposed in recent years. Experimental results show that the proposed method performs well for text detection in complex natural scenes, demonstrating the promising effectiveness of the proposed module for improving the performance of text detection.

Key words: text detection; natural scene; graph convolutional networks(GCN); topological feature

场景文本检测已广泛应用于各种应用领域, 如在线教育、产品搜索、即时翻译和视频场景解析。随着深度学习的蓬勃发展, 文本检测算法在规则的文

本检测任务中的表现得到很大的提升。然而, 由于自然场景中文本图像的多样性, 以往的文本检测方法在检测任意形状和多方向文本的任务中的表现仍有

收稿日期: 2023-01-28

基金项目: 国家自然科学基金资助项目(U20A6003); 国家自然科学基金广东联合基金资助项目(U1801263, U1701262, U2001201); 广东省信息物理融合系统重点实验室项目(2020B1212060069); 佛山市重点领域科技攻关项目(2020001006832)

作者简介: 郑侠聪(1996-), 男, 硕士研究生, 主要研究方向为计算机视觉、人工智能, E-mail: 413169248@qq.com

通信作者: 黄国恒(1985-), 男, 副教授, 博士, 主要研究方向为计算机视觉、机器学习和模式识别等, E-mail: kevinwong@gdut.edu.cn

提升空间。目前在场景文本检测任务中的方法大致可以分为基于分割和基于回归两大类。而在复杂文本检测任务和弯曲文本检测任务中使用锚点框策略的基于回归的方法往往能获得更好的表现。

通过卷积神经网络回归计算文本锚点框的几何特性,再将文本锚点框的几何属性生成节点特征,再构造局部图,并将局部图送到图神经网络来挖掘文本锚点框之间的深层连通关系,这种方法对自然场景文本检测任务性能提升的有效性已在用于任意形状文本检测的深度关系推理图网络(Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection, DRRG)中被证实^[1]。为了用锚点框的特征构造局部图,本文在锚点框预测模块使用改进的VGG-16作为骨干网络生成锚点框,并将锚点框的几何属性与骨干网络提取出的特征融合生成原始节点特征。根据文本图像中文本区域的自身性质,同一文本实例中的文本组件通常顺序连接形成链,对应了拓扑结构中的0维Betti数(即节点的联通数)的性质。然而传统的图卷积网络往往只能注意到邻接节点的信息而忽略了图自身的拓扑特征,在自然场景文本检测任务中使用传统图卷积网络的方法会导致锚点框漏检和文本实例划分出错等问题。针对这个问题,本文提出的融合拓扑特征的图卷积模块提取出了图中的拓扑特征,并与原始节点特征融合,然后送入图卷积网络进行锚点框连接关系的判断,有效提升了文本检测的效果。

1 相关工作

1.1 基于回归的文本检测算法

文本检测任务可以看成是特殊的目标检测任务,但它与一般目标检测算法相比存在着特殊之处。在一般的目标检测算法中,检测出的判定框只需与目标重叠率高于一个阈值就可认为是检测有效,而在文本检测任务中则需要正确检出且覆盖整个文本长度才能认为是检测有效。受目标检测算法^[2-4]的启发,基于回归的方法通常通过直接回归文本边界框或连接文本组件的几何信息来推断文本区域。受Faster RCNN^[3]的启发,连接文本建议网络(Connection Text Proposal Network, CTPN)使用垂直锚回归机制检测小规模文本锚点框^[5],并使用循环神经网络来利用前后文的信息进行文本位置的预测。TextBoxes^[6]通过使用长卷积核来改进单目标检测(Single Shot Detector, SSD)^[4],从而提高了水平文本检测的效果,TextBoxes++^[7]则进一步优化了卷积内核,并修改了

网络的输出,使其能够从多个角度检测文本。然而,这些方法无法检测弯曲文本。为了解决这个问题,TextSnake^[8]将文本实例表示为一系列重叠的圆。而为了更好地检测任意形状的文本,TextDragon^[9]在TextSnake^[8]的基础上改为使用一系列带角度的锚点框来表示文本区域,并首先检测文本的局部区域,然后根据它们的几何关系对这些文本锚点框进行分组。然而以上这些方法都只是单纯的通过锚点框浅层的几何特征来进行文本区域的判断,在本文的方法中则对锚点框之间的连接关系进行推理。

1.2 连接关系推理

基于回归的方法往往在长文本和复杂形状文本的检测任务中表现得更为出色,然而这些方法在检测任务中的表现很大程度上依赖于锚点框文本实例划分中的连接方法是否有效。CTPN^[5]中采取直接将相邻文本框连接起来的策略,这种方法虽然简单易实现,但只能用在水平文本检测中。而在PixelLink^[10]的方法中则借助了卷积神经网络(Convolutional Neural Networks, CNN),在连接元素之间进行了一个连接关系二分类预测。但这种借助CNN的方法很大程度受到卷积核的限制,对距离稍远的锚点框进行连接关系的预测结果不佳。为了避免这一问题,本文选择基于图卷积网络的锚点框连接策略。

2 方法

2.1 总体框架

本文方法的框架如图1所示,由锚点框预测模块和融合拓扑特征的图卷积网络模块组成。锚点框预测模块使用融合了特征金字塔网络^[11](Feature Pyramid Networks, FPN)的VGG-16作为主干网络来提取特征,将用于文本成分预测和图形神经网络。在获得文本组件的几何信息后,生成局部图以粗略地建立文本组件之间的连接。之后在融合拓扑特征的图卷积网络模块中提取局部图的拓扑特征,与局部图的原始节点特征融合,并将其输入图卷积网络^[12](Graph Convolutional Networks, GCN),以进一步推断锚点框之间的连接关系。

2.2 锚点框预测模块

锚点框预测模块的结构如图2所示,图中 H 为图片的高, W 为图片的宽,除了得到锚点框的8维预测结果外,还将其与上采样后得到的32维特征融合得到一个40维的特征 F_g ,该特征将用作后续的图卷积网络阶段的输入。锚点框预测模块中的下采样部分

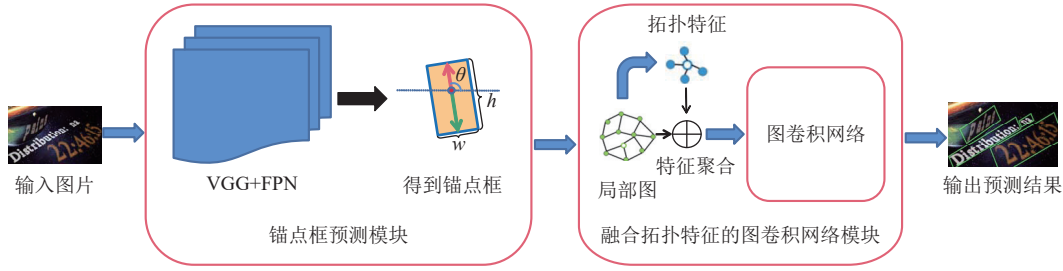


图1 方法总览

Fig.1 Overview of our method

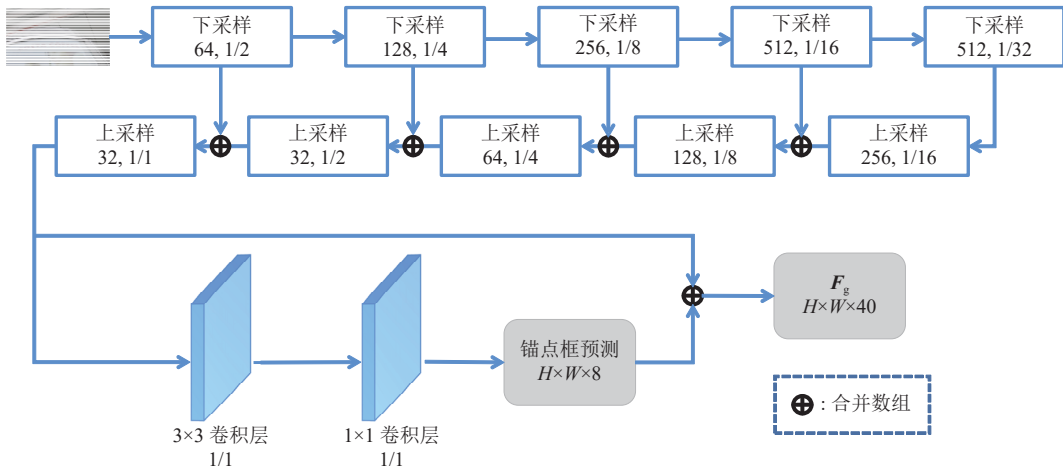


图2 锚点框预测模型

Fig.2 Anchor box prediction module

是一个VGG-16网络。每个上采样的基本单元由2个跟随着ReLU激活函数的卷积层和1个反卷积层构成,其中2个卷积层的卷积核分别是 3×3 和 1×1 。在本文的方法中,每个锚点框 C 由一组其几何特征的数据表示,具体为 $C = \{x, y, h, w, \sin \theta, \cos \theta\}$,式中: x 和 y 为文本锚点框的位置坐标, h 和 w 分别为文本锚点框的高度和宽度, $\sin \theta$ 和 $\cos \theta$ 为文本组件的倾斜角度。其中 w 由 h 线性变化而来,公式表示为

$$w_i = \begin{cases} w_{\min}, & h_i \leq 2w_{\min} \\ h_i/2, & 2w_{\min} < h_i < 2w_{\max} \\ w_{\max}, & 2w_{\max} \leq h_i \end{cases} \quad (1)$$

式中: w_i 和 h_i 分别为第 i 个锚点框的宽和高, $w_{\min}$ 和 $w_{\max}$ 根据经验分别设置成8和24。

在Textsnake^[8]和傅里叶轮廓嵌入网络^[13](Fourier Contour Embedding Net, FCENet)等方法中使用的文本中心区域(Text Center Region, TCR)和文本区域(Text Region, TR)概念可有效提高文本检测的精确度。如图3所示,本文通过将TR区域的两端收缩 $0.5w$ 得到TCR区域的两端,使用Textsnake^[10]中的方法生成中心线,再将中心线沿垂直方向延伸 $0.3h$ 得到TR区域。

图像在经过锚点框预测模块后可以得到一个8维的预测,其中4维为TR/TCR区域的预测结果,另外4维是 h_1 、 h_2 、 $\cos \theta$ 、 $\sin \theta$,其中 h_1 和 h_2 分别为 h 在中心线之上的值和 h 在中心线以下的值。此外,在进行最终预测之前需要对TR/TCR分类分数进行softmax操作,并且需要对 $\sin \theta$ 和 $\cos \theta$ 的预测值进行正则化使它们的平方和等于1。

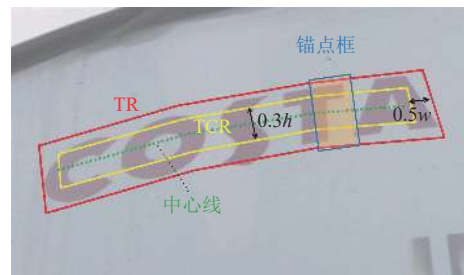


图3 TCR和TR示意图

Fig.3 Illustration of TCR and TR

文本锚点框检测模块的损失函数 L_a 由回归损失函数 L_r 和分类损失函数 L_c 两个部分相加而成,其中为了节约计算成本,在TCR区域之外的像素点都不会预测几何属性,因此 L_r 的计算只针对TCR区域中的像

素点,表示为

$$L_a = L_r + L_c \quad (2)$$

$$L_r = L_h + L_{\sin} + L_{\cos} \quad (3)$$

$$L_{\sin} = \text{smooth}_{L1}(\sin \hat{\theta} - \sin \theta) \quad (4)$$

$$L_{\cos} = \text{smooth}_{L1}(\cos \hat{\theta} - \cos \theta) \quad (5)$$

$$L_h = \frac{1}{O} \sum_{i \in O} \left(\log(h+1) \sum_{k=1}^2 \text{smooth}_{L1} \left(\frac{\hat{h}_{ki}}{h_{ki}} - 1 \right) \right) \quad (6)$$

式中: L_h 为锚点框高度损失函数, $L_{\sin}$ 和 $L_{\cos}$ 为锚点框的角度损失函数, smooth_{L1} 为L1光滑损失函数, O 是TCR中的正例元素, $\hat{h}_{ki}$ 、 $\sin \hat{\theta}$ 、 $\cos \hat{\theta}$ 为预测值, h_{ki} 、 $\sin \theta$ 、 $\cos \theta$ 分别为对应的ground-truth值。

2.3 融合拓扑特征的图卷积模块

2.3.1 局部图的生成

图神经网络经过近年来的发展,在节点分类任务中的表现已经得到了广泛的认证^[14-16]。由于文本检测任务通常只需要连接锚点框附近的其他锚点框,因此本文不会为每张任务图片都生成一个完整的图,而是只需生成多个局部图。将中枢点用 p 表示, G_p 为 p 所在的局部图, V_p 表示 G_p 中的节点。本文改良了ReLaText^[14]中的方法来生成局部图,只使用 p 的2-hop邻居来生成局部图,1-hop取 p 最近的8个节点,2-hop则取4个最近的节点。2-hop邻居将会为1-hop邻居提供辅助信息来完善上下文信息。通过计算节点间欧几里得相似度 u 来进行K最近邻算法(K-Nearest Neighbors, KNN)操作, u 的计算表示为

$$u = 1 - \frac{D(p, v_i)}{\max(H, W)}, v_i \in V_p \quad (7)$$

式中: $D(p, v_i)$ 为 p 和 v_i 之间的L2距离。为了避免相邻节点之间产生的局部图过于相似从而导致样本梯度累积,对局部图的生成设定了一个限定条件:

$$\frac{\overline{G}_a \cap \overline{G}_b}{\overline{G}_a \cup \overline{G}_b} < \zeta, a, b \in S \quad (8)$$

式中: S 为同一文本实例, $\overline{G}_a, \overline{G}_b$ 分别为 G_a, G_b 中只含节点 a, b 的1-hop邻居节点的子图。 ζ 在本文中按照经验设置为0.75。

为了获得局部图中节点的原始节点特征,将图2中得到的 F_g 送入RRoI-Align^[17]层reshape后得到 F'_g ,在此过程中会损失一部分几何属性特征,因此本文将锚点框中的几何属性用Attention Is All You Need^[18]中的方法嵌入到高维空间中,得到包含几何属性的特征 F_s 。将 F'_g 和 F_s 拼接之后得到 d 维原始节点特征向量。节点 $v \in V_p$ 的原始节点特征向量表示为 $m^{(v)}$, G_p 的原始节点特征矩阵记为 M 。

2.3.2 拓扑特征的融合

将拓扑特征融合到图神经网络中可提升预测结果在前人的工作中已得到验证^[19-20]。结合文本检测中文本图像的特点,由于0维Betti数十分契合锚点框自身的特性,本文选择0维Betti数作为要使用的拓扑特征,如图4所示。将一个局部图表示为 $G = (V, E)$,式中 V 为 n 个节点原始特征向量的集合, E 为局部图中节点的邻接矩阵。为了更好地表示出 G 的拓扑特征,首先需要将 G 划分为若干个子图 $G^{(j)}$,且 $G^{(j)}$ 满足以下条件:

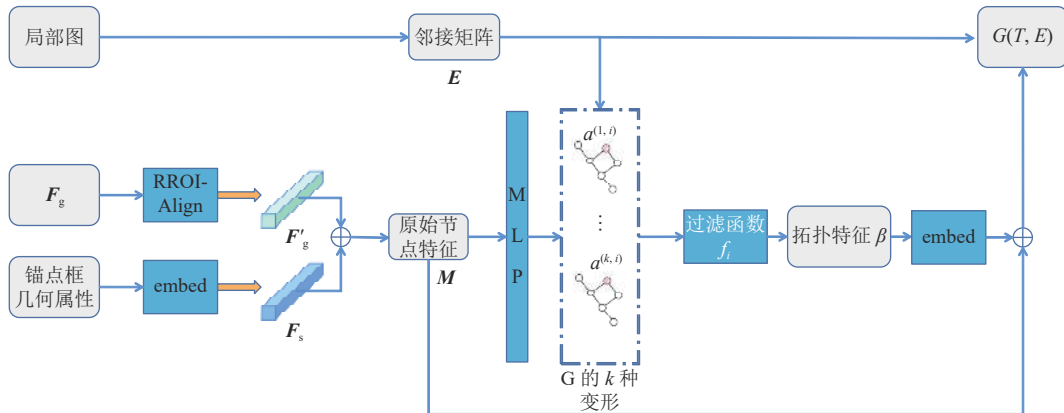


图4 拓扑特征融合模块

Fig.4 Topological feature fusion module

$$\emptyset = G^{(0)} \subseteq G^{(1)} \subseteq \dots \subseteq G^{(n-1)} \subseteq G^{(n)} = G \quad (9)$$

当一个0维Betti数在 $G^{(c)}$ 中首次出现并且在 $G^{(d)}$ 中消失时,使用一个元组 (c, d) 来表示这一拓扑持

久同源性。如图4所示,通过一个多层感知机将节点 v 的 d 维原始节点特征向量输出为 k 个值,将局部图中 n 个节点的原始特征向量映射为 $a^{(1,i)} < a^{(2,i)} < \dots < a^{(n,i)}$

的形式($1 < i < k$), $a^{(n,i)}$ 为第 n 个结点的原始特征经过第 i 种变形后得到的值。通过这种方式得到 G 的 k 种变形后,引入 k 个过滤函数 $f_i(1 < i < k)$,进一步得到满足以下条件的子图 $G^{(c,i)}$:

$$G^{(c,i)} = (V^{(c,i)}, E^{(c,i)}) \quad (10)$$

$$V^{(c,i)} = \{v \in V | f_i(m^{(c)}) \leq a^{(c,i)}\} \quad (11)$$

$$E^{(c,i)} = \{(a, b) \in E | \max\{f_i(m^{(a)}), f_i(m^{(b)})\} \leq a^{(c,i)}\} \quad (12)$$

通过子图可以统计出一系列的持久同源性元组,对元组进行append操作得到一个用来收集拓扑特征的数组 B_i , k 个过滤函数 f_i 可得到 k 个 B_i ,表示为 $\beta = \{B_1, B_2, \dots, B_k\}$ 。由于每个元组都可以与生成该元组的结点对应,即 B_i 与 V 构成双射关系,因此便可以为每个节点 $v \in V$ 分配一个由 β 嵌入到 d 维的向量 $e^{(v)}$,并与原始特征向量以残差的方式进行聚合,最终得到融合了拓扑特征的节点特征 $t^{(v)}$,表示为

$$t^{(v)} = m^{(v)} + e^{(v)} \quad (13)$$

把融合了拓扑特征的图用 $G(T, E)$ 表示,其中 T 为融合了拓扑特征的特征矩阵。

2.3.3 图卷积

本文使用图卷积网络来推导锚点框之间的深层连接关系,受Wang^[21]等的启发,本文中的图卷积层表示为

$$Y^{(l)} = \sigma((T^{(l)} \oplus LT^{(l)})W^{(l)}) \quad (14)$$

式中: $T^{(l)} \in J^{N \times d_{in}}$, $Y^{(l)} \in J^{N \times d_{out}}$; $J^{N \times d_{in}}$ 和 $J^{N \times d_{out}}$ 分别为 $N \times d_{in}$ 和 $N \times d_{out}$ 矩阵; d_{in} 和 d_{out} 分别为节点特征的输入和输出维度; N 为节点个数; $\oplus$ 为矩阵的串联操作; W 为权重矩阵; σ 为一个非线性激活函数,本文选择ReLU作为激活函数。 L 为对称归一化的拉普拉斯算子,表示为

$$L = D^{-1/2} \hat{E} D^{-1/2} \quad (15)$$

式中: D 为对角矩阵, $\hat{E}$ 为添加了自连接的邻接矩阵,表示为

$$\hat{E} = E + I_N \quad (16)$$

$$D_{ii} = \sum_j \hat{E}_{ij} \quad (17)$$

式中: I_N 为 $N \times N$ 的单位矩阵。本文的锚点框深层连接关系推理模块由1个批归一化层和4个图卷积层构成,最后通过全连接层输出节点与其1-hop邻居之间是否连接的结果来进行锚点框连接关系的推导。

3 实验

3.1 实验数据集

为了检验本文方法对弯曲文本以及多方向文本的检测能力,本文采用Total-Text^[22]数据集和CTW1500^[23]数据集作为实验数据集。

Total-Text^[22]数据集由1 255张训练集图像以及300张测试集图像组成,总共包含11 459个文本行。数据集包含水平、多向和弯曲各种情况的文本实例,大部分为英文文本,包含少量中文文本,采用多边形标注。

CTW1500^[23]数据集由1 000张训练集图像以及500张测试集图像组成,包含大量弯曲文本。每行标注包含32个数字,前4个数字为该弯曲文本在整张图上的矩形框坐标值;剩下的28个值(14个点)为与左上角坐标形成的差值。

3.2 实验细节

为了提高模型的鲁棒性,训练过程分为2个阶段:在SynthText^[24]数据集上的预训练阶段以及在特定数据集上的微调阶段。在预训练阶段对图像的文本区域进行随机裁剪操作,并将其大小调整为 512×512 ,每次训练所选取的样本数设置为12,采用Adam优化器对模型进行2轮的训练,学习率为0.000 1。在微调阶段,对数据进行随机裁剪操作后将其大小调整为 640×640 ,每次训练所选取的样本数设置为4,使用SGD优化器对模型进行600轮训练。微调阶段的初始学习率为0.01,每迭代100轮后学习率减少0.002。此外,一些常见的数据增强技术也被应用到训练中,如旋转、裁剪、颜色变化和翻转等。

本实验平台操作系统为Ubuntu 20.04。图像处理单元为NVIDIA GeForce RTX 3060显卡,显存为12 G,CUDA版本为11.0。深度学习框架为python 3.7搭配pytorch 1.9。

3.3 实验结果与分析

对于弯曲和多方向文本的部分实验效果如图5所示,从图5(c)和图5(e)对比可以看出,由于加入了拓扑特征,联系了锚点框之间的上下文信息,对文本区域的判断更加完整精确。此外通过对比图5(d)和图5(f)可以发现融入拓扑特征后模型对于文本实例的划分有了明显改善。

通过与近年来提出的几种文本检测方法进行对比,验证了本文方法的有效性。本文采用精确率(Precision)、召回率(Recall)和F分数(F-Score)3种指标



图 5 实验效果对比图

Fig.5 Comparison of experimental results

作为评价标准。精确率可以表示预测为正的样本中有多少是真正的文本区域,召回率可以表示图片中有多少文本区域被正确预测了,但精确率和召回率都只能关注到测试性能的一个方面, F 分数则是对它们进行了综合考虑。这3种评价标准的计算方式为

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

$$F - \text{Score} = (1 + \beta^2) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{20}$$

式中:TP为重叠度(Intersection over Union, IoU)大于0.5的预测结果数,FP为IoU小于0.5的预测结果数,FN为没有被检测出来的文本数量,超参数 β 表示Precision和Recall在性能评估中的相对重要程度,本文中 β 设置为1。

实验结果数据如表1和表2所示。可以看出在CTW1500^[23]数据集中,本文提出的方法在召回率、精确率和 F 分数这3个评价指标均产生了有竞争力的结果,值得一提的是本文方法的 F 分数在几个对比方法中是最高分。而在Total-Text^[22]数据集中的召回率和 F 分数均是最高分。可以看出本文方法的有效性和鲁棒性。

表 1 CTW1500实验数据对比

Table 1 Comparison experimental results of five methods on the CTW1500 database

方法	召回率	精确率	F 分数
TextSnake	0.853	0.679	0.756
MSR	0.790	0.841	0.815
CES	0.761	0.787	0.774
TextField	0.798	0.830	0.814
PSENet	0.797	0.848	0.822
本文方法	0.823	0.833	0.828

表 2 Total-Text实验数据对比

Table 2 Comparison experimental results of five methods on the Total-Text database

方法	召回率	精确率	F 分数
Total-Text	0.680	0.780	0.730
E2E	0.752	0.825	0.786
CTD	0.738	0.765	0.752
MSR	0.725	0.814	0.767
本文方法	0.771	0.807	0.788

此外,为了验证拓扑特征对文本检测有效性的提升,本文继续在CTW1500和Total-Text数据集上进行消融实验,实验结果如表3、表4所示。消融实验以仅使用了VGG-16+FPN的锚点框推断骨架网络作为baseline,分别设置加入图神经网络的模型和同时加入图神经网络以及拓扑特征的模型作为对照组。从消融实验的结果来看,加入图神经网络和拓扑特征后对模型在多个指标上都产生了较为明显的性能提升,其中CTW1500数据集上的召回率指标更是比baseline提升了3%的性能。

表 3 CTW1500上的消融实验结果

Table 3 Ablation experimental results on the CTW1500 database

方法	召回率	精确率	F 分数
Baseline	0.793	0.814	0.803
Baseline+图神经网络	0.812	0.811	0.812
Baseline+图神经网络+拓扑特征	0.823	0.833	0.828

表 4 Total-text上的消融实验

Table 4 Ablation experimental results on the Total-text database

方法	召回率	精确率	F 分数
Baseline	0.749	0.789	0.768
Baseline+图神经网络	0.758	0.795	0.769
Baseline+图神经网络+拓扑特征	0.771	0.807	0.788

但同时,在实验的过程中也发现了本文方法中

可能存在的一些问题。如图6所示,在一些含有极短文本任务的图像中有时会出现对短文本的漏检情况。推测是文本实例过短导致生成的锚点框数量过少,使生成的局部图中节点过少,最终导致无法很好地推断出锚点框之间的连接关系。在未来的工作中将会继续研究这个问题。



图6 极短文本漏检示意图

Fig.6 Missing detection of extremely short text

4 结束语

本文提出了一种嵌入拓扑特征的自然场景文本检测方法。首先通过骨干网络生成一系列文本锚点框,利用锚点框的几何属性和位置信息,将锚点框作为图的节点生成多个局部图。然后提取出局部图中的拓扑特征并与原始节点特征相融合,得到融合了拓扑特征的局部图。最后将融合了拓扑特征的局部图送入图卷积网络,推断节点之间的深层连接关系,根据推断出的连接关系对锚点框进行连接,最终得到文本区域。本文对拓扑特征提取模块做了消融实验对比,在2个公开数据集上验证了拓扑特征对模型性能的提升作用。此外,将本文方法与近年来的一些文本检测算法进行了性能对比,证明本文方法具有一定竞争力。

虽然本文提出并验证了拓扑特征可以有效提高文本检测的效果,但仍存在局限性。下一步的工作主要是研究如何在文本区域较短、锚点框和节点较少的情况下仍然能有效进行文本区域的判断。

参考文献:

- [1] ZHANG S X, ZHU X B, HOU J B, *et al.* Deep relational reasoning graph network for arbitrary shape text detection[EB/OL]. arXiv:2003.07493. (2020-08-30)[2023-05-12]. <https://doi.org/10.48550/arXiv.2003.07493>.
- [2] ROSS B, GIRSHICK, JEFF D, TREVOR D, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 580-587.
- [3] GIRSHICK R. Fast R-CNN[EB/OL]. arXiv:1504.08083. (2015-09-27)[2023-05-12]. <https://doi.org/10.48550/arXiv.1504.08083>.
- [4] LIU W, ANGUELOV D, ERHAN D, *et al.* SSD: single shot multi box detector[C]//Proc of the 2016 European Conference on Computer Vision. Amsterdam: Springer, 2016: 21-37.
- [5] ZHI T, HUANG W, TONG H, *et al.* Detecting text in natural image with connectionist text proposal network[C]//Proc of the 2016 European Conference on Computer Vision. Amsterdam: Springer, 2016: 56-72.
- [6] LIAO M, SHI B, BAI X, *et al.* TextBoxes: a fast text detector with a single deep neural network [C]//Proc of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 186-196.
- [7] LIAO M, SHI B G, BAI X. TextBoxes++: a single-shot oriented scene text detector [J]. *IEEE Trans on Image Processing: a Publication of the IEEE Signal Processing Society*, 2018, 27(8): 3676-3690.
- [8] LONG S B, RUAN J Q, ZHANG W J, *et al.* Textsnake: a flexible representation for detecting text of arbitrary shapes[C]//Proc of the European conference on computer vision. Munich: Springer, 2018: 20-36.
- [9] WEI F, HE W H, YIN F, *et al.* Textdragon: an end-to-end framework for arbitrary shaped text spotting[C]//Proc of the IEEE/CVF International Conference on Computer Vision, Long Beach: IEEE, 2019: 9076-9085.
- [10] DAN D, LIU H F, LI X L, *et al.* Pixellink: detecting scene text via instance segmentation[C]//Proc of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 296-308.
- [11] LIN T Y, DOLLAR P, GIRSHICK R, *et al.* Feature pyramid networks for object detection[C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE, 2017: 2117-2125.
- [12] DEFFERRARD M, BRESSON X, VANDERGHEYNST P, *et al.* Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in Neural Information Processing Systems. Barcelona: MIT, 2016: 29.
- [13] ZHU Y Q, CHEN J Y, LIANG L Y, *et al.* Fourier contour embedding for arbitrary-shaped text detection[C]//Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3123-3131.
- [14] MA C X, SUN L, ZHONG Z Y, *et al.* ReLaText: exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks[EB/OL].

- arXiv:2003.06999. (2020-03-16)[2023-05-12]. <https://doi.org/10.48550/arXiv.2003.06999>.
- [15] KOHLI H, AGARWAL J, KUMAR M. An improved method for text detection using Adam optimization algorithm [J]. *Global Transitions Proceedings*, 2022, 3(1): 230-234.
- [16] LIU Z, FANG Y, HUANG C, *et al*. GraphXSS: an efficient XSS payload detection approach based on graph convolutional network[J]. *Computers & Security*, 2022, 114: 102597.
- [17] MA J Q, SHAO W Y, YE H, *et al*. Arbitrary-oriented scene text detection via rotation proposals [J]. *IEEE Trans on Multimedia*, 2018, 20(11): 3111-3222.
- [18] ASHISH V, NOAM S, NIKI P, *et al*. Attention is all you need [J]. *Advances in Neural Information Processing Systems*, 2017, 30(10): 6000-6010.
- [19] GAO H, XIANG Y, SUI Y, *et al*. Topological graph convolutional network based on complex network characteristics[J]. *IEEE Access*, 2022, 10: 64465-64472.
- [20] JIANG W. Graph-based deep learning for communication networks: a survey[J]. *Computer Communications*, 2022, 185: 40-54.
- [21] WANG Z, ZHENG L, LI Y, *et al*. Linkage based face clustering via graph convolution network[C]//Proc of the IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 1117-2225.
- [22] CHENG C K, CHAN C S, LIU C L. Total-text: toward orientation robustness in scene text detection [J]. *International Journal on Document Analysis and Recognition(IJDAR)*, 2020, 23(1): 31-52.
- [23] YUAN T L, ZHU Z, XU K, *et al*. A large chinese text dataset in the wild [J]. *Journal of Computer Science and Technology*, 2019, 34(3): 509-521.
- [24] GUPTA A, VEDALDI A, ZISSERMAN A, *et al*. Synthetic data for text localisation in natural images[C]//Proc of the IEEE CONFERENCE on Computer Vision and Pattern Recognition. Las Vegas : IEEE, 2016: 2315-2324.
- (责任编辑: 杨耀辉 英文审核: 费伦科)

(上接第70页)

- [15] DAI C. Attentional feature fusion[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 3560-3569.
- [16] WOO S, PARK J, LEE J Y, *et al*. CBAM: convolutional block attention module[C]// European Conference on Computer Vision(ECCV). Munich: Springer, 2018: 3-19.
- [17] YI D, LEI Z, LIAO S, *et al*. Learning face representation from scratch [J]. *Computer Science*, 2014, 29(1): 51-59.
- [18] STAN L, DONG Y, ZHEN L, *et al*. The casia nir-vis 2.0 face database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2013: 348-353.
- [19] HUANG G B, RAMESH M, BERG T, *et al*. Labeled Faces in the Wild: a database for studying face recognition in unconstrained environments[EB/OL]. (2008-10-01) [2023-03-10]. https://people.cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf
- [20] ZHENG T, DENG W, HU J. Cross-pose lfw: a database for studying corsspose face recognition in unconstrained environments[EB/OL]. arXiv: 1708.08197(2017-08-28) [2023-03-10]. <https://arxiv.org/abs/1708.08197v1>.
- [21] MOSCHOLOU S, PAPAIOANNOU A, SAGONAS C, *et al*. Agedb: the first manually collected, in-the-wild age database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2017: 51-59.
- [22] TANG L, YUAN J, MA J. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network [J]. *Information Fusion*, 2022, 82: 28-42.
- [23] MA J, ZHANG H, SHAO Z, *et al*. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. *IEEE Transactions on Instrumentation and Measurement*, 2020, 99: 1.
- [24] PANETTA, KAREN, QIANWEN W, *et al*. A comprehensive database for benchmarking imaging systems[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(3): 509-520.
- (责任编辑: 张玮欣 英文审核: 费伦科)