基于 CTPN 神经网络对营业执照文字检测模型

邵慧敏,张太红*
(新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830001)

摘 要:对于复杂背景图片的文字识别,首先要做的就是定位目标文字的位置,即文字检测。想要文字识别率高,那对文字检测的准确度的要求就非常高了。传统的 RPN(region proposal network)神经网络在文字检测领域的研究已经很成熟,但 RPN 神经网络在营业执照水平文字检测的准确度上不是很理想。而基于 CTPN(connectionist text proposal network)神经网络的文字检测模型明显提高了营业执照水平文字检测的正确率,但用于项目中的话,准确率还是远远不够的。该文是以最新的营业执照作为研究对象,由于检测的图片易受光照和采集设备的影响,加上营业执照的背景比较复杂,所以能够准确地检测到目标文字的位置就非常具有挑战性。文中是通过 CTPN 神经网络模型来检测出营业执照中水平文字所在的位置,用矩形框来标注,也就是横向水平检测。目前开源的 CTPN 模型,都是基于某种数据集来训练的,所以对营业执照的文字检测效果就很差,因此该文使用 2 000 张营业执照图像作为实验数据,进行 10 000 迭代训练 CTPN 模型,最终能够准确地检测到营业执照中目标文字的位置,供项目使用。

关键词:营业执照;文字检测;Tensorflow;Opencv;CTPN
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2021)01-0094-04
doi:10.3969/j.issn.1673-629X.2021.01.017

Text Detection Model for Business License Based on CTPN Neural Network

SHAO Hui-min,ZHANG Tai-hong*
(School of Computer and Information Engineering,Xinjiang Agricultural University,Urumqi 830001,China)

Abstract:For text recognition of complex background images,the first thing to do is to locate the location of the target text,that is,text detection. To let the text recognition rate is high,the accuracy of text detection requirements are quite high. The traditional RPN (region proposal network) neural network has been very mature in the field of text detection,but the accuracy of RPN neural network in text detection at the level of business license is not ideal. While the text detection model based on CTPN (connectionist text proposal network) neural network significantly improves the accuracy of text detection at the level of business license,but the accuracy is far from enough when applied to the project. We focus on the latest business license as the research object. Since the detected pictures are susceptible to the influence of lighting and acquisition equipment,and the background of the business license is complex,it is highly challenging to accurately detect the location of the target text. The position of horizontal text in the business license is detected by the CTPN neural network model,which is marked by a rectangular box,that is,horizontal detection. The CTPN model of open source is based on some data sets to train,so the text detection effect of business license is poor. We will use the 2 000 business license image as the experimental data,the 10 000 iteration training CTPN model,to finally accurately detect the location of the target text in the business license for the use of the project.

Key words:business license;word detection;Tensorflow;Opencv;CTPN

0 引 言

营业执照是工商行政管理部门发给工商企业和个体经营者能够从事某些生产经营活动的证明,是证明某个企业具有一定资格的重要依据[1-2]。文本图像信息是人们获取外部信息的主要来源。在现代科学研究、军事技术、医学、工农业生产等领域,越来越多的人使用图像信息来识别和判断事物并解决实际问题。虽然从图像中获得文字信息非常重要,但更重要的是对文字图像进行处理,从图像中获取所需要的信息,因此在当今科学技术高速发展的时代,对文字图像的处理技术就有了更高的要求,能够更加快速准确地检测人们所需的图像文本信息[3-6]。

收稿日期:2020-02-04 **修回日期**:2020-06-09
基金项目:新疆维吾尔自治区重大科技专项(2017A01002-5)
作者简介:邵慧敏(1992-),女,硕士研究生,研究方向为智能算法;通讯作者:张太红(1965-),男,教授,研究方向为数据库技术。

目前,文字检测方法主要包括基于文本框回归的分类、基于分割的回归以及分割和回归结合的方法[7-8]。虽然近些年基于深度学习的文字检测方法已经取得巨大进步,但是文字作为一种具有其独有特色的目标,其字体、颜色、方向、大小等呈现多样化形态,相比一般目标检测更加困难[9-12]。一个模型在某个开源的数据集上得到了很好的效果,用这个方法直接换到另外的数据集上也许效果就不是很好,甚至是比较差的。因为很多模型是针对某项数据集来调整参数进行不断优化的,所以它极大依赖于数据,深度学习它有没有学到本质的东西,这个问题还值得深度探讨[13-17]。神经网络模型在文字检测方面已经有了研究,例如区域文本框网络(RPN),只是 RPN 进行的文字检测很难准确地进行水平检测。RPN 是通过直接训练来定位图像中的文本行,但是通过文本行来预测图像中的文本出现错误的可能性很大,因为文本是一个没有明确的封闭边界的序列。令人欣喜的是,Ren 提出了 anchor 回归机制允许 RPN 可以使用单尺度窗口检测多尺度的对象,这个想法的核心是通过使用一些灵活的 anchors 在大尺度和纵横比的范围内对物体进行预测[18-22]。其研究结果表明,根据 CTPN 方法,建立营业执照文字检测神经网络模型,能够准确地对营业执照的文字进行水平检测。

1　CTPN 神经网络简介

CTPN 神经网络模型主要包括三个部分:卷积层、双向 LSTM、全连接层。底层使用 VGG16 来提取特征,由一个 $W * H * C$ 的 conv5 的 feature map,使用大小为 3 * 3 的空间窗口,在最后一层卷积(VGG16 的 conv5)的 feature map 上滑动窗口。每行中的顺序窗口通过 BLSTM(bi-directional long short-term memory)循环连接,其中每个窗口的卷积特征(3×3×C)作为 BLSTM 的输入,再实现双向 BLSTM,增强关联序列的信息学习,再将 VGG 最后一层卷积层输出的 feature map 转化为向量形式,用于接下来的 BLSTM 训练。然后将 BLSTM 的输出再输入至 FC 中,最终模型输出:$2k$ 个 anchor 的文本/非文本分数、$2k$ 个 y 坐标、k 个 side_refinement(侧向细化偏移量)。该模型设计的 CTPN 神经网络模型如图 1 所示。

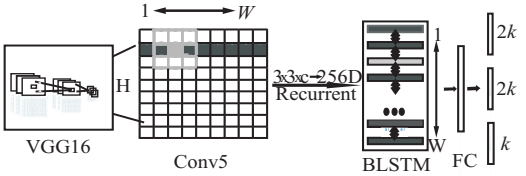


图 1　CTPN 神经网络模型

CTPN 神经网络是一个完整的卷积网络,可以允许输入任意大小的图像。CTPN 通过在 CNN 的 feature map 上密集地移动窗口来检测文本行,输出的是一系列的适当尺寸(固定宽度 16 像素,长度是可以根据情况调整的)的文本 proposal。给每个 proposal 设计了 k 个垂直 anchor 用来预测每个点的 y 坐标。这 k 个 anchor 具有固定 16 个像素的水平位置,但垂直位置在 k 个不同的高度上变化。此次使用 10 个 anchors,高度在 11 ~ 273 个像素变化,垂直坐标是通过一个 proposal 边界框的高度和 y 轴的中心计算得到的。有关预测 anchor 边界框的相对垂直坐标的计算公式如下:

$$\begin{cases} v_c = (c_y - c_y^a)/h^a \\ v_h = \log(h/h^a) \\ v_c^* = (c_y^* - c_y^a)/h^a \\ v_y^* = \log(h^*/h^a) \end{cases} \tag{1}$$

其中,$V = \{v_c, v_h\}$,$V^* = \{V_c^*, V_h^*\}$ 分别为预测的坐标和实际的坐标 GT(ground truth)。c_y 和 h^a 是 anchor box 的 y 轴的中心坐标和高度。

CTPN 的三个输出都被一起连接到全连接层上。这三个输出同时预测文本/非文本分数,垂直坐标和 side-refinement 的偏移量。采用 k 个 anchor 对它们三个分别预测,依次在输出层产生 $2k$、$2k$ 和 k 个参数(CTPN 固定了水平位置,只预测垂直位置)。利用多任务学习来联合优化模型参数,目标函数如下:

$$L(s_i, v_j, o_k) = \frac{1}{N_s}\sum_i L_s^{\mathrm{cl}}(s_i, s_i^*) + \frac{\lambda_1}{N_v}\sum_j L_s^{\mathrm{cl}}(s_j, s_j^*) + \frac{\lambda_2}{N_0}\sum_k L_0^{\mathrm{re}}(o_k, o_k^*) \tag{2}$$

每一个 anchor 是一个训练样本。I 是一个 anchor 在一个小部分的序列。s_i 是 anchor i 预测为一个真文本的概率,s_i^* 是 GT 的概率,j 是用于 y 坐标回归的有效 anchor 集合中的 anchor 的索引,v_j 和 v_j^* 是第 j 个 anchor 的 y 轴方向的预测值和 GT 的 k 个 side-anchor 的标号,side-anchor 为一系列离 GT 文本行框从左至右在水平距离范围内的 anchors。o_k 和 o_k^* 是 x 轴上第 k 个 anchor 预测的和 GT 的偏移量。L_s 为区分文本/非文本的 Softmax 损失,L_v 和 L_o 都为回归损失,其中 λ_1 为损失权重,根据经验设置为 1.0 和 2.0。

偏移量计算公式如下:

$$\begin{cases} O = (x_{\mathrm{side}} - c_x^a)/w^a \\ O = (x_{\mathrm{side}}^* - c_x^a)/w^a \end{cases} \tag{3}$$

其中,O 表示在 X 方向的归一化的偏移量,c_x 表示 anchor 的中心,x_{side} 表示预测的中心,w 表示 anchor 的宽度。针对文本/非文本的分类,二进制的标签被分给每一个正 anchor(文本)和负 anchor(非文本),正负 anchor 是由 IoU 与 GT 边界重叠计算得到的。正的

anchor 被定义为:IoU 与 GTbox 的重叠大于 0.7 的或者最高(集是一个很小的文本 pattern 也会被分为一个正的 anchor)的 anchor,负的 anchor 是 IoU 小于 0.5 产生。

2 估测模型训练过程

2.1 实验数据集

实验数据是笔者用手机拍摄及扫描的,总共收集大约 2 000 张营业执照数据集,采集日期是 2018 年 12 月初-至今。由于营业执照含有持有者的个人信息,所以收集起来比较困难。

2.2 数据预处理

2.2.1 图像采集

手机拍照或者扫描得到营业执照的图片。

2.2.2 图像预处理

营业执照的图像背景噪声大,所以首先利用 Opencv 对图像进行灰度化、矫正处理,再用 labelimg 对 2 000 张数据进行标注,得到 xml 格式的数据集,然后再转成 VOCdevkit 数据集,用于训练 CTPN 模型。

2.3 训练过程

该模型使用随机梯度下降(SGD)对现有的 CTPN 进行训练。因为牵扯到大量数据的计算训练,所以选用的服务器是适合于大规模运算的 Google Cloud Platform 的 GPU 服务器,所用数据集为 VOCdevkit,并进行 10 000 次迭代训练。与 RPN 神经网络相同的是训练样本为 anchors,每一个 anchor 是一个训练样本。对每个预测来说,水平位置和 k 个 anchors 的位置是固定的,这个是由输入图像在 conv5 的 feature map 上窗口的位置预先计算得到的,生成的文本 proposals 是由文本分数值大于 0.7(通过使用 NMS)的 anchor 生成的。通过使用垂直 anchor 和 fine-scale 策略,detector 可以处理各种比例和纵横比的文本行,进一步节省了计算量和时间。在迭代训练过程中,生成的 total_loss、model_loss 如图 2、图 3 所示。

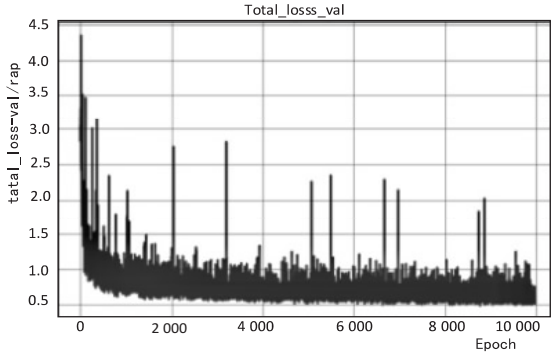


图 2 total_loss

目前,Mean Average Precision 特别适用于预测目标位置及类别的算法,因此它对评估定位模型、目标检测模型和分割模型非常有用。在计算 mAP 之前先要了解 Precision 和 Recall 也就是精确率和召回率,精确率主要用于衡量模型做出预测的精准度,召回率主要用于衡量模型对 Positives 的检测程度。

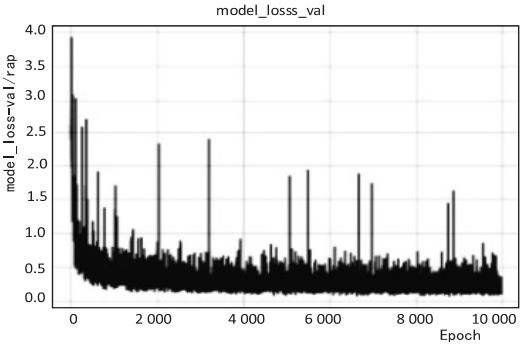


图 3 model_loss

$$\text{Preision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

其中,TP = True Positive,TN = True Negative,FP = False Positive,FN = False Negative。随着 Recall 从 0 到 1 之间的提升,AP(average precision)可以由计算 11 个不同 Recall 阶层最大 Precision 的评价值而得到。该文使用如下方式计算 AP:

$$\text{AP} = \frac{1}{11}\sum_{r\in\{0,0.1,\cdots,1\}} \text{Pinterp}(r) \tag{6}$$

$$\text{Pinterp}(r) = \max_{\tilde{r}\geqslant r}(r)$$

模型训练完成后进行测试,得到的 AP 如图 4 所示。

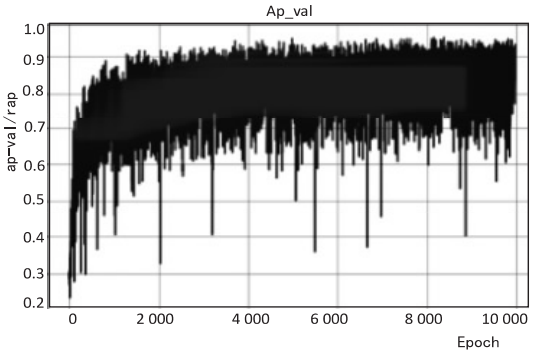
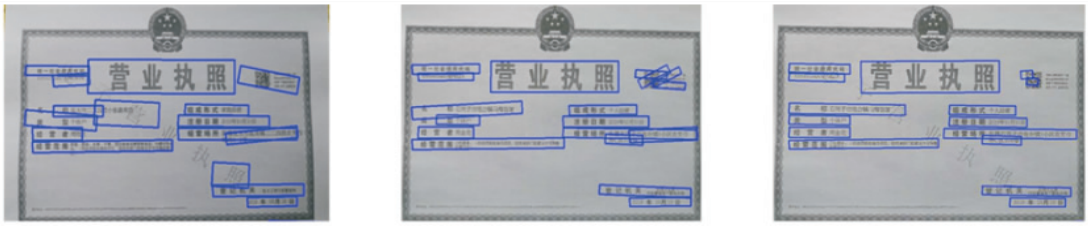


图 4 AP

3 实验结果及其对比

本次实验首先进行数据采集,其次对采集数据进行预处理,并分析所研究的营业执照文字中所需要检测的位置,然后运用 python 语言结合 Tensorflow 框架、Opencv 等第三方工具包构建 CTPN 神经网络,再根据评价指标对模型进行参数优化,最后确定模型并与现有的方式进行对比分析。经过多次实验后,选取了其中的一个样本进行对比分析,此次是将 RPN 和未训练的 CTPN 及训练后的 CTPN 对营业执照图像的文字检测结果进行对比,检测结果如图 5 所示。



(a)RPN　　(b)未训练的 CTPN　　(c)已训练的 CTPN

图 5　检测结果

由于营业执照复杂的背景,噪声干扰,字体大小不一致,颜色多样,且需检测的文字都是水平检测,所以基于 RPN 的方法对营业执照图像中文字的检测不具有良好的鲁棒性,经测试有很多的文字都未被检测到。像营业执照中的经营场所和经营范围出现多行文字的情况时,每一行文字没有被分开检测,这对后续的文字识别率有很大的影响,因为 OCR 识别只能识别单行的文字。没有用营业执照数据集训练的 CTPN 的检测率也不理想,部分文字未被检测到,同样在出现多行文字时没有将每一行进行分开检测,但检测效果比 RPN 要好。而经过训练后的 CTPN,检测的准确率大大提高,也解决了出现多行文字时每一行文字被分开检测,准确率达到项目使用的要求。

4　结束语

CTPN 利用文字序列的特点降低了检测难度,使其能够对背景复杂的营业执照图像进行高精度检测。而系统的不足之处是对拍摄角度、曝光度及像素较低的图像的检测率较低,将会在后期的研究中对其进行改进。目前数据集较少,还需要再不断收集数据,并且在用 labelimg 标注数据时,要避免勾图的框过大,保证文字被完整框住即可,使用这样标注的数据集进行训练,会更有利于提高 CTPN 模型的文字检测率。

参考文献:

[1] 闫永霖. 电子营业执照及其在工商全程电子化登记管理中的应用[J]. 财经界,2019(27):234-235.
[2] 靳振伟. 基于 CTPN 的网店工商信息提取系统的研究和实现[J]. 现代信息科技,2018,2(11):27-28.
[3] 王梦迪,张友梅,常发亮. 基于边缘检测和特征融合的自然场景文本定位[J]. 计算机科学,2017,44(9):300-303.
[4] 郭芬红,谢立艳,熊昌镇. 自然场景图像文字检测研究综述[J]. 计算机应用,2018,38(S1):173-178.
[5] 邱晓欢,吴啟超. 一种基于改进 EAST 网络和改进 CRNN 网络的火车票站名识别系统[J]. 南方职业教育学刊,2019,9(6):81-88.
[6] 马胜蓝. 基于深度学习的文本检测算法在银行运维中应用[J]. 计算机系统应用,2017,26(2):184-188.
[7] 谷振峰. 基于深度神经网络的自然场景文本检测方法的研究[D]. 天津:天津师范大学,2018.
[8] 杨国亮,王志元,张　雨,等. 基于垂直区域回归网络的自然场景文本检测[J]. 计算机工程与科学,2018,40(7):1256-1263.
[9] 胡　胤. 基于深度学习的自然场景文字检测方法研究[D]. 广州:广东工业大学,2018.
[10] 高友文,周本君,胡晓飞. 基于数据增强的卷积神经网络图像识别研究[J]. 计算机技术与发展,2018,28(8):62-65.
[11] 王柯力,袁红春. 基于迁移学习的水产动物图像识别方法[J]. 计算机应用,2018,38(5):1304-1308.
[12] 戈嘉宇,刘为嵩. 基于深度学习的身份证识别系统的设计与实现[J]. 电子世界,2020(2):109.
[13] GHANBARI-ADIVI F,MOSLEH M. Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE)[J]. Neural Computing and Applications,2019,31(12):8971-8983.
[14] WU Xianyu,LUO Chao,ZHANG Qian,et al. Text detection and recognition for natural scene images using deep convolutional neural networks[J]. Computers,Materials & Continua,2019,61(1):289-300.
[15] XU Yongchao,WANG Yukang,ZHOU Wei,et al. TextField:learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing,2019,28(11):5566-5579.
[16] 王雪冰,姜道义,张海洋. 基于卷积神经网络的文字识别优化方法研究[J]. 中国石油大学胜利学院学报,2019,33(4):39-42.
[17] 杨　捷,刘进锋. 利用 CTPN 检测电影海报中的文本信息[J]. 电脑知识与技术,2018,14(25):213-215.
[18] 张　勋,陈　亮,朱雪婷,等. 基于区域卷积神经网络 Faster R-CNN 的手势识别方法[J]. 东华大学学报:自然科学版,2019,45(4):559-563.
[19] ZHONG Zhuoyao,SUN Lei,HUO Qiang. An anchor-free region proposal network for Faster R-CNN-based text detection approaches[J]. International Journal on Document Analysis and Recognition,2019,22(3):315-327.
[20] YE Yangyang,ZHANG Chi,HAO Xiaoli. ARPNET:attention region proposal network for 3D object detection[J]. Science China Information Sciences,2019,62(12):40-42.
[21] 沈伟生. 基于 R2CNN 的自然场景图像中文本检测方法[J]. 无线互联科技,2019,16(2):107-109.
[22] 曹长玉,郑佳春,黄一琦. 基于区域卷积网络的行驶车辆检测算法[J]. 集美大学学报:自然科学版,2019,24(4):315-320.