有长方形的文字框而影响分割,于是首先对横向切割的像素做了一次遍历,如果有连续 5 个或以上像素的投影值都相同(对于正常图片,这种情况几乎不会出现),就将这些投影值都设置为 0.然后再依次进行把所有的满投影的值设置为 0、处理图像边缘和利用阈值切割的办法来得到切分的图片,其经过处理后的投影图很容易被切分,如图 10 所示.

Fig.9 Complicated binarization example

图 9 比较复杂的二值图

Fig.10 Modified horizontal projection

图 10 处理后的水平投影

把图表变成了易于检测的文字块以后,记录下行数 *m* 和列数 *n*.把所有图片按序号送入 tesseract OCR 中进行识别即可,识别后将结果进行拼接,即可得出最后的结果.

2.2 图片中的文字块检测

文字块检测和表格检测采用同样的方式获取数据和有针对性地进行训练,不同的是数据标记的方式,因为文字种类较为多样,我们选取了 100 张具有代表性的图片进行训练并只对训练样本的单行和相同字体的参数进行标记,从而增加了模型识别的鲁棒性,提高了模型的召回率.

训练过程采用与表格检测类似的方法.因为每个预测器只对一种字体和尺度的短文字块进行检测,所以模型的精度和召回率都很高,设置阈值为 0.85,并对检测结果使用非极大值抑制处理,即可得到我们需要的检测器,检测效果如图 11 所示.

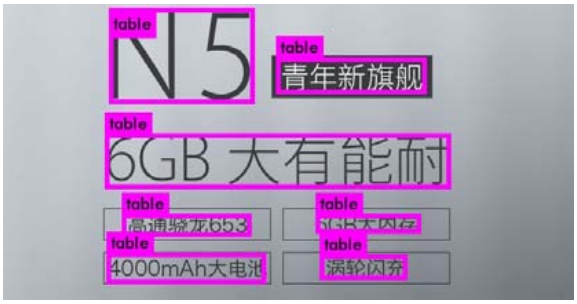

Fig.11 Results of text block detection

图 11 文字块的检测结果

检测得到文字块后,利用 OpenCV 切割所有的文字块.类似地,我们采用图像灰度化、直方图的均衡化、图像二值化以及中值滤波等方法对文字块进行处理,使得文字更加清晰并易于识别.然后把所有图片按从左至右和从上至下的方式送入 tesseract OCR 中进行识别,识别后拼接即得到最后的结果.

3 实验结果

本文的训练数据和测试数据全部来源于京东公开的商品介绍图片.本文使用 mAP 来评估检测精度,使用 Recall(召回率)来评估检测覆盖率.其中,AP 定义为每个类别根据召回率和正确率绘制的 P-R 图下的面积,mAP

则为多个类别 AP 的均值;Recall 定义为系统检测到的目标数目与测试集中所有的目标数目的比值.本文标注了 100 张含有表格的参数图片和 100 张含有文字块的参数图片,共包含 137 个表格块和 443 个文字块,测试集采用了 30 张表格图片和 30 张文字块图片,共包含 40 个表格块和 90 个文字块.实验使用初始的 YOLO 版本、更换 Darknet-19 网络的 YOLO 版本、添加 Batch Normalization 层的 YOLO 版本和使用上述全部改进的 YOLOv2 分别在微调阶段采用迭代了 20 000 次的权重进行对照实验.实验结果显示,使用预训练的方法来初始化 YOLOv2 模型中的网络参数具有良好的性能.可以看到,除了添加 Batch Normalization 后的表格区域检测 mAP 略有下降外,检测效果整体呈上升趋势,说明对模型的改进是有效的.最终的版本对表格区域的检测 mAP 高于 90%,对文字块的检测 mAP 均高于 85%,且具有较高的召回率,检测结果见表 1.同时,检测每张图片的时间约为 0.03s,远低于传统方法中版面分析等步骤耗费的时间,在保证识别速度的同时也大大提升了识别的精度.

Table 1 Detection results

表 1 检测结果

检测任务	表格区域检测		文字块区域检测	
评测指标	mAP(%)	召回率(%)	mAP(%)	召回率(%)
YOLO	83.65	87.50	77.31	82.22
YOLO+Darknet-19	87.36	95.00	80.06	87.78
YOLO+Darknet-19+BN	85.20	95.00	81.10	90.00
YOLOv2 (modified)	90.42	95.00	85.22	94.44

在文字识别阶段,本文调用开源 OCR 工具 tesseract,表格和文字块识别结果样例分别如图 12 和图 13 所示.对于表格参数和文字块参数同时出现的情况,我们的系统使用表格检测区域抑制与其重叠的文字块检测区域,可以将表格和文字块正确区分并同时检测出来.

Fig.12 Recognition results of table region

图 12 表格区域的识别结果

由于训练模型时使用的数据集较小,对各种情形的覆盖度不全,有时会存在误检和漏检的情况,如图 14 所示.本文中的方法只根据表格和文字块训练了两个模型,考虑到商品参数形式的多样化,对于极不相似的表格和文字块可以分别训练模型,这可以在一定程度上提高模型的检测准确率.

Fig.13 Recognition results of text block

图 13 文字块的识别结果

Fig.14 Failure detection examples

图 14 有误检和漏检的情况

4 总结与展望

本文提出了一种将深度学习检测算法与传统 OCR 技术相结合的方法,解决了传统方法中对图片的复杂背景检测率低的缺点.使用 YOLOv2 算法训练了表格检测和文字块检测两类模型对图片中存在商品参数的区域进行检测,结果使用表格区域抑制文字区域的生成,然后经过一系列图像处理方法后得到标准的易于识别的单行文字块,再使用 tesseract 进行简单的文字提取,最后按照参数的对应位置将提取出的文字排列为正确的格式即得到识别结果.

目前,CNN 也可以应用在文字识别领域中,得益于卷积神经网络强大的特征提取能力及其鲁棒性,如果使用这种方法来提取图片中的商品参数,不仅不需要依赖传统的 OCR 工具,甚至也不需要对检测出的图片做对比度增强等处理.整个检测流程可以简化为目标检测、图片分割、文字识别这 3 个阶段(甚至可以在目标检测提取到的特征图上直接分割并识别,只使用卷积神经网络实现端对端的任务流程),其中,第 1 个和第 3 个阶段都使用卷积神经网络来完成,这可以使系统的鲁棒性和识别精度得到更大的提升.

References:

[1] Liu Z, Wu QY. Research on forms registration in general forms processing system. Ruan Jian Xue Bao/Journal of Software, 1996,7(7):409–414 (in Chinese with English abstract). http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=19960704&journal_id=jos

[2] Li XY, Gao W. A robust method for unknown structure form analysis. Ruan Jian Xue Bao/Journal of Software, 1996,10(11):1216–1224 (in Chinese with English abstract). http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=19991118&journal_id=jos

[3] Zhang QH. On automatic recognition of form data. Journal of Xi'an University of Science & Technology, 2000,20(4):1001–7127 (in Chinese with English abstract).

[4] Zheng YF, Liu CS, Ding XQ, Pan SY. A form frame-line detection algorithm based on directional single-connected chain. Ruan Jian Xue Bao/Journal of Software, 2002,13(4):790–796 (in Chinese with English abstract). http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20020446&journal_id=jos

[5] Fang J, Gao LC, Qiu RH, Tang Z. Automatic table boundary detection and performance evaluation in fixed-layout documents. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013,49(1):45–53 (in Chinese with English abstract).

[6] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11):2278–2324. [doi: 10.1109/5.726791]

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the CVPR. 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

[8] Chen PH, Lin CJ, Schölkopf B. A tutorial on ν-support vector machines. Applied Stochastic Models in Business & Industry, 2005, 21(2):111–136. [doi: 10.1002/asmb.537]

[9] Felzenszwalb P, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(9):1627–1645. [doi: 10.1109/TPAMI.2009.167]

[10] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the CVPR. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

[11] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(9):1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

[12] Girshick RB. Fast R-CNN. In: Proc. of the ICCV. 2015. 1440–1448. http://ieeexplore.ieee.org/document/7410526/

[13] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,39(6):1137 1149.

[14] Redmon J, Farhadi A. YOLO 9000: Better, faster, stronger. In: Proc. of the CVPR. 2017. 7263–7271. http://ieeexplore.ieee.org/document/8100173/

[15] Smith R. An overview of the tesseract OCR engine. In: Proc. of the ICDAR. 2007. 629–633. [doi: 10.1109/ICDAR.2007.4376991]

附中文参考文献:

[1] 刘真,吴泉源.通用表格处理系统中定位方法的研究.软件学报,1996,7(7):409–414. http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=19960704&journal_id=jos

[2] 李星原,高文.一种鲁棒性的结构未知表格分析方法.软件学报,1996,10(11):1216–1224. http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=19991118&journal_id=jos

[3] 张群会.表格数据自动识别技术研究.西安科技学院学报,2000,20(4):1001–7127.

[4] 郑冶枫,刘长松,丁晓青,潘世言.基于有向单连通链的表格框线检测算法.软件学报,2002,13(4):790–796. http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=20020446&journal_id=jos

[5] 房婧,高良才,仇睿恒,汤帜.版式电子文档表格自动检测与性能评估.北京大学学报(自然科学版),2013,49(1):45–53.

丁明宇(1996－),男,吉林白山人,硕士生,主要研究领域为深度学习,计算机视觉.

卢志武(1978－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为机器学习,计算机视觉.

牛玉磊(1992－),男,博士生,CCF 学生会员,主要研究领域为计算机视觉,机器学习.

文继荣(1972－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为互联网大数据管理,信息检索.