软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(4):1039−1048 [doi: 10.13328/j.cnki.jos.005408]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度学习的图片中商品参数识别方法*

丁明宇, 牛玉磊, 卢志武, 文继荣

(大数据管理与分析方法研究北京市重点实验室(中国人民大学 信息学院),北京 100872)

通讯作者: 卢志武, E-mail: luzhiwu@ruc.edu.cn

摘 要: 计算机计算性能的提升使得深度学习成为了可能.作为计算机视觉领域的重要发展方向之一的目标检测也开始结合深度学习方法并广泛应用于各行各业.受限于网络的复杂度和检测算法的设计,目标检测的速度和精度成为一个 trade-off.目前电商领域的飞速发展产生了大量包含商品参数的图片,使用传统方法难以有效地提取出图片中的商品参数信息.针对这一问题,提出了一种将深度学习检测算法和传统 OCR 技术相结合的方法,在保证识别速度的同时大大提升了识别的精度.所研究的问题包括检测模型、针对特定数据训练、图片预处理以及文字识别等.首先比较了现有的目标检测算法,权衡其优缺点,然后使用 YOLO 模型完成检测任务,并针对 YOLO 模型中存在的不足进行了一定的改进和优化,得到了一个专用于检测图片中商品参数的目标检测模型,最后使用 tesseract 完成文字提取任务.在将整个流程结合到一起后,该系统不仅有着较好的识别精度,而且是高效和健壮的.最后讨论了优势和不足之处,并指出了未来工作的方向.

关键词: 目标检测;图像切割;光学字符识别;商品参数;深度学习

中图法分类号: TP391

中文引用格式: 丁明宇,牛玉磊,卢志武,文继荣.基于深度学习的图片中商品参数识别方法.软件学报,2018,29(4):1039−1048. http://www.jos.org.cn/1000-9825/5408.htm

英文引用格式: Ding MY, Niu YL, Lu ZW, Wen JR. Deep learning for parameter recognition in commodity images. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1039−1048 (in Chinese). http://www.jos.org.cn/1000-9825/5408.htm

Deep Learning for Parameter Recognition in Commodity Images

DING Ming-Yu, NIU Yu-Lei, LU Zhi-Wu, WEN Ji-Rong

(Beijing Key Laboratory of Big Data Management and Analysis Methods (School of Information, Renmin University of China), Beijing 100872, China)

Abstract: The improvements of computing performance make deep learning possible. As one of the important research directions in the field of computer vision, object detection has combined with deep learning methods and is widely used in all walks of life. Limited by the complexity of the network and the design of the detection algorithm, the speed and precision of the object detection becomes a trade-off. At present, the rapid development of electronic commerce has produced a large number of pictures containing the product parameters. The

* 基金项目: 国家自然科学基金(61573363); 北京市科委类脑计算专项(Z171100000117009); 中国人民大学预研委托项目(15XNLQ01); 中国人民大学拔尖创新人才培育资助计划

Foundation item: National Natural Science Foundation of China (61573363); Beijing Brain Research Project of Beijing Municipal Science & Technology Commission (Z171100000117009); Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01); Outstanding Innovative Talents Cultivation Funded Programs of Renmin University of China

丁明宇、牛玉磊为共同第一作者,对本文贡献相同.

本文由“多媒体大数据处理与分析”专题特约编辑赵耀教授、李波教授、华先胜研究员、文继荣教授、蒋刚毅教授、常冬霞副教授推荐.

收稿时间: 2017-04-29; 修改时间: 2017-06-26; 采用时间: 2017-10-13; jos 在线出版时间: 2017-12-01

CNKI 网络优先出版: 2017-12-04 06:46:57, http://kns.cnki.net/kcms/detail/11.2560.TP.20171204.0646.009.html

traditional method is hard to extract the information of the product parameters in the picture. This paper presents a method of combining deep learning detection algorithm with the traditional OCR technology to ensure the detection speed and at the same time greatly improve the accuracy of recognition. The paper focuses the following problems: The detection model, training for specific data, image preprocessing and character recognition. First, existing object detection algorithms are compared and their advantages and disadvantages are assessed. While the YOLO model is used to do the detection work, some improvements is proposed to overcome the shortcomings in the YOLO model. In addition, an object detection model is designed to detect the product parameters in images. Finally, tesseract is used to do the character recognition work. The experimental results show that the new system is efficient and effective in parameter recognition. At the end of this paper, the innovation and disadvantage of the presented method are discussed.

Key words: object detection; image segmentation; optical character recognition; product parameters; deep learning

随着电子商务的发展,几乎所有的商品数据都可以在互联网中找到.但是还有大量数据是以表格或者文字块的方式存在于图片中,导致商品信息的利用率不高.利用图片中的商品数据可以进行商品图片检索、商品真实性验证以及相似商品匹配等.如何准确、快速地识别出图片中的商品参数,成为一个非常值得研究的问题.由于商品参数一般以表格和文字块两种方式存在于图片中,提取商品参数首先要进行表格和文字块定位,然后再进行文字识别.与传统的目标检测问题相比,商品参数识别问题更为复杂,不仅需要对表格和文字进行位置检测,还要对区域进行提取并进一步识别其文字信息.同时,考虑到已有的目标检测算法着重于目标物体的颜色、纹理等特征的提取,与表格、文字等目标物体的特征有较大差别,传统的目标检测算法不能直接应用于商品参数识别问题中.综上,图片中的商品参数提取由于其问题的复杂性与目标表格、文字的特殊性,有必要采取新的目标检测方法解决这一问题.

1 相关工作

早期的表格定位和识别方法往往有着很大的局限性,刘真等人[1]提出了表格的四角定位法,但是需要预先设置表格模型,人力成本高且应用面窄;李星原等人[2]提出以表格线为导引的矩形块抽取方法,但只能抽取被框线包围的表项内容;张群会[3]提出了利用二值化和投影来识别表格数据,其投影同样基于表格框线,无法识别没有框线的表格;郑治枫等人[4]提出基于有向单连通链的表格框线检测算法来检测表格框线.上述方法的通用性不强且定位准确率不高.房婧等人[5]提出了基于文本布局特征分析的表格定位方法,其检测是针对电子文档,根据多条水平行和竖直列的相交程度来判断表格位置,但仍然难以识别不规则或空缺较多的表格.

针对日益增长的图片中信息提取的需求和传统方法识别流程复杂且对于复杂背景和不规范场景识别准确率低的缺点,加之京东、淘宝等商家的商品图片数目是数以亿计的且还在不断增长中,检测速度也是选择模型和评估系统性能的一个重要指标.据此,本文提出了将 CNN[6]和传统 OCR 技术结合起来完成大规模图片中商品参数识别任务的方法,其中,整个模型采用端对端的方法训练,通过识别图片中的表格参数和图片中的文字块参数两类方法来加以实现,不仅提高了检测速度,简化了识别的流程,也提高了检测系统的鲁棒性和检测精度.

1.1 相关检测算法

早期的目标检测使用人工提取的特征,如 LPB、HOG[7]、SIFT 等特征,随后利用 SVM[8]、Adaboosting 等分类器进行检测,其检测效果最好的一个代表是 DPM 算法[9],但是人工提取的特征在通用性方面有着较大的缺陷,往往只适用于特定领域,且对复杂度高的背景的处理难度较大;在 DPM 算法经历了很多年的平台期后,Girshick 等人[10]在 2014 年提出的 R-CNN 方法使用选择性搜索提取候选框替代了传统方法中的滑动窗口,同时使用 CNN 提取的特征替代了传统目标检测中人工设计的特征,检测效果提升显著;SPP-net[11]在最后的卷积层和全连接层中间加入一个 SPP 层,将候选框在原图的位置映射在尺寸为 $a\times a$ 的特征图上,然后在 SPP 层进行空间金字塔采样生成固定维度的特征,解决了 R-CNN 中存在的问题;Fast R-CNN[12]采用了 multi-task loss,把区域的回归任务也放入到网络的训练之中,免去了训练 SVM 分类器的步骤,实现了整个网络端到端的训练方式,Fast R-CNN 与 R-CNN 和 SPP-net 相比,训练步骤简单,不需要把提取的特征保存到磁盘,可以更新所有层的参数,在大幅度提高训练速度的同时实现了检测精度的提升,但其 region proposal 过程耗时太久,而使用 selective search

无法实现真正意义上的端对端训练;Faster R-CNN[13]中使用区域建议网络(region proposal network)代替 selective search 方法,实现了和后续的检测网络共享特征,使得产生候选区域的步骤几乎不耗时.Faster R-CNN 中的区域建议网络和检测网络共享卷积特征,不仅大大加快了检测速度,更新的 RPN 所生成的区域质量更高,检测精度也得到了进一步的提高.

1.2 YOLO模型

YOLO 模型使用了回归的思想,可以直接在输入图像的网格划分中回归出待检测目标的类别和目标定位,大大加快了检测的速度,同时依然保持着较高的检测精度.首先,将输入图像划分为每边 S 块的网格,如果目标的中心点落在某个网格里,那么这个网格就负责检测对应的目标.每一个网格通过回归的方式预测 B 个区域的位置并为每个区域预测一个得分.这个得分反映了区域内包含一个目标的置信度和预测区域的精确度,用 $\Pr(Object)\times IOU_{pred}^{truth}$ 来表示,其中, IOU_{pred}^{truth} 表示目标区域与预测区域面积的交并比.如果网格中没有目标,预测得分为 0,否则,预测得分等于 IOU 的值.这样,每个区域需要 5 个预测值,分别是归一化的 x,y,w,h 和得分,x 和 y 代表区域的中心点对于 $cell$ 左上角的偏移量,w 和 h 代表区域相对于全图的宽和高,它们都介于 0~1 之间.此外,不论每个格子中有多少个区域,都只为 C 个类预测一个属于此类的后验概率 $\Pr(Class_i|Object)$, 测试时将上述得分和条件概率相乘即可得到一个区域的特定类的评分,即

$$\Pr(Class_i|Object)\times\Pr(Object)\times IOU_{pred}^{truth}=\Pr(class_i)\times IOU_{pred}^{truth},$$

输出一个 $S\times S\times(B\times(4+1)+C)$维的向量,同时完成了目标分类和定位的任务.

在网络设计方面,采用了 GoogLeNet 的设计思想,包含 24 个卷积层和 2 个全连接层,但采用 1×1 的归约层后接 3×3 的卷积层来代替其中的 inception modules,其网络结构如图 1 所示.

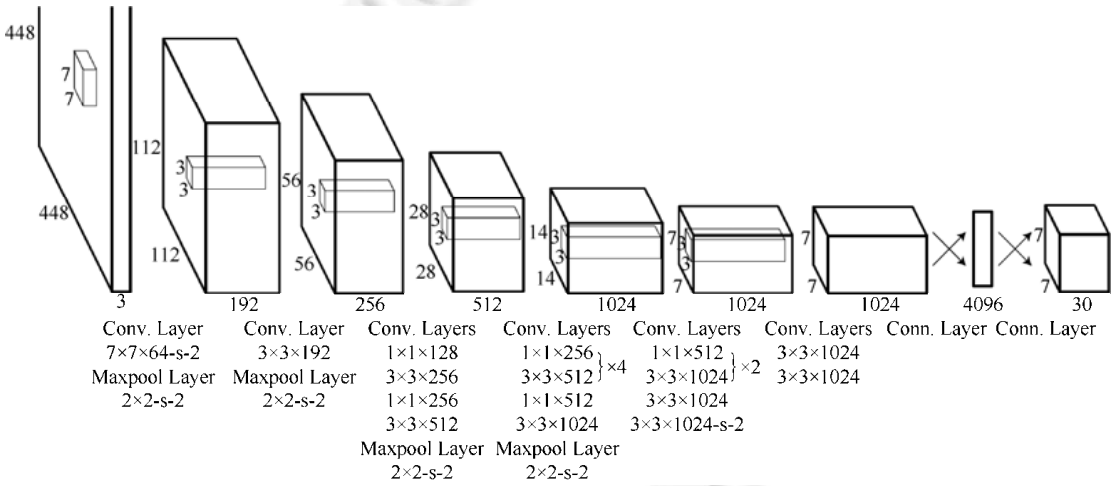

Fig.1 Network architecture of YOLO

图 1 YOLO 的网络结构

2 识别系统

目标检测的新方法不仅在学术界倍受关注,在工业界也发挥了很大的作用,考虑到检测大规模商品参数图片时对速度的要求,本文采用 YOLOv2[14]模型对图片中的表格和文字块进行定位,简化了传统文字识别的版面分析和特征工程等步骤,本文实现的系统包括图片中商品参数的定位、识别以及还原成表格的整个流程.

图片中的商品参数分为两种形式,第 1 种是以表格的形式存在,含有表头以及对应的参数值,需要首先定位表格的位置,然后把表格切分后进行识别;第 2 种是以文字块的形式存在,直接定位文字位置然后加以识别即可.无论哪一种参数,首先都需要对参数位置进行定位,本文采用 YOLOv2 模型同时对一张图片进行表格检测和文

字块检测,然后将结果统一.为了防止表格中的文字被重复检测,本文使用高置信度的阈值来过滤表格检测的结果,降低其假阳性比率,然后使用非极大值抑制的方法,即使用检测出的表格位置来抑制与其 *IOU* 大于 0.1 的文字块位置的选择,然后输出包含表格位置和文字块位置两类的最终检测结果.之后,利用 OpenCV 工具切割出商品参数区域的图片,然后对其进行灰度处理、二值化处理以及对比度增强等操作以使其易于识别.为了降低对 OCR 方法功能的需求,对于表格中的商品参数需要增加一步额外的处理,即将表格中的每项分别切分出来,从而便于后面识别对应文字.本文主要使用了投影分割的方法来完成表格项的切割.

当所有的商品参数都转变成了易于识别的单行的清晰文字后,利用 OCR 工具即可识别出对应文字,最后将表格中识别出的文字拼接成图片中表格的样子,将文字块中识别出的文字按序排列即可,文字识别阶段采用开源工具 tesseract[15].系统的识别流程如图 2 所示.

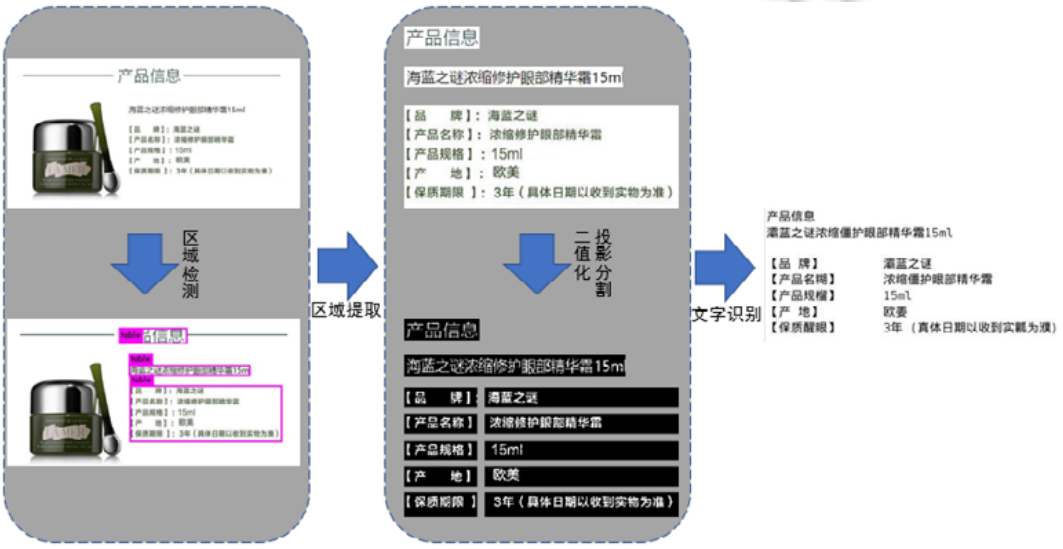

Fig.2 Flowchart of parameter recognition

图 2 商品参数的识别流程

2.1 图片中的表格检测

2.1.1 区域检测

YOLO 模型有着很大的局限性,主要表现在定位误差和召回率低两个方面.为了弥补 YOLO 中存在的问题,对 YOLO 模型进行改进得到了 YOLOv2,在保证速度的同时大幅度提升了检测精度,其主要改进点如下所示.

(1) 新的全卷积网络.YOLOv2 中构建了一个新的网络 Darknet-19 用于分类,滤波器大多采用 3×3 的大小,后面跟随着池化层,并使特征图的数量加倍.Darknet-19 包含 19 个卷积层和 5 个最大值池化层.

(2) 添加 Batch Normalization 层.在训练过程中,本文的方法在每一个卷积层后添加批量归一化项,提高了网络的收敛性同时减少了对其他正则化方法的依赖,舍弃了 dropout 优化后依然可以防止过拟合.

(3) 使用固定区域进行卷积.YOLO 模型中含有全连接层,在 YOLOv2 中去除了全连接层并使用固定区域进行预测,把预测任务放到了提取的区域中,这可能导致 mAP 的微小下降,但可以极大地提升模型的召回率.

(4) 维度聚类.在 Faster R-CNN 中使用的固定区域是手动确定的大小和比例,如果采用更有代表性的先验区域维度,网络的学习则会更加容易.YOLOv2 中对数据集中的标记进行 *K*-means 聚类,每个网格中区域的个数为聚类个数 *k*,区域的维度则使用 *k* 个聚类中心的宽高维度来表示.

(5) 细粒度特征.在 YOLOv2 中则采用类似 ResNet 中 identity mapping 的方法把浅层特征图(26×26)连接到深层特征图.通过添加一个 passthrough 层把高分辨率的特征图和之后的特征图联系到一起,把特征划分到多个 feature map 里,使得特征图的通道数提高了 4 倍.

模型的损失函数由 5 部分组成,整体定义如下.

$$Loss = \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2\right]+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right]+$$
$$\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}(C_i-\hat{C}_i)^2+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{noobj}(C_i-\hat{C}_i)^2+\sum_{i=0}^{S^2}1_i^{obj}\sum_{c\in classes}(p_i(c)-\hat{p}_i(c))^2,$$

其中,1_{ij}^{obj} 用来表示第 i 个网格中的第 j 个区域是否负责这个 *object*(即 *object* 的中心落在网格中,且此区域与真实框的 *IOU* 最大),如果是,就取 1,否则为 0,式子的前两项表示定位的预测损失;第 3 项表示含有 *object* 的区域的置信度预测损失,即 *IOU* 误差;第 4 项表示不含有 *object* 的区域的置信度预测损失(*IOU* 误差);第 5 项中 1_i^{obj} 表示是否有目标的中心落在网格中,用来表示分类损失.只有当某个网格中有 *object* 时才对分类误差进行惩罚,只有当某个区域预测器对某个 *object* 负责的时候,才会对坐标误差进行惩罚.

为了精准检测商品图片中存在的表格区域,我们需要使模型只针对这一类数据进行训练.训练的数据全部来自于京东中含有表格参数的商品图片,我们人工选取了 100 张具有代表性的图片并对其中参数的位置进行了人工标注,标注的坐标进行了归一化后生成了区域中心坐标(x,y)和区域的宽高(w,h)以用于后续检测.训练检测模型时,我们采用了 YOLOv2 在 Imagenet 数据集上预训练的模型来初始化网络的权重,网络共迭代 45 000 次,权值衰减率和动量分别为 0.000 5 和 0.9,用得到的模型对图片进行检测.为了保证检测的高准确率并防止误检,检测模型中采用 thresh 为 0.85,用高阈值过滤后的模型输出,在召回率几乎没有降低的情况下大大降低了假阳性比例.检测效果如图 3 所示.

Fig.3 Results of table region detection

图 3 表格区域检测结果

2.1.2 区域切割

得到区域位置后对截取到的图片进行切割,把表格的每个表头和每个参数全部分离,生成易于检测的单行短文字用于之后的光学字符识别.我们采用投影分割的方法来对商品参数区域进行分割.

区域分割的第 1 步是将 RGB 三通道彩色图像转化为灰度图像,我们使用加权平均的办法,取图像中红色 R、绿色 G、蓝色 B 这 3 种颜色的像素值,根据一定的权值加权平均,这样就把三通道的图片变成了单通道,即得到了灰度图像.这里取权重 $W_R=0.299$,$W_B=0.114$,$W_G=0.587$, 计算公式如下所示:

$$R=B=G=\frac{W_R\times R+W_G\times G+W_B\times B}{W_R+W_G+W_B}.$$

本文采取了直方图均衡化的方法来提升灰度图的对比度,它通过累计分布函数变换来进行直方图的修正,把图片的直方图的分布改变为均匀密度分布,更有利于图像二值化的阈值选取.其原理图如图 4 所示.

我们对图像进行二值化处理,使用 OTSU 算法得到一个自适应阈值.然后使用了中值滤波的方法来去掉二值图的噪点,即把每个位置的像素值用其相邻的 8 个像素点的中值来代替.

经过上述处理后,文字和背景基本可以清晰地分离出来.我们发现,对于不同的背景颜色和文字颜色,二值图的呈现方式也有所不同.有的是白底黑字,而另外的可能是黑底白字,以便于后续的表格切割,需要统一所有

图片的文字和其他的二值化颜色.我们采用基于颜色分量的判断方法,一般地,字体颜色浅的图片中红色 R<蓝色 B,而字体颜色深的图片中的蓝色 B>红色 R,经过这一步判断后,我们对相反的图片进行了反色处理,使得生成的二值图大部分都为黑底白字,示例结果如图 5 所示.

我们使用投影分割法对二值图进行分割,同时对二值图作水平和垂直方向的投影,根据投影的空白间隔的位置即可判断出应该切割的具体位置.纵向只需要分离表头和表项,为了防止因为字体不连贯而导致的多余的切割,我们设置了切割的阈值,保证不会在文字的缝隙中进行切割.本文中纵向切割的阈值为 8,横向切割的阈值为 3,考虑到图片边缘一般是没有文字的,因此投影量为 0,为了防止对边缘位置的误切和噪点的影响,我们首先从上下左右 4 个方向对图片的投影连续小于 4 的位置进行填充,填充值为 40,保证边缘不会被切割,两个方向的投影图如图 6 和图 7 所示,上述二值图被切割后的示意图如图 8 所示.

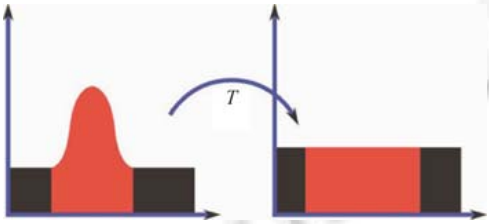

Fig.4 Illustration of histogram equalization

图 4 直方图均衡化原理

Fig.5 Binarization of table region

图 5 二值化的表格区域

Fig.6 Horizontal projection

图 6 水平投影

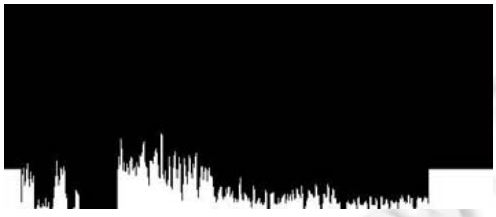

Fig.7 Vertical projection

图 7 竖直投影

Fig.8 Segmentation results of table region

图 8 表格区域的分割结果

对于生成的二值图不为黑底白字的情况,我们采用部分反向的方法来判断,如果投影的值等于图片的边长,那么这个二值图可能和我们想要的相反,则需要采取把所有的满投影的值设置为 0,然后再处理图像边缘并利用阈值切割的方式来解决,这样,即使图片为白底黑字,依然可以切割到正确的位置.对于更加复杂的情况,如图 9 所示,表格中既含有黑底白字也含有白底黑字,图片的投影不满足横向切割的阈值,我们发现这样的图片往往会

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.

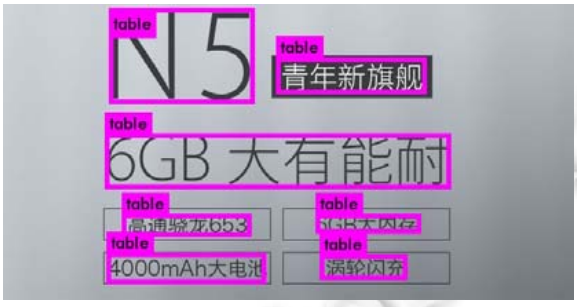

Fig.11 Results of text block detection

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.