计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(10):248−254 [doi: 10.15888/j.cnki.csa.007522]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

文本检测与识别在细粒度图片分类中的应用①

姜　倩, 刘　曼

(北京工业大学 信息学部, 北京 100124)
通讯作者: 姜　倩, E-mail: jiangqian_0411@163.com

摘　要: 细粒度的图片分类是深度学习图片分类领域中的一个重要分支, 其分类任务比一般的图片分类要困难, 因为很多不同分类图片中的特征相似度极高, 没有特别鲜明的特征用以区分, 因而需要优化一个传统的图片分类方法. 在一般的图片分类中, 通常通过提取视觉以及像素级别的特征用来训练, 然而直接应用到细粒度分类上并不太适配, 效果仍有待提高, 可考虑利用非像素级别的特征来加以区分. 因此, 我们提出联合文本信息和视觉信息作用于图片分类中, 充分利用图片上的特征, 将文本检测与识别算法和通用的图片分类方法结合, 应用于细粒度图片分类中, 在 Con-text 数据集上的实验结果表明我们提出的算法得到的准确率有显著的提升.

关键词: 文本检测; 文本识别; 光学字符识别; 图片分类; 细粒度图片分类

引用格式: 姜倩,刘曼.文本检测与识别在细粒度图片分类中的应用.计算机系统应用,2020,29(10):248–254. http://www.c-s-a.org.cn/1003-3254/7522.html

Application of Text Detection and Recognition in Fine-Grained Image Classification

JIANG Qian, LIU Man

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Fine-grained image classification is an important branch in the field of deep learning image classification. Since many different classified images are very similar in their features, and there is no particularly distinctive feature can be used to distinguish among them, it makes the classification task of fine-grained image more difficult than that of the general image. Therefore, a traditional image classification method needs to be optimized. Usually, visual and pixel-level features extraction is used in the training of the general image classification. However, direct application of this method to the fine-grained classification is not very suitable, and the effect still needs to be improved, while non-pixel-level features can be used to distinguish. Hence, we propose to combine text and visual information in the image classification, make full use of the features on the images, combine the text detection and recognition algorithms with general image classification methods, and apply it to the fine-grained image classification. In Con-text dataset, the experimental result shows that the accuracy obtained by the proposed algorithm has been significantly improved.

Key words: text detect; text recognition; OCR; image classification; fine-grained image classification

1　概述

随着互联网技术发展的越来越成熟, 人们进行交流和传递信息变得更加方便快捷, 可使用的方式也变得多样化, 更多的人使用图片或者视频来传递信息. 而文字作为人们交流对话的媒介, 是图片和视频中信息的主要表达形式, 所以文本识别的重要性不言而喻. 目前, 文本识别已广泛应用到地图搜索, 运单识别, 证件识别等各种应用中, 文本的智能化识别在带来极大的便利的同时也极大的提高了工作的效率. 在图片分类领域中, 细粒度分类作为图片分类任务中的一个极其

① 收稿时间: 2019-12-17; 修改时间: 2020-01-14; 采用时间: 2020-02-11; csa 在线出版时间: 2020-09-30

重要的分支, 虽然图片分类技术日趋成熟, 利用深度学习卷积神经网络的技术在 ImageNet[1] 比赛中图片分类准确率可以达到 99%. 但是在细粒度的图像分类中, 由于不同种类的特征比较相似, 常用的特征提取方法得到的准确率还没有达到最优, 仍有需求来打破瓶颈, 使得其准确率能够比肩成熟的图片分类. 因此, 本论文中研究了文本识别在细粒度分类中的应用, 我们将文本检测与识别的算法应用于 Con-text[2] 数据集. Con-text 是一个建筑物类图片的数据集, 包含咖啡店, 洗衣店, 餐厅等各类建筑物. 从外观上观察这些建筑物并无太大差别, 常用的提取图片特征方法不能进行有效的区分, 但建筑物外观上的文字却能够很好的表征特点, 通过分析文字可以得到该建筑物的类别, 所以考虑将外观上的文字信息作为特征的一部分. 本文提出对该数据集进行文本识别, 将自然场景下的文本识别技术应用到图片分类中, 有效的联合文本信息和视觉信息, 在很大程度上提高图片识别的准确率.

在本文中, 我们结合图片分类和文本识别技术来完成图像的细分类任务. 使用卷积神经网络对非文本图片进行图片分类, 同时应用改进后的 EAST[3] 检测算法对有文本图片进行处理, 得到文本的位置后使用 CRNN[4] 结合 CTC[5] 的方法进行文本的识别, 再将识别到的文字进行分析后分类得到对应的建筑物类别, 在一定程度上提升了分类的准确率. 如图 1 是本文的算法流程图.

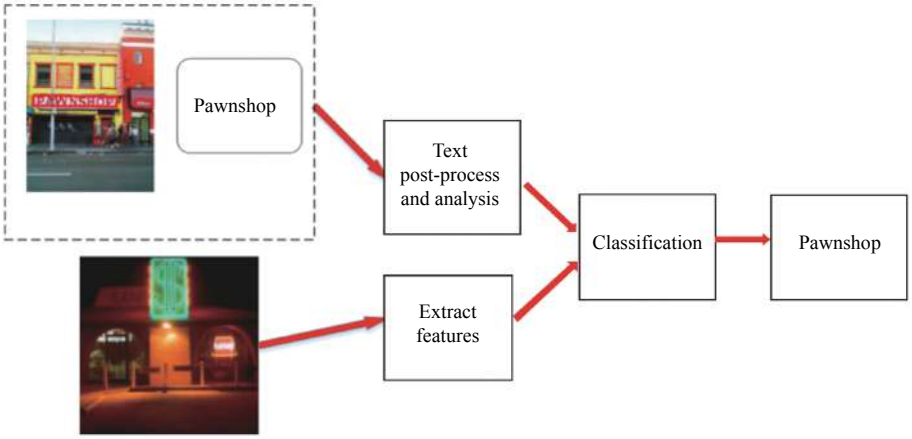

图 1 算法流程图

本文后续的内容结构如下: 第 2 节介绍文本检测与识别和图片分类中常用的算法; 第 3 节介绍本文结合文本识别和图片分类技术对 Con-text 数据集进行细分类的算法; 第 4 节进行实验结果的比较, 展示文本识别在图片细分类应用的效果.

2 相关研究

本文研究中, 文本识别在图片细分类中的应用包含文本检测技术, 文本识别技术, 图片分类技术, 以及文本分类技术.

2.1 基于深度学习的文本检测方法

在深度学习领域中, 常用于文本检测的方法一般分为 3 种: 基于边界框回归的方法, 基于图像分割的方法以及结合回归和分割的方法. 在边界框回归方法中, 核心思想是将文本当成目标进行目标检测的过程, 和常见的目标检测方法一样, 分为 Two-stage 和 One-stage 两种方法. Two-stage 方法有 R2CNN[6], CTPN[7], RRD[8], IncepText[9], LOMO[10] 等. One-stage 方法有: Seglink[11], Textboxes[12], Textboxes++[13], DMPNet[14], EAST 等. 包含 PSENet[15] 以及 CRATD[16] 等基于分割方法的核心思想是将文本和背景切割开来, 而回归和分割组合方法的核心思想类似于 Mask RCNN[17]. 以上 3 种类型的算法都各有其特点和优势, 但为了权衡各方面的性能, 基于边界框回归的方法是常用的方法.

2.2 基于深度学习的文本识别方法

在文本检测完成后, 根据预测得到的文本位置将文本区域提取出来识别. 文本识别可分为单字符识别和行识别. 在单字识别中, 切割文本行得到单个字符送到卷积神经网络训练的单字符分类器中进行预测, 连接单字符可得到目标区域识别的结果. 在行识别中, 文本检测得到的文本框从图片中截取出来后, 利用卷积循环神经网络 CRNN 训练得到一个基于文本行的预测模型. 在行文本识别的训练过程中, 有两种常用的方法, CRNN 结合 CTC 的方法以及 CRNN 结合 Attention[18]

的方法. 由于在单字符切割中有可能会出现字符粘连以及字符被切断的情况, 后续会直接影响字符识别的效果, 所以我们使用行文本识别的方法.

2.3 基于深度学习的图片分类方法

图片分类是深度学习计算机视觉领域中一个常见的任务. 从 2010~2017 年间出现大量基于深度学习卷积神经网络的算法来处理大规模的图片分类. 从最先出现的 Lenet[19] 到 Alexnet[20], GoogleNet[21], VGGNet[22], 以及 ResNet[23] 在 ImageNet 比赛中获得冠军, 图片分类方法的发展在近几年发展的相当迅速, 越来越多的人投身到深度学习方向的研究上来. 目前的深度学习模型的识别能力已经超过了人眼, 图像分类中使用的算法带来的效果已经满足了预先的期望, 但实际应用中面临着比大赛中更加复杂和现实的问题, 在细粒度分类问题中, 还未超越人类, 仍有很大的发展空间.

2.4 常见的细粒度图片分类方法和应用

细粒度图片分类在图片分类中是一个重要的研究方向, 是在区分出基本类别的基础上, 对基本类别划分得到更加精细的子类, 是处理得到一个更精确分类的任务, 如区分花的品种, 鸟的种类、狗的品种和车的款式等, 其业务需求和应用场景在工业界和实际生活中分布广泛. 现在通常使用的细粒度分类方法分为 4 种, 基于常规图像分类网络的微调方法, 基于细粒度特征学习的方法, 基于目标块的检测和对齐的方法以及基于视觉注意力机制的方法.

3 结合文本识别与图片分类的细粒度图片分类算法

在本文的研究中, 我们使用的细粒度图片分类方法中融合了文本检测和识别与图片分类的方法. 具体流程可参考图 1, 在流程图中可以看出, 研究中对有文本图像的图片进行文本检测得到包含文本区域的图片, 利用识别算法进行图像文本的行文本识别, 后处理识别得到的结果进行文字分析并分类, 同时对没有文字的图片进行图片分类, 经过以上的识别流程后图片的分类正确率有大幅提高. 在我们的研究中, 文本检测算法中改进了 EAST 方法, 使其检测结果更加准确, 在文本识别中改进 CRNN 结合 CTC 的方法, 和单字识别相比有更好的识别效果, 并设计文本分类的逻辑来优化分类结果, 同时利用 ResNet 进行非文本图片的分类, 最后叠加两个结果得到最终的正确率.

3.1 文本检测

在本文中采用优化 EAST 的方法来进行文本检测. EAST 将文本检测转换成一个目标检测的任务, 能够实现对自然场景下倾斜文本的检测, 可以对单词级别, 行级别以及任意形状的四边形文本进行检测. 在 EAST 中, 使用全卷积网络 (FCN[24]) 能够直接回归文本位置, 得到文本框的位置以及其角度后, 利用基于 NMS[25] 改进的 Locality-Aware NMS 设置合适的阈值对候选区域进行筛选, 过滤掉 score 较低和重复的文本框, 保留下来的就是经过 EAST 检测器得到的预测文本框. EAST 因为能够直接回归文本框, 所以速度相对较快, 而且准确率也有提高, 可以又快又好的检测文本.

如图 2 是 EAST 的网络结构, 从图中可以看出我们替换 PVANet[26] 为 ResNet, 使用 ResNet 进行特征提取. 在卷积部分, 经过 4 层卷积后可以得到不同尺度的特征图, 这些多尺度的特征图对实际场景中文本行的精准定位变得更鲁棒. 其中 ealy stage 用来检测小的文本行, late stage 用来检测大的文本行. 第二部分是特征融合层, 使用 U-net[27] 的方法来进行特征融合, 该部分的每一个层都进行上采样操作, 将上采样得到的特征和特征提取层中卷积后与之得到的相同尺寸特征进行融合, 通过此操作可以得到更多特征的信息. 最后是网络输出层, 输出文本得分 score 和预测框 RBOX 的信息.

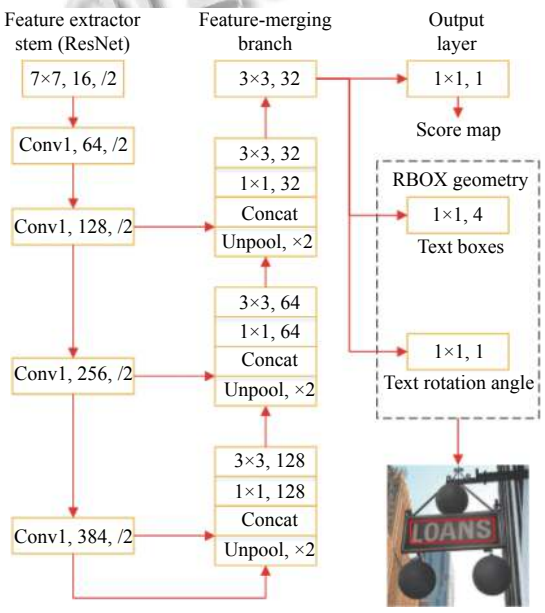

图 2 EAST 结构图

但是由于 EAST 在制作 polygon 的时候采用了 shrink_poly 的思想用于缓解标注带来的误差, 制作 mask 时对边缘有 shrink 的操作, 在一定程度上避免了不可预测的误差, 采用的方法如下:

$$l'_w = 0.7 * l_w \tag{1}$$

其中, l_w表示 mask 宽的长度, l'_w表示 mask 缩放后宽的长度.

但随之也会带来边界框预测不准的情况, 所以我们针对边缘的处理进行优化. 短边我们保持原来 0.3 shrink 的比例, 长边保持 0.1 shrink 的比例, 方法如下:

$$l'_w = 0.9 * l_w \tag{2}$$

通过此项改进后长边边缘字符被截断的情况有所改善. 并且我们在训练集中加入任意角度的数据, 加大对角度的学习, 让检测模型更加鲁棒, 同时使得在提取特征时的效果更好. 经过以上的优化后, EAST 的检测效果相比之前有大幅提高. 从表 1 中可以看出以上本文基于 EAST 作出的两个部分改进给文本检测的效果带来了明显的提升.

表 1 EAST 方法效果对比

方法	*F-score* (%)
EAST - baseline	78.20
ours	80.73

3.2 文本识别

在本文中使用 CRNN 结合 CTC 的方法来进行文本识别. 在该结构中, 先使用卷积神经网络 CNN 来提取图片的特征序列, 然后使用 RNN 对序列进行预测, 最后利用 CTC 转录层, 将预测变为最终的标签序列. 需要注意的是, 在将图片输入进模型之前, 需要将图片缩放到统一的高度. 在 CRNN 模型中, 一般采用标准的 CNN 网络模型中的卷积层和最大池化层来构造卷积层网络结构, 用于从图像中提取可以表征该类特征的序列, 这些特征序列作为循环层的输入. 在 CRNN 模型中使用深度双向循环神经网络 LSTM[28], 该循环网络与卷积层连接, 能够得到不同的序列特征以及单个字符的序列信息, 且使用的双向的 LSTM 能够得到前后的的上下文信息, 可以实现对任意长度的序列进行预测. 最后的 CTC 转录层用来接收循环层的输出, 即根据每帧预测找到具有最高概率的标签序列, 进而将标签信息映射成字符信息. 通过分析实际应用场景来训练数据, 本文训练了一个针对英文分类的 CRNN 模型, 具体结构如图 3 所示.

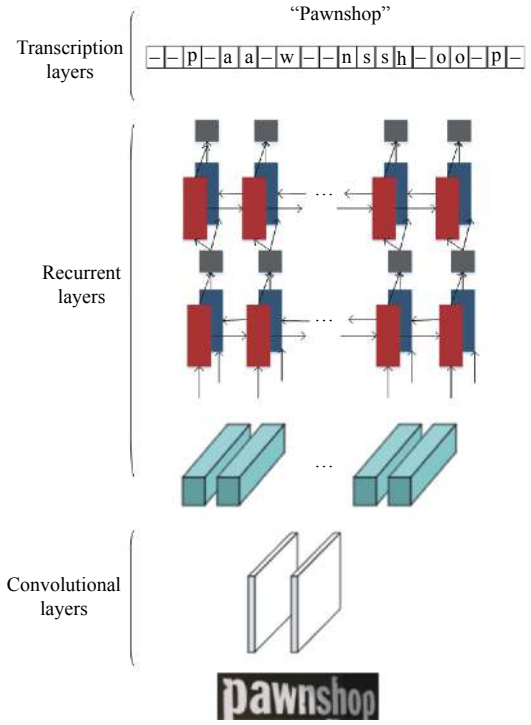

图 3 CRNN 结构图

3.3 图片分类

本文中使用 ResNet 为 baseline 进行非文本图片的分类, ResNet 网络的一个最大的特点就是引入了残差块, 通过残差网络, 我们构建很深的网络出现过拟合的情况大大减少, 而且其分类精度随之提升. ResNet 中的残差块是与其他网络结构最不相同的部分, 其学习到的是目标值和输入值的差值, 即残差. 这种跳跃式的结构, 打破了网络结构的局限性, 不再是传统的神经网络结构中前一层的输出作为后一层的输入, 而是使得网络结构中某一层的输出可以直接跨过连续的几层作为后面某一层的输入, 其意义在于使用多层网络而使得整个学习模型的错误率不降反升的难题提供了新的方向. 至此, 神经网络的层数可以不再受限于传统网络带来的约束, 除掉了局限性, 其网络层数可以达到几十层、上百层甚至千层, 且出现过拟合的情况大幅降低, 一方面提高了精度另一方面为高级语义特征提取和分类提供了可行性.

3.4 文本分类

通过文本识别后得到的文本需要进行处理后才能进行分类. 具体步骤如下:

1) 分析和理解数据. 分类之前要对不同建筑分类中出现的单词进行统计, 找到能够代表该类建筑物的

关键性词语, 即总结出每一类的主要关键词.

2) 改善识别后词语的分类逻辑. 除了完全匹配外, 根据实验结果分析得到, 认定只要识别得到的字符按顺序匹配, 能达到关键字的 50% 就判定关键词对应的分类即为该词的分类.

3) 若一张图中有多处文字从而得到多个分类结果, 取出现次数最多的分类, 若出现的次数相同, 取匹配占比最高的关键字对应的分类.

4 实验结果和分析

4.1 数据集

为验证算法的有效性, 我们使用了 Con-text 数据集, 该数据集包含 28 类街边常见建筑物, 共 24 255 张图片, 其中训练集 19 404 张, 测试集 4851 张. 数据集上有文字信息能够很好的表征建筑物的分类, 例如常见的“干洗店”, “咖啡店”, “餐馆”, “折扣店”等这些建筑物上面都会有明显的文字信息来区分. 这 28 类分别是: bakery, barbershop, bistro, bookstore, cafe, theatre, dry cleaner, computer store, country store, diner, discount house, pharmacy, funeral, hotspot, massage parlor, medical center, repair shop, motel, pawnshop, pet shop, pizzeria, tavern, repair shop, restaurant, school, steakhouse, teahouse 和 tobacco shop. 这些数据均为自然场景下拍摄的图片, 因为街边建筑物必须要醒目, 所以建筑物上面的文字字体较大且间隔也大, 这给文字检测带来了一定的难度. 同时, 这些文字带来的文字并非所有的都是有效信息, 所有这也给文本分类带来了一定的难度.

4.2 参数设置

在文本检测训练中, 使用 EAST 为 baseline, 使用随机梯度下降训练, 其中动量和权值衰减系数分别设置为 0.9 和 5×10^{-4}, 最大迭代次数为 10 万次, 学习率初始设置为 10^{-3}. 该实验在 tensorflow 中训练完成, 训练和测试图像的尺寸都为 512×512.

在文本识别的训练中, 使用 CRNN 结合 CTC 的方法, 利用 RMSProp 优化随机梯度下降训练, 其中动量和权值衰减系数分别设置为 0.9 和 5×10^{-4}. 最大迭代次数为 100 次, 学习率初始设置为 10^{-2}. 该实验在 Pytorch 中训练完成, 训练图像的尺寸都为 256×32.

在图像分类的训练中, 使用 RMSProp 优化随机梯度下降训练, 其中动量和权值衰减系数分别设置为 0.9 和 5×10^{-4}, 最大迭代次数为 10 万次, 学习率 (learning rate) 初始设置为 10^{-2}. 该实验在 Pytorch 中训练完成, 训练和测试图像的尺寸都为 224×224.

表 2 各个分类的 *AP*

方法	bak.	bar.	bis.	b.s	cafe	thea.	d.c	com.	cou.	din.	disc.	pha.	fun.	h.p
Visual Baseline	79.5	87.2	18.4	88.5	14.9	80.6	37.5	64.0	51.5	71.3	56.5	73.6	72.8	68.8
ours	85.5	88.8	52.2	89.5	78.3	86.3	79.6	76.8	74.9	77.1	72.2	82.4	79.3	77.9
方法	mas.	med.	r.s	mot.	pawn.	pet.	piz.	tav.	rep.	res.	sch.	steak.	tea.	toba.
Visual Baseline	75.5	66.7	85.5	68.7	67.9	29.1	27.1	42.9	47.1	55.6	75.0	46.8	11.7	66.7
ours	82.8	81.4	87.2	73.9	78.8	77.6	71.1	78.5	69.3	78.1	85.8	74.5	51.2	79.8

4.3 性能指标

在常见的评价指标中, 一般用 3 个评价指标, 分别为 *P*(precision, 准确率), *R*(recall, 召回率) 和 *mAP*. 其中 *mAP* 中 *AP* 表示任意一个种类的平均值, *mAP* 为所有类的平均值. 如式 (3), 式 (4), 式 (5) 分别表示了 *P*, *R* 以及的 *AP* 的计算方式.

$$P(k) = TP/(TP+FP) \tag{3}$$

$$R(k) = TP/(TP+FN) \tag{4}$$

$$AP = \sum_{i=1}^{n} P(i) * (R(i) - R(i-1)) \tag{5}$$

其中, $P(i)$ 和 $R(i)$ 表示在当前数据中的指定类的 *P* 和 *R*, *n* 表示数据集中图片的数量.

4.4 实验结果分析

针对 Con-text 数据集的测试, 我们的结果与 visual result 以及[2] 进行对比, 看表 2 可看出本文算法在在各个分类中的 *mAP*, 看表 3 可以得到单纯的图片分类算法结果以及结合文本检测与识别的联合算法得到的结果. 从表 2 中可以看出, 文字信息较少分类的 *mAP* 会比其他分类低, 例如 tea house, bistro 等. 从表 3 中可以看出我们的方法和文献 [2] 相比有明显的提高, 说明我们改进的方法有成果. 但是从总体上来看, 结合文本识别

后的联合算法比通常的图片分类算法的 *mAP* 高, 能够将结果融合到更高的精度. 这表示文本信息在分类中起到了重要的作用, 在图片的细分类任务中起到了强辅助作用.

表 3 整体分类的 *mAP*

方法	*mAP* (%)
Visual result	60.67
Karaoglu et al.[2]	71.0
ours	77.53

虽然从结果上来看, 准确率有了一定的提升, 但是仍然还有上升空间. 在文本检测和识别中, 我们采用的是 two-stage 方法, 识别强依赖于检测结果, 未来可采用 one-stage 端到端的方法尽可能的规避中间误差带来的影响,或许能在一定程度上提升 *mAP*.

5 结论

在此研究中, 我们研究了文本检测和识别的相关方法, 并将其应用到了在图片细分类中, 将 OCR 应用到了图片分类中, 提高了图片分类的准确率, 但是准确率仍旧不是很高, 还有很大的提升空间. 相信在未来会有更好的方法将文本检测与识别和图片分类算法结合来细分类图片.

参考文献

1 Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–255.

2 Karaoglu S, van Gemert JC, Gevers T. Con-text: Text detection using background connectivity for fine-grained object classification. Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain. 2013. 757–760.

3 Zhou XY, Yao C, Wen H, *et al*. EAST: An efficient and accurate scene text detector. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2642–2651.

4 Shi BG, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298–2304. [doi: 10.1109/TPAMI.2016.2646371]

5 Graves A, Fernández S, Gomez F, *et al*. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA, USA. 2006. 369–376.

6 Jiang YY, Zhu XY, Wang XB, *et al*. R2CNN: Rotational region CNN for orientation robust scene text detection. arXiv: 1706.09579, 2017.

7 Tian Z, Huang WL, He T, *et al*. Detecting text in natural image with connectionist text proposal network. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 56–72.

8 Liao MH, Zhu Z, Shi BG, *et al*. Rotation-sensitive regression for oriented scene text detection. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 5909–5918.

9 Yang QP, Cheng ML, Zhou WM, *et al*. Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden. 2018. 1071–1077.

10 Zhang CQ, Liang BR, Huang ZM, *et al*. Look more than once: An accurate detector for text of arbitrary shapes. arXiv: 1904.06535, 2019.

11 Shi BG, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3482–3490.

12 Liao MH, Shi BG, Bai X, *et al*. Textboxes: A fast text detector with a single deep neural network. Proceedings of the 31 AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4161–4167.

13 Liao MH, Shi BG, Bai X. TextBoxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing, 2018, 27(8): 3676–3690. [doi: 10.1109/TIP.2018.2825107]

14 Liu YL, Jin LW. Deep matching prior network: Toward tighter multi-oriented text detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 3454–3461.

15 Li X, Wang WH, Hou WB, *et al*. Shape robust text detection with progressive scale expansion network. arXiv: 1806.02559, 2018.

16 Baek Y, Lee B, Han D, *et al*. Character region awareness for text detection. arXiv: 1904.01941, 2019.

17 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on

Computer Vision. Venice, Italy. 2017. 2980–2988.
18 Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv: 1508.04025, 2015.
19 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]
20 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25. Lake Tahoe, NV, USA. 2012. 1106–1114.
21 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.
22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
24 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3431–3440.
25 Neubeck A, Van Gool L. Efficient non-maximum suppression. Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China. 2006. 850–855.
26 Hong S, Roh B, Kim KH, *et al*. PVANet: Lightweight deep neural networks for real-time object detection. arXiv: 1611.08588, 2016.
27 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany. 2015. 234–241.
28 Shi XJ, Chen ZR, Wang H, *et al*. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 802–810.