计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):179−185 [doi: 10.15888/j.cnki.csa.008038]

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

基于深度学习的场景文本检测与识别①

宫法明[1], 刘芳华[1], 李厥瑾[2], 宫文娟[1]

[1](中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
[2](山东电子职业技术学院 教务处, 济南 250200)
通讯作者: 刘芳华, E-mail: lfh19951219@163.com

摘　要: 针对复杂场景下文本识别流程复杂繁琐、适应性差、准确度低等缺点, 本文提出一种复杂场景下文本检测和识别的新方法. 该方法由文本区域检测网络及文本识别网络构成, 文本区域检测网络为改进的 PSENet, 将 PSENet 的骨干网络改为 ResNeXt-101, 在特征提取过程中加入可微二值化操作来优化分割网络, 不仅简化了后处理, 而且提高了文本检测的性能. 将卷积神经网络和加入聚合交叉熵损失的长短时记忆网络组成文本识别网络, 聚合交叉熵的引入提高了文本识别的准确性. 本文在两个数据集上进行验证, 实验结果表明, 两个网络模型融合后准确率最高达到 95.6%, 优于改进之前的方法. 该方法能有效地检测和识别任意文本实例, 具有很好的实用性.

关键词: 可微二值化; 聚合交叉熵; 文本检测; 文本识别

引用格式: 宫法明,刘芳华,李厥瑾,宫文娟.基于深度学习的场景文本检测与识别.计算机系统应用,2021,30(8):179–185. http://www.c-s-a.org.cn/1003-3254/8038.html

Scene Text Detection and Recognition Based on Deep Learning

GONG Fa-Ming[1], LIU Fang-Hua[1], LI Jue-Jin[2], GONG Wen-Juan[1]

[1](College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)
[2](Academic Affairs Office, Shandong College of Electronic Technology, Jinan 250200, China)

Abstract: This study proposes a new method for text detection and recognition in complex scenes to eliminate the shortcomings of a complicated text recognition process, poor adaptability, and low accuracy. This method is composed of a text area detection network and a text recognition network. The text area detection network is an improved PSENet. The backbone network of PSENet is changed to ResNeXt-101, and a differentiable binarization operation is added to optimize the segmentation network in the feature extraction process, which not only simplifies post-processing but also improves text detection. The text recognition network is formed by combining a convolutional neural network with a long short-term memory network with aggregate cross-entropy loss. The introduction of aggregate cross-entropy improves the accuracy of text recognition. Furthermore, experimental verification is carried out on two data sets, and the results show that the new method has accuracy as high as 95.6%, which is better than the previous methods. This method can effectively detect and recognize any text instances and has good practicability.

Key words: differentiable binarization; aggregate cross-entropy; text detection; text recognition

近年来, 在场景图像中浏览文本, 因为其广泛的实际应用, 如图像/视频理解、视觉搜索、自动驾驶、盲辅助等, 成为一个活跃的研究领域. 场景文本检测作为场景文本读取的关键组成部分, 对每个文本实例的边界框或区域进行定位仍然是一项具有挑战性的任务, 因为场景文本往往具有各种尺度和形状, 包含水平文

① 基金项目: 科技部创新方法工作专项 (2015IM010300)
Foundation item: Special Project for Innovation Method of Ministry of Science and Technology of China (2015IM010300)
收稿时间: 2020-11-19; 修改时间: 2020-12-21; 采用时间: 2021-01-07; csa 在线出版时间: 2021-07-31

本、多取向文本和弯曲文本. 基于分割的场景文本检测由于其在像素级上的预测结果, 可以描述各种形状的文本, 因此近年来受到了广泛的关注. 然而, 大多数基于分割的方法需要复杂的后处理, 在推理过程中造成了相当大的时间开销.

针对文本识别问题, 传统的文本识别方法[1–3] 适应性差、需要分离训练目标, 导致麻烦的预分割和后处理阶段. 在计算机行业飞速发展的今天, 自动处理算法逐渐成熟, 文本检测和识别算法[4–7] 的准确度都大大提升. 近年来出现的 CTC[8](Connectionist Temporal Classification) 和注意力显著缓解了这种训练问题, 但这两种识别模型算法实现很复杂, 可能会导致训练成本增高并降低了识别准确率.

本文的贡献在于提出了一种复杂场景下文本检测和识别的新方法, 记为 TDRNet (Text Detection and Recognition Net). 在原本检测和识别网络的基础上加以改进, 采用更高效的特征提取网络, 在文本区域检测网络中加入可微二值化进行优化, 大大简化了后处理过程. 在文本识别网络中, 使用聚合交叉熵损失函数解决序列了识别问题, 对 CTC 和注意力机制具有竞争性能, 提高了检测和识别性能. 本文将文本定位网络和文本识别网络结合提高识别准确率, 取得了较好的性能. 该方法能有效地检测和识别任意文本实例, 具有很好的实用性.

1 相关工作

文本识别通常包含 3 个部分: 首先进行图像预处理, 紧接着进行文本检测, 最后进行文本识别. 为了使图像被检测或扫描, 通常需要对输入图像进行捕获、二值化、平滑等处理, 对输入图像进行校正, 并根据文本大小对图像进行裁剪. 编辑图像后, 我们可以对文本进行检测了.

近年来, 文本检测技术的研究取得了长足的进展. 传统的特征提取方法大多采用人工, 在深入研究计算机视觉任务之后, 文本检测逐渐转向基于深度的学习方法. 目前基于深度学习方法包含两大类: 一种是从目标探测发展而来的, 例如基于候选字段的文本检测, 其基本构想是基于默认框架集创建一系列候选文本框, 再进行一系列调整、筛选, 最终通过非极大抑制 NMS (Non-Maximum Suppression) 得到最终的文本边界框, 例如为文本检测而设计 TextBoxes[9]、SegLink[10] 等网络结构模型; 一个是从语义分割发展而来的, 例如基于图像分割的文本检测. 其想法是分割网络结构, 达到像素的语义分区, 然后根据分割的结果构建一个文本行. 例如 PixelLink[11] 和 FTSN[12], 会 生成分段映射, 然后在接下来的编辑之后, 最终得到文本限制字段. 这种方法可以准确定位文本位置, 提高自然场景图像中文本检测的准确性, 但是他们的后处理算法导致了思维速度的下降. 文本识别又分为两种识别方法, 包括单字符识别和行识别. 以往的文本识别是采用 K 近邻的方法识别单字符, 在实时度要求高的系统中不适合这种计算量很大的方法. 通过广泛应用深层学习方法, 出现了许多基于深入学习的优秀识别模型, 大大提高了单字符识别的精度. 现在主要使用文本行识别. 有两个主要的方法是为了识别文字, 在最近的两年里取得了更好的结果, 分别是: CRNN OCR (Convolutional Recurrent Neural Network Optical Character Recognition) 和 Attention[13] OCR. 这两种方法在其特征学习阶段都采用了 CNN+RNN 的网络结构, CRNN 在对齐时采用了 CTC 算法, 而 Attention OCR 采用了注意力机制. 但是, 这些方法会导致很多计算和内存消耗. 因此, 解决后处理的繁琐问题的方法成为了紧急问题.

2 文本检测识别框架

本章详细介绍了基于深度学习的场景文本识别方法. 总体设计思路是先对整个图像进行分割, 然后通过阈值跟踪分割结果, 得到处理后文本区域的位置. 利用位置信息对文本区域进行切割, 并将裁剪后的文本区域发送到文本识别网络中进行识别以得到结果. 整个方法由两部分组成: 文本区域检测器 TLDNet (Text Location Detection Net) 和文本区域识别网络 TRNet (Text Recognition Net).

2.1 文本区域定位网络

识别的准确性取决于定位的准确性, 所以确保文本区域定位的准确性尤为重要. 为了确保文本区域定位网络的准确性, 本文在 PSENet[14] 的基础上对其进行改进: (1) 采用更高效的特征提取网络来保证分类的准确性. (2) 插入一个可微二值化[15] 操作放到分割网络里来一起优化, 更能区分前景和背景. 通过该两方面的改进, 确保了文本定位的准确性.

TLDNet 中的骨干网络采用了类似于 FPN[16] 和 U-Net 的思路, 因为 ResneXt 相比 ResNet 网络结构更简

单, 可以防止对于特定数据集的过拟合, 而且更简 单的网络意味着在用于自己的任务的时候, 自定义和修改起来更简单, 需要手动调节的参数少. 与 ResNet 相比, 相同的参数个数, ResneXt 结果更好, 且计算量少. 因此将该网络 ResNet[17] 换成 ResneXt-101[18], 然后将不同尺度的特征图进行融合来让最终进行回归的特征图获得不同尺度的特征信息和感受野以处理不同尺寸大小的文字实例. 接下来由骨干网络输出的特征图得到一个分隔图和一个阈值图, 二者由可微分的二值化而得到二值化图, 最后经过一系列后处理得到文本区域. 网络结构如图 1 所示.

ResneXt-101 中每个卷积组由卷积层, 池化层, 激活层构成, 该网络含有 5 个卷积组, 卷积组 2-5 借鉴了 GoogLeNet[19] 的卷积范式 split-transform-merge 思想, 在大卷积核两层加入 1×1 的卷积, 控制核个数的同时减少参数个数. 相比如 Resnet 结构简明, 大大降低了参数, 计算量小, 提高了速度和精度.

输入的图像经过不同层的采样之后获得不同大小的特征图, 之后由这些特征图构建特征金字塔, 从而构建出统一尺度的特征图 F. 这个特征图用于预测分割概率图 P 与阈值图 T, 之后将 P,T 结合得到估计的二值图 $\hat{B}$. 在训练的时候 P,B 是使用同样的表现作训练, 而 T 会使用单独的阈值图作训练. 对于分割特征图$P \in R^{H\times W}$, 使用下面的方式进行二值化处理:

$$B_{i,j} = \begin{cases} 1, & \text{if } P_{i,j} \geq t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

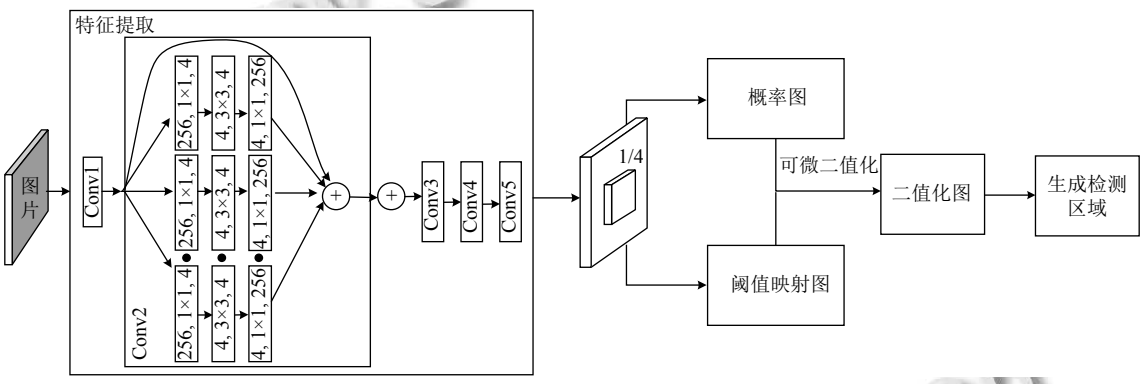

图 1 字符区域定位网络结构图

在以往基于分割的文字检测方法中, 大部分都会采用预设的阈值进行二值化用来处理网络输出的分割图, 不能随着训练过程再分割网络进行优化. 所以实验中引入了可微分的二值化函数, 将二值化过程嵌入到网络中实现了优化. F 是生成的近似二值图, T 是生成的阈值特征图, k 是放大倍数. 通过这样的方式可以有效地将文本区域与背景区域分离, 还可以减少文本之间重叠的情况.

$$\hat{B}(x) = \frac{1}{1+\mathrm{e}^{-kx}}, x = \left(P_{i,j} - T_{i,j}\right) \tag{2}$$

本文中在阈值图上应用了边界的监督并将阈值映射作为二值化的阈值. 利用对概率图映射按固定的阈值进行二值化, 得到二值映射, 进而由二值映射缩小文本区域, 最后利用偏移裁剪算法对缩小后的区域进行扩张得到最终的文本位置.

2.2 文本识别网络

本文的文本识别网络借鉴了文献 [20] 的方法, 可将网络视为编解码器结构, 编码器由特征提取网络 DenseNet[21] 和双向长短时记忆网络 (BiLSTM)[22] 构成; 引入聚合交叉熵损失 (ACE)[23] 的长短时记忆网络 (LSTM) 组成解码器. 网络结构如图 2 所示. ResNet 是每个层与前面的某层短路连接在一起, 连接方式是通过元素级相加. 而在 DenseNet 中, 每个层都会与前面所有层都相连, 即每层的输入, 在前面的所有层的输出都相连. 相比 ResNet, 这是一种密集连接. DenseNet 是直接连接来自不同层的特征图, 这可以实现特征重用, 提升效率.

双向长短记忆网络有两个 LSTM 组成, 能够同时利用过去时刻和未来时刻的信息, 本文将两个 LSTM 组成的方式由连接改为结合, 提高识别的准确率. 解

码器由加入聚合交叉熵损失 (ACE) 的长短时记忆网络构成. 长短时记忆网络 (LSTM) 的长期存储功能是有限的. 如果序列信息特别长, 经过多层之后, 初始信息就会丢失. 可以通过引入注意力机制重新计算得到当前时刻的特征, 但需要复杂的注意力来帮助注意力机制实现其功能, 进而产生额外的参数和时间, 特别是对于较长的输入序列, 缺失或多余的字符很容易导致错位问题, 混淆和误导训练过程, 从而降低识别准确度. 聚合交叉熵损失可以沿时间维度聚合每一个列别的概率, 并将累积的结果和标签标准化为所有类别的概率分布, 最后使用交叉熵来比较这两个概率的分布, 从而降低识别准确度. 本文将聚合交叉熵损失代替注意力机制, 只需要计算各类别字符出现次数, 不用考虑特征的顺序, 识别速度更快. 通过要求网络精确预测标注中每个类的字符数来最小化一般损失函数, 计算公式如下:

$$P(N_k|k,I;w) \tag{3}$$

其中, $|C|$表示类别数, $P(N_k|k,I;w)$表示在图像 I 的预测结果中, 第 k 个类别的字符出现的次数等于标签中给定次数 N^k的条件概率.

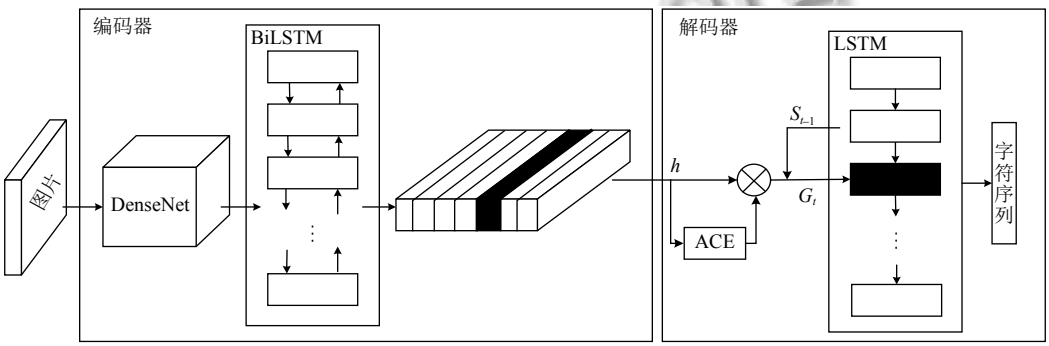

图 2 字符识别网络结构图

本文通过 CNN+BiLSTM 得到的特征序列维度为 $(T\times K)$, 其中 T 为序列长度, K 为字符类别数, 本文定义输出的特征序列张量为 Y, 第 t 个时刻的特征向量为 y^t, 第 t 个时刻第 k 个类别的预测概率为y_k^t. 整个字符序列中所有位置第 k 个类别出现的总概率为$\sum_{t=1}^{T} y_k^t$.

本文从回归问题的角度调整损失函数, 计算公式如下:

$$L(w)=\frac{1}{2}\sum_{(I,S)\in Q}\sum_{k=1}^{|c^{\in}|}\left(N^k-y^k\right)^2 \tag{4}$$

式中, T 表示预测文本长度, $|S|$表示标签文本长度, 本文用$(T-|S|)$表示字符串中空白字符的个数, 即$N_{\in}=T-|S|$.

为了防止梯度消失问题, 本文把第 k 个字符y_k的累计概率标准化为$\bar{y}_k=y_k/T$, 把字符数量N^k标准化为$\bar{N}_k=\bar{N}_k/T$. 然后, 在$\bar{y}$和$\bar{N}$之间的交叉熵可以表示为:

$$L(I,S)=-\sum_{k=1}^{|C^{\in}|}\bar{N}_k\ln\bar{y}_k \tag{5}$$

LSTM 在最后输出的概率矩阵中利用贪心搜索的方法获得最终的文本字符序列.

3 实验分析

3.1 实验准备

本文使用了 MSRA_TD500、TotalText 和 CTW1500 三个数据集进行实验. MSRA_TD500 是一个包含英语和汉语的多语言数据集. CTW1500 是一个聚焦于弯曲文本的数据集. Total-Text 是一个数据集, 其中包含各种形状的文本, 包括水平的、多取向的和弯曲的. 这 3 个数据集包含了中文和英文的数据集共 6 万张, 用于文字检测和识别. 将每个数据集, 按照 5:1 的比例分成训练集和测试集.

3.2 实验设计

训练文本区域检测模型: 本文首先用随机选取的 3 个数据集中的 5 万张图片进行预训练, 然后, 本文在相应的其他数据集上进行调整, 训练时, 批大小设为 16, 初始学习效率设为 0.007. 为了提高训练效率, 所有处理后的图像都被重新调整为 640×640, 在推理阶段,

本文保持测试图像的高宽比, 并通过为每个数据集设置合适的高度来重新调整输入图像的大小, 获得最终模型.

训练文本识别模型: 将 5 万张图片中裁剪下来的包含文本的数据集进行微调后进行了训练, 对于不规则数据集上的序列识别, 本文的实验基于 DenseNet 网络, 其中 conv1 变为 4×4, 步长为 1, conv4_x 作为输出, 并使用 ACE 损失函数最终得到文本识别模型.

3.3 主干网对比分析

随着神经网络层数的增多, 则对输入图像提取的特征将会更加抽象, 这是因为后层神经元的输入是前层神经元的累加和, 而特征的抽象程度越高, 更有利于后期的分类任务或回归任务. 但要提高模型的准确率, 都是加深或加宽网络, 但是随着超参数数量的增加, 网络设计的难度和计算开销也会增加. ResneXt 特征网络增加了基数且用平行堆叠相同拓扑结构的 blocks 代替原来 ResNet 的三层卷积的 block. 在不明显增加参数量级的情况下提升了模型的准确率, 同时由于拓扑结构相同, 超参数也减少了.

因此本文采用了更高效的特征提取网 ResneXt-101 作为主干网络提高分类效果, 为了更好的证明该文本检测网络的性能, 并在大规模 TotalText 数据集上进行测试. 事实证明, 以 ResneXt 作为主干网络的检测器比 ResNet 性能更好, 且更深的神经网络可以提高大规模图像分类和目标检测的性能. 如表 1 所示. 在相同的设置下, 将主干网络由 ResNet-50 改为 ResneXt-50 性能从 78.2% 改善到 83.6%, 提高了 5.2%, 本文又将主干深度从 50 提高到 101, 通过对比可以明显性能从 83.6% 改善到 85.8%, 提高了 2.3%. 综合发现本文的选取 ResneXt-101 作为文本检测的主干网络在精度和速度上都达到了最先进的性能.

表 1 检测框架中不同主干网络结果

主干网	准确率(%)
ResNet-50	78.2
ResneXt-50	83.6
ResneXt-101	85.8

在文本识别网络框架中采用了 DenseNet 作为骨干网络. 如表 2 所示, 与之前的主干网络相比, DenseNet 在准确率上略好与之前最好的结果.

3.4 独立性对比分析

为了验证实验中字符顺序的独立性与识别网络使用的聚合交叉熵损失的关系, 本文使用聚合交叉熵, CTC 和注意力在 3 个数据集上进行实验. 将标注的字符顺序按照不同的比例随机打乱, 如图 3 所示. 可以发现, ACE 的性能随打乱比例的增加基本保持不变, 而注意力和 CTC 的性能在不断下降. 所引入的 ACE 损失函数对于打乱的字符顺序识别结果基本一致.

表 2 识别框架中不同主干网络结果

主干网	准确率(%)
Convnet	90.1
ResNet	92.3
DenseNet	95.4

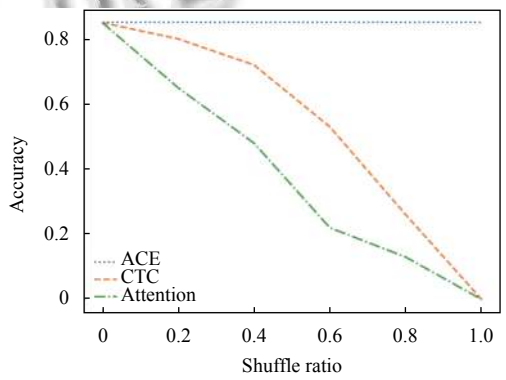

图 3 ACE, CTC 和注意力性能对比图

3.5 实验结果对比

检测与识别是判别是否达到要求的重要条件. 由于图片太多, 不利于展示, 本文选取几张代表性图片用于结果展示.

本文将改进前后的检测网络进行了可视化比较, 图片 4 左侧是 PSENet 的部分检测结果, 可以看出有漏检以及错误的定位现象. 图 4 右侧是加入了可微二值化后的部分检测结果, 比较明显的看出, 本文提出文本区域定位网络 TLDNet 可以很好的定位出复杂的曲线文本位置.

我们在两个数据集上对本文改进后的方法和之前的方法进行了比较, 如表 3 所示, 文本的方法在精度和速度上都达到了最优. 具体来说, 本文的提出方法在 TotalText 和 CTW1500 数据集上的表现比之前的方法要好. 文本的方法比 PSENet 方法要快, 并且可以通过使用 ResneXt 主干进一步提高速度. 在表 3 中文本区域定位网络 TLDNet 的准确率在两个数据集上比 PSENet 高 2.4% 和 3.4%, 但自然场景采集的图片由于采集环境等因素造成图片的模糊、反光等现象, 进而导致漏

检和定位不准确的现象发生, 本文的方法可以有效缓解该现象, 但不能完全消除. 本文可以保证在干净明亮采集环境下获取清晰的图片, 可以从本质上防止上述现象的发生.

图 4 文本区域定位结果对比图

表 3 TotalText 和 CTW1500 数据集的检测结果 (%)

方法	TotalText准确率	CTW1500准确率
PSENet.	82.7	81.2
Ours-ResneXt-50	83.5	83.2
Ours-ResneXt-101	**85.1**	**84.6**

本文将检测后的图片输入到文本识别模型中, 得到如图 5 的识别结果. 图 5 右侧为识别正确的效果图. 图 5 左侧为识别错误的效果图, 由此结果可以发现文本区域的定位直接影响文本识别的准确性, 在确保定位准备的条件下, 基本能够正确识别文本信息.

图 5 文本识别效果图

对于不规则的场景文本, 本文提供了和以前注意力机制方法进行的比较, 如表 4 所示, 所加入的 ACE 损失函数在数据集 TotalText 和 CTW1500 上表现出优异的性能, 特别是在 CTW1500 上, 准确率提高了 8.1%. 因为数据集 CTW1500 是专门用于弯曲文本识别的, 因此, 充分展示了 ACE 损失功能的优势. 同时, 两个数据集中有的图像具有非常低的分辨率, 这对语义上下文建模产生了非常高的要求, 本文中的识别模型在使用词汇时获得了最高的结果, 语义上下文可以访问. 这再次验证了所提出的 ACE 损失函数的稳健性和有效性.

表 4 TRNet 和之前的方法比较 (%)

方法	TotalText 准确率	CTW1500 准确率
Aster	85.2	86.4
TRNet	**93.8**	**94.5**

本文采用的方法减少了训练神经网络模型所需要的训练数据、计算成本等. 本文实验对预训练模型 (TDRNet) 和初始模型 (PSENet+Aster) 的准确率进行比较, 如图 6 所示.

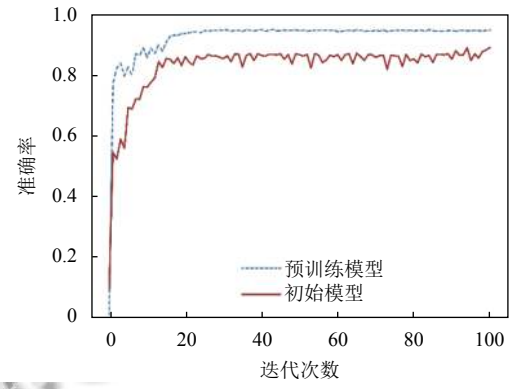

图 6 准确率对比

由表 5 可知, 将改进前的检测和识别网络结合的原始模型准备率为 90.3%, 经改进训练后 TDRNet 模型最终达到 95.6% 的识别准确率. 根据上述实验结果, 本文有以下结论: (1) 本文将使用了更高级的特征网络并在后处理过程中加入可微二值化的方法在准确度上优于其他方法. (2) 聚合交叉熵损失对于文本识别过程中字符序列预测至关重要, 具有一定的通用性. (3) 将检测和识别网络相结合, 提高了文本识别的准确性以及识别速率.

识别的准确性取决于定位算法的准确性. 所以在实际应用中, 为了提高文本区域定位算法的准确性, 尽量保证采集环境干净明亮.

表 5 网络准确率表

方法	准确率(%)
PSENet+Aster	90.3
TDRNet	95.6

4 结束语

在本文中, 我们提出了一个新的框架检测和识别任意形状的场景文本, 其中包括采用更高效的特征提取网络并在检测框架中加入了可微二值化过程分割网络, 在识别框架中基于聚合交叉熵的损失函数, 优化了检测和识别器网络结构, 简化了后处理方法, 较好地满足了复杂场景下文本定位和识别的任务要求, 实验证明, 本文的方法在 3 个标准场景文本基准测试中, 在速度和准确性方面始终优于最新的方法. 在未来, 如何实现端到端的检测和识别问题将成为下一步主要研究的工作.

参考文献

1 王德青, 吾守尔·斯拉木, 许苗苗. 场景文字识别技术研究综述. 计算机工程与应用, 2020, 56(18): 1–15. [doi: 10.3778/j.issn.1002-8331.2004-0333]

2 梁林森. 复杂背景下电力客户证件识别关键技术的研究与实现. 科技与创新, 2019, (7): 70–71.

3 黄攀. 基于深度学习的自然场景文字识别 [硕士学位论文]. 杭州: 浙江大学, 2016.

4 刘华春. 卷积神经网络在车牌识别中的应用研究. 计算机技术与发展, 2019, 29(4): 128–132. [doi: 10.3969/j.issn.1673-629X.2019.04.026]

5 李阳, 李绍彬, 解云超, 等. 基于卷积神经网络的文本检测算法研究. 中国传媒大学学报 (自然科学版), 2019, 26(1): 70–76.

6 王润民, 桑农, 丁丁, 等. 自然场景图像中的文本检测综述. 自动化学报, 2018, 44(12): 2113–2141.

7 缪裕青, 刘水清, 张万桢, 等. 自然场景图像中的中文文本检测算法. 计算机工程与设计, 2018, 39(3): 804–807, 818.

8 Graves A. 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.

9 Liao MH, Shi BG, Bai X, *et al*. TextBoxes: A fast text detector with a single deep neural network. arXiv: 1611.06779, 2016.

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

11 Deng D, Liu HF, Li XL, *et al*. PixelLink: Detecting scene text via instance segmentation. arXiv: 1801.01315, 2018.

12 Dai YC, Huang Z, Gao YT, *et al*. Fused text segmentation networks for multi-oriented scene text detection. arXiv: 1709.03272, 2018.

13 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 2204–2212.

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

15 Liao MH, Wan ZY, Yao C, *et al*. Real-time scene text detection with differentiable binarization. arXiv: 1911.08947, 2019.

16 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 936–944.

17 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

18 Xie SN, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 5987–5995.

19 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.

20 Shi BG, Yang MK, Wang XG, *et al*. ASTER: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2035–2048. [doi: 10.1109/TPAMI.2018.2848939]

21 Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2261–2269.

22 Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015. 1127–1137.

23 Xie ZC, Huang YX, Zhu YZ, *et al*. Aggregation cross-entropy for sequence recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. 2019. 6531–6540.