计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):175−180 [doi: 10.15888/j.cnki.csa.007426]

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

基于 FCN 的图像中文字目标语义分割①

刘信良, 王静秋

(南京航空航天大学 机电学院, 南京 210016)
通讯作者: 王静秋, E-mail: meejqwang@nuaa.edu.cn

摘　要: 本文提出一种基于全卷积神经网络的图像中文字目标语义分割算法和一种新的数据集制作与增广方法. 该算法首先采用改进全卷积神经网络对图像中的文字目标进行初步分割, 然后利用大津法进行二值化处理, 划分出目标的大致区域, 最后用全连接条件随机场算法进行修正, 得到最终结果. 该算法在测试集上准确率为 85.7%, 速度为 0.181 秒/幅, 为图像目标区域的进一步分析做准备.

关键词: 语义分割; 全卷积神经网络; 大津法; 全连接条件随机场

引用格式: 刘信良,王静秋.基于 FCN 的图像中文字目标语义分割.计算机系统应用,2020,29(6):175–180. http://www.c-s-a.org.cn/1003-3254/7426.html

Semantic Segmentation of Character Targets in Images Based on FCN

LIU Xin-Liang, WANG Jing-Qiu

(College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

Abstract: This study proposes an algorithm for semantic segmentation of targets in images based on fully convolutional neural networks and a new method to make and augment dataset. The algorithm primarily segments the targets from images using improved fully convolutional neural networks, OTSU method is applied to binarize images and segment the general areas of targets, finally, the fully connected conditional random field algorithm is used to correct the general areas of targets and get the final results. This algorithm achieves the accuracy of 85.7% and speed of 0.181 second per image on test set, and prepares for further analysis of targets in images.

Key words: semantic segmentation; Fully Convolutional Neural (FCN) networks; OTSU method; fully connected conditional random field

数字图像是由有限像素值组成的二维矩阵, 包含丰富的语义信息, 自动识别图像中的目标, 例如文字是当前研究的热点问题. 图像文字识别技术是指利用计算机视觉等技术分离图像中文字与背景区域, 并且准确识别文字内容的过程. 该技术可将图像中的文字转化为可供计算机识别和处理的文本信息, 减少人为参与, 提高自动化程度.

图像文字识别技术一般先分割图像中的目标文字再识别其内容, 目标文字分割的效果影响文字内容识别的准确度. 林孜阳等[1]提出了一种结余连通域的目标文字分割算法, 通过连通域阈值分析, 将文本块联通, 从而实现对难提取的文字分割. 易小波[2]提出了一种图像二值化分割处理的方法, 通过对图像特征的研究, 选取适当的阈值对图像进行二值化, 从而达到分割的效果. 郑泽鸿等[3]提出将 AP 聚类算法用于字符分割, 根据类中心对特征点进行归类得到分割结果. 上述算法运行速度快、对特定场景分割准确度高, 但需要人为设定大量参数, 且通用性较差、鲁棒性不足, 在实际应用中仍然存在较多的局限性.

近些年, 深度学习中的卷积神经网络被应用到众

① 收稿时间: 2019-10-28; 修改时间: 2019-11-20; 采用时间: 2019-11-29; csa 在线出版时间: 2020-06-10

多领域, 如图像分类[4,5]、目标检测[6,7]、目标追踪[8]、图像分割[9]等, 取得了令人瞩目的成绩, 逐渐成为研究的热点. Zhang 等[10]提出一种串联全卷积神经网络(Fully Convolutional Networks, FCN)[11]的方法, 先用 FCN 模型粗分割出文字区域, 再用另一 FCN 分类器预测出每个文字区域的中心位置, 以去除粗分割时错误区域. 此方法在稀疏文本下表现较好, 但无法解决复杂文本重叠情况.

本文针对图像中文字逐点标注复杂费时的问题, 设计一种简单的数据标注与增广方式, 并且改进 FCN 模型, 结合大津法[12]及全连接条件随机场[13], 实现对较复杂图像中目标文字的语义分割, 为后续文字识别做准备.

1 图像文字语义分割算法

语义分割是指自动分割并识别图像中具有特定语义的目标物体的过程. 如图 1 所示, 将人和自行车作为两种语义类别, 在图像中自动分割出属于这两类的所有物体, 并标定其类别, 其余部分当作背景. 图像中目标文字分割也可以当作语义分割问题求解, 将文字当作唯一的语义类别, 其余当作背景, 从而将文字从图像中分割出来.

1.1 图像文字语义分割算法流程

本文提出了一种基于全卷积神经网络的图像文字语义分割的算法, 并采用大津法和全连接条件随机场进行修正处理. 如图 2 所示, 首先利用 FCN 模型对输入图像进行初步分割, 再用大津法对其二值化处理, 最后使用全连接条件随机场修正, 得到精细的语义分割结果.

1.2 基于 FCN 的图像初步分割

FCN 模型是用于任意尺寸图像输入的语义分割模型, 此模型包括图像特征提取编码和标记图像解码生成两个部分. 图像特征提取编码是对图像高级语义特征抽象的过程, 通过多层卷积和池化操作, 删除冗余信息, 提取出图像的本质信息. 标记图像解码生成是对语义特征重建的过程, 通过上采样操作恢复图像的原始尺寸, 并得到每个像素所属的类别.

如图 3(a) 所示, 本算法中的 FCN 特征提取基础网络为 VGG16 网络[14], 利用 13 个 3×3 的卷积层和 5 个最大池化层提取图像中抽象的语义特征, 通常这些特征是整张图像的全局特征. 同时, FCN 将 VGG16 网络的全连接层用 1×1 卷积层替换, 以解决全连接层神经元个数必须固定的缺点, 从而实现任意尺寸图像输入. 由于图像特征提取过程经过了 5 个最大池化, 原始图像被缩小了 32 倍, 故需要用反卷积进行上采样, 恢复成原始图像的大小. 如图 3(b) 所示, 第 5 个池化层输出的特征图经过反卷积扩大 2 倍后与第 4 个池化层输出的特征图结合, 将其结果反卷积扩大 2 倍后与第 3 个池化层输出的特征图结合, 最后反卷积扩大 8 倍生成预测标签图像. 通常特征提取中间步骤的特征图包含着丰富的浅层特征, 如边缘、纹理等, 将此浅层特征与最终提取的深层抽象特征结合可以更加准确地分割物体.

图 1 语义分割示例

图 2 算法流程图

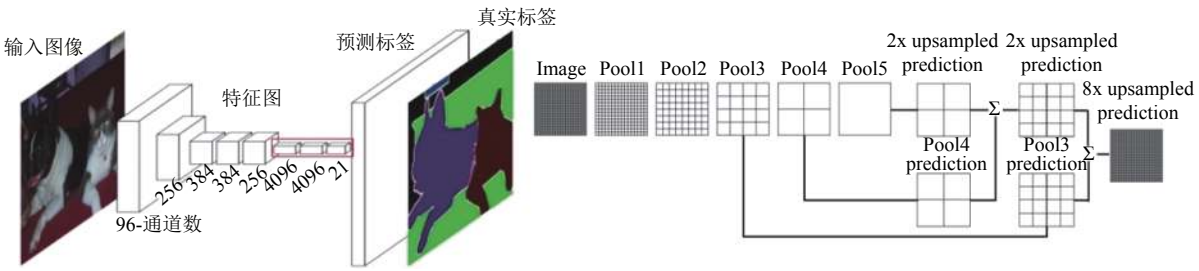


图 3 FCN 整体结构

本文改进原有的 FCN 模型, 采用均方差损失函数, 定义为式 (1), 其中y_i为预测的标签值, $\hat{y}_i$为真实的标签值. 采用此损失函数目的是利用回归的方式, 使得最终特征图的目标值接近 255, 背景值接近 0, 通过前景背景较大的差异值准确分割出文字区域大体位置.

$$loss = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

1.3 大津法二值化

大津法能够根据阈值将图像分为目标和背景两部分, 通过计算目标和背景的类间方差, 遍历所有灰度值寻找最佳阈值使得类间方差最大. 对于带有文字的图像, 假设属于文字的像素点数占整幅图像的比例为 ω_o, 平均灰度为 μ_o; 背景像素点数占整幅图像的比例为 ω_b, 平均灰度为 μ_b. 类间方差 g 如式 (2) 计算:

$$g = \omega_o\omega_b(\mu_o - \mu_b)^2 \tag{2}$$

遍历 0~255 各灰度值, 计算并寻找类间方差的极大值即为文字目标与背景分割的最佳阈值.

1.4 全连接条件随机场修正

在本文中, 对于一幅带有文字的图像, 每个像素点 i 具有像素值 I_i, 对应的类别标签为 x_i, 以每个像素作为节点, 像素之间的关系作为边, 构成了一个全连接条件随机场. 其吉布提能量可以表示为:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{ij}\psi_p(x_i, x_j) \tag{3}$$

式中, $\psi_u(x_i)$为一元势能, 来自大津法二值化的输出, 只与像素 i 自身相关, 表述了像素 i 与其所属类别的差异度:

$$\psi_u(x_i) = -\log P(x_i) \tag{4}$$

二元势能$\psi_p(x_i, x_j)$表达式为:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j)(\omega_1\nabla_1 + \omega_2\nabla_2) \tag{5}$$

$$\nabla_1 = \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2}\right) \tag{6}$$

$$\nabla_2 = \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2}\right) \tag{7}$$

式中, ∇_1为外观核, p 表示位置信息, I 表示颜色信息, σ_α为用于控制位置信息的尺度, σ_β为用于控制颜色相似度的尺度, 通常在边界位置颜色有较明显的变化, 因此位于边界两侧的像素不同类别的可能性较大; ∇_2为平滑度核, σ_γ为控制位置信息的尺度, 此项只与位置信息有关, 用于平滑去除一些孤立的小区域, 通常这些区域不是文字区域.

本文提出的图像文字语义分割算法融合了改进的 FCN 模型、大津法和全连接条件随机场, 相比传统数字图像处理算法, 此算法需要设定的参数少且能够适用复杂场景的文字分割.

2 实验

2.1 数据集准备

数据集图像中的目标文字主要有两种类型: 光学字符 (文字是由摄像设备输入) 和合成字符 (通过软件将字符合成到图像上). 原版 FCN 模型的数据集需要逐点标注目标的轮廓并给出目标的类别, 对于图像集制作起来复杂、耗时, 因此本文提出一种新的数据集制作方法: 对原始数据集图像用最小四边形标记文字区域, 即四边形包围区域设为目标, 像素值为 255, 其余区域设为背景, 像素值为 0, 如图 4(a)、图 4(b) 所示; 为扩充数据集, 让分割结果更加准确, 制作模拟合成字符图像, 即准备一批无文字的背景图像, 将文字集随机生成并合成到图像中, 文字的类型包含中文、英文和数字, 颜色随机生成, 如图 4(c)、图 4(d) 所示. 数据集包含训练集, 验证集和测试集, 其中训练集有 10 000 张图片, 验证集和测试集各有 1000 张图片.

(a) 原始数据集图 (b) 文字标记图

(c) 模拟合成图 (d) 随机生成文字

图 4 数据集示例

2.2 FCN 模型训练

本次实验的 FCN 模型是用 Tensorflow[15]框架搭建, 实验平台为个人电脑, 硬件为: i7-8700k CPU, 16 GB 内存, RTX 2070 GPU.

训练时, 所有图像和标签都被放缩为 224×224, 以 32 张为一批送入模型中训练. 应用指数衰减的学习率, 初始学习率设置为 0.001, 衰减系数设置为 0.9. 优化策略采用随机梯度下降, 共训练了 50 000 次, 耗时约 11 个小时, 其损失函数值的趋势如图 5 所示, 从图中可以看出损失函数值随着训练的进行不断下降, 最终稳定在一定数值.

2.3 实验结果及分析

如图 6(a)、图 6(b) 所示, 将 FCN 初步分割结果转化为 RGB 格式, 亮色代表此区域为文字的概率值大, 暗色代表此区域为背景的概率值大. 从图中可以看出, 亮色区域与图像中文字区域基本对应, 越亮的部分代表着此区域是文字的概率越大; 相反, 暗色区域与图像中的背景相对应. 接着, 用大津法对 FCN 初步分割结果进行二值化处理, 计算得到使得目标/背景的类间方差最大的阈值分别为 107、91、93. 如图 6(c), 可以看出大津法大致将文字区域划分出来, 但是划分结果较粗糙, 存在错误标记区域. 最后, 使用全连接条件随机场进行结果修正, 如图 6(d) 所示, 可以看出经过全连接条件随机场后处理的图像能够提升分割的结果, 将大津法处理后的粗糙结果细化, 更正了误识别区域, 为后续文字内容识别做好准备. 然而, 此算法仍存在一定漏检和误检的问题, 分别如图 6 中第 2 行与第 3 行矩形框所示.

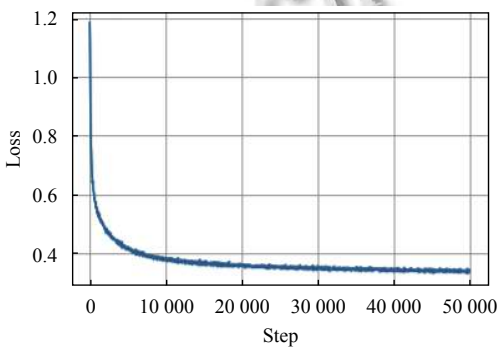
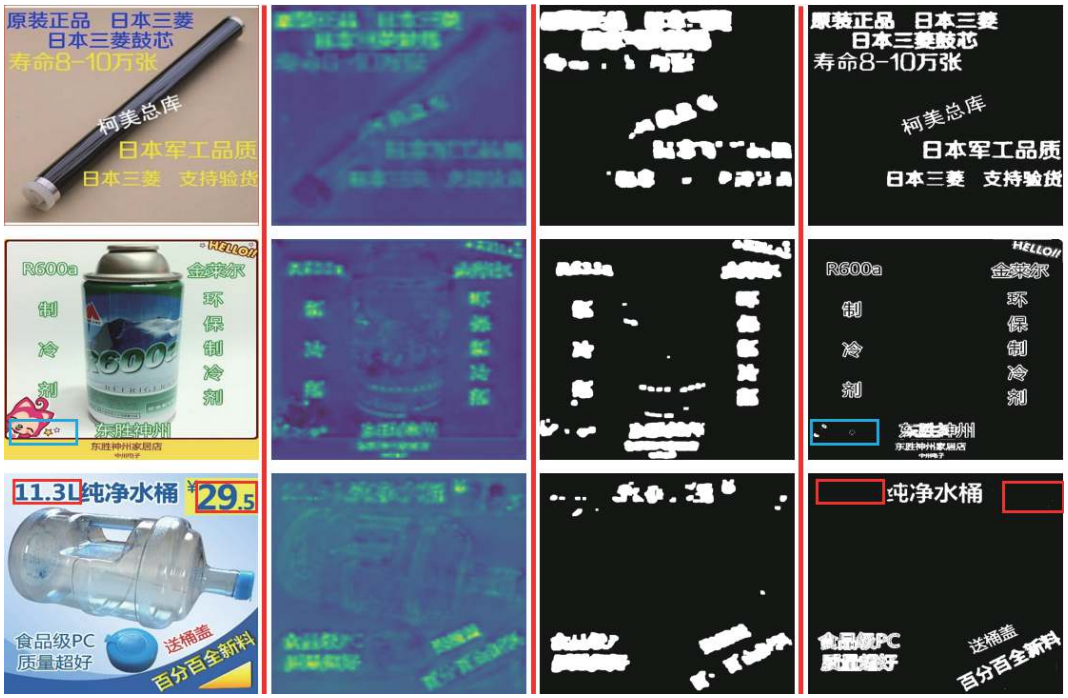


图 5 训练时损失函数曲线



(a) 原始图像 (b) FCN 分割结果 (c) 二值化结果 (d) 最终结果

图 6 实验结果

图 7 为本文提出的算法与“FCN+全连接条件随机场”算法对比结果. “FCN+全连接条件随机场”算法的结果记为对照组, 即采用 FCN 模型进行初步分割 (采用交叉熵损失函数), 再用全连接条件随机场进行后处

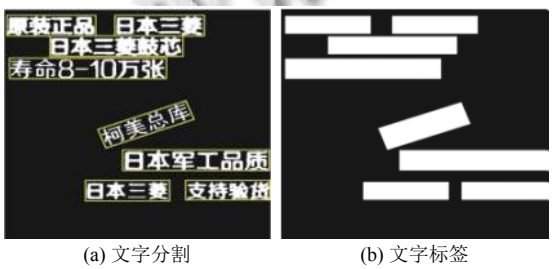
理. 图 7(a)、图 7(b) 分别为两种算法的初步分割图, 相较本文算法, 对照组初步分割的结果较粗糙, 有较多孤立的小片区域, 且文字区域不够明显. 最终分割结果如图 7(c)、图 7(d), 对照组能够分割出文字, 但是出现一些漏分割现象, 如图中矩形框所示, 且精细程度不如本文提出的算法.

(a) 对照组初分割图 (b) 本实验初分割图 (c) 对照组结果图 (d) 本实验结果图

图 7 对比结果

由于验证集和测试集的标签类型如图 4(b) 所示, 无法使用传统的图像分割的评判标准, 如像素准确度、分割交占比 (IOU) 等, 本文提出一种新的评判标准, 评判步骤如下:

(1) 先用 OpenCV 对文字分割结果用最小矩形框集合, 如图 8(a) 所示, 矩形框内部即为目标文字区域.



(a) 文字分割 (b) 文字标签

图 8 评价结果图

(2) 对真实标签的每个文字区域 (如图 8(b)), 按照式 (8) 计算其对预测结果中每个文字区域的 C_i, 选取最大 $C_{\max}$, 并将取最大值时两个文字区域的索引值记录到集合 I 中. 若 $C_{\max}\geq 60\%$ 表示预测正确; 若 $10\%\leq C_{\max}<60\%$ 预测错误; 若 $C_{\max}<10\%$ 则表示漏检; 若预测结果中文字区域的索引值未出现在 I 中, 则此区域为误检.

$$C=\frac{S_p\cap S_t}{S_t\cup S_t} \tag{8}$$

其中, S_p 表示预测的文字区域, S_t 表示真实文字区域.

(3) 错误总数 (N_{err}) 为预测错误个数 (N_e) 与漏检个数 (N_m) 之和, 准确率计算如下:

$$acc=\frac{N_c}{N_a} \tag{9}$$

其中, N_c 表示预测正确的总个数, N_a 为验证集/测试集文字区域的总个数

根据此项判断标准, 实验结果如表 1 所示, 可以看出整体的实验效果较好, 没有出现过拟合的情况; 同时, 可以发现结果中误检数量较漏检数量多, 可知本次实

验提出的算法容易将类似文字的图案 (背景) 识别为文字 (目标). 本文算法的速度为 0.181 s/幅, 即 5.5 fps, 运行速度较快.

表 1 图像目标文字分割准确率分析

参数	验证集	测试集
图像数	1000	1000
真实区域个数 (N_a)	19 514	18 571
预测正确个数 (N_c)	17 036	15 915
预测错误个数 (N_e)	1641	1529
漏检个数 (N_m)	837	1127
误检个数 (N_l)	733	975
准确率 (acc)(%)	87.3	85.7

3 结论

本文提出了一种基于改进全卷积神经网络的图像目标分割算法, 此算法使用 FCN 模型进行初步分割, 再利用大津法进行二值化, 最后使用全连接条件随机场进行修正. 此算法在准确度和速度上都取得了较好的效果, 在测试集上可以达到 85.7% 的准确度以及 0.181 s/幅的速度.

参考文献

1 林孜阳, 穆雪, 吴凯锋, 等. 基于连通域的快速文字图像分割算法. 计算机光盘软件与应用, 2014, (22): 89–90.

2 易小波. 身份证图像识别系统中文字分割的研究. 企业技术开发, 2003, (11): 19–21.

3 郑泽鸿, 黄成泉, 梁毅, 等. 基于 AP 聚类的中文字符分割. 智能计算机与应用, 2018, 8(1): 65–67, 71.

4 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.

5 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.

6 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 21–37.

7 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

8 Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4293–4302.

9 Chen LC, Papandreou G, Kokkinos I, *et al*. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

10 Zhang Z, Zhang CQ, Shen W, *et al*. Multi-oriented text detection with fully convolutional networks. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4159–4167.

11 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

12 Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62–66. [doi: 10.1109/TSMC.1979.4310076]

13 Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with gaussian edge potentials. http://arxiv.org/abs/1210.5644.

14 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556.

15 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv: 1603.04467, 2016.