计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(10):1−10 [doi: 10.15888/j.cnki.csa.006539]

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

基于特征融合网络的自然场景文本检测①

余　峥, 王晴晴, 吕　岳

(华东师范大学 计算机科学与软件工程学院, 上海 200062)
通讯作者: 余　峥, E-mail: henuyz@163.com

摘　要: 目前, 基于深度学习的自然场景文本检测在复杂的背景下取得很好的效果, 但难以准确检测到小尺度文本. 本文针对此问题提出了一种基于特征融合的深度神经网络, 该网络将传统深度神经网络中的高层特征与低层特征相融合, 构建一种高级语义的神经网络. 特征融合网络利用网络高层的强语义信息来提高网络的整体性能, 并通过多个输出层直接预测不同尺度的文本. 在 ICDAR2011 和 ICDAR2013 数据集上的实验表明, 本文的方法对于小尺度的文本, 定位效果显著. 同时, 本文所提的方法在自然场景文本检测中具有较高的定位准确性和鲁棒性, F 值在两个数据集上均达到 0.83.

关键词: 深度学习; 自然场景; 文本检测; 特征融合; 文本边界框

引用格式: 余峥,王晴晴,吕岳.基于特征融合网络的自然场景文本检测.计算机系统应用,2018,27(10):1–10. http://www.c-s-a.org.cn/1003-3254/6539.html

Scene Text Detection Based on Feature Fusion Network

YU Zheng, WANG Qing-Qing, LYU Yue

(School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China)

Abstract: At present, scene text detection based on deep learning has achieved good performance in complex background. However, it is difficult to precisely detect text with small scale. To solve this problem, this study proposes a deep neural network based on feature fusion, and a new neural network with senior semantic is constructed by combining the high-level feature and low-level feature of traditional deep neural network. Strong semantic information of the high layer network is utilized to improve the overall performance of the neural network, and the feature fusion network directly predicts text with multiple scales through multiple output layers. Experimental results on ICDAR2011 and ICDAR2013 datasets show that proposed method is significantly effective in detecting small scale text. Meanwhile, the proposed method has high accuracy and robustness in scene text detection, and the F-measure achieves 0.83 on both datasets.

Key words: deep learning; natural scene; text detection; feature fusion; text bounding boxes

1　概述

随着互联网和多媒体技术的发展, 越来越多的信息载体以图像的形式存在. 自然场景图像中的文字作为一种极其重要的信息来源, 捕获和识别这些文字有助于理解和分析图像, 因此, 自然场景图像中的文本检测成为当下热门的研究话题之一. 目前文本检测技术在现实生活中有着广泛的应用, 例如, 手机设备上的拍照翻译软件, 可以拍摄异国街道或路牌上的文字, 将一种语言实时翻译为另一种语言, 提供导游帮助; 公安机关的高速监控设备, 可以抓拍识别高速公路上行驶汽车的车牌号码, 智能化收集违章车辆信息[1]. 除此之外, 文本检测技术在图像检索[2]、视频字幕提取[3]等领域

① 基金项目: 上海市自然科学基金 (17ZR1408200)
Foundation item: Natural Science Foundation of Shanghai (17ZR1408200)
收稿时间: 2018-02-02; 修改时间: 2018-02-28; 采用时间: 2018-03-06; csa 在线出版时间: 2018-09-28

也存在广泛的应用. 因此, 对自然场景图像中的文本检测进行研究具有重要的理论意义和实用价值.

由于自然场景图像中背景错综复杂, 以及文字所处的位置可能存在逆光、遮挡和模糊等现象, 准确检测出场景中的文字成为一项具有挑战性的工作. 同时, 自然场景中的文字具有字体多样、颜色多变、分布不一的特点, 文本检测技术需要具有较强的鲁棒性.

传统的自然场景文本检测方法主要依赖于手动创建图像的特征, 利用机器学习的方法判别出文字的位置, 此类方法存在计算量大、检测过程复杂等缺点. 近年来, 随着深度学习的发展, 基于深度学习的方法在文本检测中取得显著的效果, 这些方法简单高效, 利用单个神经网络便能检测到不同尺度的文本. 但是, 大多数的神经网络在检测小尺度的文本上不能取得很好的效果. 因此, 本文基于传统深度神经网络, 在保证网络层次结构不变的前提下, 提出将网络中的高层特征与低层特征进行融合, 构建一种高级语义的神经网络用于自然场景文本检测.

为了验证高层特征与低层特征不同融合方式对网络性能的影响, 本文提出三种特征融合网络, 分别为相邻两层特征融合网络、相邻三层特征融合网络和最高层特征融合网络. 特征融合网络在层次结构上是金字塔结构, 通过自底向上和自顶向下的连接方式将不同层的特征进行融合. 特征融合后的网络具有多个输出层, 每个输出层都具有较强的语义信息并能检测不同尺度的文字. 本文在 ICDAR2011 和 ICDAR2013 两个标准数据集上进行了实验, 实验表明本文提出的特征融合网络可以有效地检测出小尺度的文本, 并具有较高的定位准确性和鲁棒性.

2 相关研究

自然场景文本检测是从具有复杂背景的图像中检测出文字的位置. 目前自然场景文本检测方法主要分为三类: 基于滑动窗口的文本检测方法、基于连通域的文本检测方法和基于深度学习的文本检测方法.

2.1 基于滑动窗口的文本检测方法

基于滑动窗口的文本检测方法使用多尺度的滑动窗口去扫描图像, 搜索图像中文字出现的位置. 基于文字的特征, 运用一个预训练的文字分类器, 判别窗口内是否存在文字. 其中文献[4]使用滑动窗口结合方向直方图 (Histogram Of Gradient, HOG) 特征建立文本置信图, 然后使用随机蕨 (random ferns) 过滤掉图中的非文本区域. 文献[5]结合多尺度滑动窗口利用 AdaBoost 算法, 将多个弱文本分类器组合成强文本分类器, 过滤掉图中的非文字区域. 这类方法的主要缺陷是需要对整张图像进行穷尽式的扫描, 计算量大、消耗时间.

2.2 基于连通域的文本检测方法

基于连通域的文本检测方法是利用文字区域具有相同的颜色和结构等特征来生成文本连通域, 然后根据连通域的大小, 宽高比等先验知识来获得文字区域. 文献[6]提出使用笔画宽度变换 (Stroke Width Transform, SWT) 算子提取出字符笔画的边缘图, 再结合几何推理恢复出字符的形态, 该算子可以有效地提取复杂背景图像中不同尺度的文本. 文献[7]率先提出最大稳定极值区域 (Maximally Stable Extremal Regions, MSER) 算法检测文字, 该算法能有效地提取候选文本连通域, 然后通过形态学操作和连通域的形状来确定文本区域. 为解决 MSER 算法检测结果存在较多嵌套区域的问题, 文献[8]采用 MSCR (Maximally Stable Color Regions) 算法与 MSER 算法相结合提取候选字符区域, 依据字符区域的颜色一致性和几何邻接关系对字符进行合并, 最终得到文本区域. 基于连通域的方法降低了扫描图像的计算复杂度, 但这类方法应用了大量的自定义规则和参数, 并且很容易生成大量的非文字候选字符和重复的文字候选字符. 为了消除无效的候选字符, 该类方法还需要设计一个字符级别的分类器过滤掉无效的候选文字, 使得检测复杂度增大.

2.3 基于深度学习的文本检测方法

近年来, 随着深度学习的发展, 越来越多的研究倾向于使用深度神经网络来解决文本检测问题. 文献[9]率先提出使用卷积神经网络 (Convolutional Neural Network, CNN) 训练一个文本分类器. 卷积神经网络通过提取图像的深层特征来区分文本和非文本, 训练过程简单高效. 基于卷积神经网络的强分类性能, 文献[10]首先使用 MSER 算子提取图像中的候选文字连通域, 然后使用 CNN 分类器过滤掉 MSER 产生的无效连通域, 该方法大幅度地提高了传统检测文本的性能.

随着深度神经网络在目标检测中的发展, 先后涌现出一系列的目标检测方法, 例如, R-CNN (Regions with CNN)[11], Fast R-CNN[12], Faster R-CNN[13], SSD (Single Shot multibox Detector)[14]. 其中, SSD 通过单个卷积神经网络直接预测目标的边界框并且得到相应类

别的概率.

受 SSD 直接预测目标的边界框的启发, 文献[15]将 SSD 应用于文本检测, 并提出一个用于文本检测的神经网络 TextBoxes, TextBoxes 利用网络层中的特征图 (feature map) 直接输出文本的边界框和置信度. 其网络结构, 如图 1 所示. 该网络是一个全卷积神经网络, 网络结构里有多个输出层 (conv4_3, conv6_2, conv7_2, conv8_2, conv9_2, conv10_2, conv11_2). 这些输出层是网络中的卷积层, 也是网络结构中的关键组成部分, 可以在其特征图上预测文本出现的概率和文本边界框. 网络最后使用非极大值抑制算法聚集所有的 Textbox 层输出的文本框, 得到最终的文本位置.

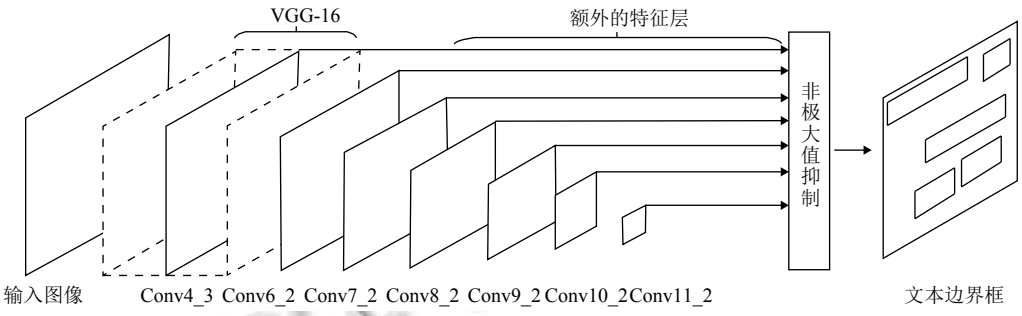


图 1 TextBoxes 的网络结构图

TextBoxes 的网络模型可以端到端进行训练, 不仅训练过程简单, 而且检测速度快. TextBoxes 可以在不同分辨率的特征图上预测文字的位置, 与以往的文本检测方法相比, 它的处理过程简单, 不需要设计启发式的规则, 使得文本检测更加高效. 但是它不能较好地预测小尺度文本. 因此, 本文将提出新的方法来提高网络对小尺度文字的定位准确率, 进一步提高网络的性能.

3 基于特征融合网络的自然场景文本检测

TextBoxes 的网络模型具有金字塔特征层次结构, 网络高层的语义信息比较强, 低层语义信息比较弱. 由于网络低层特征图表达能力不足, 所以不能较好地预测小尺度的文本. 为了解决该问题, 提高低层特征图的表达能力, 使网络能在不同分辨率的特征图上都能检测到对应尺度的文本, 本文提出将网络高层的特征与低层的特征进行融合得到新的特征图, 在新的特征图上预测文字的位置.

3.1 特征融合

特征融合是指提取和综合目标的两种或多种特征, 提高同一类别的目标识别率. 一般是将不同的特征向量组合起来, 组成一个新的特征向量, 然后采用分类器进行判别分类. 在神经网络中, 将网络高层特征和低层特征进行融合, 可以使用融合特征图的方式. 将特征图进行融合一般有两种方式, 分别是元素求和方式和元素点积方式.

神经网络中的特征图相当于二维矩阵, 使用元素求和方式和元素点积方式必须要求两个矩阵的大小一致. 由于高层和低层输出层对应的特征图大小不一致, 不能直接进行融合. 为了融合高层特征和低层特征, 本文对网络高层输出的特征图使用一个反卷积操作, 将网络高层特征图的尺度大小处理成与低层特征图一致. 反卷积操作类似于双线性差值, 可以有选择地对特征图进行放大. 在神经网络中, 使用反卷积层实现反卷积操作, 反卷积层输出的特征图大小的计算公式为:

$$o=(i-1)*s+k-2p \tag{1}$$

其中, i 表示反卷积层输入特征图的大小, k 表示卷积核的尺寸, s 表示步长大小, p 表示填充边距. 网络高层的特征图通过反卷积层设置相应的参数, 便可得到与低层一样大小的特征图.

假设网络高层特征图为 $\mathbf{A}$ ($n\times n$ 矩阵), 低层特征图为 $\mathbf{B}$ ($m\times m$ 矩阵), 高层特征图 $\mathbf{A}$ ($n\times n$ 矩阵) 通过反卷积操作得到新的特征图 $\mathbf{A}'$ ($m\times m$ 矩阵). 将两个相同尺度的特征图 $\mathbf{A}'$和 $\mathbf{B}$ 进行融合, 使用元素求和方式, 即两个矩阵对应元素求和, 融合后的特征图为 $\mathbf{T}_1$:

$$\mathbf{T}_1=\mathbf{A}'+\mathbf{B}=\begin{Bmatrix} a'_{11}+b_{11} & a'_{12}+b_{12} & \cdots & a'_{1m}+b_{1m} \\ a'_{21}+b_{21} & a_{22}+b_{22} & \cdots & a'_{2m}+b_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a'_{m1}+b_{m1} & a'_{m2}+b_{m2} & \cdots & a'_{mm}+b_{mm} \end{Bmatrix} \tag{2}$$

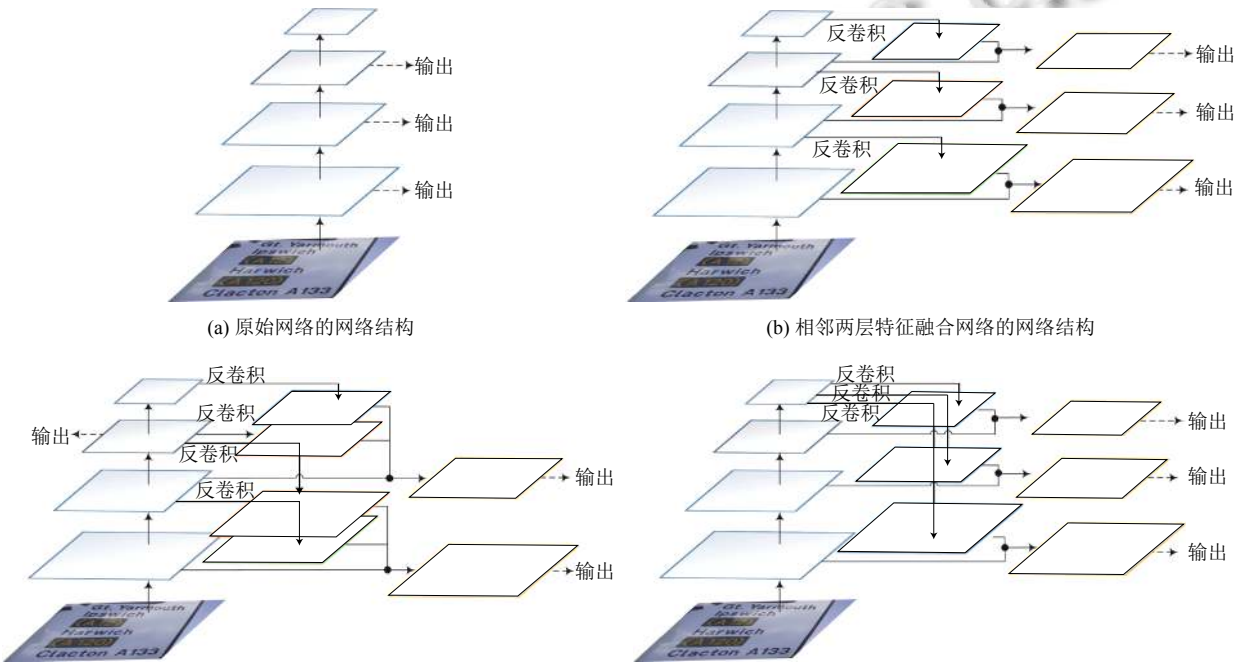
使用元素点积方式融合两个特征图, 即两个矩阵对应元素相乘, 融合后的特征图为 $\mathbf{T}_2$:

$$\mathbf{T_2} = \mathbf{A}' \cdot \mathbf{B} = \begin{Bmatrix} a'_{11}b_{11} & a'_{12}b_{12} & \cdots & a'_{1m}b_{1m} \\ a'_{21}b_{21} & a'_{22}b_{22} & \cdots & a'_{2m}b_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a'_{m1}b_{m1} & a'_{m2}b_{m2} & \cdots & a'_{mm}b_{mm} \end{Bmatrix} \quad (3)$$

研究表明[16], 点积计算能得到更好的精度, 获得更好的融合效果, 因此, 本文采用元素点积方式实现特征图的融合.

3.2 特征融合网络的结构

原始网络的输出层是网络中独立的卷积层, 网络中特征图经过卷积核计算越来越小, 特征图语义信息越来越强, 如图 2(a) 所示. 虽然, 网络的每个输出层都可以通过特征图预测文字的位置, 但是, 网络中低层输出层语义信息表达能力弱, 无法准确检测到小尺度的文本. 为了增强网络低层输出层的语义信息, 本文运用特征融合方式, 将网络高层的特征图与低层的特征图进行融合, 并提出三种特征融合网络, 分别为相邻两层特征融合网络、相邻三层特征融合网络以及最高层特征融合网络.



(a) 原始网络的网络结构

(b) 相邻两层特征融合网络的网络结构

(c) 相邻三层特征融合网络的网络结构

(d) 最高层特征融合网络的网络结构

图 2 原始网络的网络结构与特征融合网络的网络结构对比图

特征融合网络在结构上有两种连接方式, 一种是自底向上的连接方式, 一种是自顶向下的连接方式. 自底向上是网络的前向传播过程, 特征图的大小经过卷积层后会逐渐变小, 整个网络在层次结构上是金字塔结构. 自顶向下的连接采用反卷积, 将反卷积的结果与自底向上生成的相同大小的特征图进行融合. 特征融合后的网络利用高层特征的强语义信息, 提高网络低层的语义信息. 网络通过融合不同层的特征达到预测效果, 并在每个融合后的特征层上预测文字.

以 TextBoxes 中 Conv4_3 和 Conv6_2 两层特征进行融合为例, 在 Caffe 深度学习框架下, 网络的连接方式, 如图 3 所示. 低层的 Conv4_3 层, 先连接一个 1×1 的卷积层, 目的是减少特征图的通道数, 进而降低计算复杂度, 该操作并不会对特征图的大小产生影响. 高层的 Conv6_2 层经过反卷积操作后, 特征图大小与 Conv4_3 层一致. 接着对两层特征使用 BatchNorm 层对数据进行标准化, 消除数据间的量纲关系, 避免梯度更新导致数值问题, 同时可以加快收敛速度寻找最优解. 最后使用 Eltwise 层的 product 操作, 对特征图采用元素点积方式进行融合, 融合后的结果作为新的输出层, 预测文字的位置和置信度.

本文提出三种特征融合网络, 选择不同的组合方式将高层特征与低层特征进行融合. 相邻两层特征融合网络是指原始网络低层的特征图与最近邻的高层特

征图进行融合的网络, 如图 2(b) 所示, 原始网络高层的特征图经过反卷积操作, 得到与低层尺度一样的特征图, 然后两个相同尺度的特征图进行融合得到新的特征图, 网络在新的特征图上输出文字的位置.

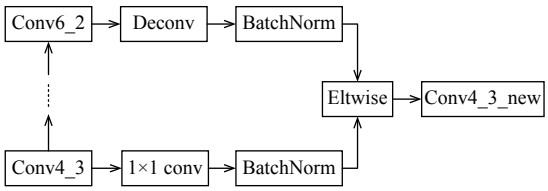


图 3 Caffe 框架下网络层的连接方式

相邻三层特征融合网络是指原始网络低层的特征图与近邻的两层特征图进行融合的网络, 如图 2(c) 所示. 其中, 近邻的两层特征图都来自于网络的高层特征图, 融合后的特征图来自于原始网络的三层特征图. 如果较高层的输出层没有两个近邻的特征层可以融合, 则输出层保持不变.

最高层特征融合网络表示原始网络中语义信息最强的特征图分别与其他输出层的特征图进行融合的网络, 如图 2(d) 所示, 新的输出层来自于低层特征与最高层特征的融合.

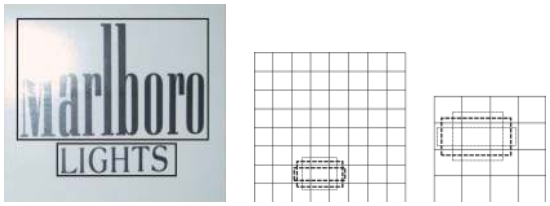
3.3 特征融合网络的采样策略

特征融合网络在训练时仅仅需要输入图像和图像中文本的真实标签框 (ground truth). 由于网络的输出是预测文本框与默认框 (default box) 的偏移坐标以及文本的置信度, 因此, 网络在训练过程中, 需要建立真实标签框和默认框之间的关系, 并对默认框进行标注.

特征融合网络在每个输出层上采用滑动窗口的模式生成默认框, $N\times N$ 的特征图有 $N\times N$ 个特征点, 每个特征点可以对应多个不同横纵比的默认框. 本文使用 jaccard 重叠率作为匹配指标对默认框进行标注, jaccard 重叠率越高表明样本相似度越高, 两个样本越匹配. 给定默认框 A 和真实标签框 B, 默认框与真实标签框的 jaccard 重叠率表示 A 与 B 的交集面积与并集面积的比值:

$$J(A,B)=\frac{A\cap B}{A\cup B} \tag{4}$$

本文将 jaccard 重叠率大于或等于 0.5 的默认框作为匹配的默认框, jaccard 重叠率小于 0.5 的默认框作为不匹配的默认框. 其中, 匹配的默认框作为正样本, 不匹配的默认框作为负样本. 如图 4(a) 所示, 文本“Marlboro”的真实标签框为图中的上方的实线框, 文本“LIGHTS”的真实标签框为图中的下方的实线框. 在图 4(b) 和 4(c) 中可以看到一些虚线框, 虚线框表示特征图上的默认框. 其中, 有两个加粗的虚线框匹配文本“LIGHTS”, 有一个加粗的虚线框与文本“Marlboro”相匹配, 因此, 标注匹配的默认框作为正样本, 不匹配的默认框作为负样本.



(a) 输出图像 (b) 8×8特征图 (c) 4×4特征图

图 4 特征融合网络的特征图

通过样本标注阶段后, 默认框中会产生大量的负样本, 这会导致正负样本的数量不均衡, 进而导致模型不稳定, 预测效果差. 为了解决该问题, 本文将默认框中的负样本通过置信度损失进行排序, 选择置信度损失值较高的默认框作为网络训练的负样本, 使训练的正负样本的比例保持在 1:3, 这样可以稳定网络的训练.

3.4 特征融合网络目标函数

特征融合网络的目标函数源自于 TextBoxes 的目标函数, 特征融合网络能处理默认框与文本的真实标签框是否匹配. 假设一张图像中存在第 i 个默认框和第 j 个真实标签框, $x_{ij}=1$ 表示第 i 个默认框与第 j 个真实标签框相匹配, 如果不匹配, 则 $x_{ij}=0$.

特征融合网络的目标损失函数是定位损失与置信度损失的加权和:

$$L(x,c,l,g)=\frac{1}{N}(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)) \tag{5}$$

其中, x 表示匹配结果矩阵, c 表示置信度, l 表示预测位置, g 表示文本的真实位置, N 表示默认框匹配真实标签框的个数; 其中, 权重系数 α 设置为 1; 定位损失 L_{loc} 是预测位置和真实位置的 L2 损失:

$$L_{\text{loc}}(x,l,g)=\frac{1}{2}\sum_{i,j}x_{ij}\|l_i-g_j\|_2^2 \tag{6}$$

置信度损失 L_{conf} 是二分类的 softmax 损失:

$$L_{\text{conf}}(x,c)=-\sum_{ij}x_{ij}\log(c_i)-\sum_i(1-\sum_j x_{ij})\log(1-c_i) \tag{7}$$

3.5 多尺度文本检测

特征融合网络在层次结构上仍然是金字塔结构, 网络在新的输出层上预测文本框的位置和置信度. 在每个输出层的特征图上定义一系列固定大小的默认框, 输出层输出文本的置信度和相对于默认框的偏移坐标. 假设图像和特征图的大小分别是 $(w_{\mathrm{im}}, h_{\mathrm{im}})$ 和 $(w_{\mathrm{map}}, h_{\mathrm{map}})$, 在特征图中 (i, j) 位置对应一个默认框 $b_0=(x_0, y_0, w_0, h_0)$, 输出层的输出为 $(\Delta x, \Delta y, \Delta w, \Delta h, c)$, 其中 $(\Delta x, \Delta y, \Delta w, \Delta h)$ 表示预测文字边界框相对于默认框的偏移坐标, c 表示文字的置信度. 预测的文字边界框为 $b = (x, y, w, h)$, 其中:

$$\begin{cases} x = x_0 + w_0 \Delta x \\ y = y_0 + h_0 \Delta y \\ w = w_0 + \exp \Delta x \\ h = h_0 + \exp \Delta y \end{cases} \tag{8}$$

x, y 表示预测的文本框的左上角的横纵坐标, w, h 为文本框的宽和高. 为了预测不同横纵比的文本边界框, 特征图上每一个特征点可以关联多个横纵比的默认框. 本文使用 6 种横纵比的默认框去预测文本边界框:

$$a_r = \{1, 2, 3, 5, 7, 10\} \tag{9}$$

此外, 由于网络中不同的输出层对应的特征图尺度不一样, 输出层可以预测不同尺度的文字. 假设网络中有 m 个输出层, 每个输出层对应一个特征图, 每个特征图中默认框的尺度为:

$$S_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1), k \in [1, m] \tag{10}$$

每个默认框的宽度和高度分别为:

$$w_k^a = S_k \sqrt{a_r} \tag{11}$$

$$h_k^a = S_k / \sqrt{a_r} \tag{12}$$

其中, $S_{\min}$, $S_{\max}$ 分别表示最低层和最高层的默认框的尺度. 从公式 (10) 可以看出, 低层输出层预测小尺度的文字, 高层输出层预测大尺度的文字.

输出层的默认框在不同的特征图上有着不同的尺度, 在同一个特征图又有着不同的横纵比, 相应的, 整个网络可以通过多个输出层预测不同尺度和不同形状的文本. 最后, 网络使用非极大值抑制算法聚集输出层输出的所有文本框, 选择置信度较高的文本框作为文本检测结果.

3.6 非极大值抑制算法

非极大值抑制算法 (Non-Maximum Suppression, NMS) 的本质是搜索局部极值点, 抑制非极大值元素, 该算法被广泛应用在目标检测的后处理中, 主要目的是排除多余的检测结果, 得到目标的最佳位置.

文本检测中普遍使用非极大值抑制算法去除冗余文本框, 因为它简单高效, 主要步骤如下:

(1) 将文本检测结果 (预测文本框) 按照置信度的值从高到低排序;

(2) 将第一个文本框作为当前抑制的文本框;

(3) 非极大值抑制. 将其他文本框作为被抑制文本框, 计算当前抑制文本框与被抑制文本框的面积交叠率 *(IOU)*. 如果交叠率高于阈值 α, 剔除该文本框.

(4) 如果只剩最后一个文本框, 则算法结束; 否则, 按照之前排列好的顺序, 取下一个未被抑制的文本框作为抑制文本框, 执行步骤 (3).

(5) 算法结束后, 选择置信度高于阈值 β 的文本框作为最终文本检测结果.

其中, 两个文本框的面积交叠率的计算方法如公式 (13) 所示, $area(A)$ 和 $area(B)$ 分别为文本框 A 和文本框 B 的面积:

$$IOU = \frac{area(A) \cap area(B)}{area(A) \cup area(B)} \tag{13}$$

使用非极大值抑制算法后, 文本检测的结果, 如图 5 所示. 图 5(a) 表示输入图像, 图 5(b) 表示通过网络检测后预测的文本框的位置及置信度, 图 5(c) 表示使用非极大值抑制算法后文本检测的最终结果.

(a) 输入图像 (b) 预测文本框 (c) 文本检测结果

图 5 使用非极大值抑制算法后文本检测结果

4 实验结果和分析

4.1 数据集

为验证网络的有效性, 本文在两个公开的场景文本检测数据集上评估网络的性能: ICDAR2011 和 ICDAR2013. 其中 ICDAR2011 数据集包含 229 张训练图像和 255 张测试图像, ICDAR2013 数据集包含

229 张训练图像和 233 张测试图像.

4.2 网络参数设置

本文的网络使用随机梯度下降 (Stochastic Gradient, SGD) 的方法训练, 其中动量 (momentum) 和权值衰减系数 (weight decay) 分别设置为 0.9 和 5×10^{-4}. 最大迭代次数为 12 万次, 学习率 (learning rate) 初始设置为 10^{-3}, 迭代 6 万次后, 学习率调整为 10^{-4}. 整个实验在深度学习框架 Caffe 平台上进行, 训练和测试图像的尺寸都为 700×700, 每个训练模型使用一个 Titan X GPU 大约训练 50 小时.

4.3 性能指标

在自然场景文本检测算法里, 涉及三个评价指标, 分别为准确率 (P)、召回率 (R) 和 F 值 (F).

准确率表示检测正确的文本框数量与算法检测出的文本框数量的比值, 召回率表示检测正确的文本框数量与数据集中真实文本框数量的比值. 准确率和召回率是一对矛盾的度量. 一般来说, 准确率高时, 召回率往往偏低; 而召回率高时, 准确率往往偏低. 所以, 准确率和召回率都不能唯一的评价算法的性能. 为了综合评价算法的性能, 一般使用准确率和召回率的调和平均数 (F 值) 来衡量算法的优劣. 准确率、召回率和 F 值, 三个评价指标的计算公式分别如公式 (14)、公式 (15)、公式 (16) 所示:

$$P=\frac{Match(G,D)}{|\mathrm{D}|} \tag{14}$$

$$R=\frac{Match(G,D)}{|\mathrm{G}|} \tag{15}$$

$$F=2\times\frac{P\times R}{P+R} \tag{16}$$

其中, $Match(G, D)$ 表示检测正确的文本框数量, D 表示算法检测出的文本框数量, G 表示数据集中真实文本框数量.

4.4 实验分析

为了确定文本检测中后处理算法 (非极大值抑制算法) 中交叠率和置信度选取的最佳阈值, 本文首先在 ICDAR2013 数据集上, 对原始网络的文本检测结果进行实验分析.

如图 6 所示, 为非极大值抑制算法中交叠率 α 和置信度 β 采用不同值进行组合下的文本检测性能. 从图中可以看出, 当交叠率 α 和置信度 β 分别取值为 0.5 和 0.6 时, 文本检测性能达到最高并趋于稳定. 因此, 本文的实验中, 非极大值抑制算法中的交叠率 α 和置信度 β 分别取值 0.5 和 0.6. 在后续的网络性能对比中, 本文均使用该阈值进行实验对比.

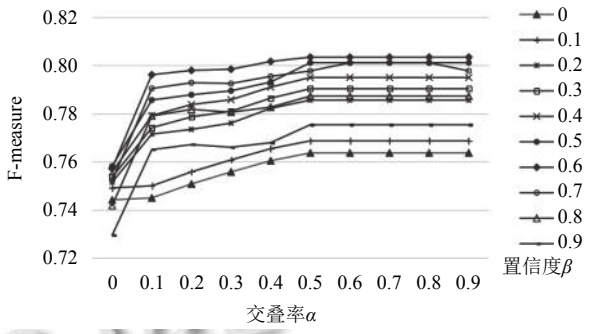


图 6 不同交叠率 α 和置信度 β 下的文本检测性能

本文提出了三个特征融合网络, 分别为相邻两层特征融合网络、相邻三层特征融合网络以及最高层特征融合网络. 本文在 ICDAR2013 数据集上验证提出的特征融合网络的性能, 在输入图像为单尺度的条件下, 与原始网络 (Fast TextBoxes)[15]进行实验对比.

如表 1 所示, 本文提出的三个特征融合网络中, 相邻两层特征融合网络和最高层特征融合网络在 F 值上分别得到 2% 和 3% 的提升, 而相邻三层特征融合网络的 F 值与 Fast TextBoxes 相比下降 1%.

表 1 原始网络与特征融合网络实验对比结果

方法	准确率	召回率	F 值	时间 (s)
Fast TextBoxes[15]	0.86	0.74	0.80	0.09
相邻两层特征融合网络	0.85	0.80	0.82	0.10
相邻两层特征融合网络	0.82	0.76	0.79	0.11
最高层特征融合网络	0.86	0.81	0.83	0.11

此外, 本文的方法与 Fast TextBoxes 相比, 在召回率上提升较高, 三个特征融合网络在召回率上分别提升了 6%、2% 和 7%. 这是因为特征融合后, 网络低层输出层的特征图的语义信息得到增强, 能准确预测出小尺度的文字, 总体的召回率得到提升. 如图 7 所示, 原始网络 (Fast TextBoxes) 对于检测小尺度文字并不理想, 不能准确检测出小尺度文字, 而本文采用不同层特征图进行融合的方式, 能有效地检测出小尺度文字.

从时间性能上比较, 本文提出的特征融合网络在时间性能上与原始网络相比存在微小的差异, 微小的差异来源于特征融合中反卷积的计算, 但并不影响现实应用.

(a) 原始网络实验结果 (b) 特征融合网络实验结果

图 7 原始网络和特征融合网络实验结果对比

相邻三层特征融合网络与相邻两层特征融合网络相比较, 在准确率和召回率上均有所下降. 此外, 在训练过程中, 多层特征进行融合存在计算量大、消耗内存的情况, 因此本文没有采用三层以三层以上的特征融合网络.

本文所提出的三种特征融合网络中, 最高层特征融合网络的性能最好. 由于最高层的语义信息比较强, 高层的语义特征融合至其他层后, 使网络在各个层级上都具有丰富的语义, 性能上取得显著的提升, 并且不牺牲速度和内存. 因此, 之后的实验中, 本文使用最高层特征融合网络作为最佳的特征融合网络, 与常用的自然场景文本检测方法进行比较.

表 2 和表 3 分别展示了最高层特征融合网络与其他方法在 ICDAR2011 和 ICDAR2013 数据集上的实验结果. 从表中可以看出, 本文的方法在 ICDAR2011 和 ICDAR2013 数据集上, F 值都达到 0.83, 比原始网络 (Fast TextBoxes) 的 F 值的提高了 3%, 比之前最好的方法提高了 2%. 本文方法最大的优势在于召回率得到显著的提升, 在 ICDAR2011 数据集上, 本文方法比之前最好的方法 Text Flow 在召回率上提升了 4%; 在 ICDAR2013 数据集上, 本文方法比之前最好的方法 FCN 在召回率上提高了 5%, 这主要因为小尺度文本检测的召回率得到提升. 综上所述, 本文的方法相比于之前的方法, 能有效地检测出小尺度文本, 文本检测的整体性能有显著的改善.

由上述实验结果可知, 本文方法在自然场景文本检测上能够有效地检测出文字的位置. 图 8 展示了使用本文的最高层特征融合网络检测文本成功和失败的图例. 检测成功的图例 (图 8(a)) 显示出本文方法具有较高的定位准确性和鲁棒性, 能有效地从复杂背景中检测出不同大小和不同形状的文字. 对于检测失败的图例 (图 8(b)), 图像中的文字极其模糊或者文字与背景具有较低的对比度, 即使人眼也很难识别出图像中的文字区域.

表 2 在 ICDAR2011 数据集上的实验结果

方法	准确率	召回率	F 值
SFT-TCD[17]	0.82	0.75	0.73
Yin et al.[18]	0.86	0.68	0.76
MSERs-CNN[10]	0.88	0.71	0.78
Zhang et al.[19]	0.84	0.76	0.80
Fast TextBoxes[15]	0.86	0.74	0.80
Text Flow[20]	0.86	0.76	0.81
最高层特征融合网络	0.86	0.80	0.83

表 3 在 ICDAR2013 数据集上的实验结果

方法	准确率	召回率	F 值
Text Spotter[21]	0.88	0.65	0.75
Iwrr2014[22]	0.86	0.68	0.76
Text Flow[20]	0.88	0.71	0.78
Zhang et al.[19]	0.84	0.76	0.80
Fast TextBoxes[15]	0.86	0.74	0.80
FCN[23]	0.86	0.76	0.81
最高层特征融合网络	0.86	0.81	0.83

(a) 检测成功示例图

(b) 检测失败示例图

图 8 本文方法检测文本示例图

5 结论与展望

本文提出了一种基于特征融合的深度神经网络, 该网络将高层特征与低层特征相融合, 利用网络高层的强语义特征增强低层输出层的语义信息, 使整个网络的输出层都具有较强的表达能力. 特征融合后的网络能在不同的输出层上预测不同尺度以及不同形状的

文字. 本文在两个公开的数据集上验证了特征融合网络的性能, 实验结果表明本文提出的特征融合网络对小尺度的文字, 定位效果显著. 其中, 本文提出的最高层特征融合网络能取得最佳的检测效果, 具有较高的定位准确性和鲁棒性, 并优于常用的自然场景文本检测方法, F 值在 ICDAR2011 和 ICDAR2013 两个数据集上均达到了 0.83. 本文的特征融合网络只支持单尺度的图像输入, 在一定程度上限制算法性能的提升. 因此, 下一步的工作, 我们将尝试把改进后的网络改为多尺度输入的网络. 网络将会从以下两方面进行修改, 一方面是改变网络中卷积层的卷积核大小, 建立输出层中不同大小的特征图之间的整体关联性, 使网络能支持多尺度图像输入. 另一方面, 使用其他方式放大高层的特征图, 例如, 反池化操作, 即记录池化过程中最大激活值所在的坐标位置, 然后上采样得到放大的特征图, 使网络中融合的特征图能自适应进行变化而不依赖于固定计算. 接下来的工作, 我们将尝试用这两种方法, 进一步提高网络的性能.

参考文献

1 陈利. 车牌识别系统设计与实现. 现代电子技术, 2012, 35(15): 142–144.

2 胡二雷, 冯瑞. 基于深度学习的图像检索系统. 计算机系统应用, 2017, 26(3): 8–19.

3 王琦, 陈临强, 梁旭. 视频中的字幕提取. 计算机工程与应用, 2012, 48(5): 177–178, 216.

4 Ozuysal M, Fua P, Lepetit V. Fast keypoint recognition in ten lines of code. Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA. 2007. 1–8. [doi: 10.1109/CVPR.2007.383123]

5 Lee JJ, Lee PH, Lee SW, *et al*. AdaBoost for text detection in natural scene. Proceedings of 2011 International Conference on Document Analysis and Recognition. Beijing, China. 2011. 429–434. [doi: 10.1109/ICDAR.2011.93]

6 Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 2963–2970. [doi: 10.1109/CVPR.2010.5540041]

7 Neumann L, Matas J. A method for text localization and recognition in real-world images. Proceedings of the 10th Asian Conference on Computer Vision. Queenstown, New Zealand. 2010. 770–783. [doi: 10.1007/978-3-642-19318-7_60]

8 易尧华, 申春辉, 刘菊华, 等. 结合 MSCRs 与 MSERs 的自然场景文本检测. 中国图象图形学报, 2017, 22(2): 154–160. [doi: 10.11834/jig.20170202]

9 Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 512–528

10 Huang WL, Qiao Y, Tang XO. Robust scene text detection with convolution neural network induced MSER trees. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 495–511. [doi: 10.1007/978-3-319-10593-2_33]

11 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

12 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]

13 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

14 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands. 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

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

16 Fu CY, Liu W, Ranga A, *et al*. DSSD: Deconvolutional single shot detector. arXiv:1701.06659, 2017.

17 Huang WL, Lin Z, Yang JC, *et al*. Text localization in natural images using stroke feature transform and text covariance descriptors. Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia. 2013. 1241–1248. [doi: 10.1109/ICCV.2013.157]

18 Yin XC, Yin XW, Huang KZ, *et al*. Robust text detection in natural scene images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970–983. [doi: 10.1109/TPAMI.2013.182]

19 Zhang Z, Shen W, Yao C, *et al*. Symmetry-based text line detection in natural scenes. Proceedings of 2015 IEEE

Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 2558–2567.

20 Tian SX, Pan YF, Huang C, *et al*. Text flow: A unified text detection system in natural scene images. Proceedings of 2015 International Conference on Computer Vision. Santiago, Chile. 2015. 4651–4659.

21 Neumann L, Matas J. Real-time scene text localization and recognition. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2012. 3538–3545. [doi: 10.1109/CVPR.2012.6248097]

22 Zamberletti A, Noce L, Gallo I. Text localization based on fast feature pyramids and multi-resolution maximally stable extremal regions. In: Jawahar CV, Shan SG, eds. Computer Vision - ACCV 2014 Workshops. Cham: Springer, 2014. 91–105. [doi: 10.1007/978-3-319-16631-5_7]

23 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. [doi: 10.1109/CVPR.2016.451]