中国科学 信息科学
SCIENTIA SINICA Informationis

基于深度学习的场景文字检测与识别

白翔, 杨明锟, 石葆光 and 廖明辉

Citation: 中国科学: 信息科学 **48**, 531 (2018); doi: 10.1360/N112018-00003

View online: http://engine.scichina.com/doi/10.1360/N112018-00003

View Table of Contents:http://engine.scichina.com/publisher/scp/journal/SSI/48/5

Published by the 《中国科学》杂志社

中国科学: 信息科学 2018 年 第 48 卷 第 5 期: 531–544
SCIENTIA SINICA Informationis
人工智能优青专刊 · 论文

基于深度学习的场景文字检测与识别

白翔*, 杨明锟, 石葆光, 廖明辉

华中科技大学电子信息与通信学院, 武汉 430074
* 通信作者. E-mail: xbai@hust.edu.cn

收稿日期: 2018–01–02; 接受日期: 2018–03–12; 网络出版日期: 2018–05–11

国家自然科学基金 (批准号: 61733007, 61222308, 61573160) 和数字出版技术国家重点实验室开放课题 (批准号: F2016001) 资助项目

摘要 场景文字检测与识别是一种通用文字识别技术, 已成为近年来计算机视觉与文档分析领域的热点研究方向. 其被广泛应用于地理定位、车牌识别、无人驾驶等领域. 相对于传统的文档文字检测和识别, 场景文字在字体、尺度、排布、背景等方面变化更加剧烈, 深度学习技术也由于卓越的性能成为该领域的主流方法. 本文主要回顾了作者基于深度学习在此领域取得的代表性成果, 并对此领域未来研究趋势进行了展望.

关键词 深度学习, 场景文字, 文字检测, 文字识别, 计算机视觉

1 引言

文字, 在自然界中无处不在, 是人类间信息传递与交互的主要途径之一. 近年来, 场景 OCR (optical character recognition) 技术, 即从图像中检测与识别文字已成为计算机视觉、文档分析等领域的热点研究方向, 得到了来自学术界与工业界的强烈关注. 做为一种通用技术, 场景 OCR 无需对特殊场景进行定制, 能够识别任意场景图片中的文字, 如交通标识、屏幕、票据、街景及商品等. 场景 OCR 目前已广泛应用于证件照识别、票据识别、信息内容安全审核等方面, 具有极其重要的研究与应用价值. 最近, 深度学习技术已经在 OCR 领域里发挥了主导作用, 基于深度学习的 OCR 技术在文字识别的精度和效率两个方面都取得了显著的提升. 基于深度学习的 OCR 技术最先由 University of Oxford VGG 研究组率先提出, 即刻就引起了国内外同行的高度关注. 鉴于此, 本文重点介绍作者在此领域基于深度学习框架最近所取得的几个代表性成果, 并对此领域的发展趋势和有潜力的研究方向进行了展望, 希望有助于对深度学习及 OCR 技术感兴趣的读者.

引用格式: 白翔, 杨明锟, 石葆光, 等. 基于深度学习的场景文字检测与识别. 中国科学: 信息科学, 2018, 48: 531–544, doi: 10.1360/N112018-00003
Bai X, Yang M K, Shi B G, et al. Deep learning for scene text detection and recognition (in Chinese). Sci Sin Inform, 2018, 48: 531–544, doi: 10.1360/N112018-00003

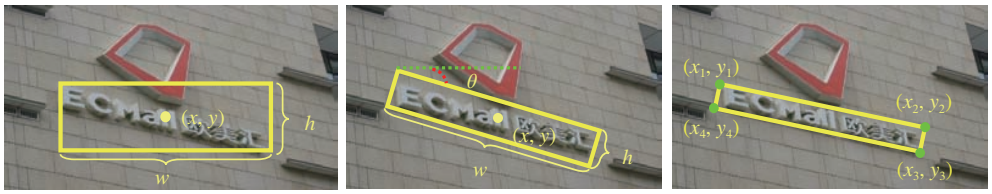

图 1 (网络版彩图) 包围盒的不同表示方式

Figure 1 (Color online) Visualization of different detection targets

2 场景文字检测与识别的定义

场景文字相关研究包含多个子问题. 其中, 文字检测和文字识别是最核心最基本的两个任务.

文字检测. 是从整幅的输入图像中定位文字的位置. 文字的位置由包围盒表示. 一般情况下, 水平的包围盒由 4 个矩形的坐标 (x, y, w, h) 组成. 在一些问题中还需要检测多方向的包围盒, 此时包围盒通常由 (x, y, w, h, θ) 5 个参数或者 $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ 8 个参数表示, 其中 θ 是包围盒和水平方向的角度, (x_i, y_i) $(i = 1, 2, 3, 4)$ 是四边形的 4 个顶点坐标, 如图 1 所示. 文字定位的挑战主要来自于自然场景的复杂的变化. 环境中存在许多视觉元素和文字相像, 例如砖墙、栅栏等物体具有和文字类似的纹理. 此外, 由于需要检测的往往是单词, 文字检测算法往往还涉及连接字母组成单词 (连词), 或者将整行文字分成单词 (分词) 的算法.

文字识别. 识别问题是从只包含文字的图像中识别出机器可读的字母序列. 该问题的难点之一在于其输出空间是长度不定的序列. 而一般的图像分类问题里, 输出空间的维度是固定的. 此外, 字体、光照、颜色、尺度的等问题也给识别造成了困难.

在实际应用中, 场景文字的检测和识别往往需要串联在一起使用. 一般先由检测器检测到文字的位置, 在这些位置上识别出文字内容. 能同时检测到文字位置并对其进行识别的方法被称作是端到端文字识别方法.

关于场景文字检测与识别的具体定义及评价方法可参考最近的两篇综述论文 [1, 2], 本文不再详细说明.

3 深度学习方法归类

3.1 检测

相比于传统的光学字符识别 [3], 自然场景文字图像的前景文字和背景物体的变化很大, 光照情况也相当复杂. 因此检测自然场景图像中的文字更具挑战.

根据检测的直接目标的不同, 以前的自然场景文字检测方法可以粗略地分为 3 类. (1) 基于局部文字的方法: 这类方法 [4~6] 首先检测字符或者文字的局部, 然后将它们聚合成单词. 其中一个代表性的方法便是文献 [5], 它首先通过分类极值区域提取出文字的局部区域, 然后使用穷举搜索的方法将它们组合起来, 形成单词区域的包围盒. (2) 基于单词的方法: 这类方法 [7~9] 通过类似于目标检测中的方法 [10~13], 它们将单词视为物体, 直接输出单词区域的包围盒. 在代表性文献 [8] 中, 作者们使用了一个基于 R-CNN [12] 的检测框架. 首先, 通过候选区域产生器产生大量的单词候选区域; 然后, 用一个随机森林分类器过滤掉部分单词候选区域; 最后, 使用一个卷积神经网络回归候选区域的精确的包围盒. (3) 基于文本行的方法: 这类方法 [14, 15] 首先检测出文本行, 然后将文本行分词得到单词的包围盒. 其

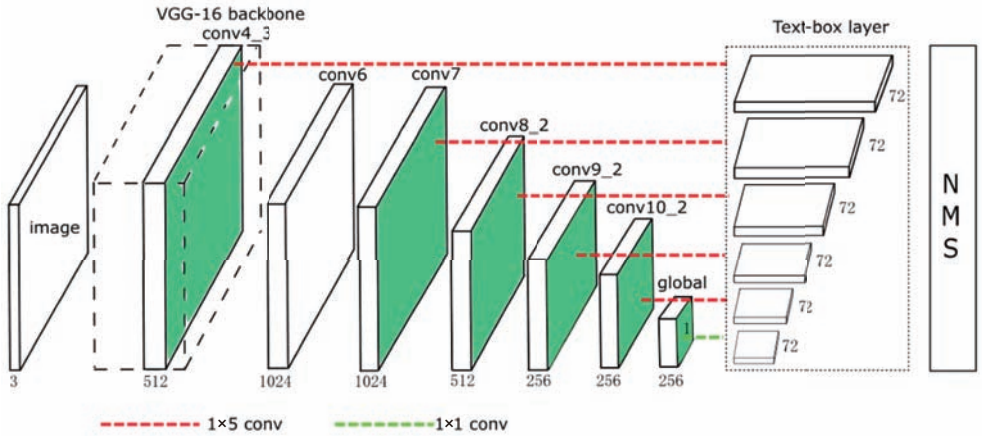

图 2 (网络版彩图) TextBoxes 的网络框架示意图 [7]: 由一个简洁的全卷积网络和标准的非最大值抑制操作组成

Figure 2 (Color online) Architecture of TextBoxes: a simple fully convolutional network and a standard non-maximum suppression

中的代表性工作是文献 [14], 它使用全卷积网络 [16] 得到文本行的分割结果, 再对文本行进行分词.

接下来, 介绍我们提出的两种自然场景文字检测的方法. 它们分别属于基于单词的方法和基于局部文字的方法.

3.1.1 基于文字框的端到端的深度神经网络文字检测器

传统的自然场景文字检测方法 [5, 8] 倾向于采取多个步骤, 比如 (1) 提取出候选的字符或单词区域; (2) 过滤候选区域; (3) 改善候选区域. 由于步骤较多, 需要花费很大的精力去调整参数和设计复杂的规则, 使每个部分能够协调工作, 这同时也限制了检测速度. 于是, 受最近的目标检测工作 [11] 的启发, 我们提出了一个端到端可训练的自然场景文字检测器, 被称作 TextBoxes[1)] [7]. 它仅需要一个网络的前向传播步骤和一个标准的非最大值抑制操作, 无需复杂的后处理操作, 就能快速准确地在自然场景图片中检测出文字区域.

TextBoxes 采用全卷积网络结构 [17], 通过预测文字包围盒的置信度和该包围盒与默认包围盒 (预先设计的初始包围盒) [11] 的坐标偏置, 直接在多个特征层输出单词包围盒的坐标信息. 由于传统的卷积神经网络的卷积核尺寸是正方形的, 所以其对应的感受野也是正方形, 但是文字区域一般呈长方形. 针对文字的这个特点, 我们对网络进行了针对性的设计. 一方面, 采用了细长型的卷积核来取代传统的正方形的卷积核, 以此使得网络的感受野更适合于检测文字; 另一方面, 也调整了默认包围盒的长宽比, 使用了许多长宽比很大的默认包围盒, 这使得默认包围盒更接近于真实的文字包围盒, 减小了网络的回归难度. 更具体地, 如图 2 所示, 它在多个尺度的特征层上 (绿色特征层) 的每个位置, 通过细长型的卷积预测出一个 72 维的向量, 该向量表示 12 个默认包围盒对应的文字框的置信度 (二维: 文字的概率和背景的概率) 和坐标偏置 (四维: 包围盒的中心坐标、长和宽的偏置). 最后, 对得到的所有包围盒进行非最大值抑制操作, 滤去重叠度比较大的包围盒, 得到检测结果.

得益于针对文字图像精心设计的卷积尺寸和默认包围盒, TextBoxes 在多个权威数据集上 (包括 ICDAR 2011 [18], ICDAR 2013 [19] 等) 性能优越. 实验表明, 它兼具高性能和高效率的优势. 与此同时, 得益于端到端的网络结构, TextBoxes 简单实用, 无需繁琐的手动调参.

1) https://github.com/MhLiao/TextBoxes.

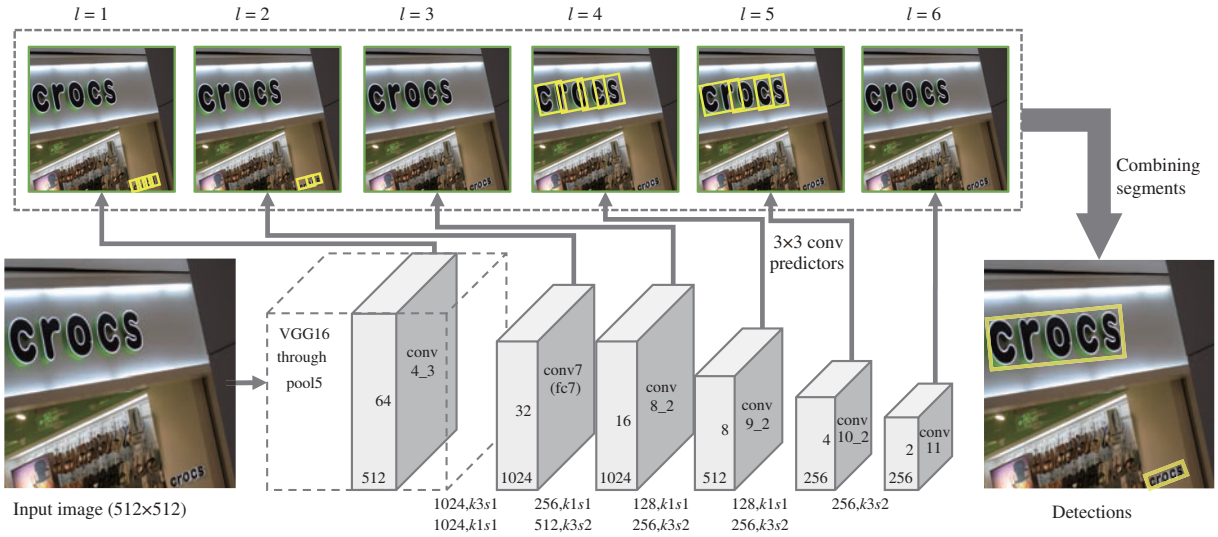

图 3　(网络版彩图) SegLink 的网络框架示意图 [20]. 特征层之间的卷积滤波器用 "(滤波器维度), k (卷积核尺寸), s (卷积步长)" 来表示. 首先, 文字片段 (黄色包围盒) 及其连接关系 (未显示) 通过多个特征层预测 (以 l 为编号); 然后, 通过一个结合算法将它们联结为一个单词的包围盒

Figure 3　(Color online) Architecture of SegLink. Convolutional filters between the feature layers (some have one more convolutional layer between them) are represented in the format of "(#filters), k (kernel size), s (stride)". Segments (yellow boxes) and links (not displayed) are detected by the predictors on multiple feature layers (indexed by l), then combined into whole words by a combining algorithm

3.1.2　基于文字片段的文字检测器

大多数的自然场景文字检测方法是针对拉丁文字, 比如英文、法文等, 它们的相邻单词之间被可见的空白区域隔开. 然而, 对于非拉丁文字, 比如中文、日文等, 它们的相邻单词之间往往没有可见的间隔, 只能通过语义来进行分词. 因此, 非拉丁文在图像中的包围盒一般是以文本行的形式来表达的. 由于普通卷积神经网络的感受也是有固定的范围, 直接检测具有极端的长宽比的文本行包围盒对基于卷积神经网络的文字检测器来说是一个极大的挑战. 然而, 由于文字的片段的长宽比是有限的, 能够容易地被普通的卷积神经网络检测出来. 因此, 我们提出了一个使用片段及其连接关系的多方向文字检测器, 称之为 SegLink[2) [20]].

SegLink 的核心思想是不直接检测整个单词或者文本行, 而是先检测单词或者文本行的局部区域, 再将这些局部区域连接起来形成一个完整的单词或者文本行. 它将文字检测任务分解成两个子任务: 检测文字片段和预测片段之间的连接关系. 其中, 片段是具有方向的矩形包围盒, 它们覆盖着单词或者文本行的一部分; 片段之间的连接关系是指两个片段是否属于同一个单词或文本行. 更具体地, 如图 3 所示, 一个基于文献 [11] 的网络被用来多尺度、密集地预测文本片段及其连接关系, 其中片段之间的连接关系包括层内连接和跨层连接, 以适应多尺度的文字. 最后, 一个基于深度优先搜索的结合算法用来处理所有的片段及其连接关系, 将属于同一单词或者文本行的片段连接合并成单词或者文本行的包围盒.

SegLink 具有极强的通用性, 既能用于拉丁文字也能用于非拉丁文字, 而且不受限于文字目标的长度和方向. 与此同时, 它兼具高性能与高效率, 能满足实际应用的需求.

2) https://github.com/bgshih/seglink.

表 1 ICDAR 2013 数据集的文字检测结果 a)

Table 1 Text detection results on ICDAR 2013 dataset

Method	Accuracy (%)	Recall (%)	F-measure (%)	Frames per second
Zhang et al. [15]	88	74	80	<0.1
Zhang et al. [14]	88	78	83	<1
Jaderberg et al. [8]	88.5	67.8	76.8	<1
Tian et al. [21]	**93.0**	**83.0**	**87.7**	7.1
TextBoxes [7]	88.0	74.0	81.0	11.1
TextBoxes Multi-scale [7]	89.0	**83.0**	86.0	1.4
SegLink [20]	87.7	**83.0**	85.3	**20.6**
SSTD* [22]	89.0	86.0	88.0	7.7
Wordsup* [23]	93.3	87.5	90.3	2
He et al.* [24]	92.0	81.0	86.0	1.1

a) Our methods and the state-of-the-art results are highlighted in bold.

表 2 ICDAR 2015 数据集的文字检测结果 a)

Table 2 Text detection results on ICDAR 2015 incidental text dataset

Method	Accuracy (%)	Recall (%)	F-measure (%)
HUST_MCLAB	47.5	34.8	40.2
NJU_Text	72.7	35.8	48.0
StradVision-2	**77.5**	36.7	49.8
Zhang et al. [14]	70.8	43.0	53.6
SegLink [20]	73.1	**76.8**	**75.0**
EAST* [25]	83.3	78.3	80.7
SSTD* [22]	80.0	73.0	77.0
Wordsup* [23]	79.3	77.0	78.2
He et al.* [24]	82.0	80.0	81.0

a) Our methods and the state-of-the-art results are highlighted in bold.

3.1.3 性能评估

如表 1 和 2 所示, TextBoxes 和 SegLink 的性能和效率在 ICDAR 2013 [19] 和 ICDAR 2015 [26] 两个标准数据集上领先于同时期的其他算法. 表 1 和 2 中带有 “*” 的为之后发表的工作. 他们在回归包围盒的表示、网络的结构等方面做出了有效的改进, 进一步提升了检测的性能.

3.2 识别

文字识别是从文字图片中识别文字序列的过程. 通常, 输入是用检测到的文字框裁剪出的图片, 目标输出是该检测框中所包含的文字序列. 文字识别是自然场景文字检测识别系统的重要一环, 其性能决定了系统的总体性能. 和检测问题一样, 识别问题同样也会受到自然场景中的嘈杂背景、光照不均、字体变化、文字排布不规律等问题的挑战. 此外, 和一般的图片识别问题不同, 文字识别的输出是长度不确定的序列, 而非维度固定的标签.

由于该问题的应用价值以及挑战, 近年来有大量的方法和系统被提出. 具有代表性的工作有文

献 [27~32]. 先前的方法大致可以分为两类, 一类是基于字符的识别方法, 另一类是基于整词的识别方法. 基于字符的识别方法一般由字符检测、字符识别, 以及字符组合 3 个步骤组成. 该类方法的模块较多, 模型训练较为复杂. 同时也需要大量的字符级别的标注数据, 获取这些数据耗费大量的人力.

基于整词的识别方法将单词作为整体去识别, 跳过了字符级别的检测和识别步骤. 该类方法往往从单词图片中提取全局特征, 并且对目标单词进行向量表示. 识别被建模成从全局特征到目标单词向量的回归过程. 此类方法的代表性工作是 Jaderberg 等 [8] 提出的方法. 他们用深度卷积神经网络对单词直接分类. 英文中大约有 9 万个常用单词, 每个都被作为一个单独的类别. 该网络的输入是整张单词图片, 输出是单词的类别概率向量, 其中概率最高的类别就对应了所识别的单词. 基于整词识别的网络识别准确率高, 但模型复杂度较高, 需要大量的学习样本.

基于对已有方法的观察和总结, 我们提出了基于序列的文字识别方法, 接下来会逐个介绍. 此类方法基于深度卷积网络, 具有精确的识别性能. 但同时避免了过高的模型复杂度, 并且又能够识别任意的、未在字典中出现的字母或数字序列, 有很强的实用价值.

3.2.1 基于图像的文字序列识别算法

我们提出了一种新型的深度神经网络模型, 名为卷积循环神经网络 (convolutional recurrent neural network, CRNN)[3] [33]. 该网络的结构如图 4 所示, 网络在底端接收图像输入, 经过多层的卷积神经网络进行特征提取及抽象, 得到丰富的卷积特征. 这一步被称为特征提取. 接下来, 该特征图被转换成特征序列, 特征图的每一列被从左至右地提取出, 形成一段特征序列. 该序列的长度等同于特征图的宽度. 序列中每个向量的维度等于特征图的高度乘以深度.

由于卷积神经网络特征的局部性, 特征提取得到的特征序列按照从左至右的顺序描述了输入图像的一个个局部区域. 为了增加这些区域的描述能力, 采用长短程记忆网络 (long-short term memory, LSTM) [34] 对特征序列进行了分析. LSTM 是常用的循环网络结构, 其内部结构如图 5 所示. 它具备记忆过往输入的特征, 因此可以捕捉到序列中单个方向上的长距离相关性. 我们设置了两个方向相反的 LSTM, 并且将它们的输出合并, 得到一个双向长短程记忆网络 (bidirectional LSTM). 这个网络可以同时分析自左向右以及自右向左两个方向上的长距离相关性, 从而让其输出的特征向量包含了丰富的上下文信息. 这个步骤被称为序列分析.

最后, 双向长短程记忆网络输出的特征向量被用来识别每个序列帧上的符号. 该符号包含所有需要识别的字母, 以及一个特殊的 “空白” 符号 (接下来用 “-” 指代). 在序列解码步骤中, 重复出现的字母会被合并成为单个字母, 然后所有 “空白” 符号会被去除. 例如 “-s-t-aattee-” 经过解码后会得到单词输出 “state”.

卷积循环网络是一个端到端 (end-to-end) 的网络模型. 它将特征提取、序列分析、序列解码 3 个算法模块, 作为不同的网络层集成在了同一个网络中. 在测试时, 该模型接收图像作为输入, 在顶端直接输出字母序列. 同时, 该模型的训练只需要图片和对应的单词标注即可, 不需要字符级别的标注等.

我们将该模型在多个国际公开标准数据集上进行的测试, 其识别准确率在大多数指标都达到了最高. 识别结果在表 3 中汇总. 其中, 在 IIIT5k [27], SVT [35], ICDAR2003 [36] 上, CRNN 的无词库准确率分别达到 81.2%, 82.7%, 91.9%, 领先或接近同时期的其他基于深度神经网络的方法. 有词汇表准确率全部超过 95%, 接近完全正确. 此外, CRNN 还具有模型参数少的优势. 它总共有大约 8.3 M 个参数, 远远少于文献 [8] 提出的方法 (490 M 个参数). 由于 CRNN 具有优异的识别精确度, 且易于训练和部

3) https://github.com/bgshih/crnn.

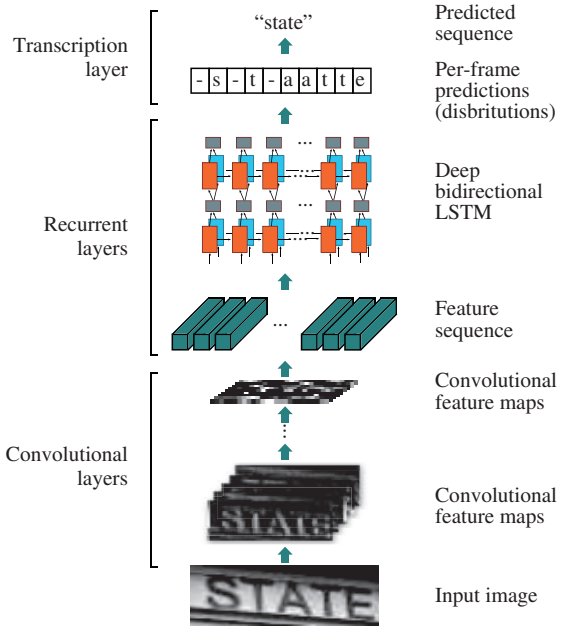

图 4 (网络版彩图) 卷积循环神经网络结构示意图 [33]

Figure 4 (Color online) Network structure of CRNN

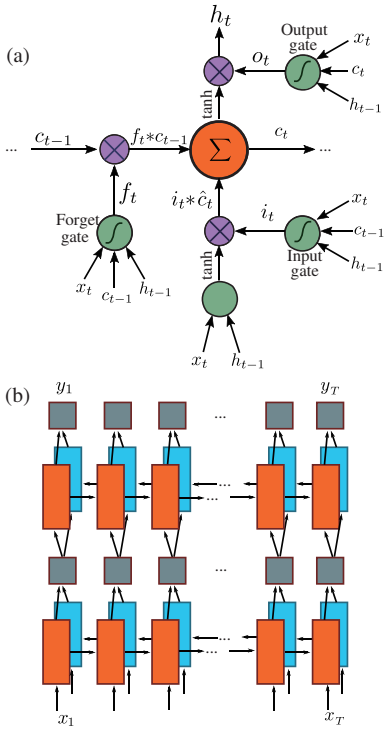

图 5 (网络版彩图) (a) 长短程记忆结构与 (b) 双向长短程记忆网络 [33]

Figure 5 (Color online) (a) Structure of LSTM and (b) bidirectional LSTM network

表 3 CRNN 在不同数据集、词汇表上的识别结果 a)

Table 3 Recognition results on different datasets with different lexicons

Method	Lexicon (%)							
	IIIT5k [27]			SVT [35]		ICDAR2003 [36]		ICDAR2013 [19]
	50	1k	None	50	None	50	None	None
Bissacco et al. [32]	–	–	–	90.4	78.0	–	–	87.6
Bai et al. [29]	85.6	72.7	–	81.0	–	90.3	–	–
Jaderberg et al. [8]	97.1	92.7	–	95.4	80.7	**98.7**	**93.1**	**90.8**
CRNN	**97.8**	**95.0**	**81.2**	**97.5**	**82.7**	**98.7**	91.9	89.6

a) The state-of-the-art results are highlighted in bold.

署, 它具有很强的实用性.

3.2.2 基于图像的不规则文字序列识别算法

自然场景中, 形状不规则的艺术文字十分常见. 图 6 展示了两种常见情形: 视角扭曲和曲形排布. 这两种情形都会对识别造成困难. 为了应对该挑战, 我们提出将文字图像先做矫正, 得到规则的文字之后再进行识别. 具体地, 我们提出一个结合了矫正和识别两项功能的神经网络 [37], 该网络包含矫正网络和识别网络两个部分. 矫正网络基于 Jaderberg 等 [38] 提出的空间变换网络实现. 网路的结构如图 7

图 6 (网络版彩图) 不规则文字

Figure 6 (Color online) Samples of irregular text. (a) Perspectively distorted text; (b) curved text

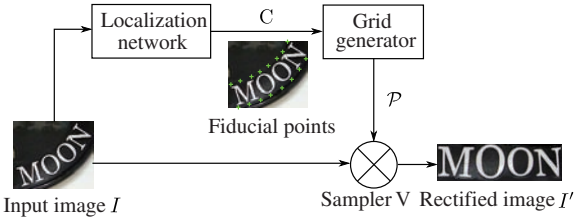

图 7 (网络版彩图) 文字矫正网络结构图 [37]

Figure 7 (Color online) Structure of the text rectifying network

所示, 它由定位网络、网格产生器、采样器 3 个部分组成. 定位网络会在文字的上下边沿定位固定数量的 "控制点", 这些控制点描述了文字的排布形状. 接下来, 网格产生器会根据控制点的位置计算采样点的位置. 最后, 采样器根据采样点位置对图片进行重新采样, 得到一张校正后的图片. 识别网络则可以由任意可导的端到端网络实现, 例如前述的 CRNN. 实际中我们采用了一种序列到序列的聚焦模型来实现文字的识别. 由于矫正网络和识别网络都是端到端可训练的网络, 可以将它们串接起来, 形成一个整体的端到端可训练的网络. 矫正网络可以解决不规则文字识别问题. 串联了矫正网络的识别器在多个标准数据集上都达到了更高的识别精确度. 特别是在非规则文字数据集 SVT-Perspective [39] 和 CUTE80 [40] 上, 该方法相比无矫正方法取得了约 5% 的显著结果提升. 并且, 矫正网络不需要额外的人工标注, 仅仅通过和识别网络的端到端训练就能够自动地学习出文字的矫正方法.

4 应用: 文字信息帮助图片细粒度分析

文字识别曾广泛应用于文档图像和数字合成图像, 比如文件扫描、名片扫描, 以及银行卡号码识别等. 随着深度学习的复兴, 自然场景下的文字识别得到了巨大的进步. 包含在自然场景图像中的文字的高层语义信息使其在网络安全、场景理解、地理位置定位、机器人导航、人机交互和智能交通等领域有着重要的应用前景. 接下来, 本文将主要介绍文字识别在细粒度图像分析方面的应用. 区别于通用图像分析任务, 细粒度图像分析 (fine-grained image analysis) 的类别更加精细, 它要求模型能够对视觉相似度极高的同一大类下的不同子类物体进行区分. 这些细粒度图像的差异往往只体现在细微

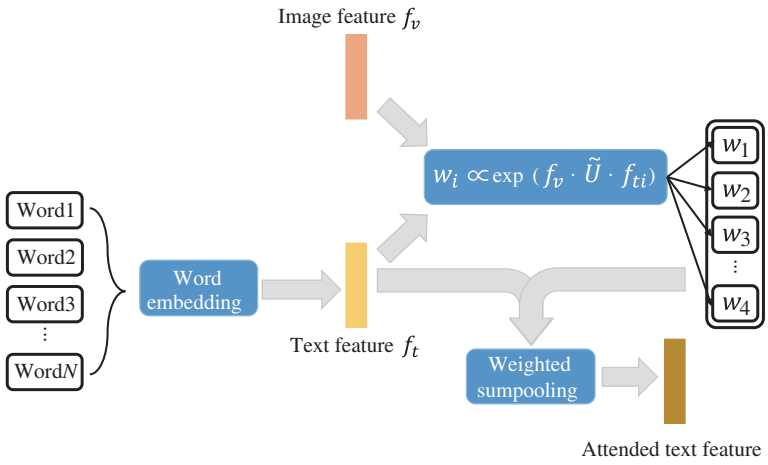

图 8 (网络版彩图) 注意力机制的框架示意图 [44]

Figure 8 (Color online) Structure of the proposed attention model

之处, 因此传统的细粒度分析模型主要集中在发现图像中有区分度的局部区域信息 [41~43]. 但是这些方法往往只利用了图像区分度不大的视觉信息. 对于某些细粒度图像类别, 比如店铺或者商品的子类 (比如饮料、化妆品等), 这类图片中包含大量和图片类别紧密相关的文字, 并且其蕴藏的丰富语义相对于视觉特征更具区分性. 基于这个现象, 我们尝试利用图像中的文字信息更好地进行这类细粒度图像分析, 提出了一个新型网络模型, 名为 FIAT (fusion of image features and attended text features) [44].

首先利用开源的文字检测算法 TextBoxes [7] 和文字识别算法 CRNN [33] 来提取图像中的文字. 然后利用词嵌入 (word embedding) 的方式将提取的文字编码成向量, 作为每个词的特征表示. 由于每张图片中的文字数目不一定相同, 所以通过平均的方式将一张图片中所有文字的词向量聚合起来, 从而得到该图片的文字特征表示. 对于图像的视觉特征表示, 采用了广泛使用的 GoogLeNet [45] 对其进行提取. 接着, 将文字特征和视觉特征串联融合进行最后的图像分类. 实验表明, 在结合文本信息之后, 许多细粒度分类或者检索任务都能得到很好的提升.

上文中提到的平均的方法让所有文字对于图像正确识别拥有同样的重要性, 然而我们注意到, 由于检测识别算法的一些缺陷, 一些图片中的文字会被错误识别, 或者正确得到的文字中, 有些与图像的类别相关性不大. 也就是说不同文字对图像最终正确分类的贡献度应该是不一样的, 错误的或者相关度小的文字权重应该小一些, 反之, 正确的并且相关度大的文字权重应该大一些. 所以我们引入了在机器翻译、图像说明中广泛使用的注意力机制来计算每个文字的权重, 最后将所有文字的词向量加权求和作为文字特征表示, 如图 8 所示. 该方法有效地过滤了相关性小的文字, 从而使分类或者检索结果得到了进一步的提升. 完整的流程图如图 9 所示, 在离线得到图像中的文字之后, 文字特征表示、图像视觉特征提取、注意力机制、特征融合、分类器 5 个算法模块被集成在了同一个网络当中, 所以可以端到端的训练和测试.

为了验证提出的方法的有效性, 在两个数据集上做了测试, 分别是 Con-Text [46] 和我们自己收集的 Drink Bottle 数据集 [44]. 前者有 28 个场景类别, 共 24255 张图片, 包括咖啡厅、书店和药店等. 后者有 20 个饮料瓶类别, 共 18488 张图片, 包括苏打水、可口可乐和伏特加等类别. 表 4 分别展示了该方法在两个数据集上面的结果. 可以看到, 加入文字信息之后, 识别结果能得到大大的提升. 此外, 图片中包含文字越多的类别, 其提升的效果越明显.

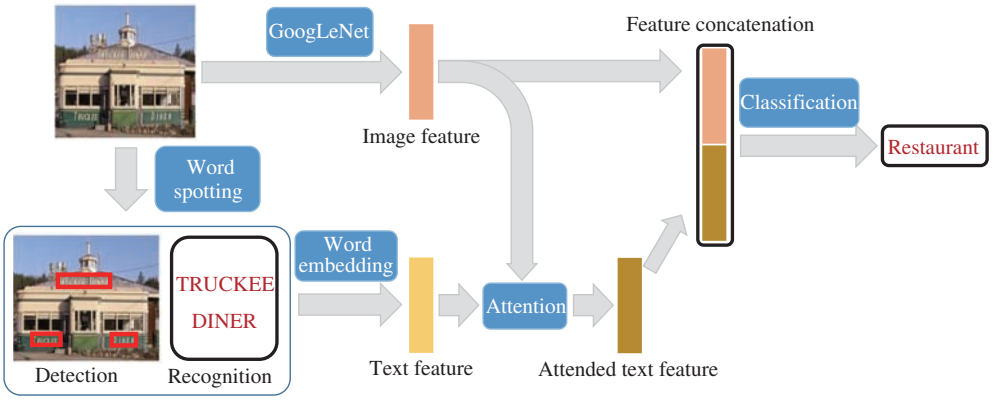

图 9 (网络版彩图) **FIAT 的网络框架示意图** [44]

Figure 9 (Color online) Structure of the proposed FIAT

表 4 Con-Text 和 Drink Bottle 数据集上的分类结果 a)

Table 4 Classification results on Con-Text dataset and Drink Bottle dataset

Method	Mean average precision (%)	
	Con-Text	Drink Bottle
Con-Text [46]	39.0	–
Words matter [47]	77.3	–
Visual baseline (FIAT)	61.3	63.1
FIAT	**79.6**	**72.8**

a) The state-of-the-art results are highlighted in bold.

5 未来研究展望

虽然场景文本检测与识别技术已经取得了显著的进展, 但该领域仍有很多研究方向值得去深入探索.

(1) 曲形文本检测与识别. 曲形排列的文本形状难以用一个矩形包围盒去覆盖, 无论是检测还是识别都比文本行更具有挑战性. 尽管有少数学者对此展开了初步的研究 [48], 但其性能仍远远难以满足实际需求.

(2) 多语言混合的端到端文本识别. 多语种文本混合是一种常见的情况, 但对文本识别系统提出了更高的要求. 除检测和识别外, 仍需要文本的语种鉴别过程. 另外, 不同语种的文本往往具备完全不同的识别规律, 难以用一个统一的识别框架去实现. 如何将语种鉴别、文本检测、不同语种的识别模型统一到一个算法框架中去是一个值得探索也是非常有挑战性的问题. 值得注意的是, 已经有学者在 ICDAR 2017 上组织了相关比赛4) [49].

(3) 弱监督或半监督的文本检测与识别. 现有的文本检测与识别方法, 与深度学习在其他场景应用的情况相似, 往往需要大量的标注. 但是文本图片的形式多样且类别繁多, 导致文本图像的标注工作是一项繁重的任务. 研究如何在样本不充足的条件下的文本检测与识别方法是一项极具意义的任务.

(4) 文本图片的自动生成. 文本图片的自动生成不但能解决文本训练样本难以标注的问题, 而且有

4) ICDAR2017 competition on multi-lingual scene text detection and script identification. http://rrc.cvc.uab.es/?ch=8.

许多潜在的其他应用场景. 例如, 合成著名书法家的毛笔字图像[50]. 最近较流行的生成式模型 GAN[51] 不能直接处理文字存在形变的情况, 因此这一问题仍有较大研究空间.

(5) 融合文本与图像视频的语义理解. 由于文本中包含丰富的语义, 将文本语义与图像视频内容进行融合具备较大的潜力, 能够给许多重要应用问题如商品搜索、图像分类等带来显著的性能提升[44].

6 结束语

场景文本检测与识别是 OCR 领域最通用最重要的研究问题, 最近 10 年来涌现了大量新理论、新方法. 本文回顾了近两年来作者在此领域结合深度学习的重要进展, 并对该领域未来的研究与发展方向进行的总结, 希望能对读者的研究工作有所帮助.

参考文献

1 Zhu Y Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends. Front Comput Sci, 2016, 10: 19–36

2 Ye Q X, Doermann D S. Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intel, 2015, 37: 1480–1500

3 Mori S, Suen C Y, Yamamoto K. Historical review of OCR research and development. Proc IEEE, 1992, 80: 1029–1058

4 Huang W L, Qiao Y, Tang X O. Robust scene text detection with convolution neural network induced mser trees. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 497–511

5 Neumann L, Matas J. Real-time scene text localization and recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 3538–3545

6 Yao C, Bai X, Liu W Y, et al. Detecting texts of arbitrary orientations in natural images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 1083–1090

7 Liao M H, Shi B G, Bai X, et al. TextBoxes: a fast text detector with a single deep neural network. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017

8 Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks. Int J Comput Vision, 2016, 116: 1–20

9 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016

10 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intel, 2017, 39: 1137–1149

11 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016

12 Ross G, Jeff D, Trevor D, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014

13 Girshick R B. Fast R-CNN. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015

14 Zhang Z, Zhang C Q, Shen W, et al. Multi-oriented text detection with fully convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016

15 Zhang Z, Shen W, Yao C, et al. Symmetry-based text line detection in natural scenes. In: Proceedings of Computer Vision and Pattern Recognition, Boston, 2015. 2558–2567

16 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015

17 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
18 Shahab A, Shafait F, Dengel A. ICDAR 2011 robust reading competition challenge 2: reading text in scene images. In: Proceedings of International Conference on Document Analysis and Recognition, Beijing, 2011. 1491–1496
19 Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, 2013. 1484–1493
20 Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017
21 Tian Z, Huang W L, He T, et al. Detecting text in natural image with connectionist text proposal network. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016
22 He P, Huang W L, He T, et al. Single shot text detector with regional attention. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 3066–3074
23 Hu H, Zhang C Q, Luo Y X, et al. WordSup: exploiting word annotations for character based text detection. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 4950–4959
24 He W H, Zhang X Y, Yin F, et al. Deep direct regression for multi-oriented scene text detection. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 745–753
25 Zhou X Y, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector. In: Proceedins of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 2642–2651
26 Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition, Tunis, 2015. 1156–1160
27 Mishra A, Alahari K, Jawahar C J. Scene text recognition using higher order language priors. In: Proceedings of British Machine Vision Conference, Surrey, 2012
28 Yao C, Bai X, Shi B G, et al. Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 4042–4049
29 Bai X, Yao C, Liu W Y. Strokelets: a learned multi-scale mid-level representation for scene text recognition. IEEE Trans Image Process, 2016, 25: 2789–2802
30 Alsharif O, Pineau J. End-to-end text recognition with hybrid HMM maxout models. CoRR, 2013. ArXiv:1310.1811
31 Almazán J, Gordo A, Fornés A, et al. Handwritten word spotting with corrected attributes. In: Proceedings of IEEE International Conference on Computer Vision, Sydney, 2013. 1017–1024
32 Bissacco A, Joseph M, Netzer Y, et al. PhotoOCR: reading text in uncontrolled conditions. In: Proceedings of IEEE International Conference on Computer Vision, Sydney, 2013. 785–792
33 Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intel, 2017, 39: 2298–2304
34 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 98: 1735–1780
35 Wang K, Babenko B, Belongie S J. End-to-end scene text recognition. In: Proceedings of International Conference on Computer Vision, Barcelona, 2011
36 Lucas S M, Panaretos A, Sosa L, et al. ICDAR 2003 robust reading competitions: entries, results, and future directions. Int J Doc Anal Recogn, 2005, 7: 105–122
37 Shi B G, Wang X G, Lyu P Y, et al. Robust scene text recognition with automatic rectification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4168–4176
38 Jaderberg M, Simonyan L, Zisserman A, et al. Spatial transformer networks. In: Proceedings of Conference on Neural Information Processing Systems, Montreal, 2015. 2017–2025
39 Phan T Q, Shivakumara P, Tian S X, et al. Recognizing text with perspective distortion in natural scenes. In: Proceedings of IEEE International Conference on Computer Vision, Sydney, 2013

40 Risnumawan A, Shivakumara P, Chan C S, et al. A robust arbitrary text detection system for natural scene images. Expert Syst Appl, 2014, 41: 8027–8048
41 Yang S L, Bo L F, Wang J, et al. Unsupervised template learning for fine-grained object recognition. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, 2012. 3122–3130
42 Jia D, Jonathan K, Li F F. Fine-grained crowdsourcing for fine-grained recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 580–587
43 Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 834–849
44 Bai X, Yang M K, Lyu P Y, et al. Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks. CoRR, 2017. ArXiv:1704.04613
45 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015
46 Karaoglu S, van Gemert J C, Gevers T. Con-text: text detection using background connectivity for fine-grained object classification. In: Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, 2013. 757–760
47 Karaoglu S, Tao R, Gevers T, et al. Words matter: scene text for image classification and retrieval. IEEE Trans Multim, 2017, 19: 1063–1076
48 Liu Y L, Jin L W, Zhang S T, et al. Detecting curve text in the wild: new dataset and new solution. CoRR, 2017. ArXiv:1712.02170
49 Shi B G, Yao C, Liao M H, et al. Competition on reading chinese text in the wild. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, 2017
50 Lyu P Y, Bai X, Yao C, et al. Auto-encoder guided GAN for chinese calligraphy synthesis. CoRR, 2017. ArXiv:1706.08789
51 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, 2014

Deep learning for scene text detection and recognition

Xiang BAI*, Mingkun YANG, Baoguang SHI & Minghui LIAO

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074*, China*
* Corresponding author. E-mail: xbai@hust.edu.cn

Abstract Scene text detection and recognition is a universal text recognition technology, which has become a hot research topic in the field of computer vision and document analysis in recent years. It is widely applied in geographical positioning, license plate recognition, and driverless applications. Compared to traditional document text detection and recognition, scene text varies more dramatically in font, color, scale, layout, and background. Owing to its excellent performance, deep learning has been widely adopted in this field. In this paper, we mainly review our representative studies based on deep learning in this field and describe the future research trends in this field.

Keywords deep learning, scene text, text detection, text recognition, computer vision

Xiang BAI received his B.S., M.S., and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003, 2005, and 2009, respectively. He is currently a professor in School of Electronic Information and Communications, HUST. He is also the vice-director of National Center of Anti-Counterfeiting Technology, HUST. His research interests include object recognition, shape analysis, scene text recognition, and intelligent systems.

Baoguang SHI received his B.S. degree from School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China in 2012, where he is currently a Ph.D. candidate. He was an intern at Microsoft Research Asia in 2014, and a visiting student at Cornell University from 2016 to 2017. His research interests include scene text detection and recognition, 3D shape recognition, and facial recognition.

Mingkun YANG received his B.S. degree from School of Electronic Information and Communications, Huazhong University of Science and Technology (HUST), China in 2016. He is currently a master student in School of Electronic Information and Communications, HUST. His main research interests include fine-grained image classification and scene text recognition.

Minghui LIAO received his B.S. degree from School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2016, where he is currently pursuing his master's degree in School of Electronic Information and Communications. His main research interests include text detection and recognition.