基于循环神经网络的图像特定文本抽取方法

杨恒杰，闫铮，邬宗玲，方定邦，段放*
华侨大学信息科学与工程学院，福建 厦门 361021

摘要 光学字符识别(OCR)难以针对图像中某些特定文本进行识别，尤其在实际场景中，识别结果通常会包含大量噪声文本。针对这一问题，提出一种基于循环神经网络的双向长短时记忆-条件随机场(BLSTM-CRF)模型。首先利用 BLSTM 网络捕获 OCR 识别结果中序列的上下文信息，得到特征序列；然后结合 CRF 建立模型特征与标签的关系，进行标签预测，通过标签即可得到特定文本。实验结果表明，该方法在场景图像数据集 YNIDREAL 上可以达到 88.52%的准确率，相较于 CRF 模型，准确率提高了 16.39 个百分点，证明了本方法的可行性和稳健性。

关键词 机器视觉；特定文本抽取；光学字符识别；双向长短时记忆网络；条件随机场
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.241501

Extraction Method of Interest Text in Image Based on Recurrent Neural Network

Yang Hengjie, Yan Zheng, Wu Zongling, Fang Dingbang, Duan Fang*
College of Information Science and Engineering, Huaqiao University, Xiamen, Fujian 361021, *China*

Abstract It is difficult to recognize a certain text of interest in the image using the optical character recognition (OCR) method; particularly in natural scenes, the recognition results usually contain a large number of noisy texts. To address this problem, a model termed bidirectional long short term memory-condition random field (BLSTM-CRF) based on a recurrent neural network for extracting texts of interest is proposed in this study. First, a BLSTM network is implemented to capture the context information of the sequence obtained by the OCR method, thereby obtaining feature sequences. Second, the relationships between the model features and tags are established by introducing the CRF. Then the text of interest can be obtained through the tags. Experimental results indicate that the proposed method can achieve an accuracy of 88.52% on YNIDREAL dataset. Compared with the CRF model, the accuracy of the proposed method is improved by 16.39 percentage points, which proves the feasibility and robustness of the proposed method.

Key words machine vision; extraction of interest text; optical character recognition; bidirectional long short term memory network; condition random field
OCIS codes 150.1135; 100.3008; 100.2960; 100.4996

1 引 言

相对于颜色、大小、纹理等特征信息，图像中的文字信息包含了高层次的语义信息。这些信息广泛应用于证件录入、拍照翻译、图像理解等领域[1]。实际应用中所需的信息往往是图像中部分特定的文本信息，而在一些自然应用场景图像中，直接通过光学字符识别(OCR)技术对图像进行识别，识别出来的文本信息往往包含大量的非特定的文本(噪声文本)信息。因此如何从图像中进行特定文本抽取，仍然是业内广泛研究者关注的重点。

早期关于图像的特定文本抽取主要是通过 OCR 中的版面分析来实现，即首先利用版面分析的方法得到图像中特定的文本区域，然后对特定文本

收稿日期：2019-05-05；**修回日期**：2019-06-03；**录用日期**：2019-06-06
基金项目：福建省自然科学基金(2017J01116)、华侨大学中青年培育计划(Z16J0070)、华侨大学科研基金(605-50Y18023)、华侨大学研究生科研创新能力培育计划(17014082026)

* **E-mail**：nkfetsh@gmail.com

区域进行文字识别。其中版面分析的方法可以归结为三大类[2]:1)自顶向下分析法[3];2)自底向上分析法[4];3)混合分析法[5]。其中:自顶向下主要以整幅图像或较大的区域块为基础,逐渐细分,得到各种类别区域;自底向上的思路与自顶向上相反,视每个像素为独立的单位;混合分析法则是综合以上两种思路。这些方法对文档图像的依赖性较强,且其分析过程包含大量复杂的图像处理技术,导致不能很好地泛化到其他类型的图像上。

近年来 OCR 研究主要集中在解决背景复杂、拍摄角度和光线不均、低分辨率的自然场景图像上,其方法也逐渐应用到端到端的场景文字检测和场景文字识别[6-7]中。场景文字检测的方法主要包括基于连通域法[8-9]、滑窗法[10]和深度学习法[11],其中基于连通域与基于滑窗的方法通常计算量较小,但难点在于如何提取一个较好的特征来训练分类器以应对背景复杂、文字旋转等情况。近年来,深度学习法在手写识别[12]、目标检测[13-14]等领域取得了较好的效果。典型的代表为 Jaderberg 等[15]与 Liao 等[16]提出的基于生成文本候选区的文字检测模型,该模型在任意方向文本的检测上表现出色。Zhou 等[17]则是采用全卷积网络(FCN)的结构设计模型,在英文文本检测上达到了当时最好的检测效果。Tian 等[18]利用更深的网络 VGG16 来提取特征,结合双向长短时记忆(BLSTM)网络学习文字空间的上下文信息,有效地排除了非文字区域,从而使得文字检测更加稳健。有关场景文本识别的研究有很多。Jaderberg 等[19]通过合成的 9×10^6 数据来训练卷积神经网络(CNN),达到了出色的识别效果,但是其全连接层限定了单词识别的种类,无法识别语种之外的样本。为了解决上述问题,Liao 等[20]借鉴语义分割的思想,利用 FCN 结构直接识别字符级别的文本。Shi 等[21]设计一种基于序列建模的端到端的模型来识别文字,由 CNN 提取特征,BLSTM 对序列建模,CTC(Connectionist Temporal Classification)转录,最终取得较好的识别效果,因此该模型目前也成为了文字识别领域主流的识别框架。

上述端到端的 OCR 技术通常无法对特定文本进行识别,其整体识别结果常包含大量的噪声文本,场景图像的复杂性,导致难以通过确定的规则来排除噪声文本。例如:特定文本为自然场景下身份证图像中的姓名部分,虽然可以通过 OCR 技术来检测和识别图像中的所有文字,但是由于场景图像的复杂性,易检测出噪声文本,识别结果对应于一串可编辑文本(姓名、噪声文本等),很难确定其中究竟哪一部分为姓名部分。为了解决该问题,受自然语言处理领域序列标注任务的启发[22],本文提出一种新的特定文本抽取思路,即双向长短时记忆-条件随机场(BLSTM-CRF)模型。将该问题类比于序列标注的问题。首先,通过 BLSTM 网络捕获 OCR 输出结果序列的上下文信息,为其建立内在的联系;然后通过 CRF 显式的内容,根据整个序列的标签进行决策,得到最佳的标签结果;最后根据标签即可得到特定的文本。通过模型直接对 OCR 结果进行处理可以避免对图像进行复杂的版面分析。利用云南省普洱澜沧供电局提供的无约束自然场景下的身份证图像数据集 YNIDREAL 进行方法验证。实验结果表明,本文方法可以达到 88.52%的准确率,且相较于传统 CRF 模型,准确率提高了 16.39 个百分点,在含有大量噪声文本的图像中仍可以很好地进行抽取,由此表明该方法具有较好的稳健性。

2 基本原理

本文从序列标注的角度设计实现 OCR 特定文本的抽取。BLSTM-CRF 模型主要由两部分构成:第一部分为 BLSTM 网络,用于捕获上下文信息并编码序列;第二部分为 CRF,用于统计分析得到最后的标签。

2.1 序列标注

序列标注任务是自然语言处理(NLP)领域里典型的任务之一,通常是给定一串输入序列 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$,通过一系列算法来预测序列对应的标签序列 $\boldsymbol{y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)$,其中 n 为序列的长度。主要的序列标注任务包括词性标注(POS)、语义角色标注(SRL)以及命名实体识别(NER)等,图 1 为一个命名实体识别示例,其实体主要包括 person,location,organization。相应的标签分别为 PER,LOC,ORG,并用 B-和 I-分别代表实体标签的开始和实体标签的剩余部分,O 代表非实体标签。传统序列标注的方法大多基于线性统计模型,如:隐马尔可夫模型(HMM)或 CRF 模型[23],由于该类模型需要较多人为设计和干预,因此很难向其他任务推广。近年来循环神经网络(RNN)凭借其自身的结构特征,通过获得序列之前时刻的依赖关系,协助决策当前时刻的输出,在标注任务中得到了出色的表现[24]。

2.2 长短时记忆网络单元

经典的RNN模型在训练过程中会存在梯度消

Jack	lives	in	New	York	and	works	for	the	European	Union
B-PER	O	O	B-LOC	I-LOC	O	O	O	O	B-ORG	I-ORG

图 1 命名实体识别示例

Fig. 1 Example of name entity recognition

失和梯度爆炸等问题，导致其难以获得长距离依赖。针对该问题，文献[25]中提出 RNN 的一类变种——长短时记忆(LSTM)网络。其构成单元如图 2所示。

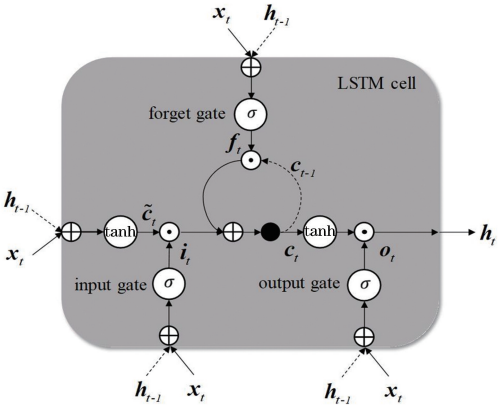

图 2 LSTM 网络单元

Fig. 2 LSTM network unit

图 2 中 $\boldsymbol{x}$ 为输入向量，$\boldsymbol{h}$ 为隐藏层输出向量，$\tilde{\boldsymbol{c}}$ 为 LSTM 单元待存储的信息向量，t 为当前时刻，$t-1$为前一时刻。序列的信息主要存储在信息传递向量 $\boldsymbol{c}$ 中，并通过三个门控单元：输入门($\boldsymbol{i}_t$)、输出门($\boldsymbol{o}_t$)、遗忘门($\boldsymbol{f}_t$)来控制信息的传递。

t 时刻的 LSTM 单元更新公式为

$$\begin{cases}\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t+\boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i)\\ \boldsymbol{f}_t=\sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t+\boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t+\boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o)\\ \tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_{cx}\boldsymbol{x}_t+\boldsymbol{W}_{ch}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c)\\ \boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t\\ \boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t)\end{cases},\tag{1}$$

式中：$\odot$为点乘操作；σ 为 sigmoid 激活函数；$\boldsymbol{W}$ 为权重矩阵，其下标表示不同门控单元对应输入 $\boldsymbol{x}$ 和隐藏层输出 $\boldsymbol{h}$ 的权重矩阵，如 $\boldsymbol{W}_{ix}$ 代表输入门 $\boldsymbol{i}$ 与输入 $\boldsymbol{x}$ 之间的权重矩阵，$\boldsymbol{W}_{fh}$ 代表遗忘门 $\boldsymbol{f}$ 与隐藏层输出 $\boldsymbol{h}$ 之间的权重矩阵；$\boldsymbol{b}$ 为相应的偏置向量，网络通过更新权重矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 来进行优化；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$ 分别为输入门、遗忘门、输出门、信息更新门的偏置向量；$\boldsymbol{c}_t$ 为信息传递向量。

2.3 双向长短时记忆网络

如图 3 所示，在前向长短时记忆网络的结构中，有

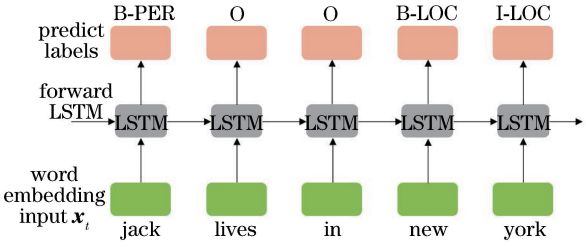

图 3 前向长短时记忆网络结构

Fig. 3 Structure of forward long short time memory network

时无法直接根据前文的信息得到“New”的标签，通过使用 BLSTM 网络[26]同时考虑序列的下文信息，很容易根据“York”这个单词的信息来确定“New”的标签应该为 B-LOC。如图 4 所示，BLSTM 网络在结构上为一个前向和后向长短时记忆网络，分别用来编码序列的上、下文信息，然后将两个网络的输出向量结合起来，得到具有上下文信息的输出。

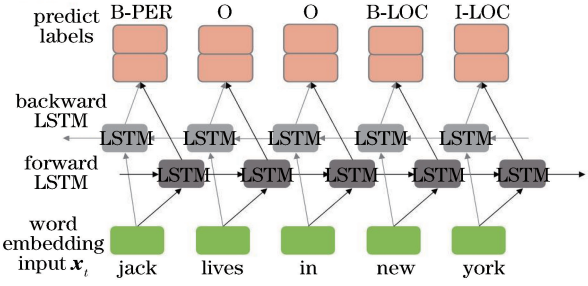

图 4 BLSTM 网络结构

Fig. 4 Structure of BLSTM network

2.4 条件随机场

传统的序列标注的任务，通常独立考虑每个标签，采用类似于集束搜索的求解标签分布的方法[27-28]进行标签预测，但这种方法没有考虑到序列之间的整体关系。CRF 为基于统计的序列标签预测模型，以整个序列的标签为单位，结合序列的整体信息，来考虑序列标签的最佳路径，这种方法可以避免一些歧义。如图 5 所示，在标准 BIO 的命名实体识别任务中，I-LOC 后面的标签为 I-ORG 是不合理的。

设 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$代表输入序列，其中 $\boldsymbol{x}_i$ 表示第 i 个单词，本文中对应为第 i 个汉字，n 为序列的长度。$\boldsymbol{y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)$代表输入序列 $\boldsymbol{x}$ 的标签，其中 $\boldsymbol{y}_i$ 为对应 $\boldsymbol{x}_i$ 的标签。定义 $\boldsymbol{Y}(\boldsymbol{x})$为输入

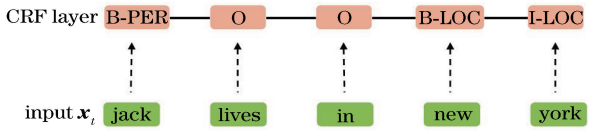

图 5 CRF 网络结构

Fig. 5 Structure of CRF network

序列 $\boldsymbol{x}$ 下所有可能的标签序列集合，$\boldsymbol{y}'$ 为 $\boldsymbol{Y}(\boldsymbol{x})$ 中可能的一组标签序列。CRF 模型就是在给定输入 $\boldsymbol{x}$ 下，预测标签 $\boldsymbol{y}$ 的条件概率为 p，公式为

$$p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{W},\boldsymbol{b})=\frac{\prod_{i=1}^{n}C_i(\boldsymbol{y}_{i-1},\boldsymbol{y}_i,\boldsymbol{x})}{\sum_{\boldsymbol{y}'\in\boldsymbol{Y}(\boldsymbol{x})}\prod_{i=1}^{n}C_i(\boldsymbol{y}'_{i-1},\boldsymbol{y}'_i,\boldsymbol{x})},\tag{2}$$

$$C_i(\boldsymbol{y}_{i-1},\boldsymbol{y}_i,\boldsymbol{x})=\exp(\boldsymbol{W}^{\mathrm{T}}_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}\boldsymbol{x}_i+\boldsymbol{b}_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}),\tag{3}$$

式中：$C_i(\cdot)$ 为配分函数；$\boldsymbol{W}_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}$ 为标签 $\boldsymbol{y}_{i-1}$ 到 $\boldsymbol{y}_i$ 的状态转移矩阵参数；$\boldsymbol{b}_{\boldsymbol{y}_{i-1},\boldsymbol{y}_i}$ 为偏置参数。给定训练样本$\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}$，模型可通过最大似然损失函数来更新 $\boldsymbol{W}$ 和 $\boldsymbol{b}$，最大似然损失函数为

$$L_{\mathrm{loss}}(\boldsymbol{W},\boldsymbol{b})=\ln p(\boldsymbol{y}_i\mid\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{b})。\tag{4}$$

2.5 BLSTM-CRF 模型

本文将双向长短时记忆网络与条件随机场组合成一个 BLSTM-CRF 模型来对 OCR 识别结果进行序列标注。如图 6 所示，首先将 OCR 识别到文字的序列 $\boldsymbol{l}=(l_1,l_2,\cdots,l_n)$ 中进行字嵌入，得到字向量 $\boldsymbol{x}'=(\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_n)$，$\boldsymbol{x}'_i$ 维度为 300 维的字嵌入向量，然后将序列 $\boldsymbol{x}'$ 送入 BLSTM 网络，得到编码上下文信息的 600 维向量 $\boldsymbol{h}_{\mathrm{concat}}$，再对 $\boldsymbol{h}_{\mathrm{concat}}$ 进行线性映射，得到各个标签(tag)的得分，并将其作为 CRF 的输入。线性映射的表达式为

$$\boldsymbol{y}_{\mathrm{score}}=\boldsymbol{W}_{\mathrm{pro}}\boldsymbol{h}_{\mathrm{concat}}+\boldsymbol{b}_{\mathrm{pro}},\tag{5}$$

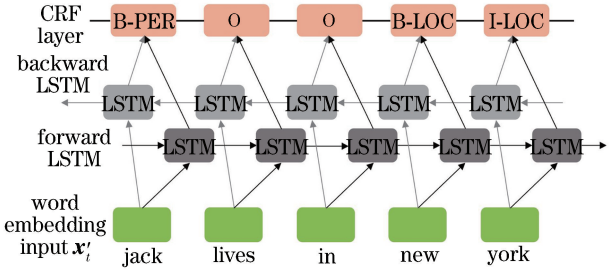

图 6 BLSTM-CRFs 模型结构

Fig. 6 Structure of BLSTM-CRFs model

式中：$\boldsymbol{y}_{\mathrm{score}}$ 为一个 $n\times k_{\mathrm{tag}}$ 的矩阵，代表长度为 n 的序列中每个字 $\boldsymbol{x}'_i$ 对应每种标签的得分，k_{tag} 为标签的种类数；$\boldsymbol{W}_{\mathrm{pro}}$ 和 $\boldsymbol{b}_{\mathrm{pro}}$ 分别为映射权重矩阵和偏置。最后将 CRF 结合整个序列的标签，从整体角度考虑标签的最终结果，得到最终序列的标签$\boldsymbol{y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)$。

2.6 场景文字检测与识别

近年来，自然场景文本检测和识别大多是基于深度学习的端到端的方法，为了验证本文针对 OCR 特定文本抽取方法的有效性，采用的 OCR 检测与识别为当下的主流框架，其中检测部分为 CTPN[18]，该模型借鉴了目标检测领域的 Faster R-CNN[29] 模型中的建议区域提取的思想，设计了针对文字特点的文本建议区域提取，在 ICDAR2013 数据集中精确率和召回率的调和平均值 F 值达到了 88%，可以有效地检测自然场景中的文字。识别部分为 CRNN[21]，该模型利用 CNN 提取序列特征，并采用 BLSTM 网络编码序列的上下文信息，最终通过 CTC 解码。这种将文字识别与序列处理相结合的思想可以解决变长标签的文本识别，如“OK”与“Congratulation”。该模型在 IIIT5K、SVT、IC03 数据集中最高准确率分别达到了 97.8%、97.5%、98.7%。

3 实验结果及分析

3.1 实验数据集

监督学习模型的训练要以大量带标签的训练集为基础，但由于缺乏标准的数据集训练本研究提出的模型，且人工标定数据耗时耗力，因此本文以身份证图像数据集作为研究对象，分析身份证的结构内容，设计算法，自动生成带标签数据，进行训练。部分生成数据如图 7 所示，其中共包含 6 种实体，name、gender、nation、birth、address、idnum，分别代表姓名、性别、民族、出生、地址、公民身份号码。本文要抽取的特定文本即为标注的 6 种实体，共生成训练集 IDTRAIN 500 份，验证集 IDVAL 100 份，如表 1 所示。

为测试本模型在自然场景下对特定文本的抽取效果，利用云南省普洱澜沧供电局提供的数据集 YNIDREAL 来进行实际测试，该数据集由数位供电局工作人员通过不同拍照设备采集，采集环境各不相同。从中筛选出 61 张背景复杂、光线不均、分辨率不一和含有大量噪声文本的代表性的作业场景样本，旨在较全面地测试本文方法在实际场景图像中的效果。图像样例如图 8 所示。

3.2 模型训练

本实验所用的深度学习框架为 Tensorflow1.8 (Google Inc)，字嵌入采用均匀分布初始化，范围为 [−0.25,0.25)，维度为 300 维，优化器为 Adam，初始学习率为 0.001，梯度裁剪为 0.5，批量大小为 64，批次内随机打乱输入数据，共迭代 40 次。实验所有操作环节在 64 位 Ubuntu18.04 LST 系统下运行，CPU 配置为8线程Corei7-7700CPU3.6GHz，显卡

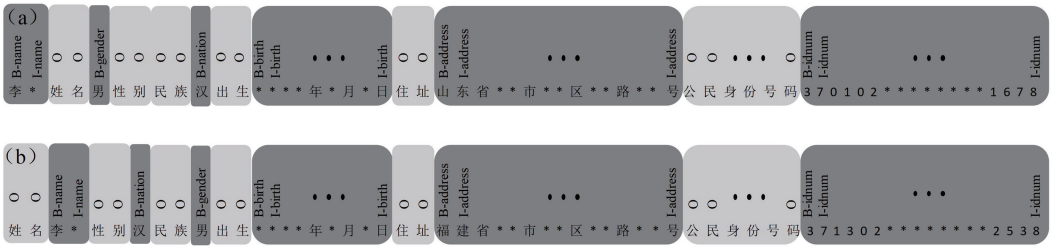

图 7 IDTRAIN 和 IDVAL 中生成的文本数据及标签示例。(a)样例 a;(b)样例 b

Fig. 7 Samples of text data and label generated in IDTRAIN and IDVAL. (a) Sample a; (b) sample b

表 1 实验数据集的分布

Table 1 Distribution of experimental data set

Item	Dataset category	Dataset type	Dataset size
Train	IDTRAIN	Text	500
Validation	IDVAL	Text	100
Test	YNIDREAL	Image	61

图 8 YNIDREAL 中的图像样本示例

Fig. 8 Samples of images in YNIDREAL

为 GTX1050,显存为 2 GB。

按照字符进行标注,实体的标签完整、识别正确计为正确识别,以图 7(a)中的“姓名”实体为例,模型需要将“李四”的起始标签 B-name 和剩余标签 I-name 全部预测正确,并以此来定义实体的正类(positive),其他情况则认为是实体的负类(negative)。序列标签的真实值(GT)中正类实体计为 True,负类实体计为 false。N_{TP}(true positive)代表模型预测为正类且实体标签真实值为正类的实体数量;N_{FP}(false positive)代表模型预测为正类且实体标签真实值为负类的实体数量;N_{TN}(true positive)代表模型预测为负类且实体标签真实值为负类的实体数量;N_{FN}(false positive)代表模型预测为负类且真实值为正类的实体数量。由此定义评价模型的指标:精确率 P、召回率 R、F1 测度值 F_1,计算公式为

$$\begin{cases} P=\dfrac{N_{TP}}{N_{TP}+N_{FP}} \\ R=\dfrac{N_{TP}}{N_{TP}+N_{FN}} \\ F_1=2\times\dfrac{P\times R}{P+R}=2\times\dfrac{N_{TP}}{2\times N_{TP}+N_{FP}+N_{FN}} \end{cases}, \tag{6}$$

式中:P 为模型预测出正确的正类占预测所有正类的比例;R 为模型预测出正确的正类占真值中所有正类的比例;F_1 为综合 P 与 R 的调和平均数,代表模型综合的效果。BLSTM-CRF 模型的训练和验证仅采用生成的数据集 IDTRAIN 和 IDVAL。图 9 显示了模型在 IDVAL 上 6 种实体的 F_1、P、R 值,图中表明,约在 15 个 epoch 训练后,模型的 F_1、P、R 已经近似收敛于 1,可见该模型在生成数据集上具有很好的表现效果。

3.3 模型测试

模型整体框架包含两部分,分别为 OCR 部分[30]和特定文本抽取部分。其中 OCR 部分已在 2.6 节所述,即首先通过对图像进行文字检测,然后进行文字识别,最后进行特定文本抽取。文本抽取分别比较了 CRF、BLSTM-CRF 模型在 YNIDREAL 数据集上的测试效果。其中 CRF 模型的人工干预特征为字符的词性和词边界。表 2 为 CRF 模型与 BLSTM-CRF 模型系统性能对比。可以看出,在相同的训练集、验证集和测试集下,BLSTM-CRF 模型的系统平均性能要优于 CRF 模型。

图 10 为本方法在 YNIDREAL 数据集的测试结

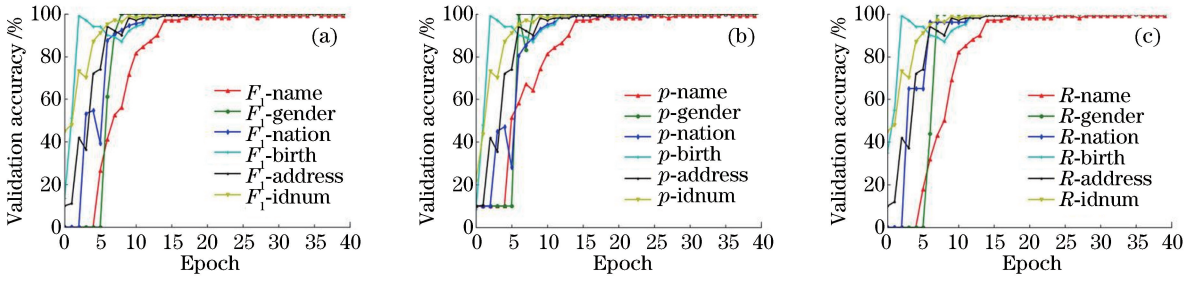

图 9 IDVAL 上 6 种实体准确率。(a) F_1 值;(b) P 值;(c) R 值

Fig. 9 Accuracy of six entities on IDVAL. (a) F_1-score; (b) P value; (c) R value

表 2 CRF 模型与 BLSTM-CRF 模型系统性能

Table 2 System performances of CRF and BLSTM-CRF models

Entity	CRF			BLSTM-CRF		
	P /%	R /%	F_1/%	P /%	R /%	F_1/%
Name	75.00	68.85	71.79	86.89	86.89	86.89
Gender	96.67	95.08	95.87	96.72	96.72	96.72
Nation	95.00	93.44	94.21	93.44	93.44	93.44
Birth	90.16	90.16	90.16	91.80	91.80	91.80
Address	90.48	93.44	91.94	93.65	96.72	95.16
Idnum	92.06	95.08	93.55	90.48	93.44	91.94
Average	89.90	89.34	89.59	92.16	93.17	92.66

果。图 10(a)中方框区域为文字检测的结果,表明除了身份证区域的信息外,其他非特定区域的噪声文本也被检测出来。图 10(b)为检测区域识别得到的可编辑文本。对比图 10(c)和图 10(d)可见,BLSTM-CRF 模型可以完整地抽取出特定文本信息,但 CRF 模型却出现姓名信息抽取不全、身份证号码抽取错误的问题。

表 3 统计了数据集上所有样本中 6 种特定文本全部完整抽取的结果以及 OCR 部分与信息抽取部分的速度。可见,文本抽取部分无论是采用 BLSTM-CRF 模型还是 CRF 模型,其抽取速度远远大于 OCR 部分的识别速度,即利用本文方法与 OCR 结合进行信息抽取时,可以忽略由信息抽取部分带来的速度损失。整体耗时仍然是由 OCR 部分主导,体现本文模型具有更强的实时适用性。另外 BLSTM-CRF 模型虽然在抽取速度上略慢于 CRF 模型,但是其抽取信息的准确率高于 CRF 模型16.39 个百分点,在有噪声文本的情况下仍能很好地将特定文本抽取出来,体现了模型的稳健性。采用 CRF 模型进行序列标注时,必须人工设置特征,实际效果受特征的限制,因此在泛化能力上,BLSTM-CRFs 模型要优于 CRF 模型。

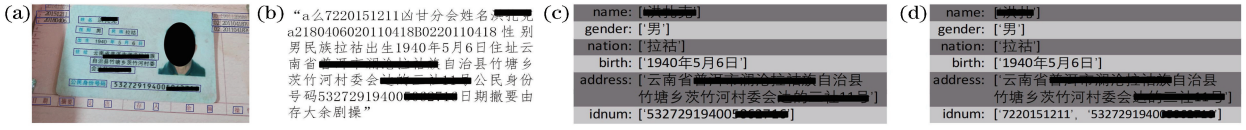

图 10 YNIDREAL 测试结果示例图。(a)文字检测结果;(b)文字识别结果;(c) BLSTM-CRF 模型特定文本抽取结果;(d) CRF 模型特定文本抽取结果

Fig. 10 Test results on YNIDREAL dataset. (a) Text detection results; (b) text recognition results; (c) result of interest text extraction using BLSTM-CRF model; (d) result of interest text extraction using CRF model

表 3 特定文本抽取完整性测试结果

Table 3 Test results of integrity of interest text extraction

Model	Succeed number	Fail number	Speed /(image · s^{-1})		Test accuracy /%
			OCR	Extraction	
CRF	44	17	0.17	97	72.13
BLSTM-CRF	54	7	0.17	82	88.52

4 结 论

借鉴自然语言处理中序列标注的思想,提出基于递归神经网络的 BLSTM-CRF 模型,对 OCR 进行特定文本抽取。仅利用生成的数据集对模型进行训练即可在 YNIDREAL 数据集上达到 88.52%的准确率,相对于仅利用条件随机场模型,效果提升了 16.39 个百分点,为 OCR 特定文本抽取提供了一个全新的思路。本文模型通过 BLSTM 编码 OCR 识别结果的上下文信息,对噪声文本有一定的过滤作用,证明了本文算法的稳健性。其中 OCR 部分与本文模型相对独立,证明方法具有一定的模块灵活性。

由于目前比较缺乏针对自然场景图像进行特定文本抽取的标准数据集,而且标准数据集对自然场景下特定文本抽取的研究至关重要,后续将考虑如

何制备一个规范的标准数据集以供研究。另外，在实验过程中发现在对一些自然场景图像进行特定文本抽取时，仍然存在由于识别错误引发的抽取失败的例子，因此在未来的工作中，如何改进并提高文本识别的准确率将是一个重点。

参 考 文 献

[1] Oliveira D A B, Viana M P. Fast CNN-based document layout analysis [C] // 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 1173-1180.

[2] Le V P, Nayef N, Visani M, *et al*. Text and non-text segmentation based on connected component features[C] // 2015 13th International Conference on Document Analysis and Recognition (ICDAR), August 23-26, 2015, Tunis, Tunisia. New York: IEEE, 2015: 1096-1100.

[3] Okun O, Doermann D, Pietikainen M. Page segmentation and zone classification: the state of the art[R]. Fort Belvoir: Defense Technical Information Center, 1999.

[4] Moll M A, Baird H S. Segmentation-based retrieval of document images from diverse collections [J]. Proceedings of SPIE, 2008, 6815: 68150L.

[5] Bukhari S S, Al Azawi M I A, Shafait F, *et al*. Document image segmentation using discriminative learning over connected components[C] // Proceedings of the 8th IAPR International Workshop on Document Analysis Systems-DAS '10, June 9-11, 2010, Boston, Massachusetts, USA. New York: ACM, 2010: 183-190.

[6] Ye Q X, Doermann D. Text detection and recognition in imagery: a survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(7): 1480-1500.

[7] Liu X Y, Meng G F, Pan C H. Scene text detection and recognition with advances in deep learning: a survey [J]. International Journal on Document Analysis and Recognition (IJDAR), 2019, 22(2): 143-162.

[8] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform[C] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 2963-2970.

[9] Neumann L, Matas J. A method for text localization and recognition in real-world images [M] // Kimmel R, Klette R, Sugimoto A. Computer vision-ACCV 2010. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2011, 6494: 770-783.

[10] Wang K, Babenko B, Belongie S. End-to-end scene text recognition [C] // 2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE, 2011: 1457-1464.

[11] Huang W L, Qiao Y, Tang X O. Robust scene text detection with convolution neural network induced MSER trees [M] // Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8692: 497-511.

[12] Fang D B, Feng G, Cao H Y, *et al*. Handwritten formula symbol recognition based on multi-feature convolutional neural network [J]. Laser & Optoelectronics Progress, 2019, 56(7): 072001.
方定邦, 冯桂, 曹海燕, 等. 基于多特征卷积神经网络的手写公式符号识别[J]. 激光与光电子学进展, 2019, 56(7): 072001.

[13] Wang X, Liu Y, Li G Y. Moving object detection algorithm based on improved visual background extractor algorithm [J]. Laser & Optoelectronics Progress, 2019, 56(1): 011007.
王旭, 刘毅, 李国燕. 基于改进视觉背景提取算法的运动目标检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011007.

[14] Zhao H, An W S. Image salient object detection combined with deep learning [J]. Laser & Optoelectronics Progress, 2018, 55(12): 121003.
赵恒, 安维胜. 结合深度学习的图像显著目标检测[J]. 激光与光电子学进展, 2018, 55(12): 121003.

[15] Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting [M] // Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8692: 512-528.

[16] Liao M, Shi B, Bai X, *et al*. Textboxes: a fast text detector with a single deep neural network [C] // Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), February 4-10, 2017, San Francisco, California, USA. USA: AAAI Press, 2017: 4161-4167.

[17] Zhou X Y, Yao C, Wen H, *et al*. EAST: an efficient and accurate scene text detector [C] // 2017 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2642-2651.

[18] Tian Z, Huang W L, He T, *et al*. Detecting text in natural image with connectionist text proposal network[M] // Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 56-72.

[19] Jaderberg M, Simonyan K, Vedaldi A, *et al*. Synthetic data and artificial neural networks for natural scene text recognition[J/OL]. (2014-12-09) [2019-05-04]. https://arxiv.org/abs/1406.2227.

[20] Liao M, Zhang J, Wan Z, *et al*. Scene text recognition from two-dimensional perspective[C] // Proceedings of the AAAI Conference on Artificial Intelligence, January 27-February 1, 2019, Hilton Hawaiian Village, Honolulu, Hawaii, USA. USA: AAAI Press, 2019, 30(1): 8714-8721.

[21] 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[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298-2304.

[22] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J/OL]. (2015-08-09) [2019-05-04]. https://arxiv.org/abs/1508.01991.

[23] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C] // Proceedings of the Eighteenth International Conference on Machine Learning, June 28-July 1, 2001, Williams College, Williamstown, MA, USA. USA: ACM, 2001: 282-289.

[24] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[25] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[26] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, 2013, Vancouver, BC, Canada. New York: IEEE, 2013: 6645-6649.

[27] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging[C] // Conference on Empirical Methods in Neural Language Processing, May 17-18, 1996, Philadelphia, PA, USA. [S.l.: s.n.], 1996.

[28] McCallum A, Freitag D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation[C] // Proceedings of the Seventeenth International Conference on Machine Learning, June 29-July 2, 2000, Stanford, CA, USA. USA: ACM, 2000: 591-598.

[29] Ren S, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C] // Neural Information Processing Systems (NIPS), December 7-12, 2015, Palais des Congrès de Montréal, Montréal Canada. Canada: NIPS, 2015: 91-99.

[30] Shi X F. CHINESE-OCR[EB/OL]. (2018-04-14) [2019-05-04]. https://github.com/xiaofengShi/CHINESE-OCR.