计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(11):35−41 [doi: 10.15888/j.cnki.csa.006647]

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

结合密集神经网络与长短时记忆模型的中文识别①

张艺玮[1], 赵一嘉[2], 王馨悦[1], 董兰芳[1]

[1](中国科学技术大学 计算机科学与技术学院, 合肥 230022)
[2](辽宁省实验中学, 沈阳 110031)
通讯作者: 董兰芳, E-mail: lfdong@ustc.edu.cn

摘　要: 文本图像识别是计算机视觉领域一项重要任务, 而其中的中文识别因种类繁多、结构复杂以及类间相近等特点很具挑战性. 为改善这一问题, 使用文本行端到端的识别模型. 首次提出利用密集卷积神经网络 (DenseNet) 提取文本图像底层特征, 同时避免手工设计、统计图像特征的繁琐; 将整行图像特征直接送入双向长短时记忆模型 (BLSTM) 进行局部相关性分析, 减少字符定位分割这一步骤; 最后采用时域连接模型 (CTC) 解码获得识别的文本信息. 实验表明所提出的模型可以高效的进行图像文本行的识别, 并对图像的多种形变具有较好的鲁棒性.

关键词: 中文识别; 端到端; 密集卷积神经网络; 双向长短时记忆模型; 时域连接模型

引用格式: 张艺玮,赵一嘉,王馨悦,董兰芳.结合密集神经网络与长短时记忆模型的中文识别.计算机系统应用,2018,27(11):35–41. http://www.c-s-a.org.cn/1003-3254/6647.html

Chinese Recognition Based on Dense Convolutional Network and Bidirectional Long Short-Term Memory Model

ZHANG Yi-Wei[1], ZHAO Yi-Jia[2], WANG Xin-Yue[1], DONG Lan-Fang[1]

[1](School of Computer Science and Technology, University of Science and Technology of China, Hefei 230022, China)
[2](Liaoning Provincial Shiyan High School, Shenyang 110031, China)

Abstract: Text recognition is an important task in computer vision. The recognition of Chinese texts is challenging because of its wide range, complicated structure, and similar classes. In order to improve this problem, an end-to-end recognition model of text is used. The proposed model uses Dense convolutional Network (DenseNet) to extract features of text images, avoiding artificial design and statistics features. Then, the features are sent to Bidirectional Long Short-Term Memory model (BLSTM) for correlation analysis of local data. This step avoids the character segmentation. Finally, the Connectionist Temporal Classifier (CTC) is used to decode the text information. Experiments show that the proposed model can effectively recognize text images, and has strong robustness to various deformed images.

Key words: Chinese recognition; end-to-end; Dense convolutional Network (DenseNet); Bidirectional Long Short-Term Memory (BLSTM); Connectionist Temporal Classifier (CTC)

文本识别[1]分为印刷体识别和手写体识别. 目前这两种识别都得到充分的研究, 并普遍认为印刷体字符识别中的关键问题已得到有效解决. 但是对于印刷体字符识别而言, 图像质量的严重下降会给识别造成极大的困难; 而关于中文字符识别[2], 需要克服的难点更多, 首先, 中文类别较多, 按照 GB2312 标准, 我们常用的一级汉字就有 3755 类; 其次, 中文字符结构复杂, 它包括偏旁、部首和字根; 还有字符间形近字比较多, 准确区分形近字也大大增加了识别难度. 所以, 实际使用的文本识别技术还有很大的提升空间.

① 收稿时间: 2018-04-11; 修改时间: 2018-05-11; 采用时间: 2018-05-15; csa 在线出版时间: 2018-09-30

中文字符的识别方法主要分为结构模式识别和统计模式识别[3]. 其中结构模式识别是早期中文识别的主要方法, 它根据字符自身的规律信息进行结构特征提取, 这些结构特征包括字符轮廓特征、骨架图像上提取到的反映字符形状的特征等[4]. 基于结构模式识别的主要优点在于匹配精度高, 区分相似字能力强; 但是由于其依赖结构特征的提取, 而特征的提取易受到干扰因素的影响, 所以这种方法的抗干扰能力较差. 随着统计理论的发展, 统计模式识别方法[5]逐渐成为中文字符识别的研究热点, 它提取将要被识别的统计特征, 然后利用某些函数对这些特征进行分类. 常见的统计特征包括网格特征、方向像素特征、穿越特征、外围特征等. 这种方法的主要优点是具有良好的抗噪声、抗干扰能力, 对字符形变也有较强的鲁棒性, 但是对细节区分能力不强.

上述传统方法都是基于手工设计、提取特征, 这个过程不仅耗费人力, 而且会积累误差和噪音, 极大地影响最后的识别效果.

近几年, 深度学习不断发展, 特别是深度卷积神经网络 (CNN)[6]等模型在模式识别及计算机视觉领域的大量突破性成果的涌现, 为中文识别带来新的活力; 2013 年富士通团队采用改进的 CNN 网络[7], 在单个汉字识别方面取得了令人瞩目的成绩.

本文在深度学习的基础上, 针对多种字符, 包括中文、英文、数字、特殊符号等, 结合密集卷积神经网络 DenseNet、双向长短时记忆模型 BLSTM 和连接时域分类 CTC 进行文本行端到端的识别. 采用 DenseNet 通过卷积、下采样等操作提取图像特征, 并将生成的特征序列传递给 BLSTM, 相对于卷积神经网络, BLSTM 使用特殊的存储记忆单元更充分的利用文本上下文特征进行建模, 最后采用 CTC 对之前的特征信息进行解码, 输出识别结果. 这个网络结构可以接受任意长度的输入序列, 不要求对文本提前分割, 在避免字符分割错误带来误差的同时, 对于字符连接信息有一定程度的记忆能力, 整体性能强, 可以进一步提高文本识别率.

1 网络结构

1.1 DenseNet

深度卷积神经网络一个很重要的参数是深度. 网络深度的提升往往伴随着网络性能的提升, 但是随着网络深度的加深, 训练参数梯度消失的问题会愈加明显, 而 DenseNet[8]的出现很好的解决了这个问题.

DenseNet 是在 Highway Networks[9,10], Residual Networks[11](ResNet) 以及 GoogLeNet[12]的基础上被提出来的. 不同于之前加深网络或者加宽网络, DenseNet 的提出者从卷积神经网络的特征序列入手, 通过对特征序列的极致利用, 简化模型参数, 同时达到更好的效果. 它的主要思想是跨层连接, 网络每一层的输入都是前面所有层输出的并集, 而该层学习到的特征序列也会被直接传给后面所有层作为输入; 在上述过程中, 信息流进行了整合, 避免了信息在层间传递丢失和梯度消失的问题.

DenseNet 一般由多个 dense block 和 transition layer 组成. 图 1 是 DenseNet 的主要结构——dense block 的示意图. 可以看出, H_4 层不仅直接用原始信息 x_0 作为输入, 同时还使用 H_1、H_2、H_3 层对 x_0 处理后的信息作为输入; 我们可以用一个非常简单的式子描述 dense block 中每一层的变换, 如式 (1):

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \tag{1}$$

式 (1) 中H_l 代表第 l 层的变换函数, 通常对应于 Batch Normalization (BN), ReLU 和 Convolution (Cnov). x_l是第 l 层的输出. 同样, 在反向传播过程中, x_0 的梯度信息不仅来自于前一层, 也包含了损失函数直接对 x_0 计算的导数.

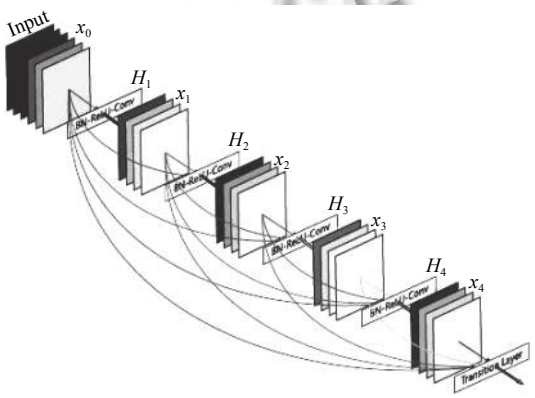


图 1 一个 5 层的 dense block 示意图

相比于普通神经网络的分类器只依赖于网络最后一层的特征, DenseNet 可以综合利用浅层特征, 加强了特征的传导和利用, 减轻梯度消失的问题, 所以更容易得到一个光滑的具有更好泛化性能的决策函数.

1.2 BLSTM

传统的递归神经网络 (RNN)[13]展开后相当于一个

多层的神经网络, 当层数过多时会导致训练参数的梯度消失问题的出现, 从而致使长距离的历史信息损失. 因此, 传统 RNN 在实际应用时, 能够利用的历史信息非常有限. 为了弥补上述缺陷, Hochreiter 等人[14]在 1997 年提出了 LSTM 单元结构, 如图 2 所示, 它由 1 个记忆细胞和 3 个门控单元组成, 记忆细胞用于存储当前的网络状态, 3 个门控单元与记忆细胞相连, 分别称作输入门、输出门和遗忘门, 它们控制信息的流动. 在信息传递时, 输入门控制输入到记忆细胞的信息流; 输出门控制记忆细胞到网络其他结构单元的信息流; 遗忘门控制记忆细胞内部的循环状态, 决定记忆细胞中信息的取舍[15]. LSTM 的这种门控机制让信息选择性通过, 使记忆细胞具有保存长距离相依信息的能力, 并可以在训练过程中防止内部梯度受外部干扰.

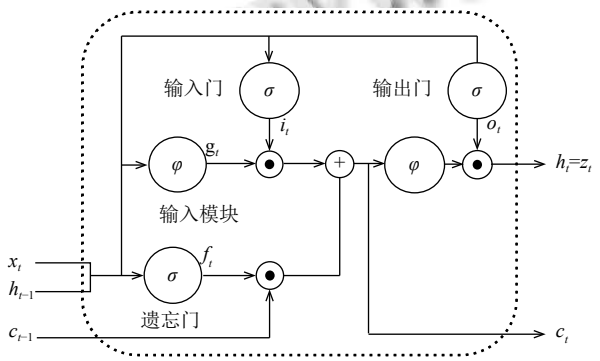


图 2 LSTM 单元结构图

已知输入为x_t, 输出为h_t; i_t、f_t、o_t、g_t、c_t分别为输入门、遗忘门、输出门、输入模块、记忆细胞的状态, LSTM 单元结构内部按照以下公式进行迭代:

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + b_i) \tag{2}$$

$$f_t = \sigma(w_{xf}x_t + w_{hf}h_{t-1} + b_f) \tag{3}$$

$$o_t = \sigma(w_{xo}x_t + w_{ho}h_{t-1} + b_o) \tag{4}$$

$$g_t = \phi(w_{xc}x_t + w_{hc}h_{t-1} + b_c) \tag{5}$$

$$c_t = f_i \cdot c_{t-1} + i_t \cdot g_t \tag{6}$$

$$h_t = o_t \cdot \phi(c_t) \tag{7}$$

上述公式中, w是权重, b是偏移, $\cdot(x)$是点乘操作, $\sigma(x)$是 sigmoid 操作, $\phi(x)$是激活函数, 通常取双曲正切函数, 它们的计算公式如下:

$$\sigma(x) = (1 + e^{-x})^{-1} \tag{8}$$

$$\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}$$

对于计算机视觉领域的很多任务, 如对模型的预测或识别, 未来信息同历史信息一样重要. 例如文本行识别, 在识别当前词时, 它之前与它之后的词语信息都会对当前词的识别有所帮助. 但是, 前文描述的模型只能单向输入, 序列无法利用未来的信息. 于是, Schuster 等人[16]提出双向 RNN (BRNN) 概念, 它的核心思想是将序列信息分两个方向输入模型中, 模型使用两个隐藏层分别保存来自两个方向的输入数据, 并将相应的输出连接到相同的输出层, 如图 3 所示.

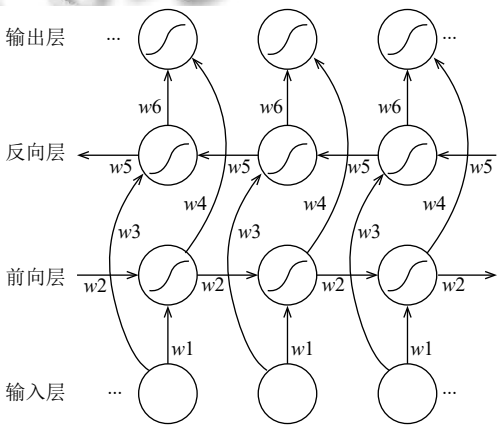


图 3 BRNN 在时间上的展开形式

图 3 中, $w1$, $w3$ 表示输入层到前向层与反向层的权重, $w2$, $w5$ 表示隐含层自身循环的权重, $w4$, $w6$ 表示前向层与反向层到输出层的权重.

BLSTM 将 BRNN 和 LSTM 这两种改进的 RNN 模型组合在一起, 即在 BRNN 模型中使用 LSTM 记忆单元, 这样可以更好的学习局部信息的相关性.

1.3 CTC

时域连接模型 CTC[17]是一种直接标记无分割序列的方法, 适合于输入特征和输出标签之间对齐关系不确定的时间序列问题. 它可以端到端地优化模型参数, 并且对齐切分的边界, 使得针对输入序列的每一帧, 网络能够输出一个标签或者空白标志 (‘-’).

CTC 网络的输出层是在给定的输入下, 计算所有可能对应的标签序列的概率, 以求出标签概率最大的序列. 对于一个长度为 T 的序列 x, 经过神经网络计算映射, 得到序列的输出 y, 定义$y^t_{\pi_t}$表示在 t 时刻标签为π_t的概率, L'^T表示在整个标签集 $L \cup \{\text{‘-’}\}$上所有长度

为 T 的序列集合, 得到公式 (10).

$$p(\pi|x)=\prod_{t=1}^{T}y_{\pi_t}^{t},\forall\pi\in L'^{T} \tag{10}$$

式 (10) 中, π 是L'^{T}中的一个序列, 并且假设网络每一时刻的输出是独立的. 接下来, 定义一个多对一的映射 β: $L'^{T}\rightarrow L^{\leqslant T}$, 后者是可能的标签序列的集合, 这个映射我们可以简单的通过删除全部的'-'和重复标签元素来实现, 如 F(行行-识别--) = F(-行识识--别别) = 行识别. 我们定义一个标签序列$l\in L^{\leqslant T}$, 它的条件概率是其对应的全部序列的概率和, 得到式 (11).

$$p(l|x)=\sum_{\beta(\pi)=l}p(\pi|x) \tag{11}$$

分类器的输出应为输入序列最有可能的标签序列, 如式 (12) 所示.

$$h(x)=\arg\max_{l\in L^{\leqslant T}}p(l|x) \tag{12}$$

根据上述公式, 目标函数最小化上述概率的负对数似然. 因为目标函数是可导的, 网络可以通过标准的 BP 方法来训练.

2 整体模型

图 4 是本文提出的模型整体架构图.

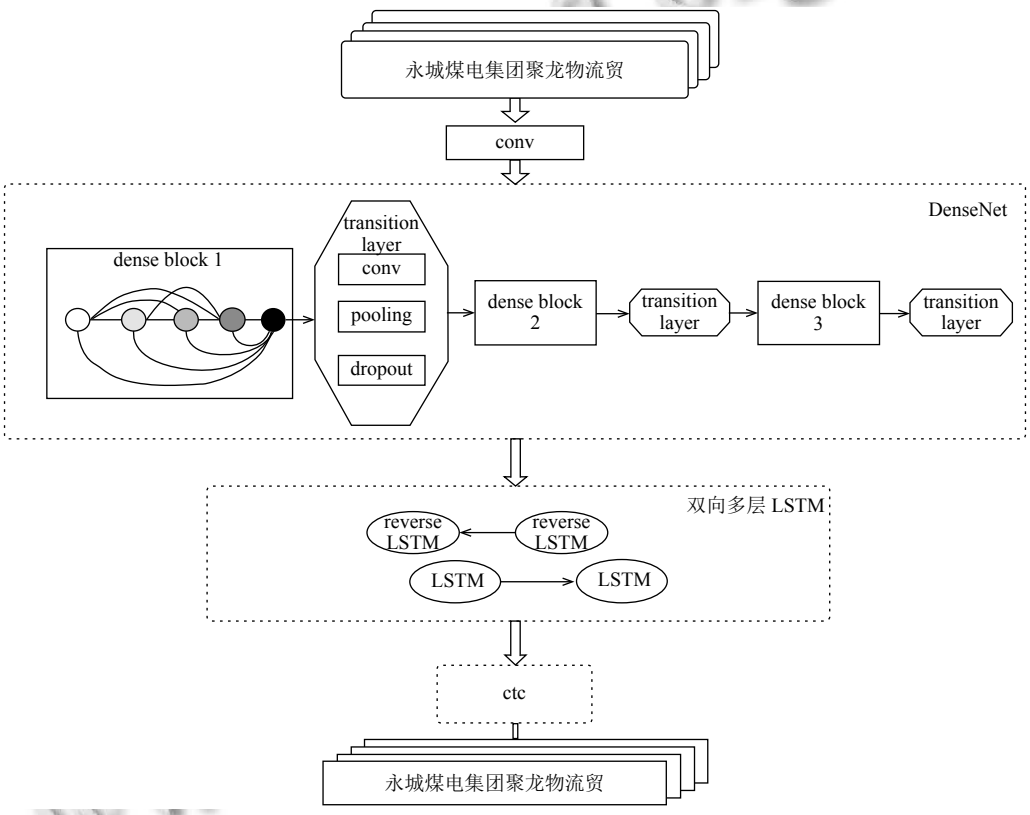


图 4 模型整体架构图

因为汉字种类繁多, 结构复杂, 简单的卷积神经网络已经很难完全提取图像细节特征, 深层网络又可能造成信息消失、参数繁多以及难收敛等问题, 所以本模型选择结构简单, 但效果突出的 DenseNet 网络结构提取图像底层特征. DenseNet 网络中的每个 dense block 中有 4 个 Bottleneck layers, 即 BN-ReLU-Conv (1×1)-BN -ReLU- Conv (3×3) 结构. 该结构中 Conv (1×1) 可以减少输入参数的数量; dense block 结构中每一层网络都设计地很窄, 只学习较少的特征序列, 这样可以减少网络参数, 提高网络效率, 达到降低冗余性的目的. DenseNet 网络中的 transition layer 用来连接 dense block, 它由 conv 层、pooling 层以及 dropout 层组成, conv 层用来决定是否压缩模型参数; pooling 层控制特征序列的大小, 因为在 dense block 内部, 特征序列的空间维度是保持不变的, 故而在两个 dense block 之间进行下采样; 最后插入 dropout 层, 它是由 Srivastava

等人[18]在 2014 年提出的防止网络过拟合的技术, 即在模型训练时按照一定的比例 (本文设置为 0.2) 随机选择某些节点不工作, 使得模型具有多模型融合的效果, 可以降低网络损失, 提升网络性能. 为了不丢失图像特征, 本文没有选择全连接层作为 DenseNet 结构的最后一层, 而是直接用图像的特征序列作为 BLSTM 的输入.

接下来, 多层的 BLSTM 提取时序信息. 将 DenseNet 提取的多维特征序列, 按照 BLSTM 层的输入要求进行转置, 然后分别送给正向 LSTM 与反向 LSTM 层进行学习, LSTM 计算提取每张图 35 列特征序列间的信息.

在 BLSTM 充分获取数据间的特征后, 运用 CTC 层对之前的训练数据强制对齐, 实现无分割序列的标签工作. 在当前输入下, CTC 层计算每一列对应到 4001(本文字符种类数 4000+1 种空白) 种标签元素上的序列概率分布, 并将这些序列按照一定的规则进行映射后, 统计每个标签序列的概率, 求出最可能的标签序列并输出. CTC 转录层可获得图像的序列描述, 即图像的最终表示方式.

每个文本行图像经过 DenseNet+BLSTM+CTC 3 个主要环节得到最终的特征表达, 如算法 1.

算法 1. 基于混合网络模型的字符识别算法

输入: 文本行图像
输出: 文本行识别结果

1. 输入大小 32×280 的文本行图像;
2. 将图像送入 DenseNet 网络: 经过 3 个 dense block 结构以及 transition layer, 输出 256 维 1×35 大小的特征序列;
3. 将步骤 2 提取的数据特征传递给双层 BLSTM 模型,计算每张图 35 列特征序列间的信息;
4. 使用 CTC 模型对文分割序列的信息进行解码, 输出文本行图像对应的最有可能的标签序列.

3 实验

3.1 实验设置

实验基于 caffe 框架. 在 Intel Core i7, 内存 8 GB, 显卡 GTX1080 机器上进行训练, 模型的初始学习率设为 0.0001, 学习率按照“multistep”方式更新.

3.2 实验数据

由于发票上存在一些比较特殊的字体, 而且发票上的印刷体字符图像存在断裂、粘连等情况, 目前没有相关已经开源的训练数据库, 所以本文的实验数据主要来自网上的印刷体图像以及自己生成大规模的数据. 实验图像生成算法如算法 2.

算法 2. 实验图像生成算法

输入: 文本编辑文件中 4000 类字符文本, 包括的中文一级汉字 3755 个
输出: 250 万张文本行图像

Begin:
1. for i=1,2,…,4000
2. 将文本编辑文件中的第 i 个字符文本读入程序
3. 使用 OpenCV 计算机视觉库生成一张 32×32 的空白图像
4. 使用 freetype 字体引擎将读入的文本按照发票上的字体形式加载到空白图像, 同时设置图像中字体的大小、图像的空白比例等
5. 将原始图像进行灰度化、二值化
6. end for
7. 读取语料库中的语句, 按顺序拼接上述步骤制作的单字图像, 以生成文本行图像
8. 对图像进行多种变换: 加入旋转变化; 加入椒盐噪声; 高斯噪声; 设置不同参数的腐蚀操作、膨胀操作; 对图像反复进行拉伸操作; 加入弹性变换; 加入模糊操作等 //降低图像质量, 拟合发票上的字符图像
9. 将图像调整至 32×280 大小, 满足网络输入要求
10. 重复 7–9 步, 共生成 250 万张图像
End

将实验数据按照 9:1 的比例分成训练集与测试集, 将图像进行自适应阈值二值化[19]后如图 5 所示.

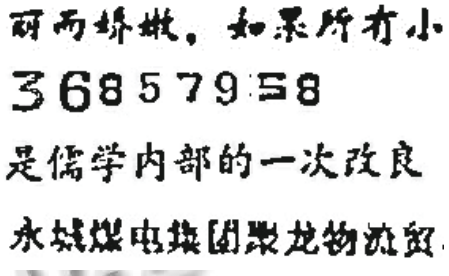


图 5 文本行图像示例

3.3 实验结果及分析

为了选取更合理的模型结构, 本文对 dense block 的设置以及 BLSTM 的层数进行了多组对比实验.

表 1 的实验是在总卷积层数相同的情况下, 改变 dense block 的结构, 观察识别效果. 实验 1 使用 4 个 dense block, 每个 dense block 内部有 6 个卷积层, 实验 2 使用 3 个 dense block, 每个 dense block 内部有 8 个卷积层, 其中引入 Bottleneck layers. 从实验结果来看, 实验 2 在识别率上表现更好, 同时因为包含 1×1 的卷积层, 参数也得到精简.

表 1 关于 dense block 的实验数据表

dense block	行识别率 (%)	字识别率 (%)
实验 1	94.35	97.94
实验 2	95.47	98.35

基于表 1 的实验结果, 选取实验 2 的 dense block 结构, 再更改 dense block 内部的 Growth rate, 即本模型中 3×3 卷积层产生的特征序列数量, 观察实验性能.

表 2 中的实验在 DenseNet-B 的基础上进行. (DenseNet-B 是指 dense block 中 1×1 卷积层产生的特征序列的数量是 3×3 卷积层的 4 倍, 我们设 3×3 卷积层的 Growth rate=k, 则 1×1 卷积层应有 $4k$ 个特征序列. 实验 3 中 Growth rate=16, 实验 4 为 32) 观察实验结果发现, 实验 4 比实验 3 的识别结果稍有提升.

表 2 关于 Growth rate 的实验数据表

Growth rate	行识别率 (%)	字识别率 (%)
实验 3	95.47	98.35
实验 4	96.68	98.61

经过多组实验, 本模型在确定 DenseNet 的 dense block 结构后, 对 BLSTM 的层数进行实验.

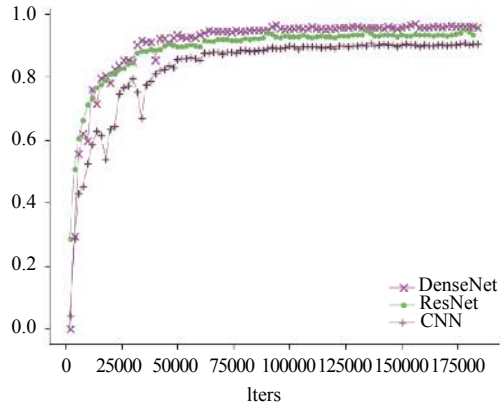
同样, 本文对比了经典的 CNN 模型和 2015 年大放异彩的 ResNet 网络结构, 如图 6 所示.

表 3 列出了 BLSTM 分别取 1 层、2 层以及 3 层时, 对实验性能的影响. 对比表明, 用两层的 BLSTM 可以取得更好的识别率, 而随着层数的增加, 模型的识别时间逐渐增加. 分析认为 2 层 BLSTM 能更好的提取文本间的信息, 1 层存在特征提取不充分的情况, 而 3 层可能出现过拟合的情况.

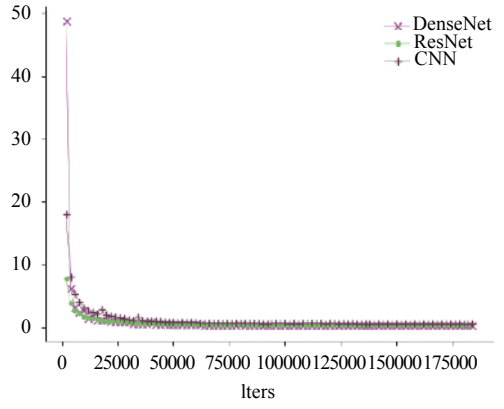
首先将本文提出的模型与经典的 CNN+BLSTM+CTC 模型进行对比, 在同样的实验数据下, 经典的 CNN 模型的最高行识别率只有 91.3%, 明显低于 DenseNet 模型的识别率, 并且收敛速度没有本文提出的模型快. 接着, 本文对 ResNet+BLSTM+ CTC 模型进行了多组测试, 最后选取结果最好的模型与本文提出的模型进行对比. 从图 6(a) 可以看出, ResNet 模型相比于经典的 CNN 模型有较大的提升, 行识别率为 95.0%, 而本文模型的行识别率达到 96.68%.

分析图 6 的 (a) 与 (b) 图, 可以看出, 随着网络迭代次数的增加, 三种模型都逐渐收敛, 其中, DenseNet 模型与 ResNet 模型收敛迅速; 当模型效果趋于稳定后, 本文提出的模型识别准确率最高, 这也充分说明了 DenseNet+ BLSTM+CTC 结构在识别率及收敛速度方面的优越性.

为了更加充分地验证本文模型的性能, 我们又与 Tesseract[20] OCR 软件进行了对比. 在开源的 OCR 引擎中, Tesseract OCR 是效果最好的. 它最先由惠普实验室开始研发, 至 1995 年时已经成为 OCR 业内最准确的三款识别引擎之一. 2005 年, Google 开始对 Tesseract 进行优化.本文充分利用 Tesseract 可以自训练识别库的优势,针对性地训练中文识别库,并利用该识别库实验.



(a) 三种模型识别率对比图



(b) 三种模型识别损失对比图

图 6 本文模型与 ResNet 模型识别性能对比图

表 3 BLSTM 层数对实验的影响

BLSTM 层数	识别率 (%)	时间 (ms/张)
1 层	98.39	8.01
2 层	98.61	11.29
3 层	98.43	14.70

表 4 两种方法实验对比

识别方法	行识别率 (%)	字识别率 (%)	时间 (ms/张)
Tesseract	24.76	82.47	530.39
本文方法	96.68	98.61	11.29

在进行 Tesseract 识别测试时发现, 对于很多文本行, Tesseract 可能出现识别错 1 个或 2 个字符的情况,

所以它针对一行文本图像完全识别正确的概率很低, 但针对行中单个字符, 它的识别率可以达到 82.47%. 对比发现, 本文提出的模型在单字识别率上提升了 16.14%, 而时间仅相当于 Tesseract 的 1/47, 效果显著.

4 结论

文本识别是一项很有挑战性的任务, 尤其是中文识别. 针对文本行图像, 本文首次提出 DenseNet+BLSTM + CTC 的端到端识别的混合架构. 利用 DenseNet 自动提取文本图像特征, 多层卷积特征融合了低层形状信息和高层语义信息, 避免了手工设计图像特征的繁琐, 减少特征计算的难度; 在分析图像信息后, BLSTM 提取字符图像间相关性特征, 并从两个方向进行分析, 前向层从前向后捕获文本演变, 后向层反方向建模文本演变, 有效的利用序列的上下文信息[21]; 最后将两个方向的演变表达融合到 CTC 中, 产生图像的序列化表达. 实验结果表明本文方法在识别率、识别时间、内存占用多方面表现优秀, 并具有无限潜能, 同样适用于其他序列标注任务.

参考文献

1 Mori S, Suen CY, Yamamoto K. Historical review of OCR research and development. Proceedings of the IEEE, 1992, 80(7): 1029–1058. [doi: 10.1109/5.156468]

2 丁晓青. 汉字识别研究的回顾. 电子学报, 2002, 30(9): 1364–1368. [doi: 10.3321/j.issn:0372-2112.2002.09.029]

3 张黔, 胡庆, 杨静宇, 等. 统计和结构模式识别方法结合的多特征印鉴真伪鉴别方法. 计算机学报, 1995, 18(3): 190–198. [doi: 10.3321/j.issn:0254-4164.1995.03.005]

4 王恺, 靳简明, 史广顺, 等. 基于特征点的汉字字体识别研究. 电子与信息学报, 2008, 30(2): 272–276.

5 杨柳. 统计模式识别在汉字识别中的应用. 内江科技, 2008, (11): 134–135. [doi: 10.3969/j.issn.1006-1436.2008.11.112]

6 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

7 Yin F, Wang QF, Zhang XY, *et al*. ICDAR 2013 Chinese handwriting recognition competition. Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, DC, USA. 2013. 1464–1470.

8 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2261–2269.

9 Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv preprint arXiv:1505.00387, 2015.

10 Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. Proceedings of the 28th Annual Conference on Neural Information Processing Systems. Montreal, Canada. 2015. 2377–2385.

11 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.

12 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.

13 Mikolov T, Kombrink S, Deoras A, *et al*. RNNLM-recurrent neural network language modeling toolkit. Proceedings of 2011 ASRU Workshop. Baltimore, MD, USA. 2011. 196–201.

14 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

15 Olah C. Understanding LSTM networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-chain.png, 2015-08.

16 Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997, 45(11): 2673–2681. [doi: 10.1109/78.650093]

17 Graves A, Fernández S, Gomez F, *et al*. 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.

18 Srivastava N, Hinton G E, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

19 Sankur B, Sezgin M. Image thresholding techniques: A survey over categories. Pattern Recognition, 2001, 34(2): 1573–1583.

20 Smith R. An overview of the tesseract OCR engine. Proceedings of the 9th International Conference on Document Analysis and Recognition. Parana, Brazil. 2007. 629–633.

21 Liwicki M, Graves A, Bunke H, *et al*. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. Proceedings of the 9th International Conference on Document Analysis and Recognition. Garching, Munich, Germany. 2007.