**CRNN: 基于神经网络的OCR系统**

**摘要**: 场景文本识别（Scene text recognition），是计算机视觉中一个长盛不衰的研究方向。与传统的OCR相比，它更富有挑战性，因为它识别的对象的前景和背景都是复杂多变的。2015年，Xiang Bai等人的论文[1]，提出了一种新颖的人工神经网络CRNN，将卷积神经网络（CNN）和循环神经网络（RNN）相结合，能够端到端地识别任意长度的文本，不需要人工提取特征。与以往的模型相比，该网络参数更少。本文的贡献在于，在理解的基础上，实现了该论文并进行了测试，在标准数据集上，我们的实现达到了与论文相近的准确率。此外，我们还探究了如何提高该神经网络训练时的收敛性。

**关键字**

文本识别，卷积神经网络，循环神经网络，机器学习

**1 引言**

OCR（Optical character recognition）即光学字符识别，最早可以追溯到1914年Emanuel Goldberg发明的盲人识字机。OCR的任务是读入一幅图片，然后将其中包含的文字输出。图片可以是扫描的文档，手写的书信，甚至可能是机器生成的验证码。OCR在邮政、银行、交通、古代文献等领域有着广泛的应用。如今，摄像头随处可得，OCR的需求也因而水涨船高。

早期的OCR系统需要对每一个字符单独进行训练，并且只对训练过的字体有效，有很大的局限性。如今，先进的OCR系统，已经能够以较高的准确度，对常用字体进行识别。但是，由于任务的挑战性和复杂性，传统的OCR系统是由多个步骤组成的。OCR系统首先需要对图片进行预处理。包括去畸、去噪、线段移除、二值化、布局分析、文本检测、字符分割等操作。这些操作都需要大量的参数调试工作和启发式的规则。而神经网络，可以自动完成这些参数调试并发现这些规则，从而简化OCR的流程。

在本文的完成的OCR系统中，我的合作者——由鸿周主要负责文本定位的工作。他实现了论文[2]的TextBoxes定位算法，将自然场景下图片中的文本检测并定位，提取出来做文本识别。TextBoxes网络实质是一个改进版的SSD网络，将SSD模型的参数（default box的长宽比、classifier的卷积滤波器、输入图像尺度等）设置为更适合于检测文本。测试时自然场景图像输入到网络，通过非极大值抑制筛选出候选检测结果。文本定位能找出场景中的文字的包围盒，并且做裁剪，如下图所示。





图. 1. 用TextBoxes做文本检测，裁剪出来的文本包围盒。

OCR可以分为基于单词的Word Recognition和基于字符的Character Recognition。我们做的是Word Recognition。假设包围盒是一个水平、狭长的图片。在获取该图片后，我们使用CRNN对其中的单词进行识别。

**2 相关工作**

目前，对于打印的拉丁文本，识别准确度依然达不到100%，即使是有清晰的图片。2009年的一项研究[3]表明，对于19世纪到二十世纪初的报纸进行逐个字符的识别，商用软件的识别准确度在81%到99%不等。而对于手写体等扭曲字符的识别，目前依然是一个研究热点。

在数据集ICDAR2013 Robust reading competition[4]比赛的第二个挑战 Focused Scene Text的Word Recognition任务中，2017年8月14号，腾讯AI实验室取得了目前为止的最高准确率。以大小写不敏感的衡量标准来看，该实验室取得了95.34%的准确率。他们使用的方法正是CNN+RNN，并配合上注意力机制。这说明CRNN是目前最好的Word Recognition方法的基础。

在CRNN出现以前，在利用人工神经网络完成OCR任务的领域，人们尝试[6]单纯使用非常深的CNN进行识别文本识别。但是那样需要构造一个非常巨大的神经网络。而使用RNN则可以借助序列前后的相关性，使得网络变得简单而有效。

**3 算法原理**

搭建CRNN，整体的网络结构如下图所示，下面将分步对其中各层进行介绍。

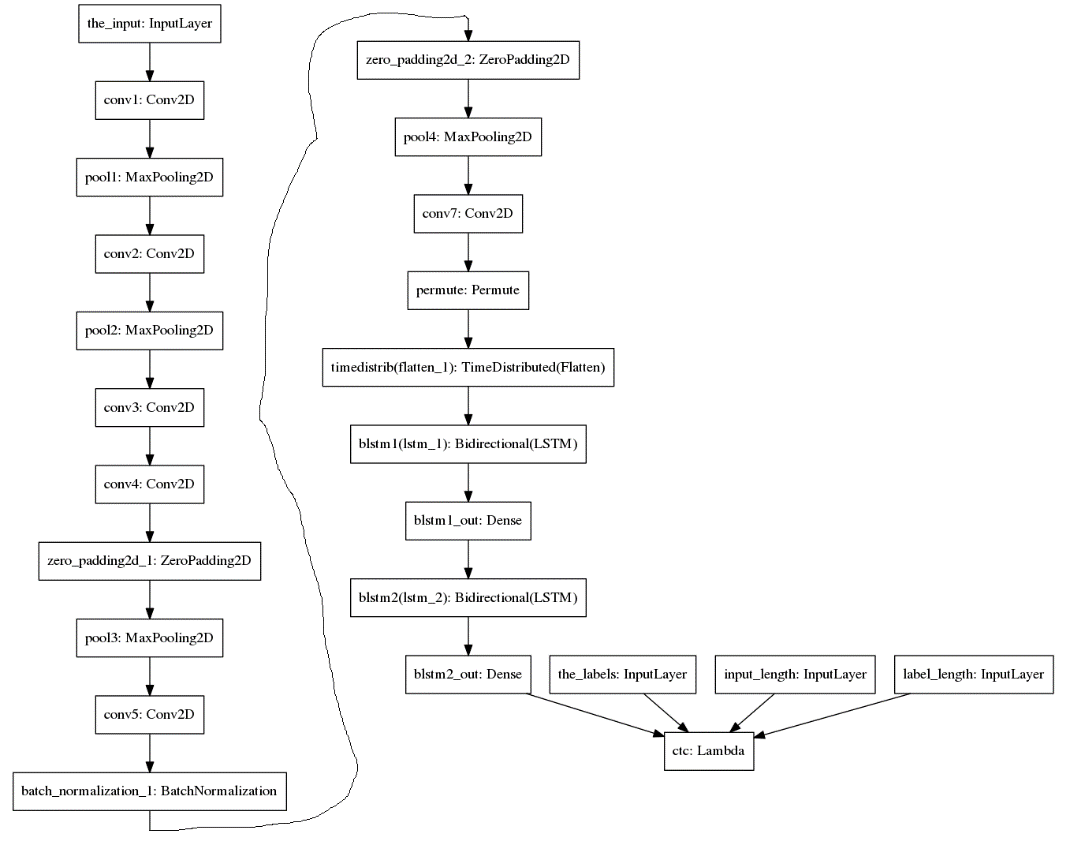


图.2. CRNN整体网络结构图

**3.1 卷积神经网络**

首先，我们假设需要识别的单词是一个水平、狭长的图片，并且有一个最大的极限长度，例如很少有超过25个字母的英文单词。我们使用与VGG深度神经网络类似的结构，来提取特征。

该网络由相间排列的卷积层和极大池化层组成，如下图所示。

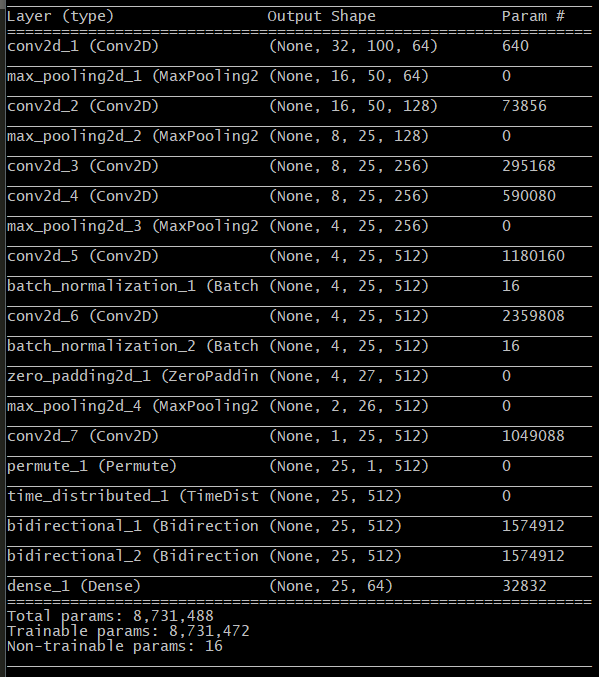


图.3. 卷积神经网络层次结构图

可以看到一幅高为32宽为100的图片，在卷积层的作用下，通道数不断增大，最终变为512。而在极大池化层的作用下，其宽度和高度不断减小，并且高度最终变为1。这样做的目的是将图片映射为由特征组成的序列，序列长度为25，序列中的每个特征向量维度为512。如下图所示，每个特征向量对应于原图中的一个狭长的感受野。感受野对应于字符的一部分。

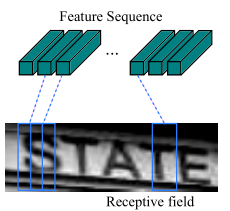


图. 4. 特征序列与感受野对应关系

此外，我们还可以看到，部分卷积层后面还添加了Batch Normalization来使得网络收敛更快，训练精度更高。

**3.2****循环神经网络**

得到了图像的特征序列后，该序列被输入到循环神经网络中，网络结构如下图所示。

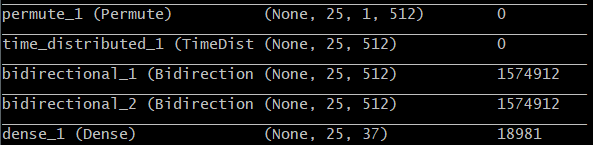


图.5. 循环神经网络层次结构图

这一步操作的核心思想是，用该特征序列，来预测与图像的文本相对应的另一个序列，序列长度为25，特征向量的维度缩小为37，即26个英文字母、10个数字和1个空白字符。RNN预测的结果，是字符类别的概率分布。

**3.2 后续处理**

RNN输出的概率分布的序列长度为25，即前面提到的一个单词可能的极限长度。而每张图片对应的单词的长度，是变化的。这两者不匹配，所以在计算最终输出结果的时候，还需要语音识别中常用的一种叫做CTC[7]的序列对齐技术，借助动态规划和极大似然法来估计最终输出的结果。下图为CTC对齐的示意图。

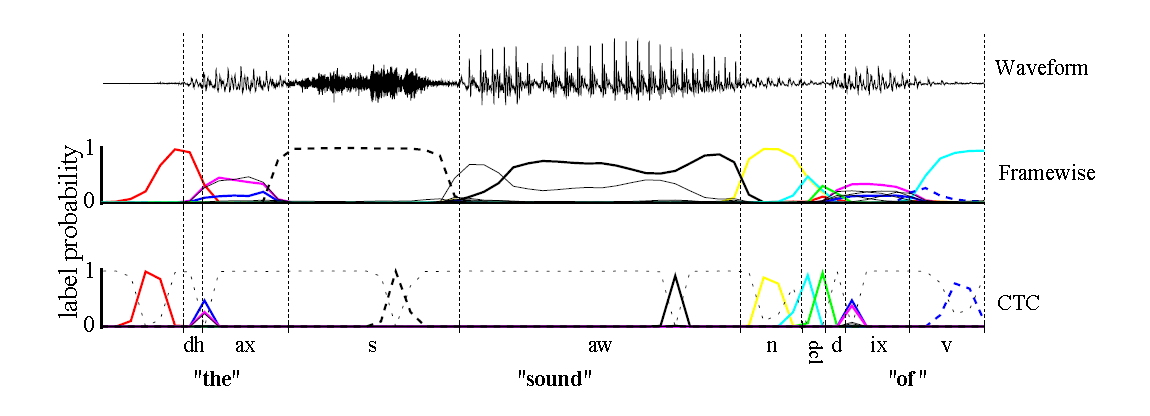


图.6. 利用CTC完成语音对齐的示意图

借助CTC，我们才能得到优化神经网络参数所需要的Loss函数，以及在测试的时候将概率分布转化为实际预测的标签。CTC技术的细节比较复杂，在这里不再赘述。Tensorflow等深度学习框架都提供了用于计算CTC的函数。

此外，对于预测得到的字符序列还可以借助语言模型，寻找最接近的单词，甚至对神经网络的参数进行反馈。由于时间所限，我们的实现中没有使用语言模型。

**4 实验结果与分析**

**4.1 实验环境与配置**

**训练：**

操作系统：Linux Mint 18.2 Sonya (GNU/Linux 4.4.0-21-generic x86\_64)

CPU: Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz \* 8

GPU: GK110GL [Tesla K20c]

Python: 2.7.12

机器学习框架：Tensorflow-gpu 1.4.1

**测试：**

操作系统：Windows 10.0.16299.192, 64bit

Python: 3.6.3

机器学习框架：Tensorflow 1.4.0

**4.2 实验方法**

**搭建网络结构**

首先我们利用tensorflow实现了CRNN神经网络。利用tensorflow的可视化工具，我们可以得到以下工作流程图。

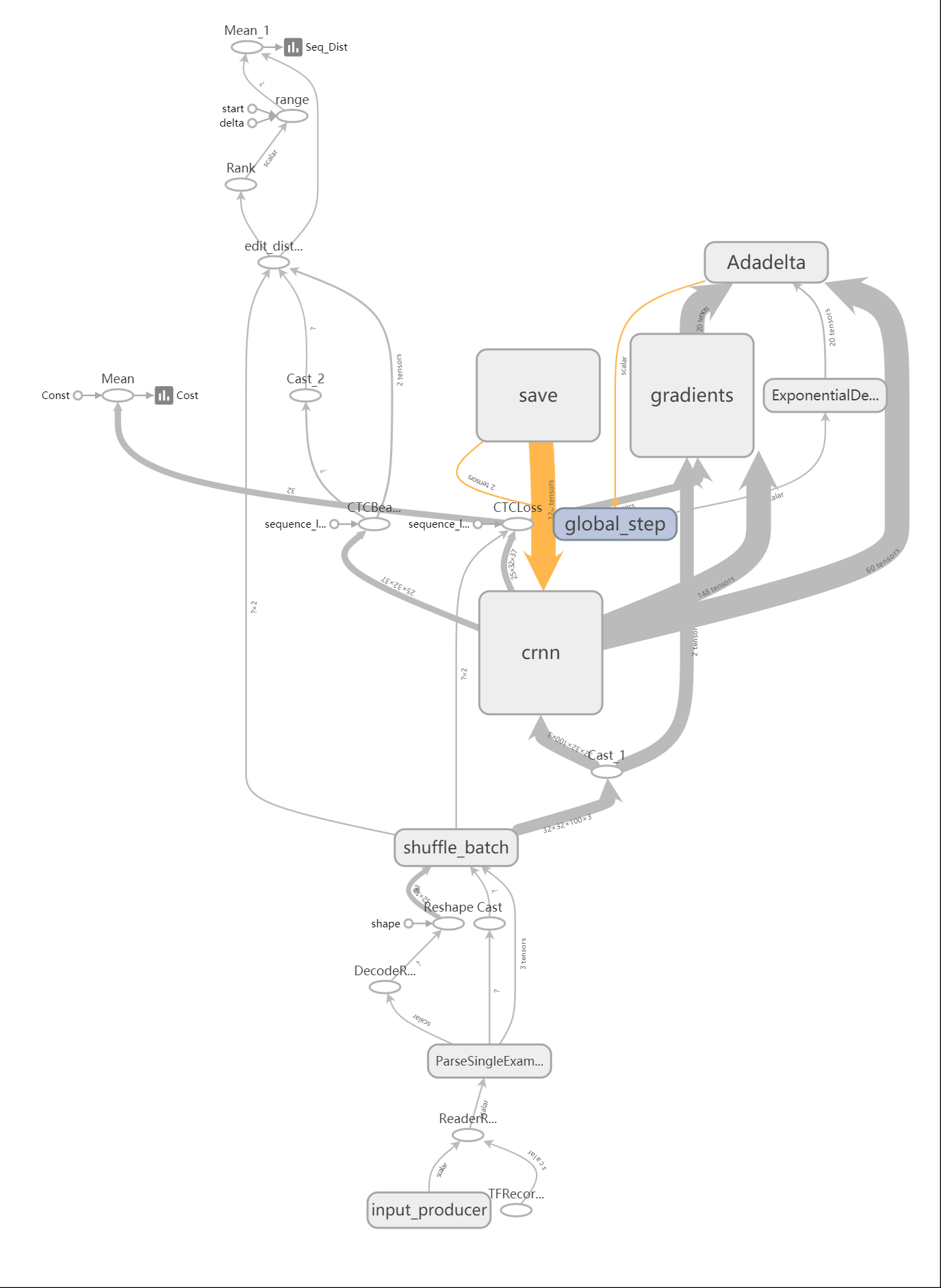


图.7. 神经网络工作流程图

可以看到，我们利用tf.nn.ctc\_loss即可在训练阶段计算loss。在测试阶段，可以使用tf.nn.ctc\_beam\_search\_decoder即可以进行预测。

**训练中使用的超参数**

接着，我们需要设置一系列的超参数和优化方法。

首先，按照经验，我们的RNN采用了LSTM，并且设置了保持概率为0.5的dropout失活处理。

和作者一样，我们的输入图片都缩放为32×100的大小。并且我们事先生成了tfrecords文件，以提高磁盘读写效率。我们也使用了Adadelta优化器。同时，参考网上的开源实现。我们使用了按指数衰减的学习率。

**生成训练数据集**

和作者一样，我们也使用了牛津大学VGG实验组合成的数据集[8]。他们使用神经网络的方法合成了大量的类似与下图的数据用于训练word recognition的网络。我们直接从网站上下载了相应的数据，并且随机抽取了100000个样本作为初始的训练数据。



图.8. VGG合成数据集样例

**训练过程**

训练过程中Loss的变化情况如下图所示：

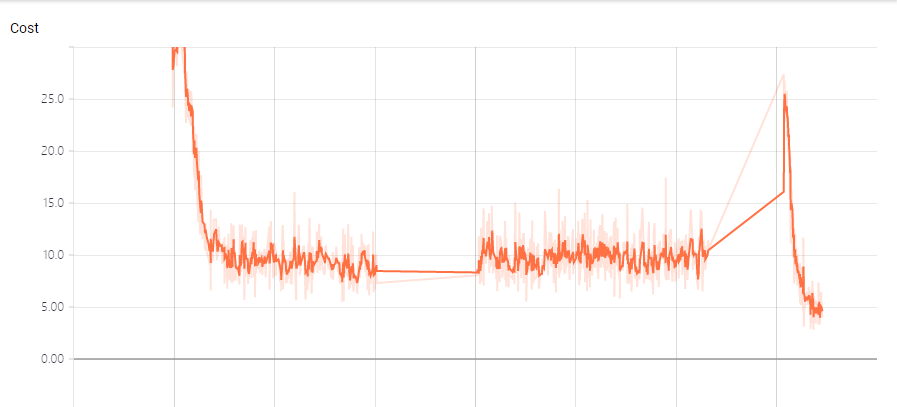


图.8. 训练过程中Loss曲线变化图

可以看到，开始的时候，Loss逐渐下降，但是到了一定的平原阶段，无论我们如何训练，Loss不再有明显的变化。通过反复调试，我们发现，CRNN很难收敛。我们的样本个数可能不够，CRNN的训练需要大量的样本才能达到更好的效果。于是我们又从VGG的数据集中随机抽取了50000个样本，加大学习率，重新训练，从而获得了更低的Loss。最终，在训练数据集上，对于每个batch，我们达到了超过90.6%的准确率。

**4.3 实验结果**

我们挑选了两个标准数据集，一个是IIIT5k[9]，另一个则是前文提到的ICDAR2013 Robust reading competition的Task3：Word Recognition的测试数据集（以下简称ICDAR13）。虽然两个数据集都提供了训练数据和测试数据。但我们没有使用训练数据，而是直接在测试数据上进行测试。

数据集ICDAR13包含1096个样本和CRNN论文的作者一样，我们去除了ICDAR13中的包含了非数字非字母符号的样本，随意还剩1017个样本。但是我们没有去除太短的单词，而CRNN论文的作者在测试时去除了长度小于3的样本。结果如下表所示，我们在两个数据集上都取得了与论文作者相近的效果。其中在IIIT5k上，我们与论文作者的准确率差别不到2%。

另外我们注意到我们的训练数据集中，标签中含有数字的样本特别少。因此，我们将两个训练数据集中的含有数字的样本去除，则在IIIT5k上，我们取得了优于论文作者的效果。

表1. 标准数据集识别准确率(%)对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实现 | IIIT5k | ICDAR2013 | IIIT5k（去除数字） | ICDAR2013（去除数字） |
| 论文作者： | 78.2 | 86.7 | None | None |
| 我们的实现： | 77.3 | 79.7 | 81.6 | 83.3 |

我们是在VGG合成训练数据集上训练的。上述结果表明，

**5 结论**

通过一番努力，我们基本实现了论文[1]，并且在标准测试数据集上获得了与论文作者相近的结果，说明了我们比较准确地实现了CRNN神经网络。通过和Textboxes技术结合，我们完成了一个较为可靠的OCR系统

**参考文献**

|  |  |
| --- | --- |
|  | Shi, B., Bai, X. and Yao, C., 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 39(11), pp.2298-2304. |
|  | G Liao, M., Shi, B., Bai, X., Wang, X. and Liu, W., 2017. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In AAAI (pp. 4161-4167). |
|  | Holley, R., 2009. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4). |
|  | Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A. and De Las Heras, L.P., 2013, August. ICDAR 2013 robust reading competition. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on (pp. 1484-1493). IEEE. |
|  | Bissacco, A., Cummins, M., Netzer, Y. and Neven, H., 2013, December. Photoocr: Reading text in uncontrolled conditions. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 785-792). IEEE. |
|  | Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. |
|  | Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J., 2006, June. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376). ACM. |
|  | Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A., 2014. Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227. |
|  | Mishra, A., Alahari, K. and Jawahar, C.V., 2012, September. Scene text recognition using higher order language priors. In BMVC 2012-23rd British Machine Vision Conference. BMVA. |