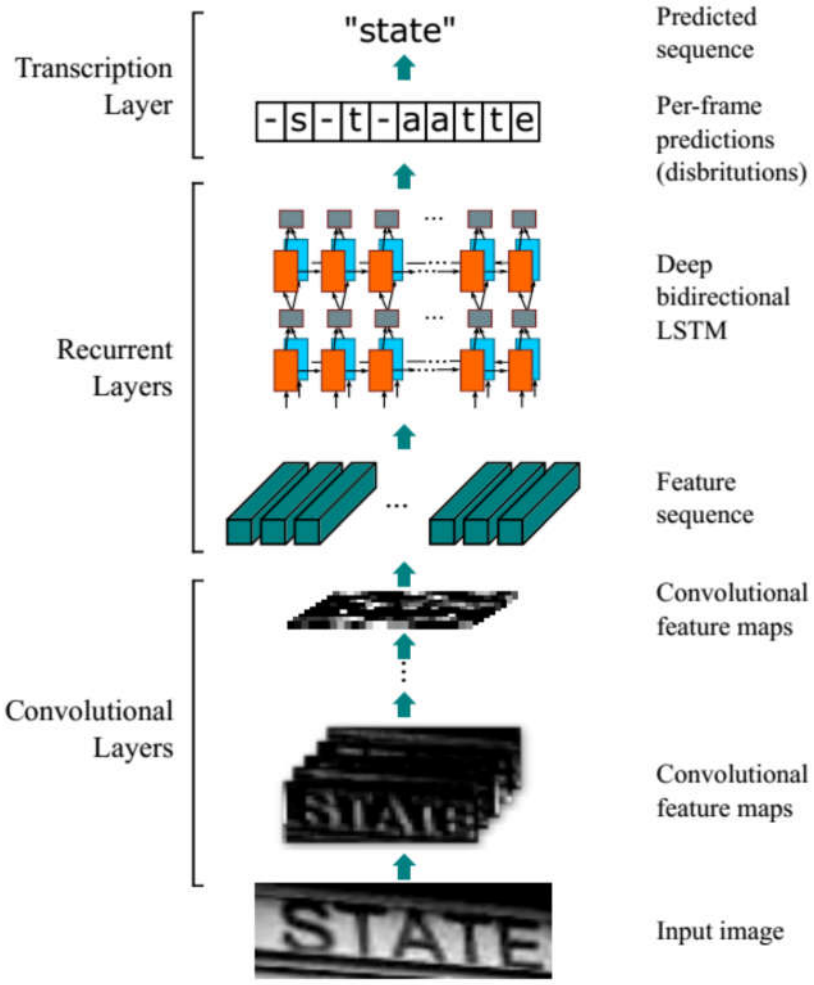
【OCR技术系列之七】端到端不定长文字识别CRNN算法详解

在以前的OCR任务中，识别过程分为两步：单字切割和分类任务。我们一般都会讲一连串文字的文本文件先利用投影法切割出单个字体，在送入CNN里进行文字分类。但是此法已经有点过时了，现在更流行的是基于深度学习的端到端的文字识别，即我们不需要显式加入文字切割这个环节，而是将文字识别转化为序列学习问题，虽然输入的图像尺度不同，文本长度不同，但是经过DCNN和RNN后，在输出阶段经过一定的翻译后，就可以对整个文本图像进行识别，也就是说，文字的切割也被融入到深度学习中去了。

现今基于深度学习的端到端OCR技术有两大主流技术：CRNN OCR和attention OCR。其实这两大方法主要区别在于最后的输出层（翻译层），即怎么将网络学习到的序列特征信息转化为最终的识别结果。这两大主流技术在其特征学习阶段都采用了CNN+RNN的网络结构，CRNN OCR在对齐时采取的方式是CTC算法，而attention OCR采取的方式则是attention机制。本文将介绍应用更为广泛的CRNN算法。

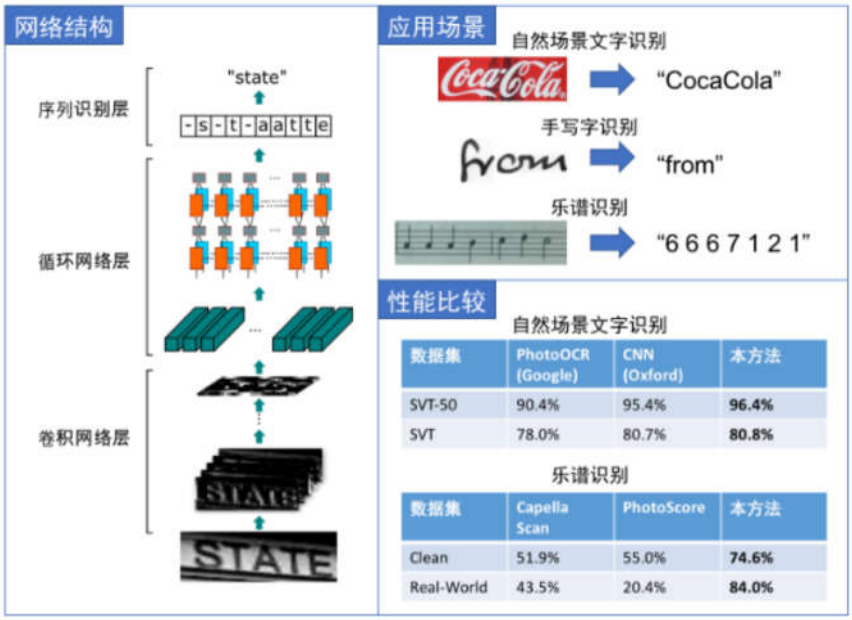


网络结构包含三部分，从下到上依次为：

1. 卷积层，使用CNN，作用是从输入图像中提取特征序列；

2. 循环层，使用RNN，作用是预测从卷积层获取的特征序列的标签（真实值）分布；

3. 转录层，使用CTC，作用是把从循环层获取的标签分布通过去重整合等操作转换成最终的识别结果；



端到端OCR的难点在哪儿呢？在于怎么处理不定长序列对齐问题！CRNN OCR其实是借用了语音识别中解决不定长语音序列的思路。与语音识别问题类似，OCR可建模为时序依赖的词汇或者短语识别问题。基于联结时序分类(Connectionist Temporal Classification, CTC)训练RNN的算法，在语音识别领域显著超过传统语音识别算法。一些学者尝试把CTC损失函数借鉴到OCR识别中，CRNN 就是其中代表性算法。CRNN算法输入100*32归一化高度的词条图像，基于7层CNN（普遍使用VGG16）提取特征图，把特征图按列切分（Map-to-Sequence），每一列的512维特征，输入到两层各256单元的双向LSTM进行分类。在训练过程中，通过CTC损失函数的指导，实现字符位置与类标的近似软对齐。

CRNN借鉴了语音识别中的LSTM+CTC的建模方法，不同点是输入进LSTM的特征，从语音领域的声学特征（MFCC等），替换为CNN网络提取的图像特征向量。CRNN算法最大的贡献，是把CNN做图像特征工程的潜力与LSTM做序列化识别的潜力，进行结合。它既提取了鲁棒特征，又通过序列识别避免了传统算法中难度极高的单字符切分与单字符识别，同时序列化识别也嵌入时序依赖（隐含利用语料）。在训练阶段，CRNN将训练图像统一缩放100×32（w × h）；在测试阶段，针对字符拉伸导致识别率降低的问题，CRNN保持输入图像尺寸比例，但是图像高度还是必须统一为32个像素，卷积特征图的尺寸动态决定LSTM时序长度。这里举个例子

现在输入有个图像，为了将特征输入到Recurrent Layers，做如下处理：

- 首先会将图像缩放到 32×W×1 大小
- 然后经过CNN后变为 1×（W/4）× 512
- 接着针对LSTM，设置 T=(W/4)， D=512 ，即可将特征输入LSTM。
- LSTM有256个隐藏节点，经过LSTM后变为长度为T × nclass的向量，再经过softmax处理，列向量每个元素代表对应的字符预测概率，最后再将这个T的预测结果去冗余合并成一个完整识别结果即可。

2018年12月 (2)
2018年10月 (1)
2018年9月 (3)
2018年5月 (1)
2018年4月 (2)
2018年2月 (6)
2018年1月 (3)
2017年12月 (4)
2017年11月 (3)
2017年10月 (1)
2017年9月 (4)
2017年8月 (3)
2017年7月 (5)
2017年6月 (4)
2017年5月 (17)
2017年4月 (1)
2017年2月 (2)
2017年1月 (5)

积分与排名

积分 - 213285
排名 - 1747

最新评论

1. Re:【OCR技术系列之四】基于深度学习的文字识别（3755个汉字）
感谢博主的无私奉献
--寒冬夜行人lee

2. Re:我的2018：OCR、实习和秋招
学长好厉害！我实习内容也是在做OCR，每天照着你的好多博客看...我是18级研究生2020年毕业然后最近也在边实习边准备秋招（但是我好菜），可不可以加学长微信交流交流呀，我的微信是Dreaminice.....
--Cocoalate

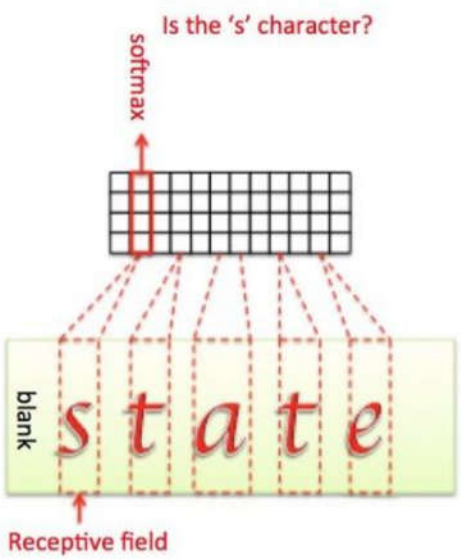
3. Re:【Keras】基于SegNet和U-Net的遥感图像语义分割
@wenny-bell我也有这样的问题，请问您解决了吗？可以加下qq讨论下，2724858160...
--嗯哼！！！

4. Re:【Keras】基于SegNet和U-Net的遥感图像语义分割
@wenny-bell我也有这样的问题，请问您解决了吗？可以加下qq讨论下，2724858160...
--嗯哼！！！

5. Re:OpenCV探索之路（十一）：轮廓查找和多边形包围轮廓
请教下，int thresh_size = (100 / 4) * 2 + 1; //自适应二值化阈值这个阈值的定义是怎么给出的呢？用你设定的参数确实得到了很好的效果，但是不知道为什么这样设置.谢谢~.....
--foxlucia

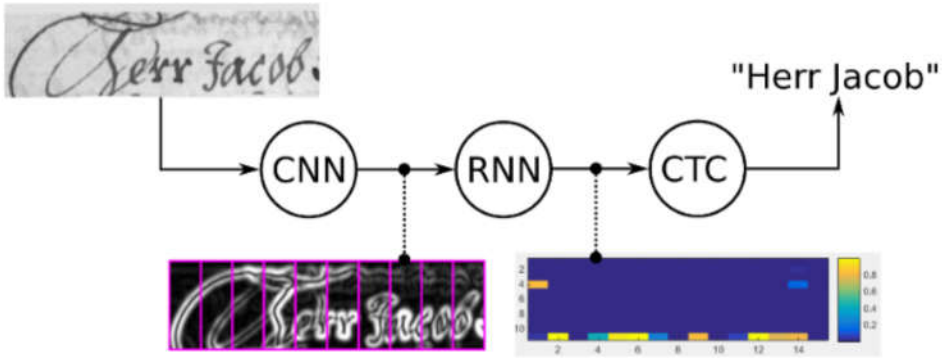
阅读排行榜

1. 卷积神经网络CNN总结(219292)
2. 基于深度学习的目标检测技术演进：R-CNN、Fast R-CNN、Faster

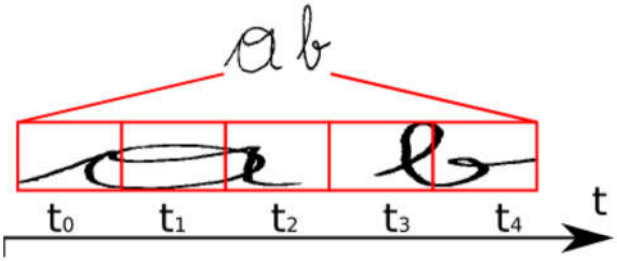


CRNN中需要解决的问题是图像文本长度是不定长的，所以会存在一个对齐解码的问题，所以RNN需要一个额外的搭档来解决这个问题，这个搭档就是著名的CTC解码。
CRNN采取的架构是CNN+RNN+CTC，cnn提取图像像素特征，rnn提取图像时序特征，而ctc归纳字符间的连接特性。

那么CTC有什么好处？因手写字符的随机性，人工可以标注字符出现的像素范围，但是太过麻烦，ctc可以告诉我们哪些像素范围对应的字符：



我们知道，CRNN中RNN层输出的一个不定长的序列，比如原始图像宽度为W，可能其经过CNN和RNN后输出的序列个数为S，此时我们要将该序列翻译成最终的识别结果。RNN进行时序分类时，不可避免地会出现很多冗余信息，比如一个字母被连续识别两次，这就需要一套去冗余机制，但是简单地看到两个连续字母就去冗余的方法也有问题，比如cook，geek一类的词，所以CTC有一个blank机制来解决这个问题。这里举个例子来说明。



如上图所示，我们要识别这个手写体图像，标签为“ab”，经过CNN+RNN学习后输出序列向量长度为5，即t0~t4，此时我们要将该序列翻译为最后的识别结果。我们在翻译时遇到的第一个难题就是，5个序列怎么转化为对应的两个字母？重复的序列怎么解决？刚好位于字与字之间的空白的序列怎么映射？这些都是CTC需要解决的问题。

我们从肉眼可以看到，t0,t1,t2时刻都应映射为“a”，t3,t4时刻都应映射为“b”。如果我们将连续重复的字符合并成一个输出的话，即“aaabb”将被合并成“ab”输出。但是这样子的合并机制是有问题的，比如我们的标签图像时“aab”时，我们的序列输出将可能会是“aaaaaaabb”，这样子我们就没办法确定该文本应被识别为“aab”还是

R-CNN(206007)
3. OpenCV探索之路（二十四）图像拼接和图像融合技术(88662)
4. CNN网络架构演进：从LeNet到DenseNet(58613)
5. OpenCV探索之路（二十三）：特征检测和特征匹配方法汇总(54127)
6. 【OCR技术系列之四】基于深度学习的文字识别（3755个汉字）(51232)
7. C++ STL快速入门(45935)
8. OpenCV探索之路（六）：边缘检测（canny、sobel、laplacian）(45154)
9. Linux编程之UDP SOCKET全攻略(41918)
10. OpenCV探索之路（十四）：绘制点、直线、几何图形(37782)

评论排行榜

1. 【OCR技术系列之四】基于深度学习的文字识别（3755个汉字）(82)
2. 【Keras】基于SegNet和U-Net的遥感图像语义分割(73)
3. OpenCV探索之路（二十四）图像拼接和图像融合技术(67)
4. 【OCR技术系列之八】端到端不定长文本识别CRNN代码实现(59)
5. 【Keras】从两个实际任务掌握图像分类(33)

推荐排行榜

1. 基于深度学习的目标检测技术演进：R-CNN、Fast R-CNN、Faster R-CNN(92)
2. 卷积神经网络CNN总结(61)
3. 我的2018：OCR、实习和秋招(22)
4. 【OCR技术系列之四】基于深度学习的文字识别（3755个汉字）(21)
5. 【Keras】基于SegNet和U-Net的遥感图像语义分割(20)
6. CNN网络架构演进：从LeNet到DenseNet(20)
7. OpenCV探索之路（二十四）图像拼接和图像融合技术(18)
8. 我在北京实习的四个月(15)
9. 【OCR技术系列之一】字符识别技术总览(13)
10. 读研以来的一些感想：名校好在哪里？(13)

“ab”。CTC为了解决这种二义性，提出了插入blank机制，比如我们以“-”符号代表blank，则若标签为“aaa-aaaabb”则将被映射为“aab”，而“aaaaaaabb”将被映射为“ab”。引入blank机制，我们就可以很好地处理了重复字符的问题了。

但我们还注意到，“aaa-aaaabb”可以映射为“aab”，同样地，“aa-aaaaabb”也可以映射为“aab”，也就是说，存在多个不同的字符组合可以映射为“aab”，更总结地说，一个标签存在一条或多条的路径。比如下面“state”这个例子，也存在多条不同路径映射为"state"：

$$B(\pi_1) = B(--stta-t---e) = state$$

$$B(\pi_2) = B(sst-aaa-tee-) = state$$

$$B(\pi_3) = B(--stta-tee-) = state$$

$$B(\pi_4) = B(sst-aa-t---e) = state$$

上面提到，RNN层输出的是序列中概率矩阵，那么$p(\pi = --stta-t---e|x,S) = \prod_{t=1}^{T} y_{\pi_t}^t = (y_-^1) \times (y_-^2) \times (y_s^3) \times (y_t^4) \times (y_t^5) \times (y_a^6) \times (y_-^7) \times (y_t^8) \times (y_-^9) \times (y_-^{10}) \times (y_-^{11}) \times (y_e^{12})$

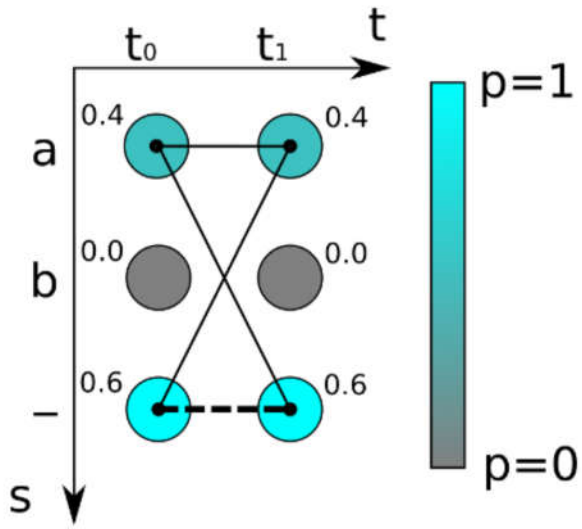
其中，y_-^1 表示第一个序列输出“-”的概率，那么对于输出某条路径π的概率为各个序列概率的乘积。所以要得到一个标签可以有多个路径来获得，从直观上理解就是，我们输出一张文本图像到网络中，我们需要使得输出为标签L的概率最大化，由于路径之间是互斥的，对于标注序列，其条件概率为所有映射到它的路径概率之和：

$$p(l|x) = \sum_{\pi \in B^{-1}(l)} p(\pi|x)$$

其中$\pi \in B^{-1}(l)$的意思是，所有可以合并成l的所有路径集合。

这种通过映射B和所有候选路径概率之和的方式使得CTC不需要对原始的输入序列进行准确的切分，这使得RNN层输出的序列长度>label长度的任务翻译变得可能。CTC可以与任意的RNN模型，但是考虑到标注概率与整个输入串有关，而不是仅与前面小窗口范围的片段相关，因此双向的RNN/LSTM模型更为适合。

ctc会计算loss，从而找到最可能的像素区域对应的字符。事实上，这里loss的计算本质是对概率的归纳：



如上图，对于最简单的时序为2的（t0t1）的字符识别，可能的字符为“ a ”，“ b ”和“ - ”，颜色越深代表概率越高。我们如果采取最大概率路径解码的方法，一看就是“--”的概率最大，真实字符为空即“”的概率为0.6*0.6=0.36。

但是我们忽略了一点，真实字符为“ a ”的概率不只是”aa” 即0.4*0.4， 事实上，“aa”, “a-“和“-a”都是代表“ a ”，所以，输出“ a ”的概率为：

```
0.4*0.4 + 0.4 * 0.6 + 0.6*0.4 = 0.16+0.24+0.24 = 0.64
```

所以“ a ”的概率比空“”的概率高！可以看出，这个例子里最大概率路径和最大概率序列完全不同，所以CTC解码

通常不适合采用最大概率路径的方法，而应该采用前缀搜索算法解码或者约束解码算法。

通过对概率的计算，就可以对之前的神经网络进行反向传播更新。类似普通的分类，CTC的损失函数O定义为负的最大似然，为了计算方便，对似然取对数。

$$O = -ln(\prod_{(x,z)\in S} p(l|x)) = -\sum_{(x,z)\in S} lnp(l|x)$$

我们的训练目标就是使得损失函数O优化得最小即可。

分类: OCR系列

Madcola
关注 - 30
粉丝 - 1334
+加关注

3 0

« 上一篇: 我的2018：OCR、实习和秋招
» 下一篇: 【OCR技术系列之八】端到端不定长文本识别CRNN代码实现

posted @ 2019-01-29 20:21 Madcola 阅读(8425) 评论(3) 编辑 收藏

评论列表

#1楼 2019-05-18 22:12 过山车小熊

求参考文献，谢谢博主！

支持(1) 反对(0)

#2楼 2019-07-02 09:02 牧童祭歌

结合博主和其他博文对CRNN有点了解了，希望博主有时间可以出一期前缀搜索算法解码或者约束解码算法的博文。

支持(0) 反对(0)

#3楼 2019-07-19 09:27 Danydai19

非常容易理解，请问博主的文章是否可以转载到公众号呢～会注明作者来源

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

相关博文:

- 端到端文本识别CRNN论文解读
- 【OCR技术系列之八】端到端不定长文本识别CRNN代码实现
- 【OCR技术系列之一】字符识别技术总览
- 【OCR技术系列之四】基于深度学习的文字识别（3755个汉字）
- 【OCR技术系列之六】文本检测CTPN的代码实现

最新新闻:
· 知乎周源发融资全员信强调工作效率：快则生慢则死
· 到处作恶的台风究竟是怎么来的？背后竟藏着这些秘密…
· 头条搜索无战事
· 小红书否认"被下架"前正在进行新一轮融资
· 三星发布1.08亿像素图像传感器ISOCELL Bright HMX，小米将首发
» 更多新闻...