图像识字

一、相关论文

2015: Convolution Recurrent Neural Network

2015: Deep-Text Recurrent Network

2017: Fucus Attention Network

2017: Single Shot Text Detector

2018: Mask Text Spotter

二、初探： Convolution Recurrent Neural Network

这篇论文来自于华科白翔老师团队，将图像特征提取(feature extraction)和序列模型(sequence modeling)整合到一个完整的网络，这个神经网络则被称为CRNN(Convolution Recurrent Neural Network)。CRNN网络的整体架构如下图所示，从下往上看，CNN从给定的图片中提取特征序列(feature sequence)，在CNN网络之上，RNN(bi-LSTM)将根据卷积层生成的特征序列做预测，最后，在CRNN之上的transcription layer将RNN层的预测结果翻译成一个label sequence。

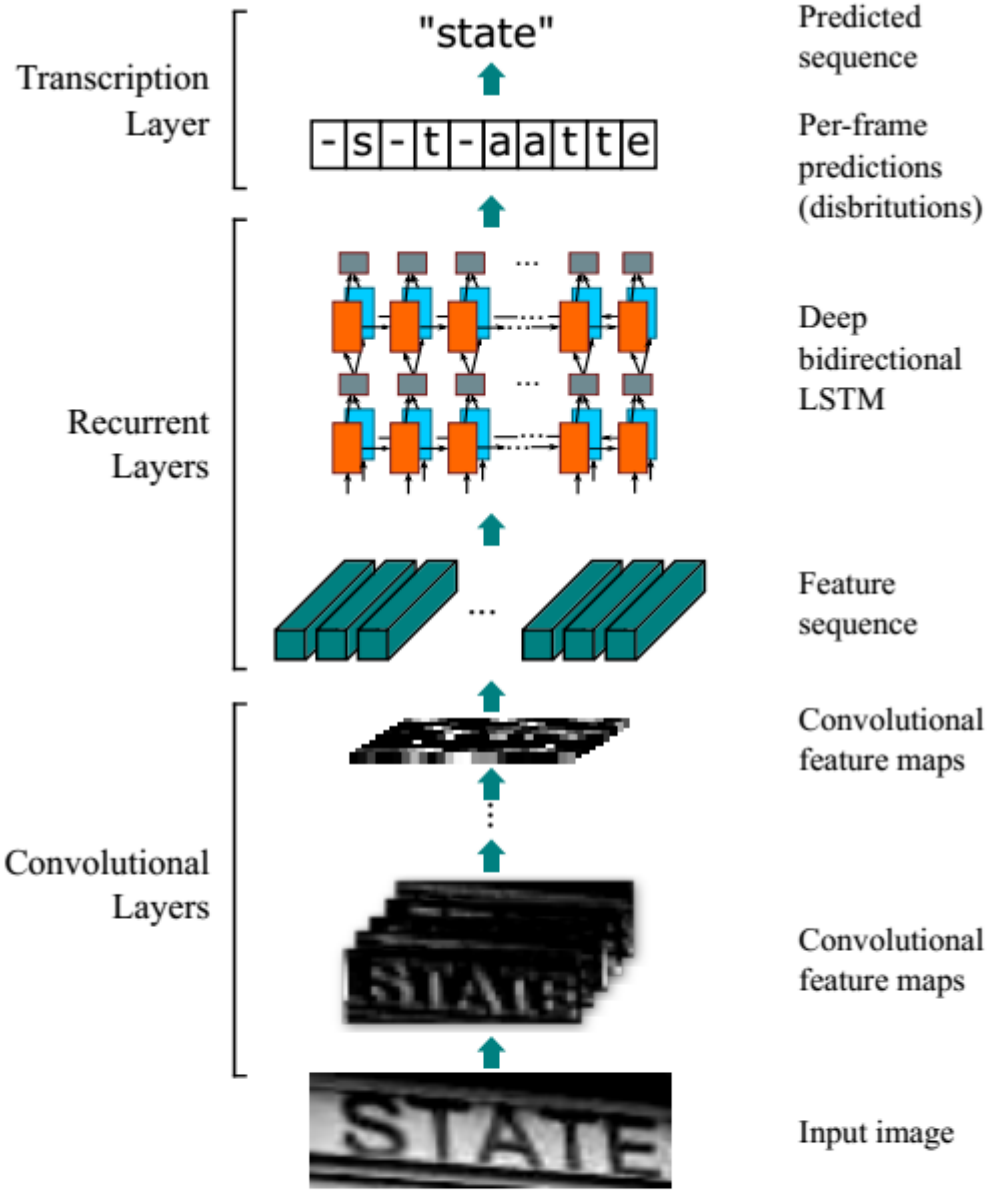

Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

在CRNN中卷积组件是由卷积和最大池化层组成，卷积层的作用是从输入图像中提取特征序列。特征序列（feature sequence）中的每一个特征向量(feature vector)是从特征图(feature maps)中从左向右按列生成，也就是说，第i-th个特征向量是由所有特征图第i列生成的特征向量连接而成，每一列的宽度被设置为一个像素。因为卷积层，最大池化和按元素的激活函数的平移不变性(translation invariant)，所以每一列的特征图对应着原图中的一个矩形区域，这个矩形区域则被称为receptive field，而且这个矩形区域是与其在特征图中相对应的从左向右的列是相同顺序的，那么就可以认为特征序列中的每一个特征向量也是与原图中的每一个receptive field相对应的，输入图像是要固定为同一height值的

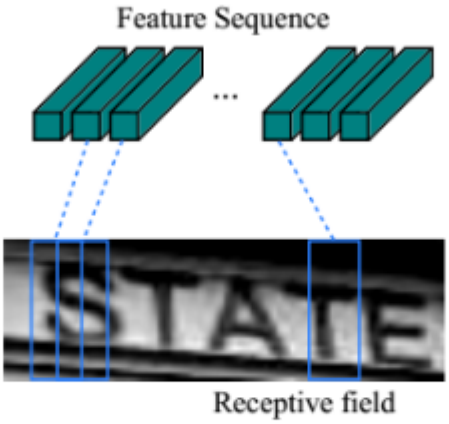

Figure 2. The receptive field. Each vector in the extracted feature sequence is associated with a receptive field on the input image, and can be considered as the feature vector of that field.

在卷积层后接上一个双向RNN(bidirectional RNN)作为recurrent layer, RNN是将每一个在卷积层生成的特征序列中的预测为一个label distribution 。选择RNN的原因一是在于RNN是有很强的能力捕捉到一个序列的上下文信息，在上述的特征提取中，可以看到一个宽字符可能有好几个连续的receptive field描述，对基于图像的序列识别使用上下文比将单个字符单独对待要更加有效，而且对于一些含糊的字符，观察其上下文信息后也会很好区分;而是因为RNN也能够后向传播(back-propagates)进行权重更新，从而使得可以将CNN和RNN连接成一个完整的网络;三是因为RNN可以处理任意长度的序列，上文中对输入图片的高进行固定，是固定每一个receptive field的大小(每一个receptive field的宽为一个像素，也是固定的)，那么这样的话，可以对任意宽度的图像进行处理。

CRNN中的RNN架构选择的是双向LSTM。

Transcription的作用是将RNN生成的预测转化成一个label sequence。论文里采用的CTC(Connectionist Temporal Classification)层中定义的条件概率，该概率是指在RNN生成的预测的条件下得到label sequence 的概率。因此就可以使用该概率的负log-likelihood作为目标函数进行训练网络，那么数据仅仅需要图片和图片相对应的label sequence即可。

假设经过recurrent layer后的输出为，每一个是在集合上概率分布，其中L包含任务中的所有lable和一个空白blank label。在序列上定义一个sequence-to-sequence的函数映射B，B是将映射到，通过先移除相同的labels，随后移除空白。那么条件概率则被定义为B将所有映射到的概率和， 其中是指在t时刻存在label 的概率。

transcription分为lexicon-free transcription 和 lexicon-based transcription，具体计算方法还得阅读其他论文，目前这一块还不是重点，得到实现时再细看。

网络架构借鉴了VGG架构，输入图片的高度值固定为32，网络训练时采用SGD，对于optimization则使用Adadelta。

Table 1. Network configuration summary. The first row is the top layer. 'k', 's' and 'p' stand for kernel size, stride and padding size respectively

Type	Configurations
Transcription	-
Bidirectional-LSTM	#hidden units:256
Bidirectional-LSTM	#hidden units:256
Map-to-Sequence	-
Convolution	#maps:512, k:2 × 2, s:1, p:0
MaxPooling	Window:1 × 2, s:2
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
BatchNormalization	-
Convolution	#maps:512, k:3 × 3, s:1, p:1
MaxPooling	Window:1 × 2, s:2
Convolution	#maps:256, k:3 × 3, s:1, p:1
Convolution	#maps:256, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:128, k:3 × 3, s:1, p:1
MaxPooling	Window:2 × 2, s:2
Convolution	#maps:64, k:3 × 3, s:1, p:1
Input	$W \times 32$ gray-scale image

总结:

- CRNN是端到端的训练网络，直接从序列标签中学习
- 对识别的对象不限定长度，仅仅需要height normalization
- 带字典和不带样本的样本都可以

三、初探： Deep-Text Recurrent Network

DTRN是来自于深圳先进技术研究院乔宇老师团队，将场景文字识别看作序列标注问题。思想和上述的白翔老师团队的类似，都是充分利用CNN和RNN的特性，CNN提取特征，RNN预测序列，将两者整合到一个统一的网络里，实现端到端的训练。

DTRN总架构如下图所示，整个系统是端到端的，它以一个带有单词的图片作为输入，直接输出相对应的单词序列，不需要预处理和后向处理等那等操作。首先由一个特殊的CNN网络架构，将输入图片编码成一个有顺序的序列(an ordered swquence)，然后使用一个RNN网络将CNN输出的序列解码成一个单词串。对于输入的图像和输出的单词串的长度是不加限制的。

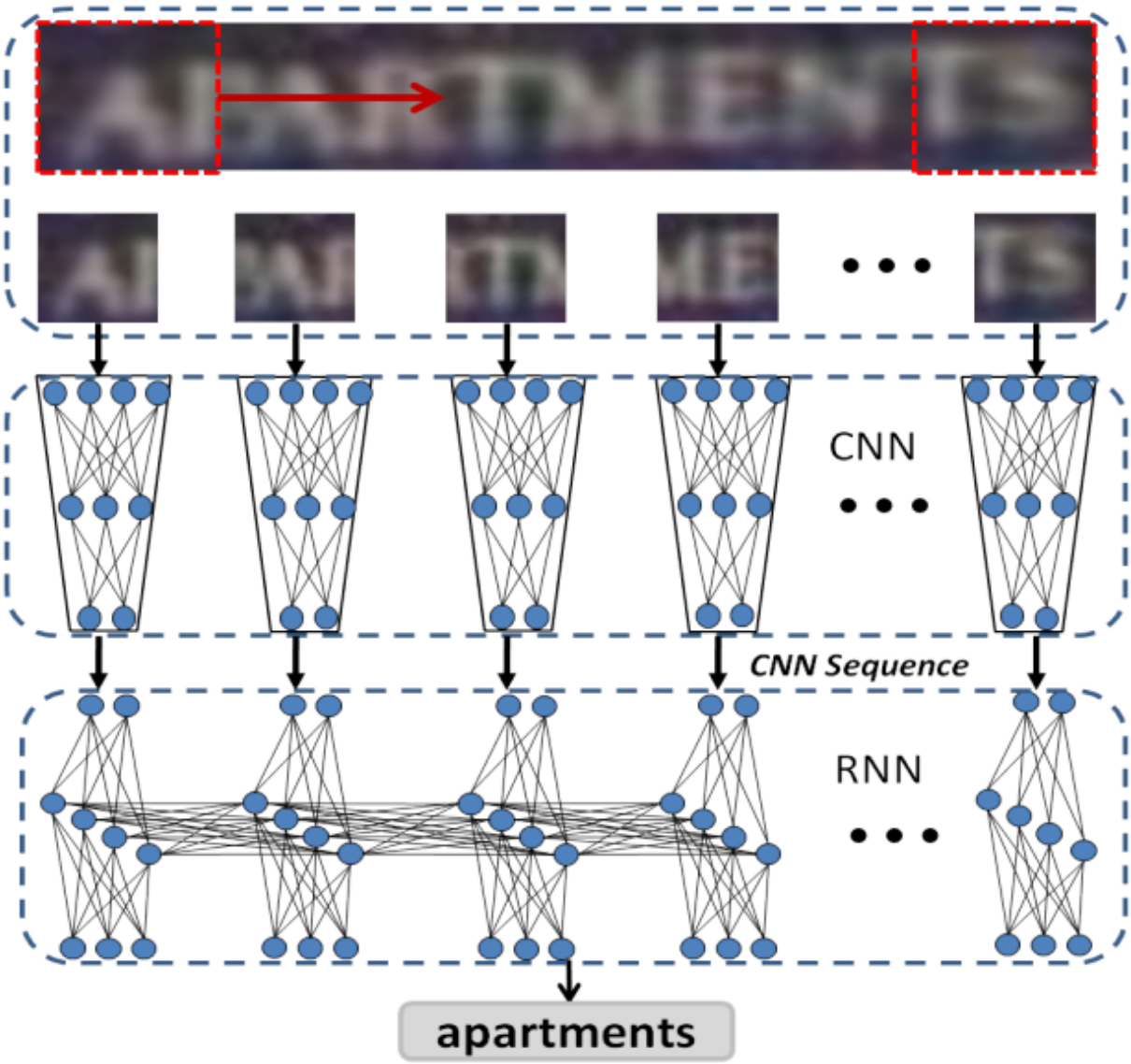

Figure 1: The word image recognition pipeline of the proposed *Deep-Text Recurrent Networks (DTRN)* model.

通过在输入图片上滑动子窗口的形式保证在CNN模型下输出一个有序的特征序列。CNN网络架构如下，CNN的输入图片大小为32×32，与滑窗的窗口大小一致，maxout CNN模型有五个卷积层，每个卷积层后面跟随two or four-group Maxout操作。对于输入的图片，将它的高度resize成32，并且保持它原先的纵横比。

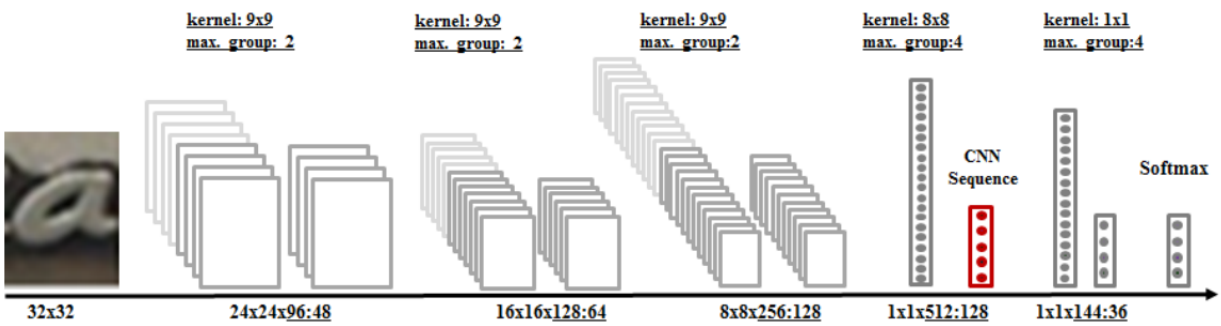

Figure 2: The structures of our maxout CNN model.

RNN选择双向LSTM架构，由于LSTM输出的长度与目标单词串的长度不一致，因此引入CTC将LSTM的序列输出匹配到目标字符串，这个和上面的RCNN类似，不再描述。

总结：

- 端到端训练
- 可以利用有用的文本信息识别高度模糊的词汇，不需要预处理和后期处理
- 深度CNN特征足可以抵抗严重扭曲的单词
- 包含词汇图片中明确的顺序信息，这是划分单词串的根本
- 不依赖预先定义的词典，可以处理未知单词和任意单词串

参考文献

https://blog.csdn.net/weixin_42111770/article/details/84838493