总结

黄飞虎

2021.1.13

学习情况

把李宏毅的cnn与rnn的视频看完，完成了实验四。

```
loading data ...
Get embedding ...
loading word to vec model ...
get words #24694
total words: 24696
Train | Len:180000
Valid | Len:20000

start training, parameter total:6415351, trainable:241351

Epoch | 1/10
Train | Loss:0.49640 Acc: 75.025
Valid | Loss:0.45211 Acc: 78.379
-----------------------------------------------
```

```
Epoch | 8/10
Train | Loss:0.36431 Acc: 83.522
Valid | Loss:0.42205 Acc: 79.847
-----------------------------------------------
Epoch | 9/10
Train | Loss:0.34834 Acc: 84.386
Valid | Loss:0.44188 Acc: 80.046
-----------------------------------------------
Epoch | 10/10
Train | Loss:0.33075 Acc: 85.252
Valid | Loss:0.43602 Acc: 79.220
-----------------------------------------------
```

论文情况

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

发表期刊： IEEE Transactions on Pattern Analysis and Machine Intelligence

论文链接：https://arxiv.org/pdf/1507.05717.pdf

本文的工作

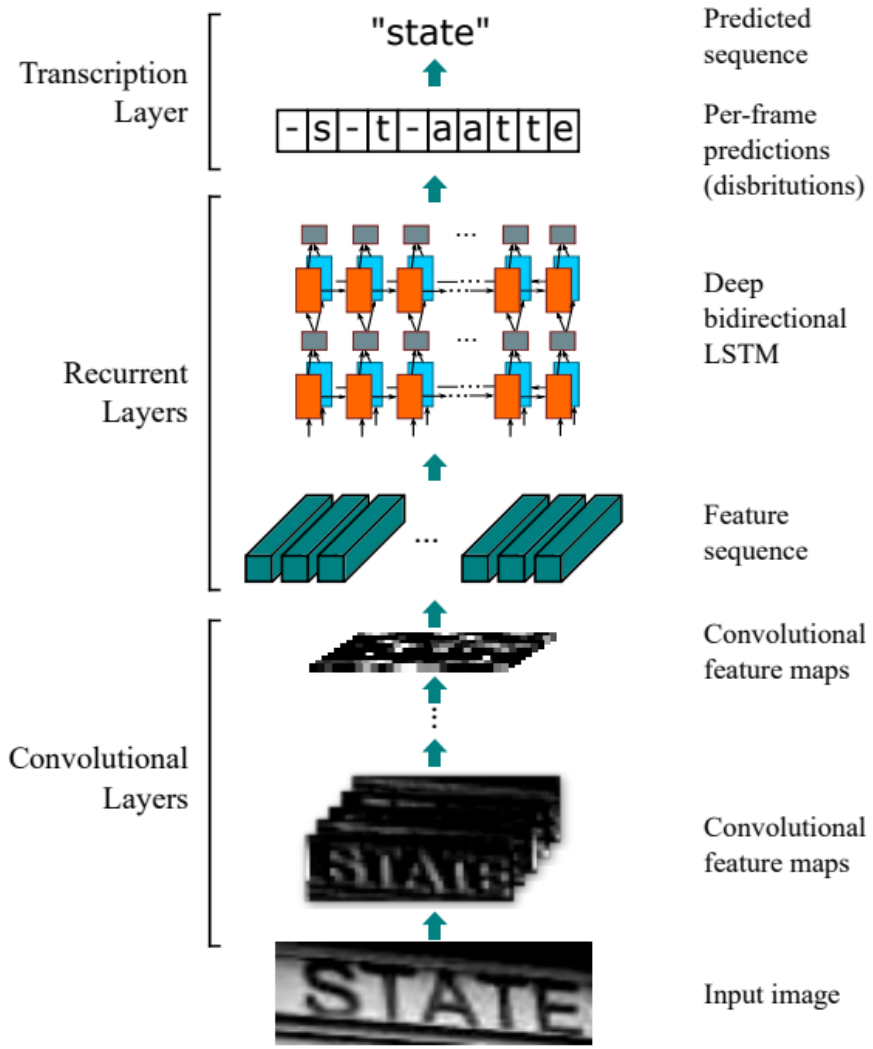
针对场景文本识别的问题，本文提出了Convolutional Recurrent Neural Network（CRNN），是DCNN与RNN的结合体。

与其他模型作了比较分析

与以前的场景文本识别相比，具有四个独特的属性：

- 1.它是端到端可训练的，与大多数现有算法相反，其中的组件是单独训练和调整的。
- 2.它自然地处理任意长度的序列，不涉及字符分割或水平缩放归一化。
- 3.它不局限于任何预定义的词典，并且在无词典和词典的场景文本识别任务中都取得了显着的表现。
- 4.它生成一个有效但小得多的模型，这对于实际应用场景更为实用。

CRNN的网络结构包括三个组成部分：卷积层、循环层和转录层。如下图所示：



转录层，它将每帧的预测转换为最终的标签序列。

循环层，预测每个帧的标签分布。

卷积层，从输入图像中提取特征序列。

实验

对于场景文本识别的所有实验，使用Jaderberg等人发布的合成数据集（Synth）作为训练数据。数据集包含8百万训练图像及其对应的实际单词。

有四个流行的基准数据集用于场景文本识别的性能评估，即ICDAR 2003（IC03），ICDAR 2013（IC13），IIIT 5k-word（IIIT5k）和Street View Text (SVT)。

比较评估

通过本文提出的CRNN模型和最新的技术，包括基于深度模型的方法，对4个公共数据集的识别精度均如下表所示。

	IIIT5k			SVT		IC03				IC13
	50	**1k**	**None**	**50**	**None**	**50**	**Full**	**50k**	**None**	**None**
ABBYY [34]	24.3	-	-	35.0	-	56.0	55.0	-	-	-
Wang *et al.* [34]	-	-	-	57.0	-	76.0	62.0	-	-	-
Mishra *et al.* [28]	64.1	57.5	-	73.2	-	81.8	67.8	-	-	-
Wang *et al.* [35]	-	-	-	70.0	-	90.0	84.0	-	-	-
Goel *et al.* [13]	-	-	-	77.3	-	89.7	-	-	-	-
Bissacco *et al.* [8]	-	-	-	90.4	78.0	-	-	-	-	87.6
Alsharif and Pineau [6]	-	-	-	74.3	-	93.1	88.6	85.1	-	-
Almazán *et al.* [5]	91.2	82.1	-	89.2	-	-	-	-	-	-
Yao *et al.* [36]	80.2	69.3	-	75.9	-	88.5	80.3	-	-	-
Rodrguez-Serrano *et al.* [30]	76.1	57.4	-	70.0	-	-	-	-	-	-
Jaderberg *et al.* [23]	-	-	-	86.1	-	96.2	91.5	-	-	-
Su and Lu [33]	-	-	-	83.0	-	92.0	82.0	-	-	-
Gordo [14]	93.3	86.6	-	91.8	-	-	-	-	-	-
Jaderberg *et al.* [22]	97.1	92.7	-	95.4	80.7*	**98.7**	**98.6**	93.3	**93.1***	**90.8***
Jaderberg *et al.* [21]	95.5	89.6	-	93.2	71.7	97.8	97.0	93.4	89.6	81.8
CRNN	**97.6**	**94.4**	**78.2**	**96.4**	**80.8**	**98.7**	97.6	**95.5**	89.4	86.7

为了进一步了解与其它文本识别方法相比，所提出算法的优点，我们提供了在一些特性上的综合比较，这些特性名称为E2E Train，Conv Ftrs，CharGT-Free，Unconstrained和Model Size，如下图所示。

	E2E Train	Conv Ftrs	CharGT-Free	Unconstrained	Model Size
Wang *et al.* [34]	✗	✗	✗	✔	-
Mishra *et al.* [28]	✗	✗	✗	✗	-
Wang *et al.* [35]	✗	✔	✗	✔	-
Goel *et al.* [13]	✗	✗	✔	✗	-
Bissacco *et al.* [8]	✗	✗	✗	✔	-
Alsharif and Pineau [6]	✗	✔	✗	✔	-
Almazán *et al.* [5]	✗	✗	✔	✗	-
Yao *et al.* [36]	✗	✗	✗	✔	-
Rodrguez-Serrano *et al.* [30]	✗	✗	✔	✗	-
Jaderberg *et al.* [23]	✗	✔	✗	✔	-
Su and Lu [33]	✗	✗	✔	✔	-
Gordo [14]	✗	✗	✗	✗	-
Jaderberg *et al.* [22]	✔	✔	✔	✗	490M
Jaderberg *et al.* [21]	✔	✔	✔	✔	304M
CRNN	✔	✔	✔	✔	**8.3M**

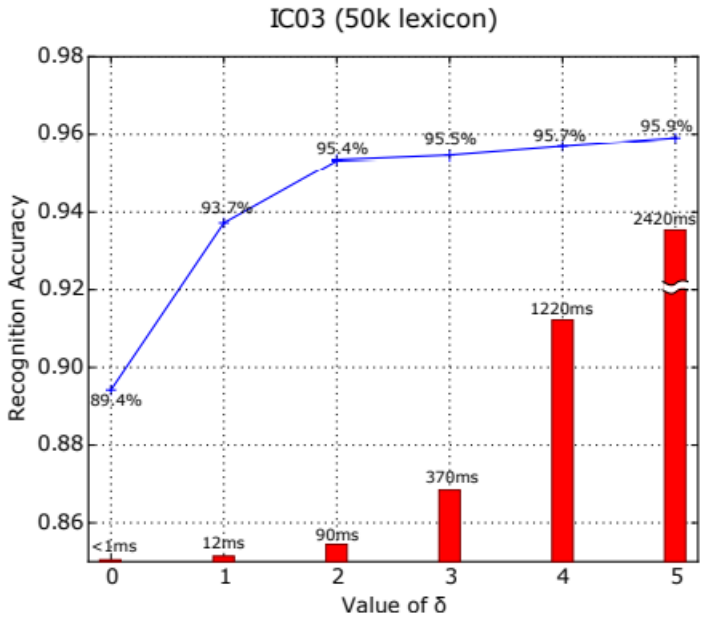
E2E Train：这一列是为了显示某种文字阅读模型是否可以进行端到端的训练。

Conv Ftrs：这一列用来表明一个方法是否使用从训练图像直接学习到的卷积特征或手动特征作为基本的表示。

CharGT-Free：这一列用来表明字符级标注对于训练模型是否是必要的。由于CRNN的输入和输出标签是序列，因此字符级标注是不必要的。

Unconstrained：这一列用来表明训练模型是否受限于一个特定的字典，是否不能处理字典之外的单词或随机序列。

Model Size：这一列报告了学习模型的存储空间。在CRNN中，所有的层有权重共享连接，不需要全连接层。



蓝线:识别精度作为函数参数δ。

红色:每个样本的词典搜索时间。

在IC03数据集上使用50k字典进行测试。

CRNN 在文字识别上的优点:

1.它是端到端的

2.能处理任意长度的序列

3.不需要预定义的字典

4.更小的模型，更少的参数（不需要全连接层）

对于序列类型的对象，CRNN所具有的优点:

1.可以直接从标签序列上进行学习（例如单词），而不需要进行另外的标注（每个字母）

2.它可以直接从图片中读取信息，无需手工设计特征

3.和 RNN 具有同太阳的优点，可产生一个序列的标签

4.只要求序列的高度标准化，不受限于序列的长度

5.文字识别任务上有出色表现

6.更少的参数，更小的存储空间

谢谢观看