DCNN_ACL2014_oxford_Kalchbrenner

2015年7月25日 星期六　　下午10:32

一、概述

DCNN，cnn in nlp的经典之作，在nlp领域仅仅比CW的晚。
主要思想就是cnn的思想，用于了解CNN挺不错的。
与标准CNN的区别是，增加了dynamtic的max-k pooling。简单说max-k pooling是在卷积之后选出最大的k个nodes，注意这k个nodes是保序的。dynamtic是指k是动态的，顶层的k小，底层的大，有个公式。

二、背景

1、其他的NN有bag-of-words的，也有recurvise、recurrent的
2、CNN之前有CW的，1989年hinton的TDNN(time delay NN)，CW可以说是TDNN的sentence model的延伸
3、名词解释：feature map的卷积窗口，在这里称为filter
4、为啥是4d-tensor？因为weight是连接的是两个矩阵的？
5、recurrent NN可以看做是recurvise NN的一个special case

三、进一步背景

1、卷积

m是卷积的weight，也叫the filter of convlution，长度m
s是输入，sequence，长度s，s中在nlp里就是每个words的序列
在一维的convlution中，最简单的表示形式（s_i和m_i这些都只看做是一个数，而不是vector、embedding什么的）

$$\mathbf{c}_j = \mathbf{m}^\mathsf{T}\mathbf{s}_{j-m+1:j}$$

这里有两种卷积：窄卷积的、宽卷积的

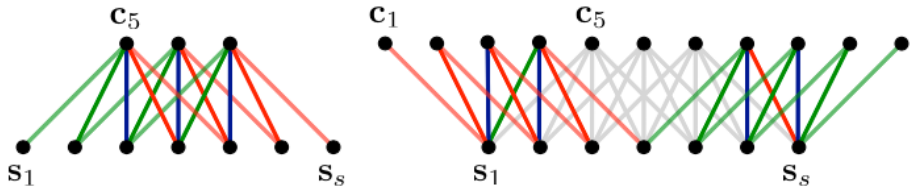


Figure 2: Narrow and wide types of convolution. The filter $\mathbf{m}$ has size $m = 5$.

A 窄卷积：

卷积的过程是相当于从第一个词开始做窗口（窗长为m），一直向

后滑窗。直到窗口末尾怼到最后一个词（窗口开头在s-m+1）？？？

可以按照上面的公式对照一下下标，上面公式中，窄卷积的情况下，下标j的范围是m到s

这个的必要条件是，s>m

我们实现的就是这种，加padding两个目的：

A 相当每个词都做一回中心词

B 应为s 不够长的情况

B 宽卷积：

按照上面的公式对照一下下标，上面公式中，宽卷积的情况下，下标j的范围是1到s+m-1，如果出现下标==0，则当做0

具体卷积过程，还需要再认真想想

对s 和 m的关系不做限制

宽卷积的好处是，卷积之后不会有空的结果，无论卷积长m和sequence长s如何变化

2、TDNN （学习之前的CNN，如CW的）

1、和传统的CNN一样，上面卷积讲的sequence s的每个词，不仅仅是一个value，而是一个embedding，记维度为d；相应的m也是 m*d的

$$\mathbf{s} = \begin{bmatrix} | & | & | \\ \mathbf{w}_1 & \dots & \mathbf{w}_s \\ | & | & | \end{bmatrix}$$

如图，很简单，w是每个词，1~s个词构成的序列是s

2、TDNN的T指的是time，指的是sequence是有一个时间维度的信息，具体没说

3、分析

卷积有缺陷：

受卷积窗口限制，高度有序(higher-order)和long-range的feature学的不好

Max-pooling有缺陷：

直接把卷积之后所有词加一起，不知道出现的特征是出现一次还是多次，不知道之间的顺序。

Ps:我感觉这个影响面非常小吧，值得因为这个加K-max-pooling吗

四、模型情况详述

1、wide conconvlution

常规的宽卷积，下标和矩阵维度详见paper

2、k-max pooling

非常简单，就是在卷积之后的layer中，选出k个最大的，而且这k个是保序的

3、dynamic k-max pooling

动态指的是，k这个参数是动态的，根据一共有多少卷积层L、目前的卷积是第几层l、句子长度来确定s，当然有个最高层的k固定值（k_top）

4、Non-linear feature function

这个没看懂，先跳过

5、多个feature map

这个就是常规的，但是他的公式没看懂，有时间看看

为啥是4d-tensor，没看懂

6、folding

由于不同的row（可能是一个sequence中的词、特征等，而不是词的embedding、特征的vector）

（注意，这样看row和colume还是有差别的，要好好研究研究、考虑考虑）

在top的full connect隐层之前，不同的row之间互相见不到，所有可能一些实际有关系的特征，在这里学不出来，到最后一层才相见，最后一层学出来的效果也不一定好。

所以要让他们提前相见，这个folding就是，在convlution layer与pooling layer之间，增加一个folding，做的是相邻的两个nodes做简单叠加 的操作。这样使得维度d的vector，变成了d/2的

这样就让他们提前相见了

我们引入folding，就是相邻

五、模型特性分析

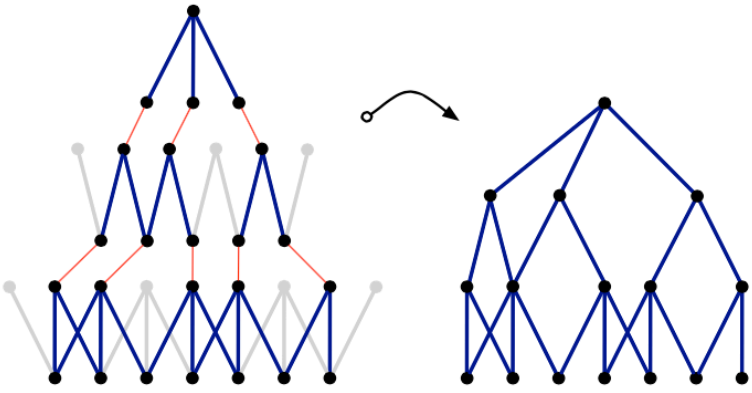
1、word和n-gram

可以学到n-gram的信息，因为是广义的卷积

可以捕捉到词的顺序信息

2、推理出feature gragh

由于过程中有polling，所以部分特征被drop掉了，留下的形成了feature graph的图状结构，如图



The cat sat on the red mat The cat sat on the red mat

后面几段没看了

六、参考资料

在论文上做的批注

Kalchbrenne
r_DCNN_...