textCNN:

1.简介

TextCNN 是利用卷积神经网络对文本进行分类的算法，由 Yoon Kim 在 “Convolutional Neural Networks for Sentence Classification” 一文 (见参考[1]) 中提出. 是2014年的算法.

2.参数与超参数

sequence\_length  
Q: 对于CNN, 输入与输出都是固定的，可每个句子长短不一, 怎么处理?  
A: 需要做定长处理, 比如定为n, 超过的截断, 不足的补0. 注意补充的0对后面的结果没有影响，因为后面的max-pooling只会输出最大值，补零的项会被过滤掉.

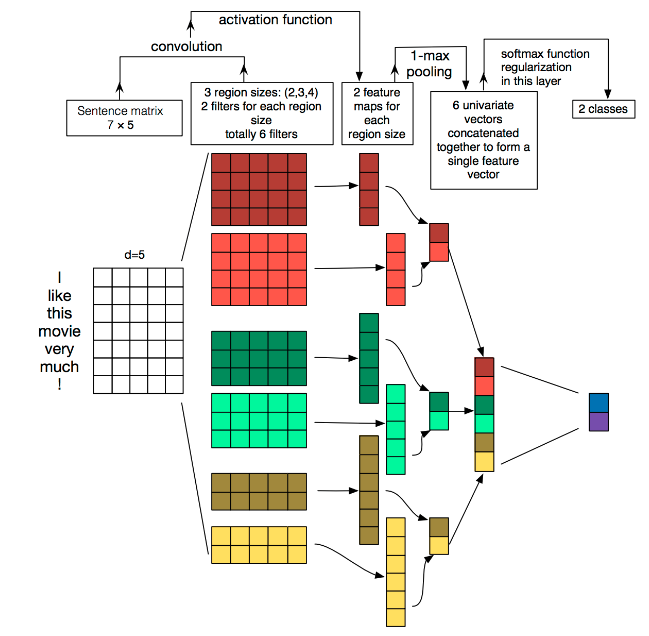
num\_classes  
多分类, 分为几类.

vocabulary\_size  
语料库的词典大小, 记为|D|.

embedding\_size  
将词向量的维度, 由原始的 |D| 降维到 embedding\_size.

filter\_size\_arr  
多个不同size的filter.

3、TextCNN常用来做短文本的分类问题，其基本网络结构如下图所示：



从上图不难看出，TextCNN主要包括嵌入层、卷积层、池化层和全联接softmax层：

Embedding：这一层是图中最左边的7乘5的句子矩阵，高为句子中单词的个数，宽为词向量的纬度，这个可以类比为图像中的原始像素点。

Convolution：这一层经过 kernel\_sizes=(2,3,4) 的一维卷积层，每个kernel\_size 有两个输出 channel。

MaxPolling：第三层是一个1-max pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示。

FullConnection and Softmax：最后接一层全连接的 softmax 层，输出每个类别的概率。

大致过程是：来了一个句子，先根据词向量得到一个二维句子矩阵，然后选择不同的过滤器(大小为filter\_size \* embedding\_len)进行卷积操作得到多个feature map, 对每个feature map进行最大池化操作，进而将其拼接起来，最后经过softmax全联接层进行分类。