

# 文本分类

Jwzheng 慧科Aij讲师

# 第三节课课程计划



- 使用keras实现word2vec(上一节课程问题回顾)
- fasttext模型
- 基于fasttext实现文本分类（ keras实现fasttext，新闻数据集 ）
- RNN，LSTM算法回顾
- textlstm模型
- textlstm代码实现（ 基于keras实现lstm，新闻数据集 ）

# embedding层

- 什么是embedding
- 词嵌入

[1,0,0,0,0,0,0,0,0,0] onehot

[0.3,0.02,0.01] 分布式表示

```
1 39191 200
2 </s> 0.002001 0.002210 -0.001915 -0.001639 0.000683 0.001511 0.0004
3 和 -0.151985 -0.284619 0.102803 -0.185771 0.461806 -0.266179 0.1981
4 物体 -0.136723 0.129971 -0.132170 -0.091457 0.475951 -0.096483 0.29
5 小 0.292008 -0.352757 -0.060755 -0.266013 0.008357 0.014113 0.15615
6 TARGET 0.106808 -0.349218 -0.171092 -0.094614 0.207454 0.011103 0.2
7 力 -0.375067 -0.006072 -0.137480 -0.026753 0.430767 -0.527297 -0.19
8 水 0.248969 0.051000 0.390238 0.199963 -0.269108 -0.126435 0.009710
9 实验 0.102817 -0.059937 0.034009 -0.305009 0.210814 0.263467 -0.150
10 电流 -0.097567 -0.384787 0.211886 -0.396669 -0.208005 0.180179 0.03
11 电路 -0.411581 -0.577620 0.179795 -0.391916 -0.020366 0.182552 -0.0
12 用 -0.167212 0.043692 0.011332 -0.396244 -0.256816 0.273450 0.12865
13 两 0.050094 -0.152268 0.094807 -0.139273 0.228088 0.108458 0.093271
14 质量 -0.253252 0.171589 0.151000 0.332677 -0.057260 0.018994 0.3434
15 解 -0.009950 -0.180089 -0.106874 -0.147667 0.147166 0.028090 0.1045
16 分析 0.131652 -0.297859 -0.017839 -0.161281 0.064406 -0.097368 0.06
17 要 0.311441 -0.310314 0.107619 -0.013610 -0.015809 -0.100576 0.1518
18 电阻 0.052539 -0.312071 0.252123 -0.303671 -0.207619 0.121251 -0.16
19 不 -0.185340 -0.276029 -0.024352 -0.266725 0.305960 -0.131068 0.032
20 大小 0.025439 -0.052366 -0.011082 -0.042531 0.146216 -0.371850 0.06
21 压强 -0.067138 -0.005838 0.298683 0.410628 -0.412239 0.272877 0.318
22 越 0.475663 -0.344053 0.222486 -0.018942 0.308269 -0.373232 -0.1021
23 液体 0.498941 0.165211 0.254356 -0.062587 0.048676 -0.179797 -0.059
24 太 0.049403 -0.220499 -0.035255 0.053958 -0.040242 -0.141764 0.0276
25 正确 0.129875 -0.273607 -0.084505 -0.089546 0.282706 0.121047 0.102
26 选 0.093668 -0.215710 -0.039160 -0.213512 0.416333 0.015223 -0.0075
27 重力 -0.064578 -0.186279 -0.171902 -0.289215 -0.451004 -0.131798 0.
28 电压 -0.218823 -0.255212 0.392569 -0.567167 -0.202134 -0.187189 0.3
29 运动 -0.130846 -0.133343 -0.232054 0.124193 0.322630 -0.169233 -0.3
30 温度 -0.216227 -0.160802 0.588604 -0.241391 0.321560 0.028865 -0.28
```

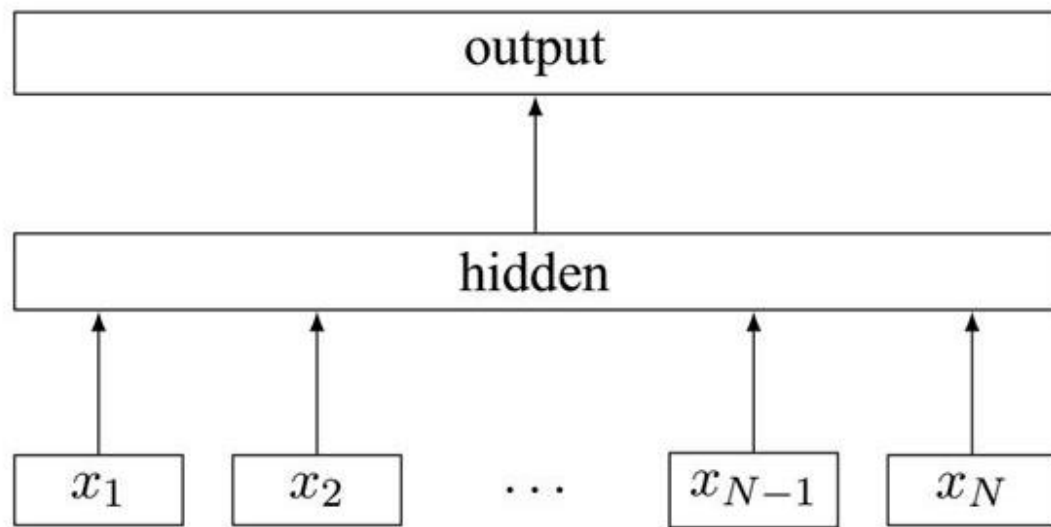
# fasttext模型

我/喜欢/你  
你/喜欢/我

慧科  
只为赋能人才

开课吧  
kaikeba.com

- 网络结构
  - 输入与输出
- 分层softmax
- N-Gram特征
  - 基于字（词，字符）的n-gram特征
  - 例子：我们/一起/去/爬山
  - $N=2$ 
    - 我们/一起 一起/去 去/爬山
  - $N=3$ 
    - 我们/一起/去 一起/去/爬山
- 优点
  - fast
  - $x_1$ :我们  $x_2$ :一起  $x_3$ :去  $x_4$ :爬山
  - embedding :
  - 100维度



**Figure 1:** Model architecture of fastText for a sentence with  $N$  ngram features  $x_1, \dots, x_N$ . The features are embedded and averaged to form the hidden variable.

# fasttext模型

- fasttext和CBOW的相同点

- 三层网络结构
- 隐含层都是对多个词向量的叠加平均
- 分层softmax

- fasttext和CBOW的不同点

- 输入输出不同
- fasttext加入了n-gram特征
- onehot 和 emdedding

- fasttext核心思想解释

- 将整篇文档的词及n-gram词向量叠加平均得到文档向量，然后使用文档向量做分类。（使用其他方法）

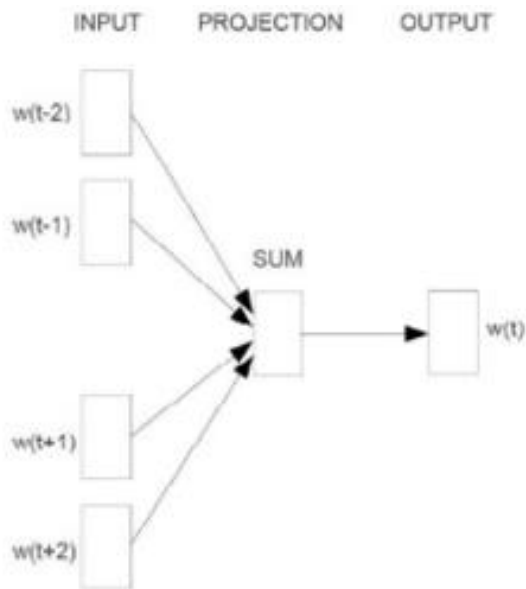
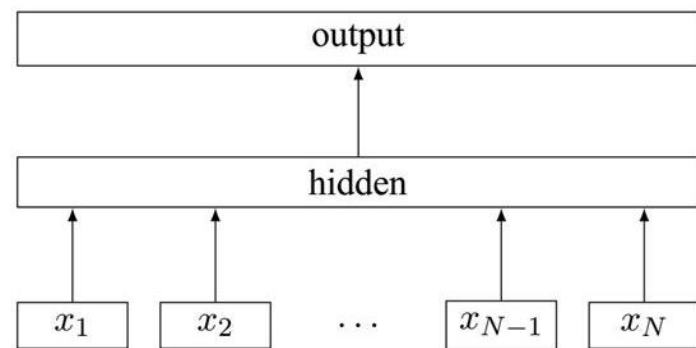


图 8 CBOW 模型



**Figure 1:** Model architecture of fastText for a sentence with  $N$  ngram features  $x_1, \dots, x_N$ . The features are embedded and averaged to form the hidden variable.

# 第四节课课程计划

- cnn算法回顾
- textcnn模型
- 新闻分类 二分类
- textcnn代码实现（基于keras）
- 基于tensorflow实现textcnn
- 多标签预测
- 包括：准确率计算，召回率计算，F值计算等等

# RNN模型

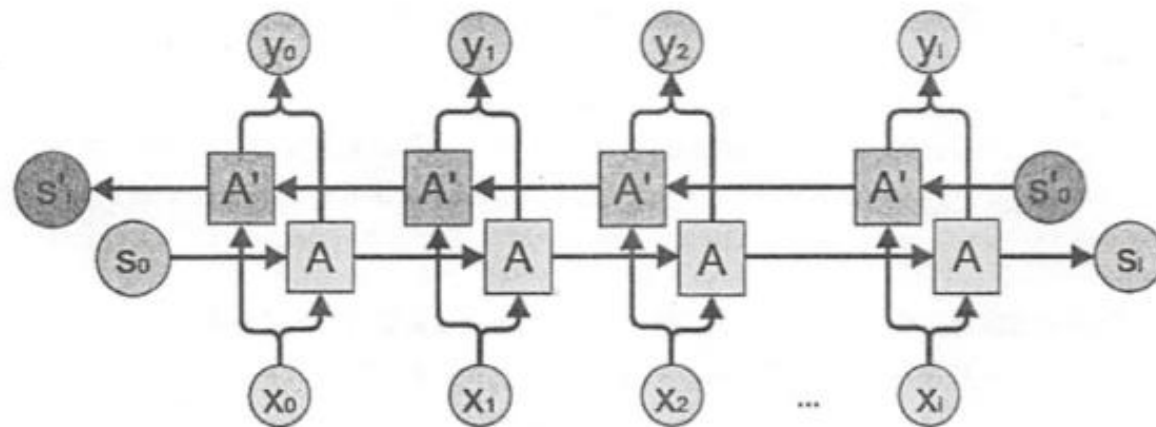
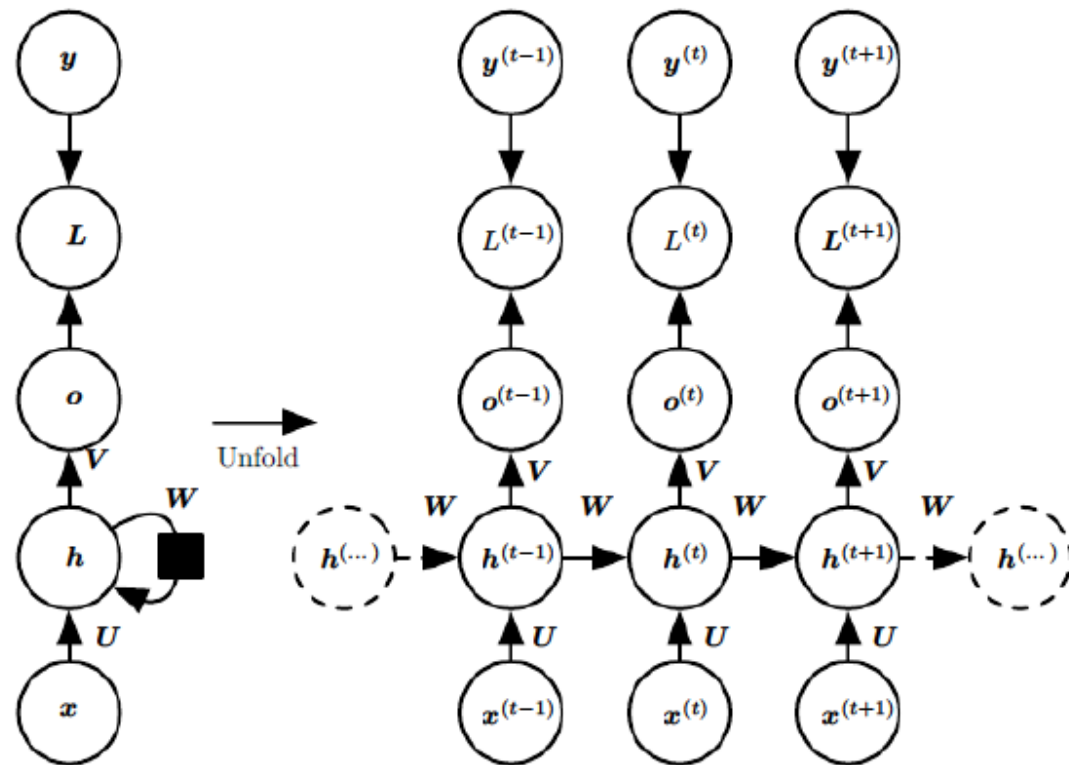
- RNN网络结构

- 不同时刻的参数共享

$$h^{(t)} = \phi(Ux^{(t)} + Wh^{(t-1)} + b)$$

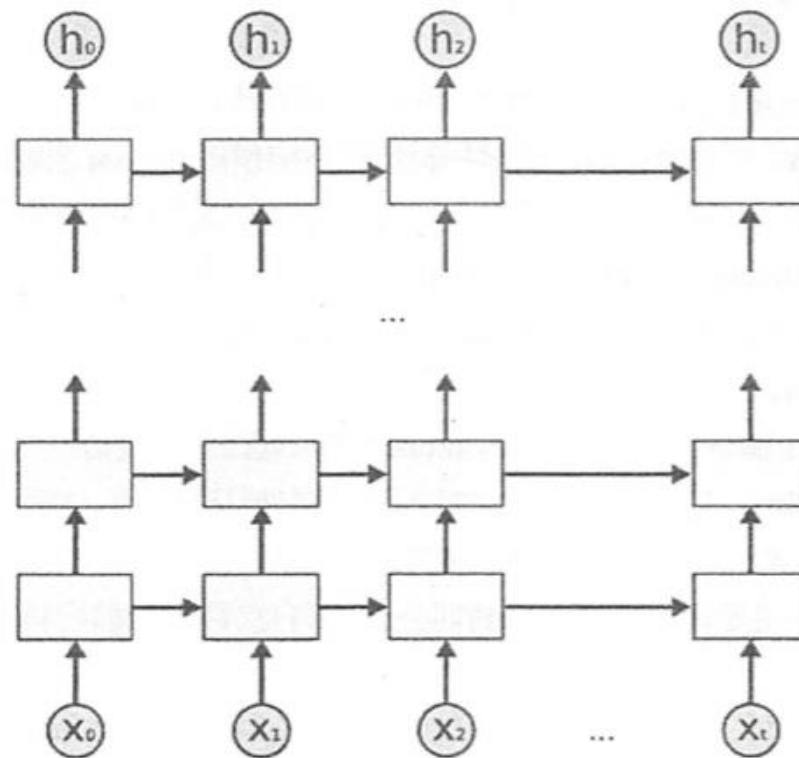
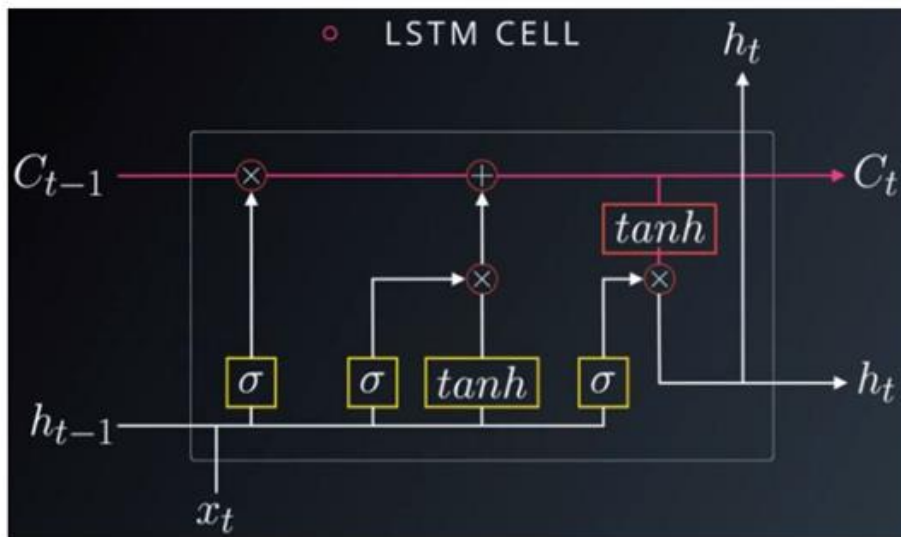
- 双向RNN网络结构

- 正向和反向不共享参数
- 双向合并策略
  - 连接求和



# RNN模型

- 深度RNN网络结构
- 深度RNN网络中的dropout
  - 在不同层之间使用
- LSTM网络结构





# TextRNN模型

- RNN用于文本分类

- BiLSTM用于文本分类

- 最后一个单词的正向和逆向concat ?

- embedding层

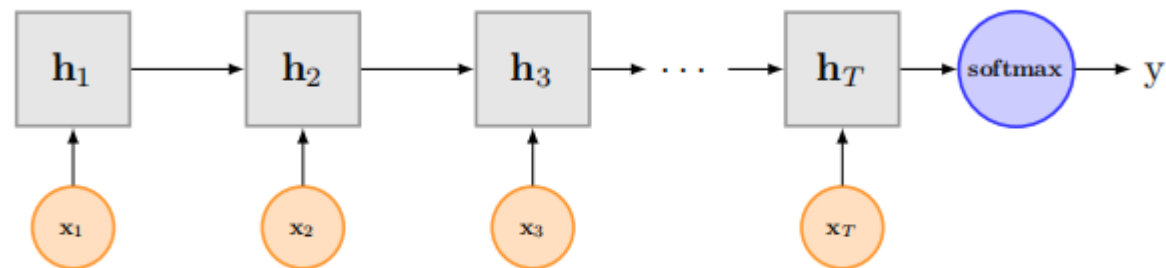
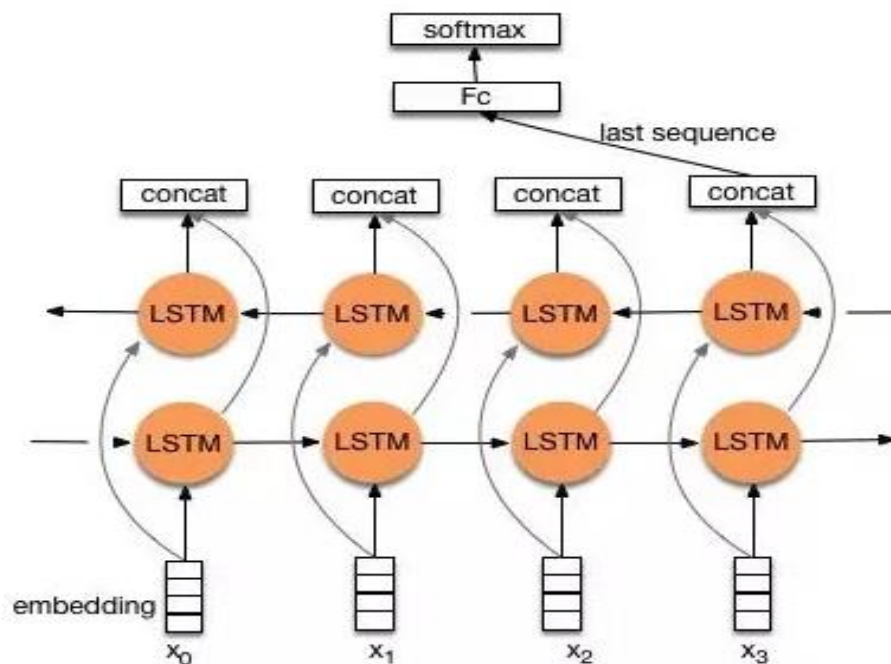
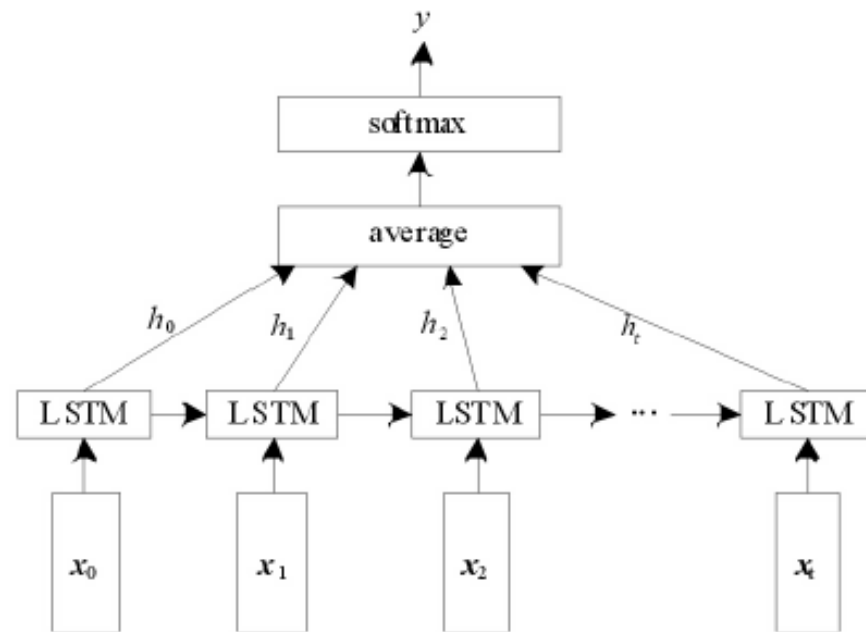
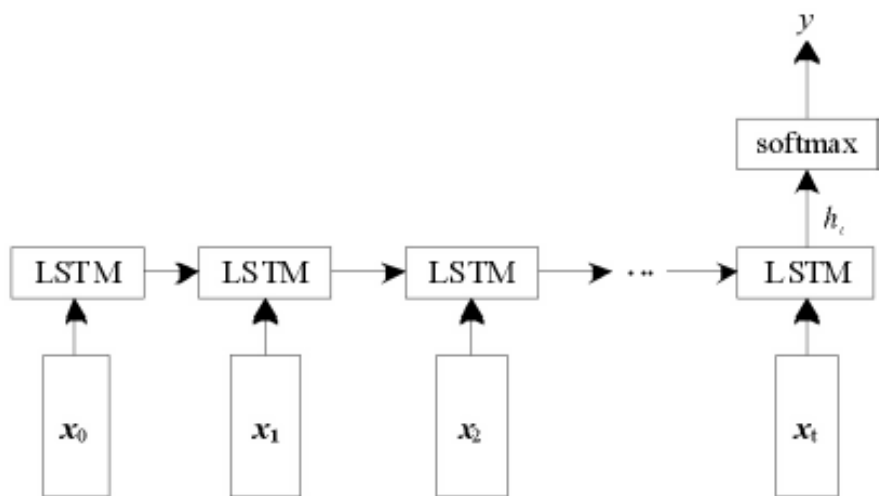


Figure 1: Recurrent Neural Network for Classification



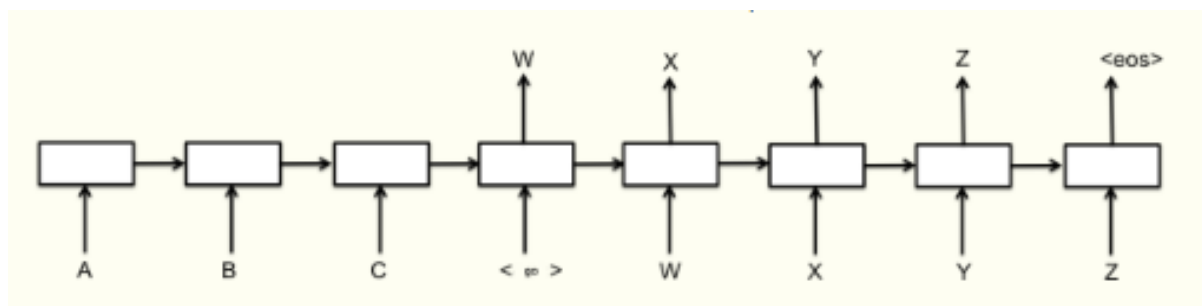
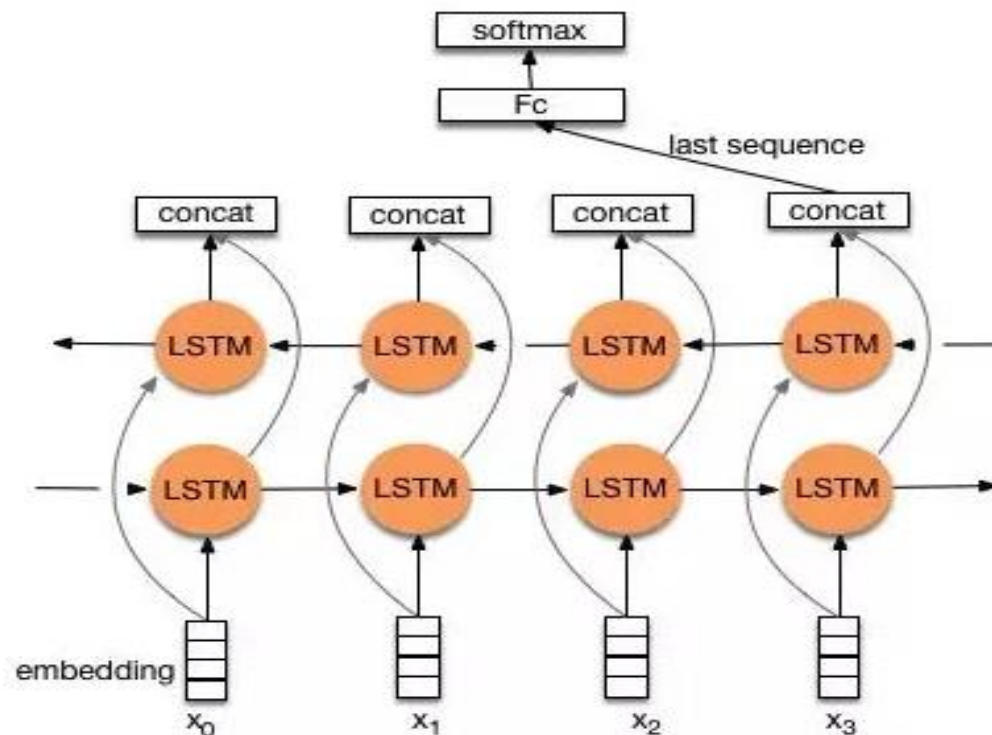
# TextRNN模型

- RNN文分类



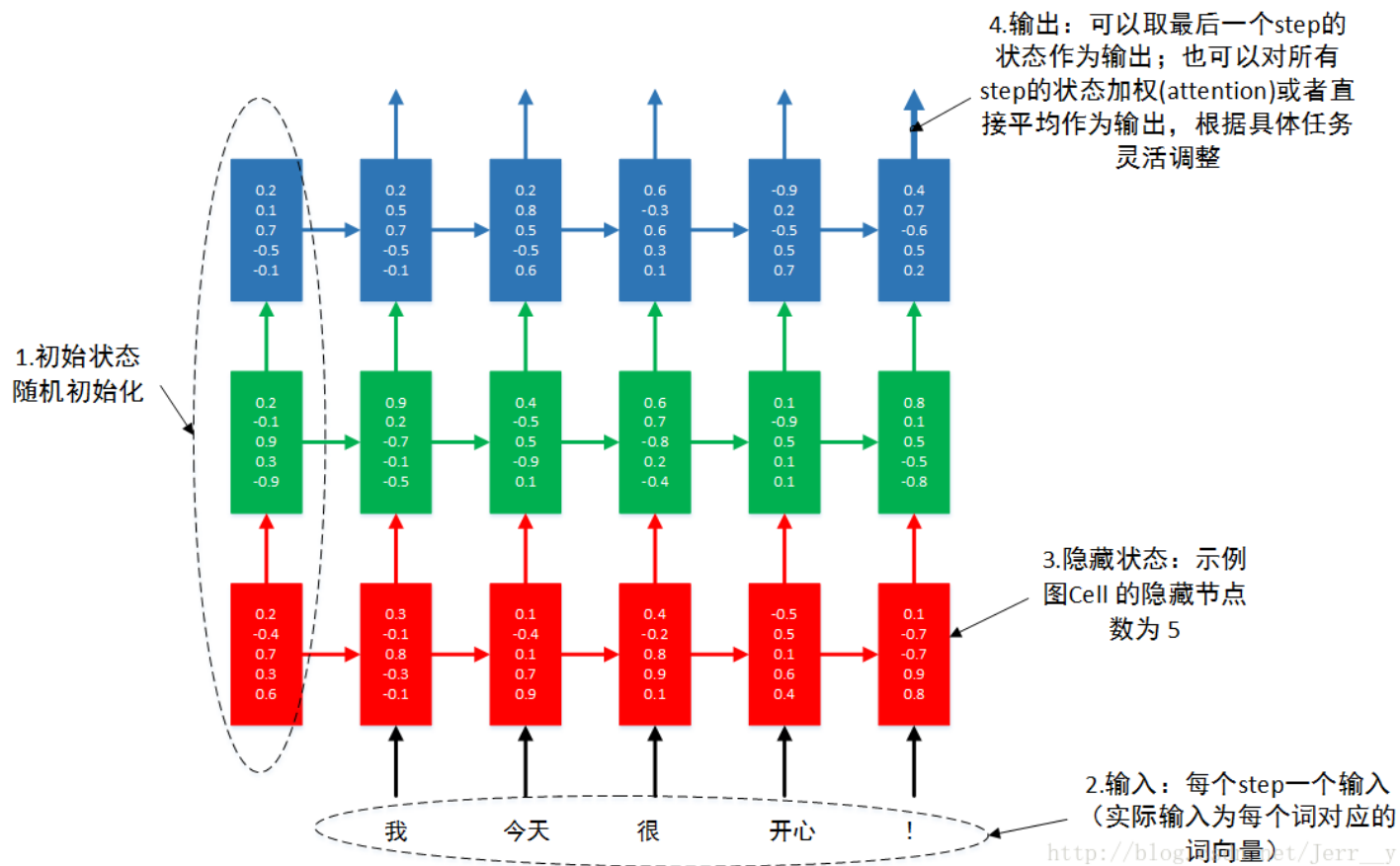
# TextRNN模型

- 为什么使用最后一个向量
  - 整个文本的压缩表示
- 其他方式
  - 使用每个神经元的输出平均后作为文本表示。
- 隐态
  - 计算文档相似度
  - 文档压缩编码
  - 机器翻译
- 端到端的模型
- RNN的文本特征提取



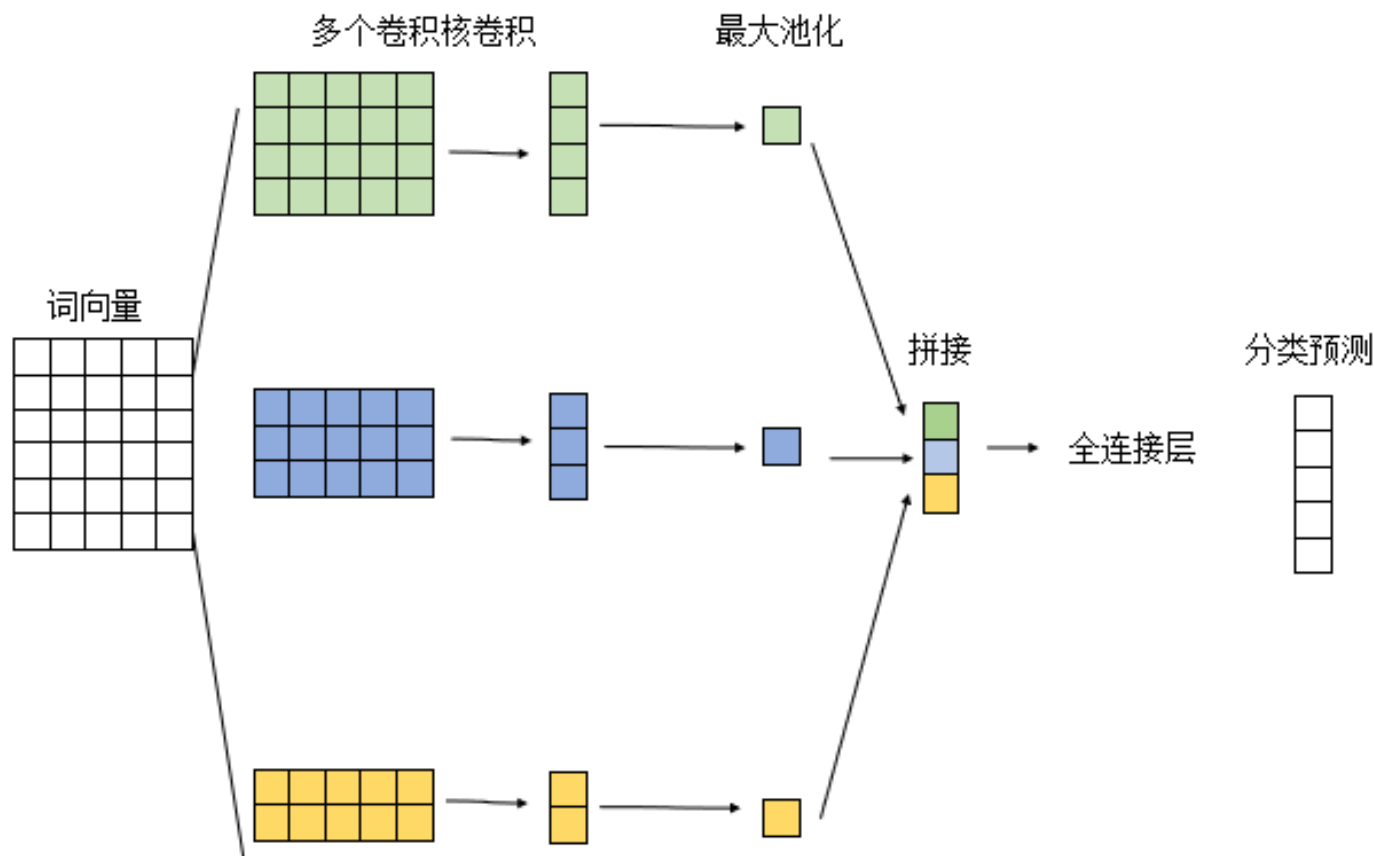
# TextRNN模型回顾

- 输入：
  - 一段中文文本（分词）
- 输出：
  - 该文本所属的类别
- 如何表示文本：
  - 词向量表示 word2vec
- 模型：
  - TextRNN, TextBiLSTM



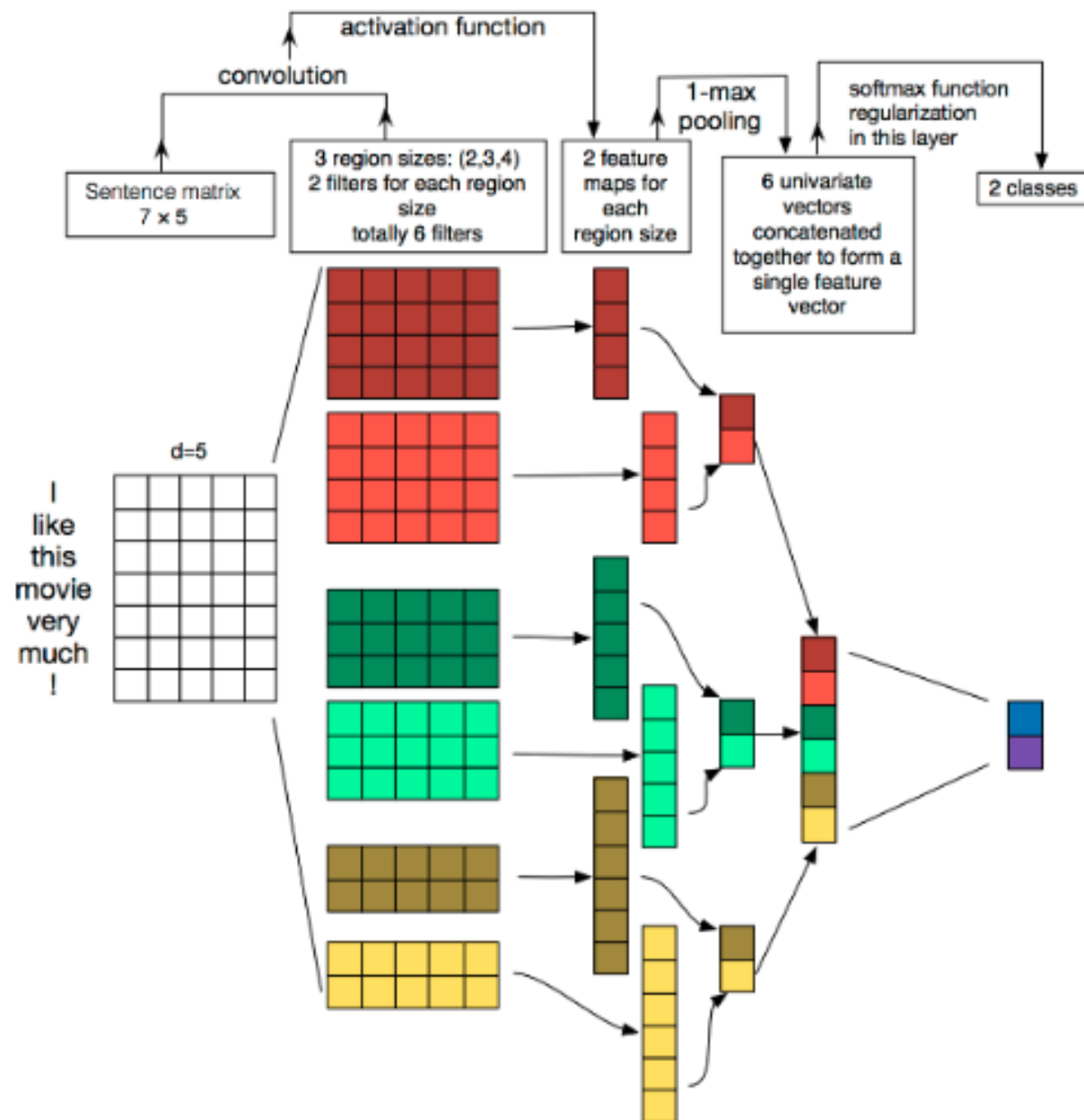
# TextCNN模型

- 输入：
  - 一段中文文本
- 输出：
  - 该文本所属的类别
- 如何表示文本：
  - 词向量表示 word2vec
- 模型：
  - TextCNN



# TextCNN模型

- 输入：
  - 一段中文文本
- 输出：
  - 该文本所属的类别
- 如何表示文本：
  - 词向量表示 word2vec
- 模型：
  - TextCNN
- 卷积核的维度的意义
  - 长 宽 高



# TextCNN模型

- 如何利用多通道
  - 使用多种不同的词向量模型
  - 图像 ( RGB )

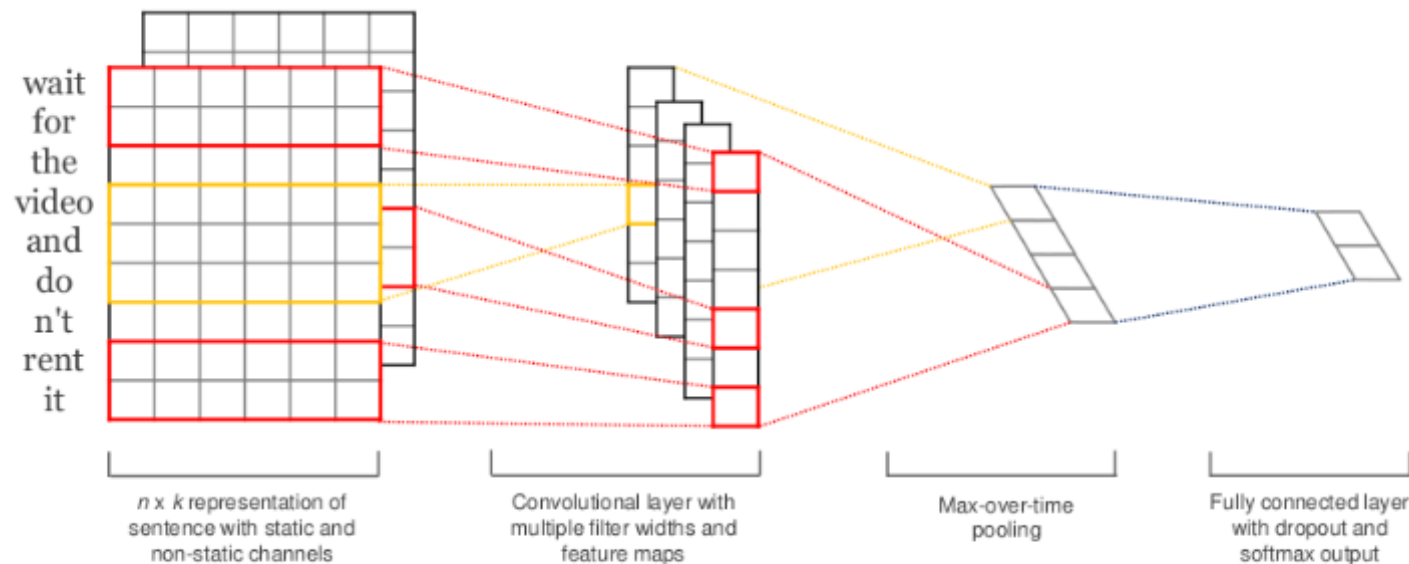


Figure 1: Model architecture with two channels for an example sentence.