慧科 只为赋能人才

开课吧 kaikeba.com

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）

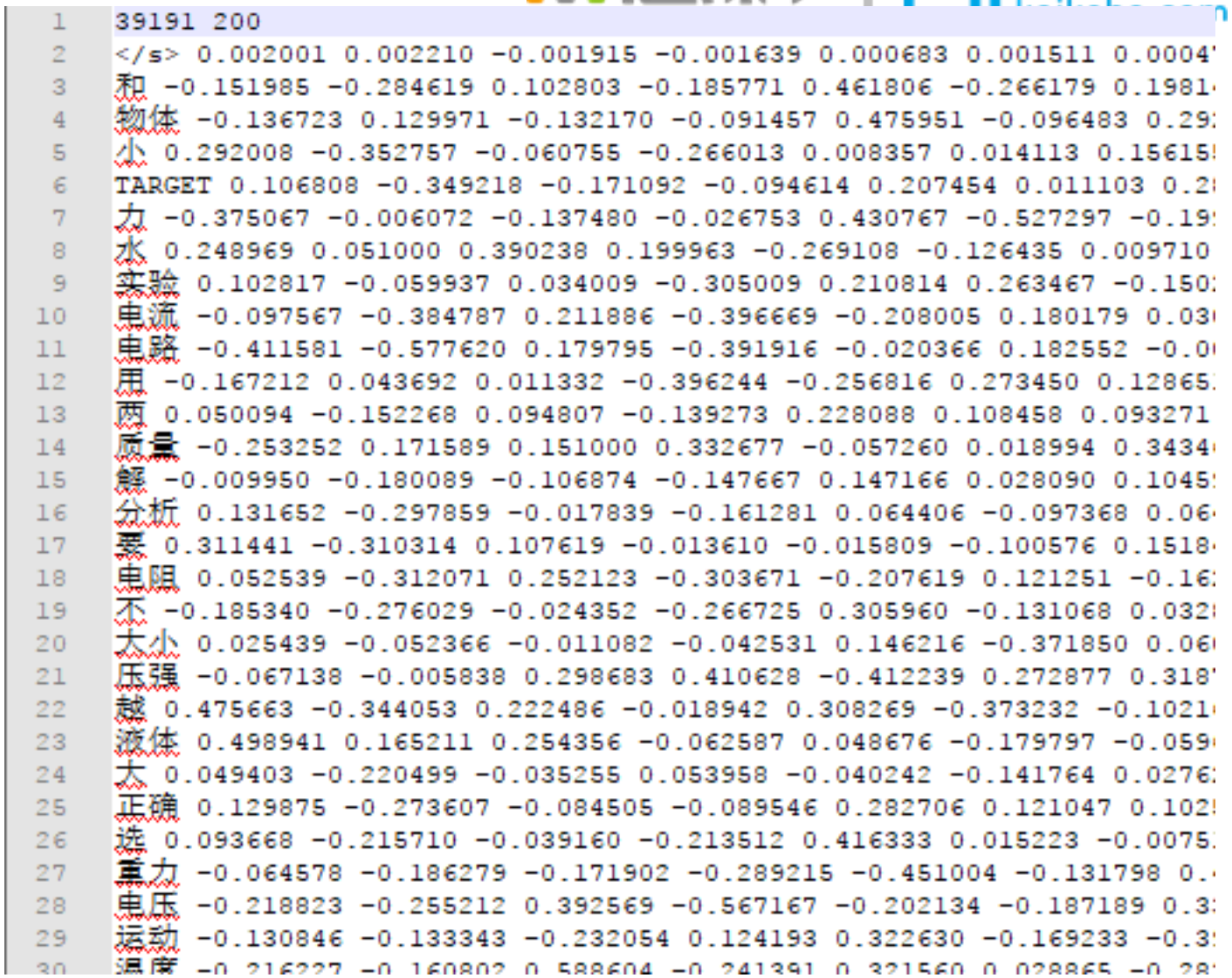
- RNN，LSTM算法回顾
- textlstm模型
- textlstm代码实现（基于keras实现lstm，新闻数据集）

embedding层

- 什么是embedding
- 词嵌入

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

  [0.3,0.02,0.01] 分布式表示



fasttext模型

我/喜欢/你
你/喜欢/我

- 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
  - x1:我们 x2：一起 x3：去 x4：爬山
  - embedding：
  - 100维度

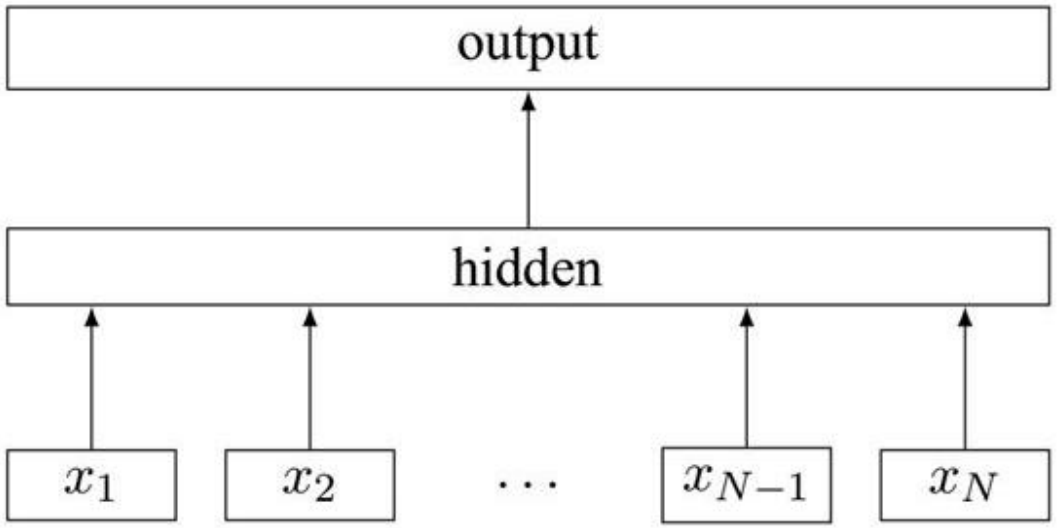


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

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）

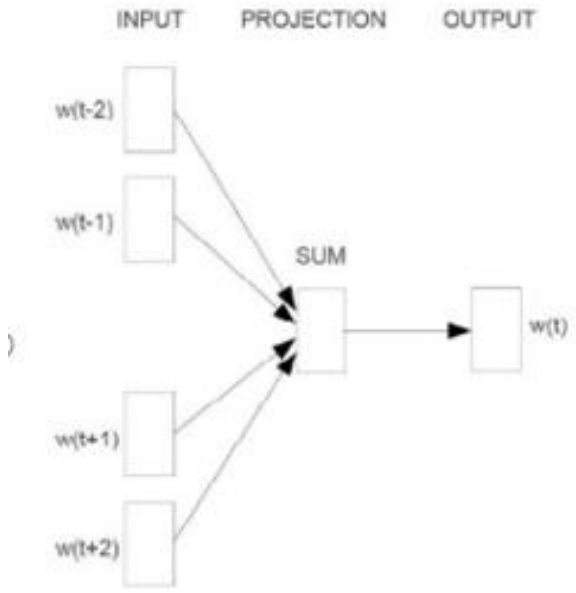


图 8 CBOW 模型

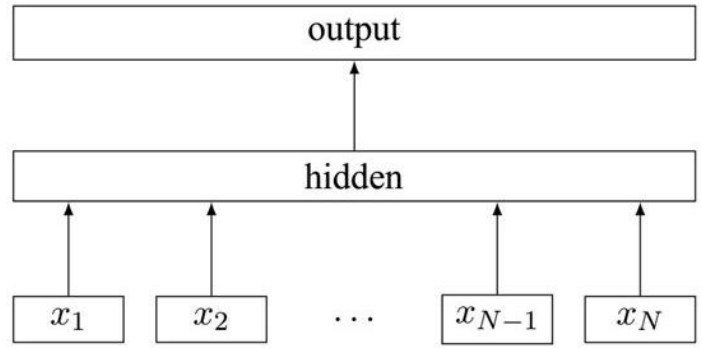


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

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）

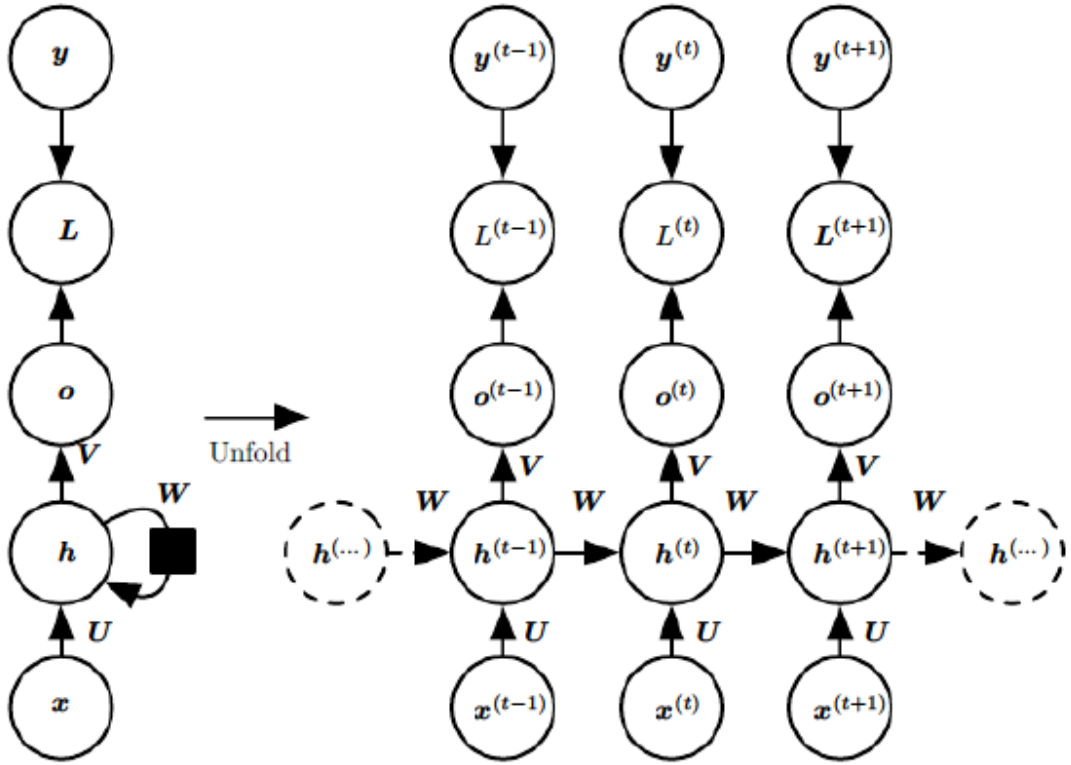
- 基于tensorflow实现textcnn
- 多标签预测
- 包括：准确率计算，召回率计算，F值计算等等

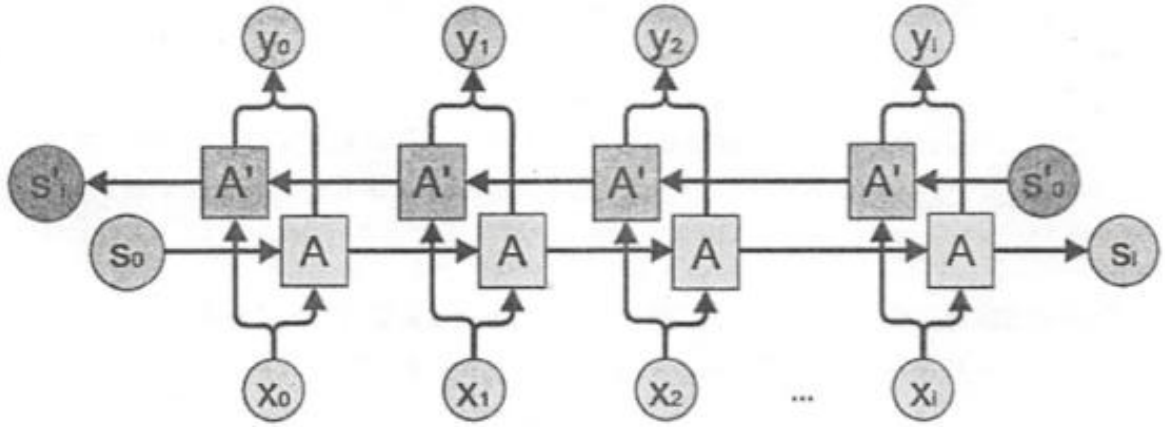
RNN模型

- RNN网络结构
  - 不同时刻的参数共享

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

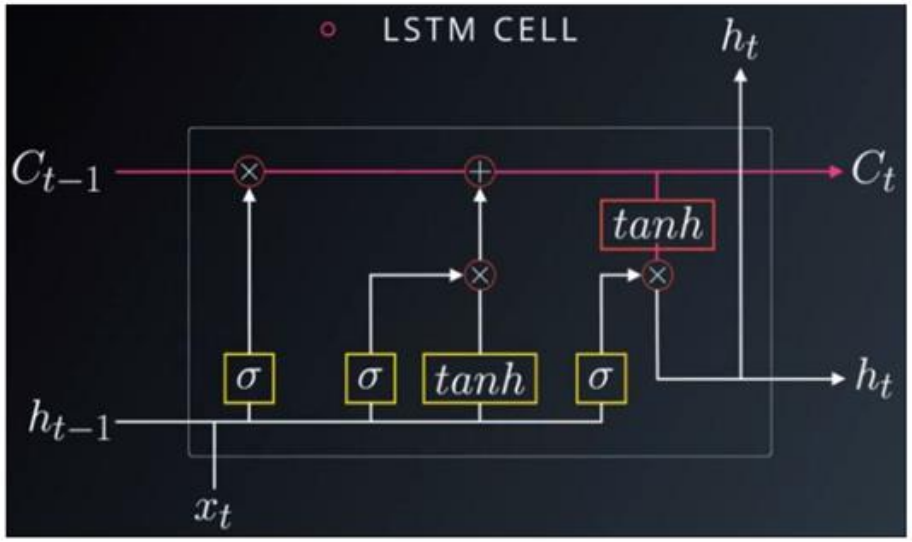
- 双向RNN网络结构
  - 正向和反向不共享参数
  - 双向合并策略
    - 连接 求和

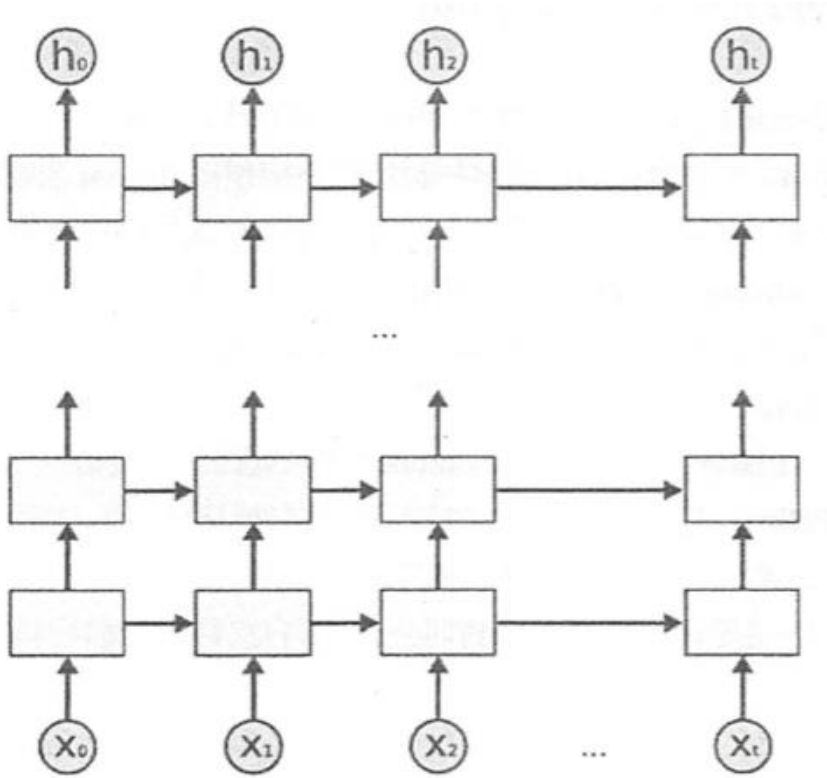




RNN模型

- 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

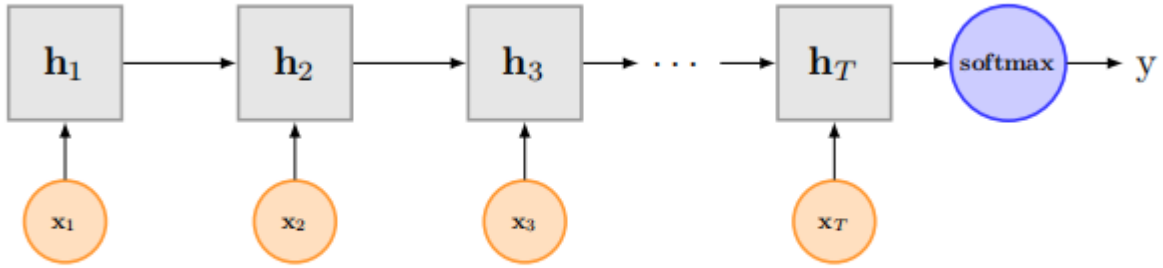
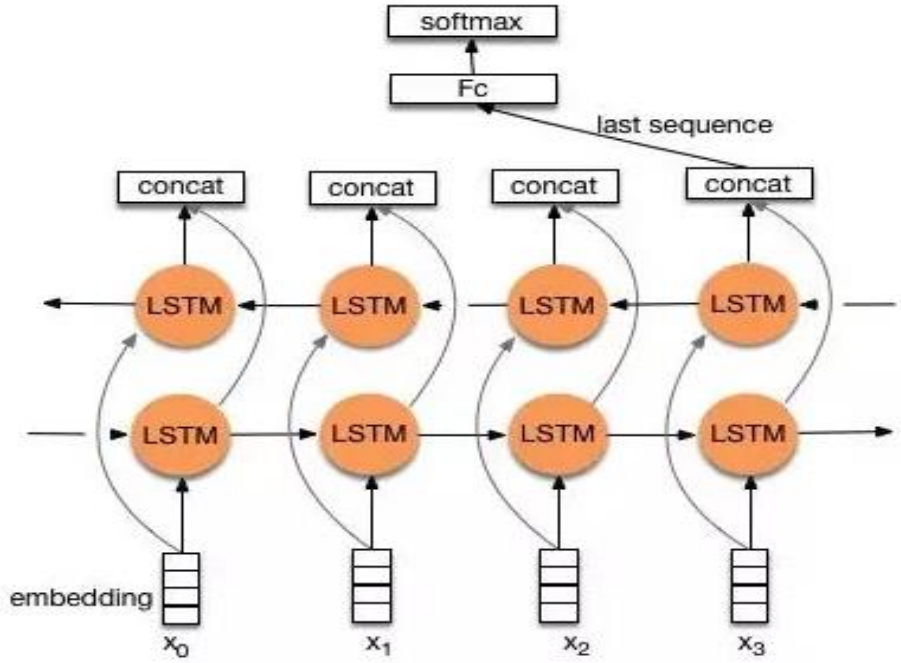


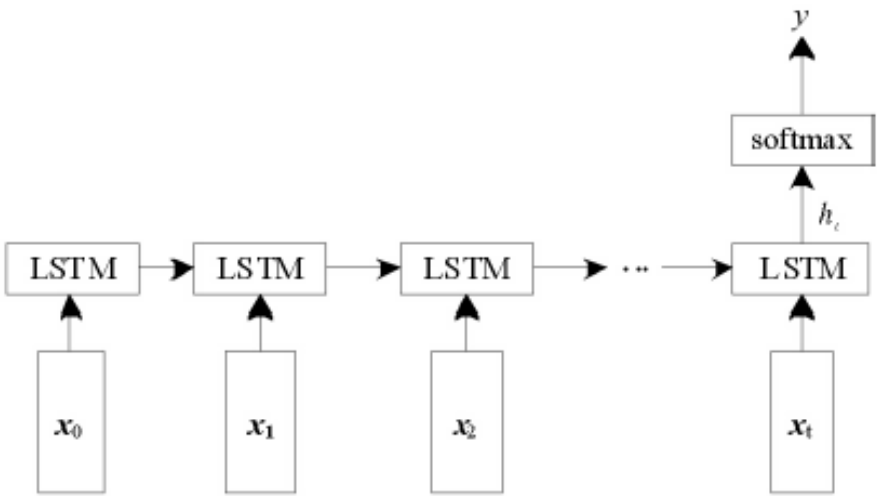
Figure 1: Recurrent Neural Network for Classification

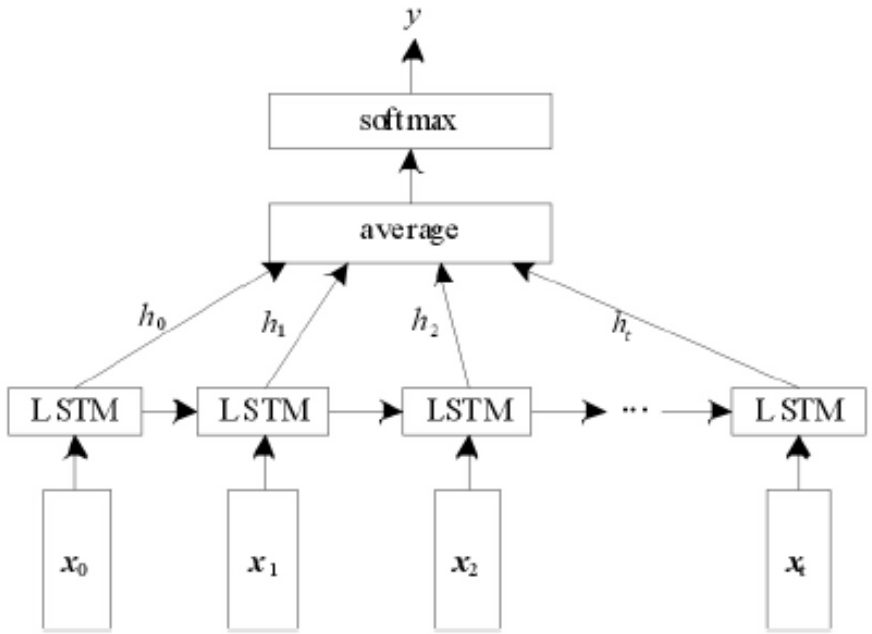
- BiLSTM用于文本分类
  - 最后一个单词的正向和逆向concat？
- embedding层



TextRNN模型

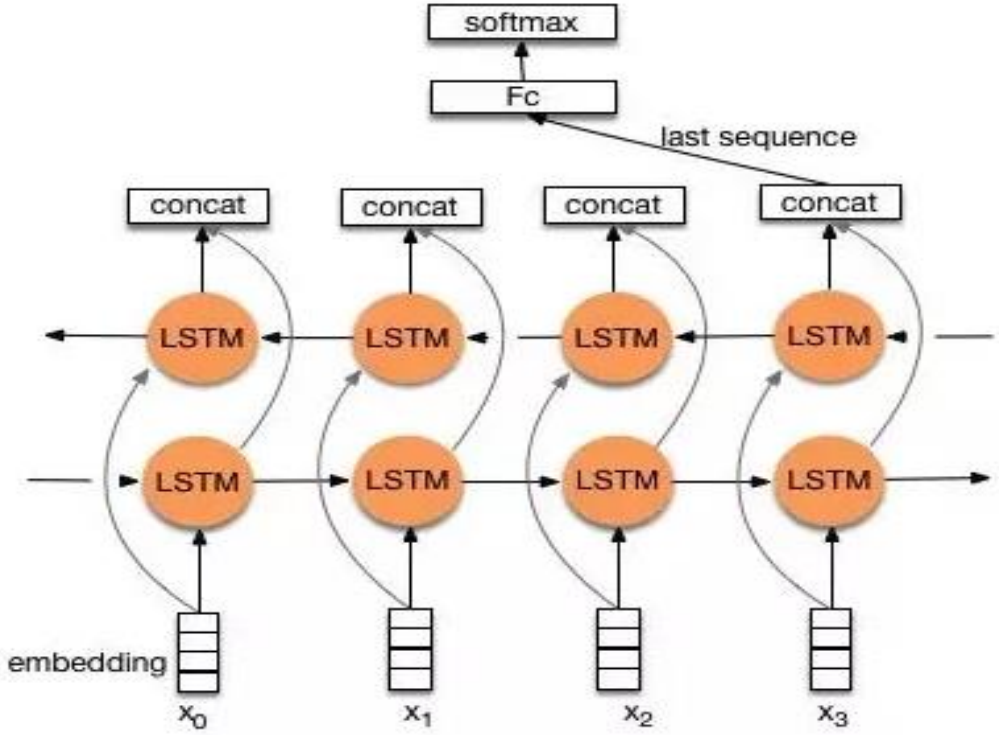
- RNN文分类

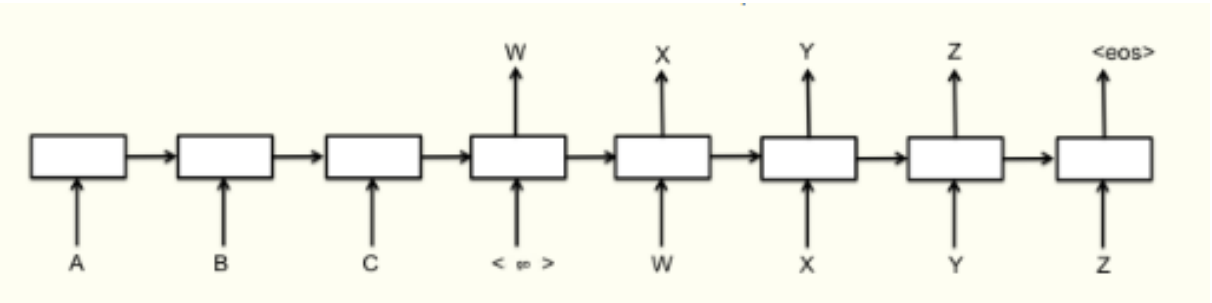




TextRNN模型

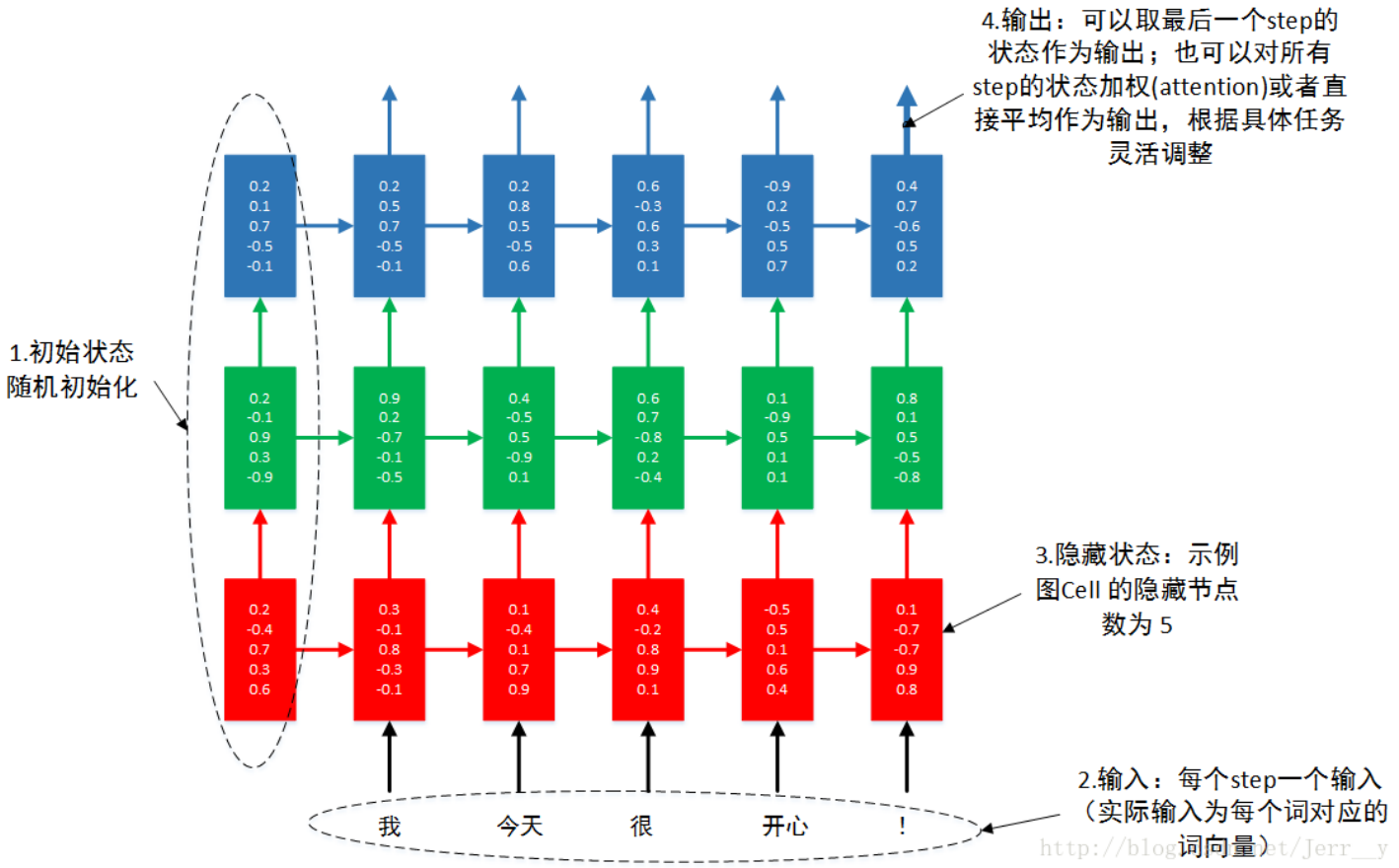
- 为什么使用最后一个向量
  - 整个文本的压缩表示
- 其他方式
  - 使用每个神经元的输出平均后作为文本表示。
- 隐态
  - 计算文档相似度
  - 文档压缩编码
  - 机器翻译
- 端到端的模型
- RNN的文本特征提取





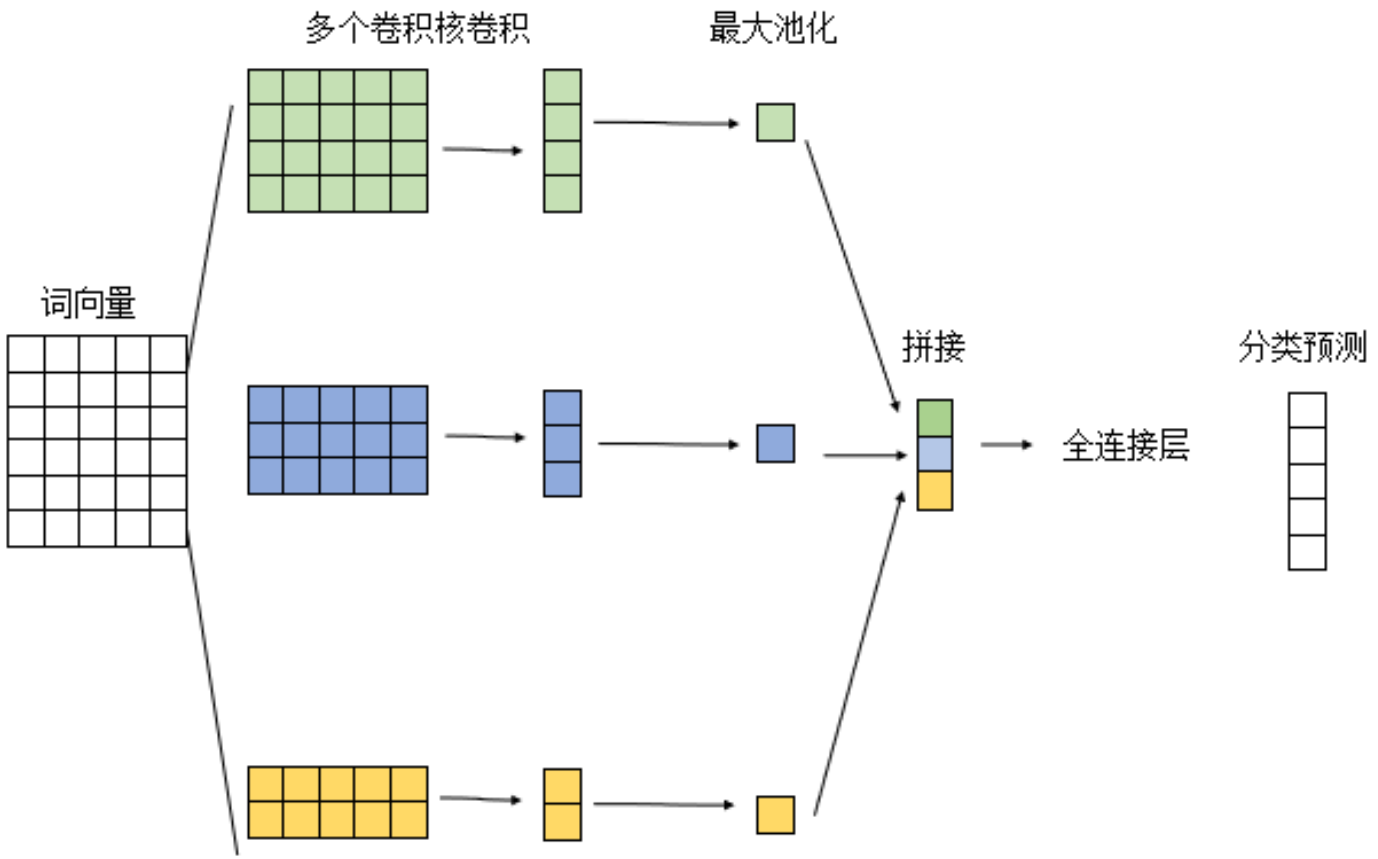
TextRNN模型回顾

- 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



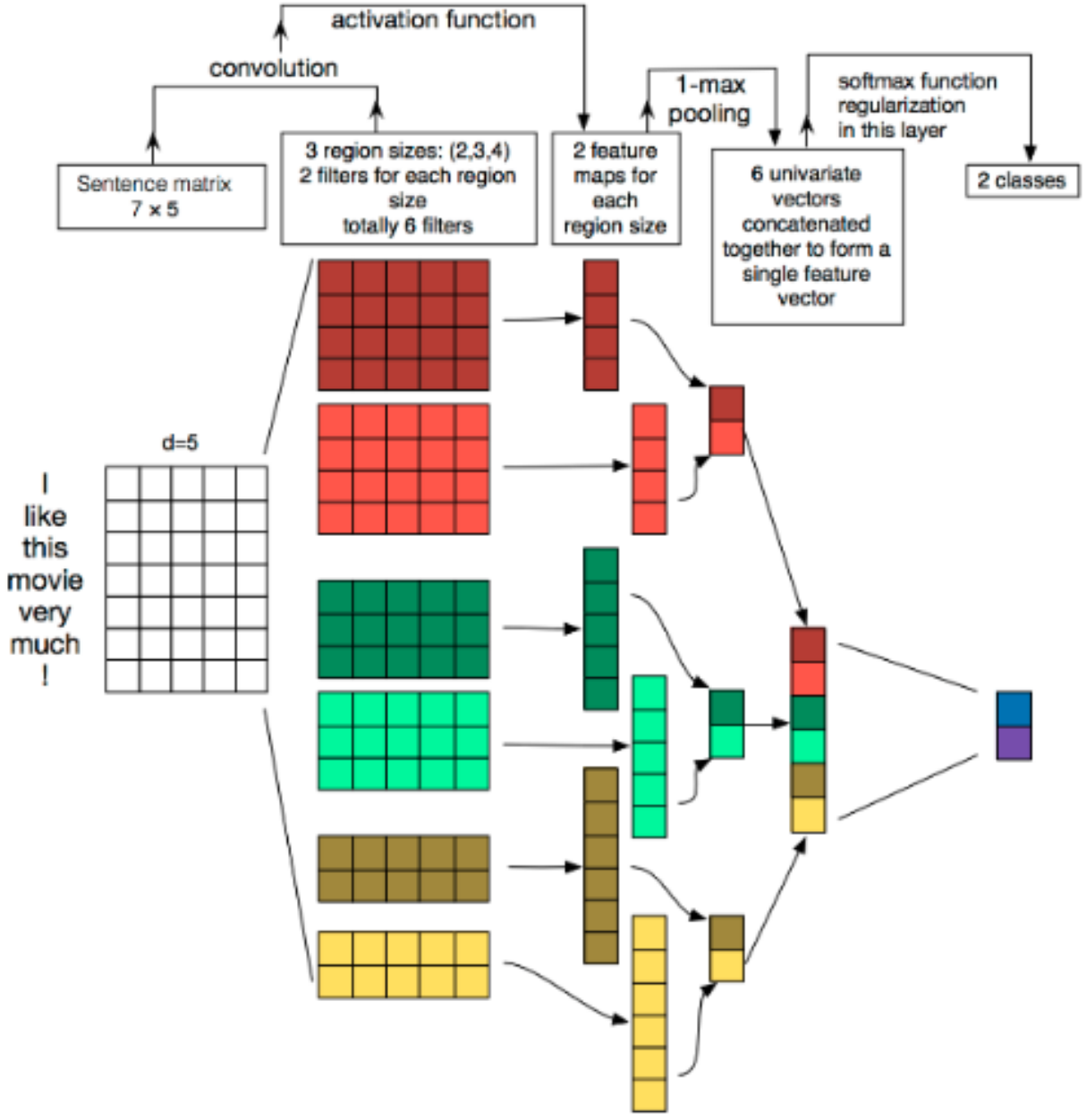
TextCNN模型

- 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）

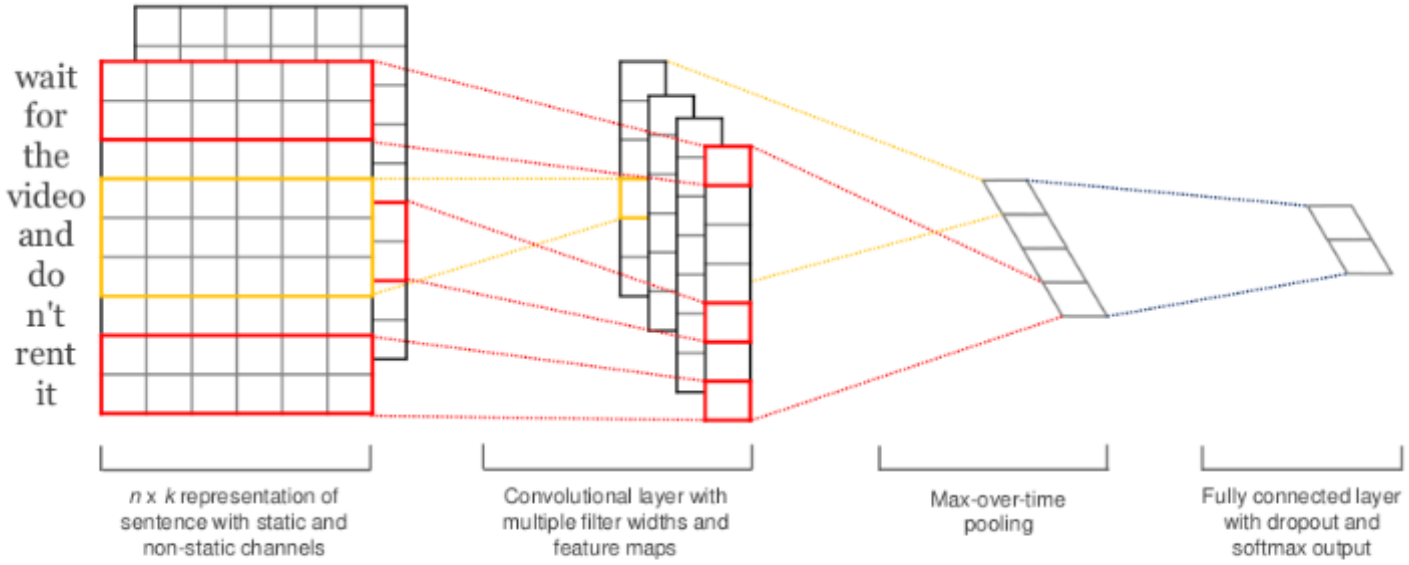


Figure 1: Model architecture with two channels for an example sentence.