- 数据准备ing
- 新闻短文本分类
- TextCNN
- RCNN

新闻短文本分类

- 文本特征提取方法：
  （1）TF-IDF
  （2）word2vec

- gbdt分类器
  （1）TFIDF: 一对多策略Accuracy 0.754，二分类策略平均Accuracy 0.836
  （2）word2vec：一对多策略Accuracy 0.642，二分类策略平均Accuracy 0.713
  （Accuracy是测试集上准确率）

- TFIDF分类准确率较高，但随着样本量的增加，词典维度增大，训练成本增大

- word2vec词向量维度固定，训练成本受样本量影响较小，但准确率较低

TextCNN

- 文本矩阵 + conv + pooling

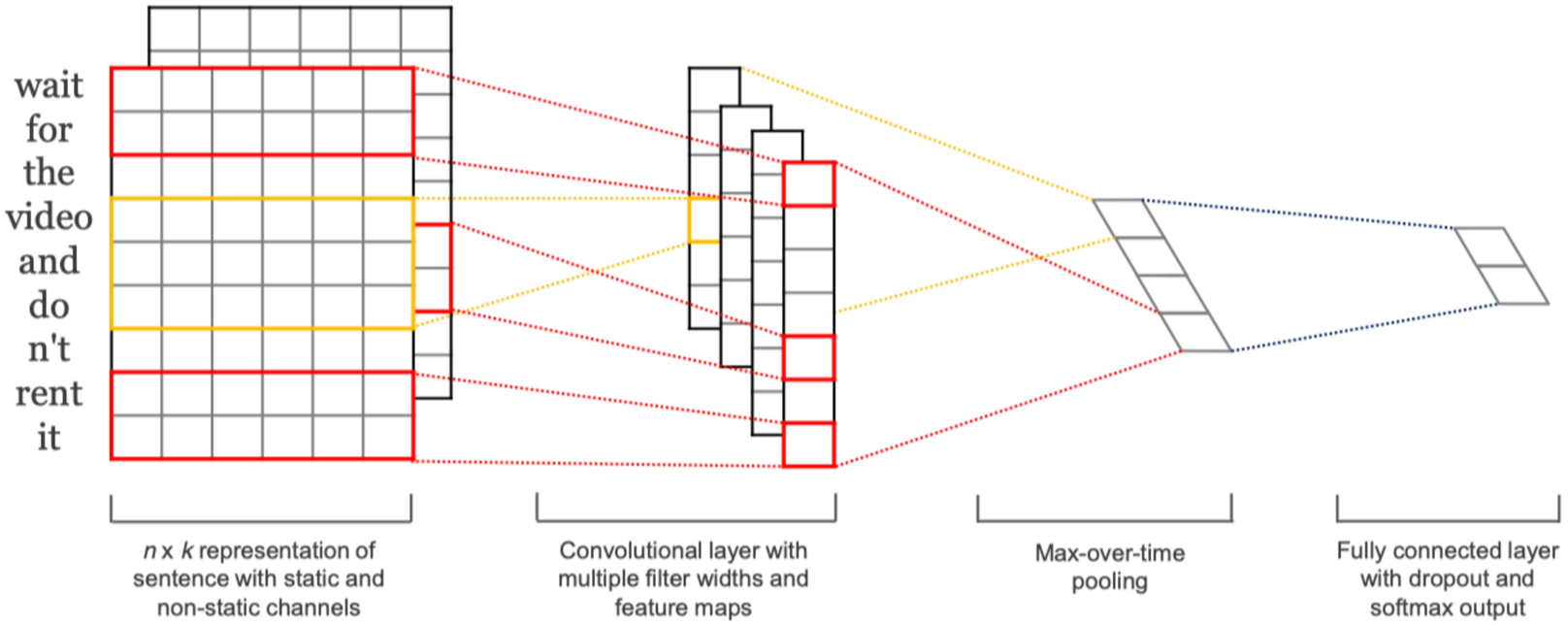
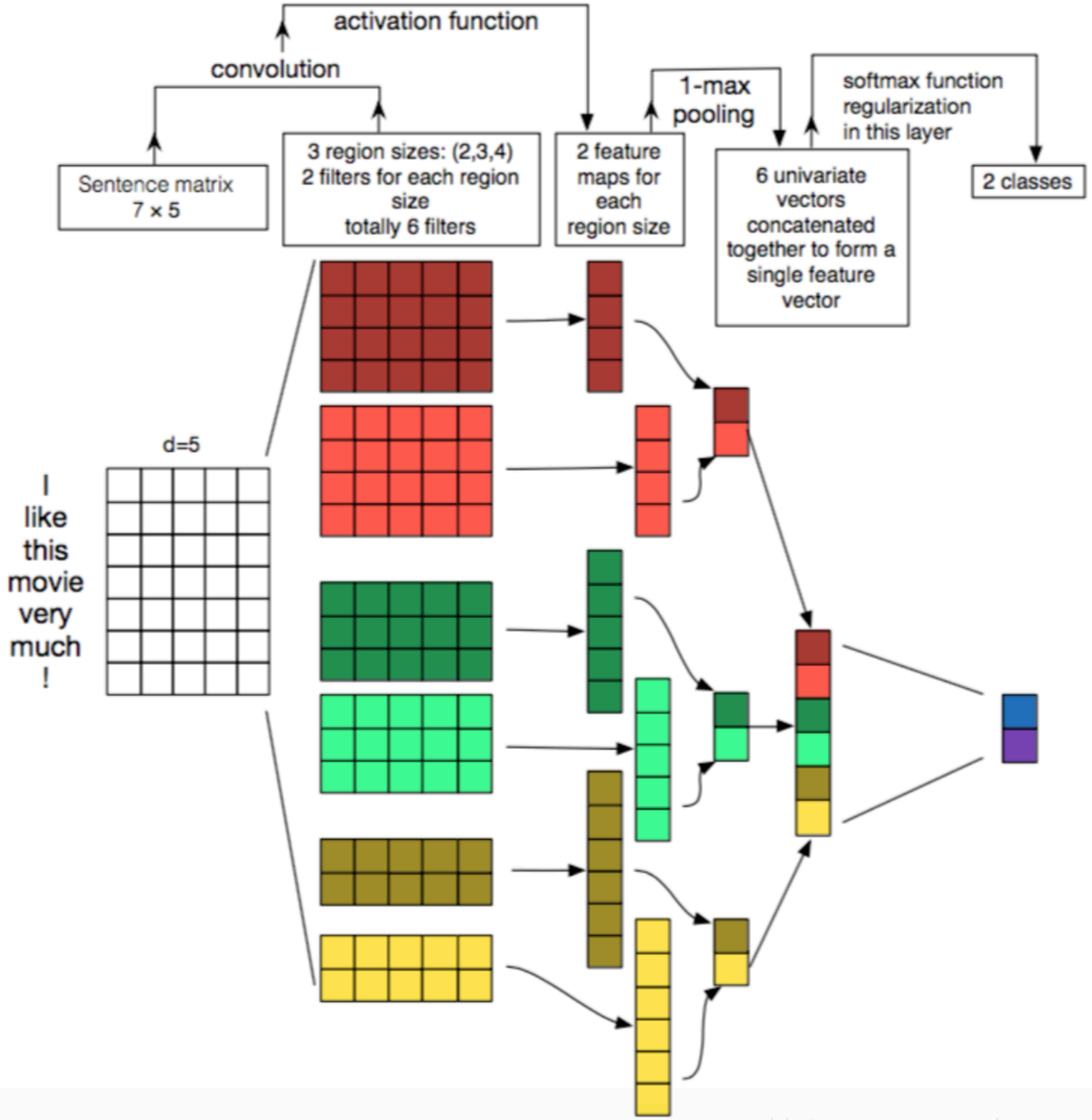


Figure 1: Model architecture with two channels for an example sentence.

TextCNN

- 重要参数
  - (1) 卷积核尺寸（1–10）
  - (2) 每种尺寸卷积核的数量（100–600）
  - (3) dropout rate



RCNN

- 文本 + 上下文表征 + MaxPooling layer

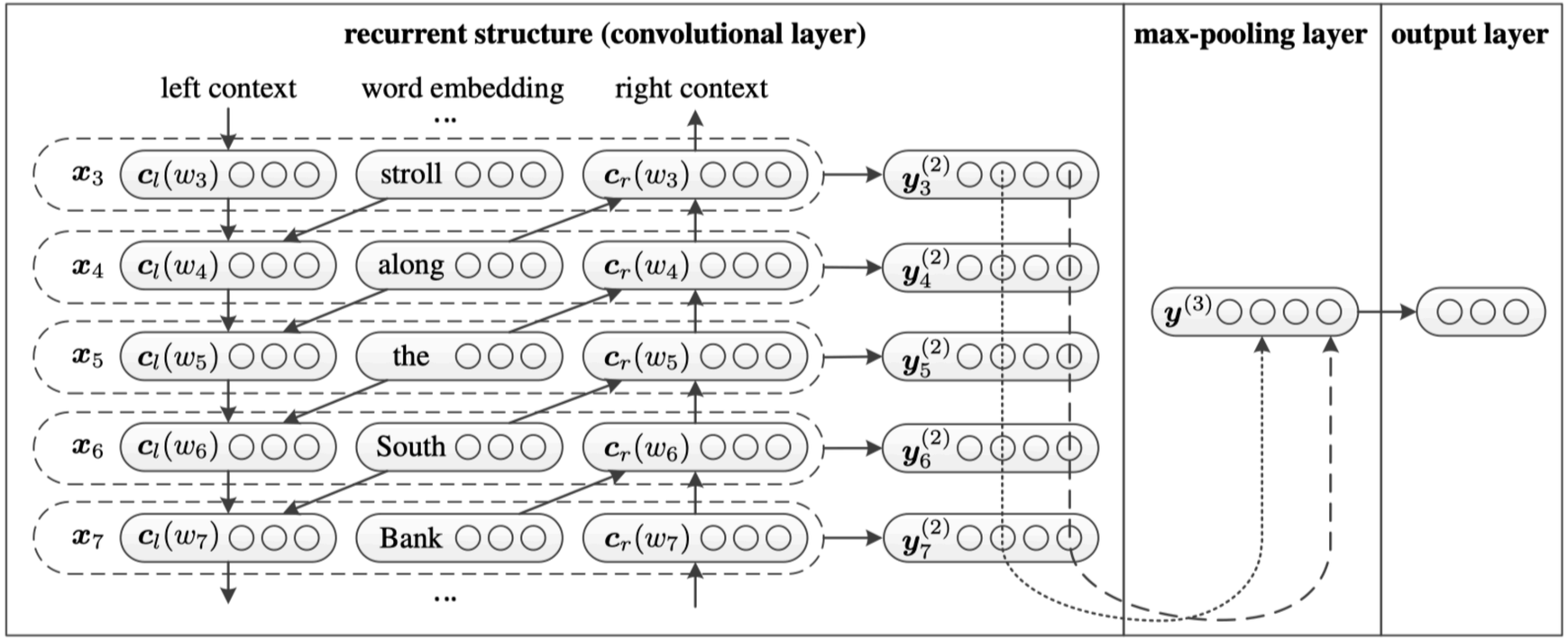


Figure 1: The structure of the recurrent convolutional neural network. This figure is a partial example of the sentence "A sunset stroll along the South Bank affords an array of stunning vantage points", and the subscript denotes the position of the corresponding word in the original sentence.