基于深度学习的文本分类

李林

2018-11-30

- 传统文本分类
- 文本的分布式表示
- 文本分类深度学习模型

文本分类的定义

Definition

给定分类体系，将文本分到某个或某几个对应的类别中。

- 二分类问题，0或1（YES or NO）
- 多分类问题，多个类别（multi-class），可拆分成二分类问题
- 多标签问题，一个文本可以属于多个分类

Example

- 垃圾邮件检测（spam or not spam）
- 情感分析（sentiment analysis）
- 新闻栏目分类
  - 类别{政治，娱乐，科技, …}
- 小说主题标签（多标签）
  - 类别{古装，武侠，爱情, …}

传统文本分类

特征工程

把文本转换成计算机可以理解的形式

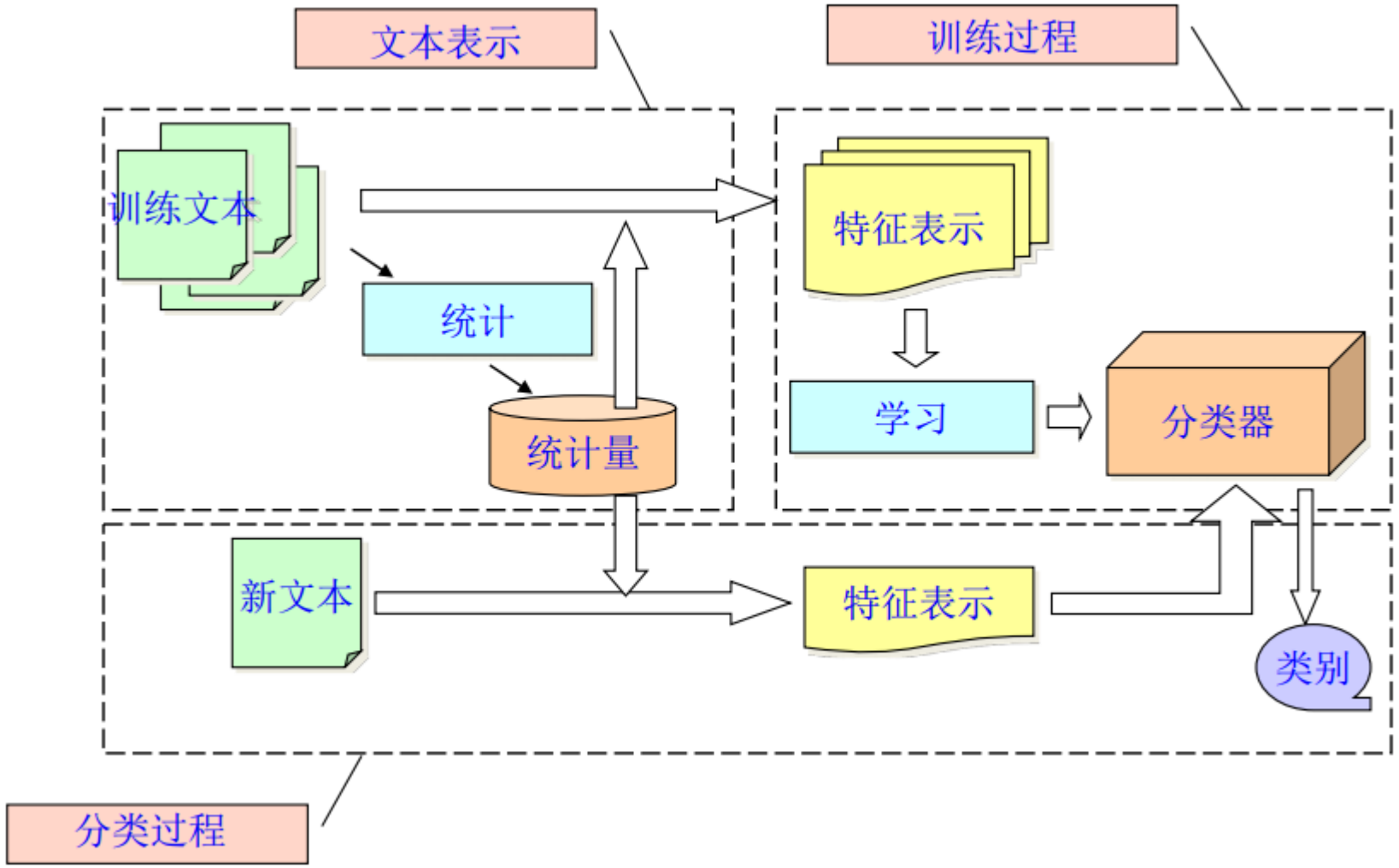
- 文本预处理
  - 去标签
  - （英文）去停用词，词根还原
  - （中文）去停用词，分词，词性标注，短语识别，…
- 特征提取（向量空间模型）
  - 特征选择（文档频率，互信息，信息增益，…）
  - 特征权重（TF-IDF及其扩展方法，计算词的重要性）
- 文本表示
  - 词袋模型（0, 0, 0, …, 1, …0, 0, 0）
  - 向量空间模型
  - LDA主题模型
  - ...

机器学习

大部分机器学习方法在文本分类领域都有所应用：

- 朴素贝叶斯
- KNN
- SVM
- 最大熵
- …

文本分类过程



传统分类中通过特征工程来得到文本的向量表示。

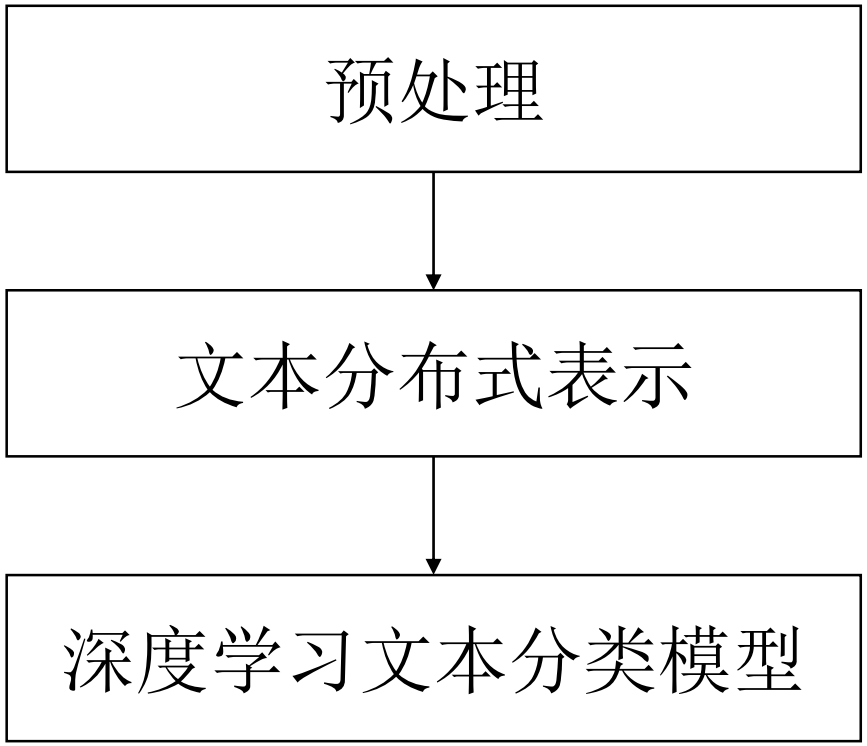
耗时耗力，泛化能力差，通常需要针对特定任务的理解。

Word Embedding

深度学习方法的基础

Word Embedding做了什么？

Word Embedding自动的将文本的内容（词）变成了连续稠密的实值向量。



Word Embedding怎么做的？

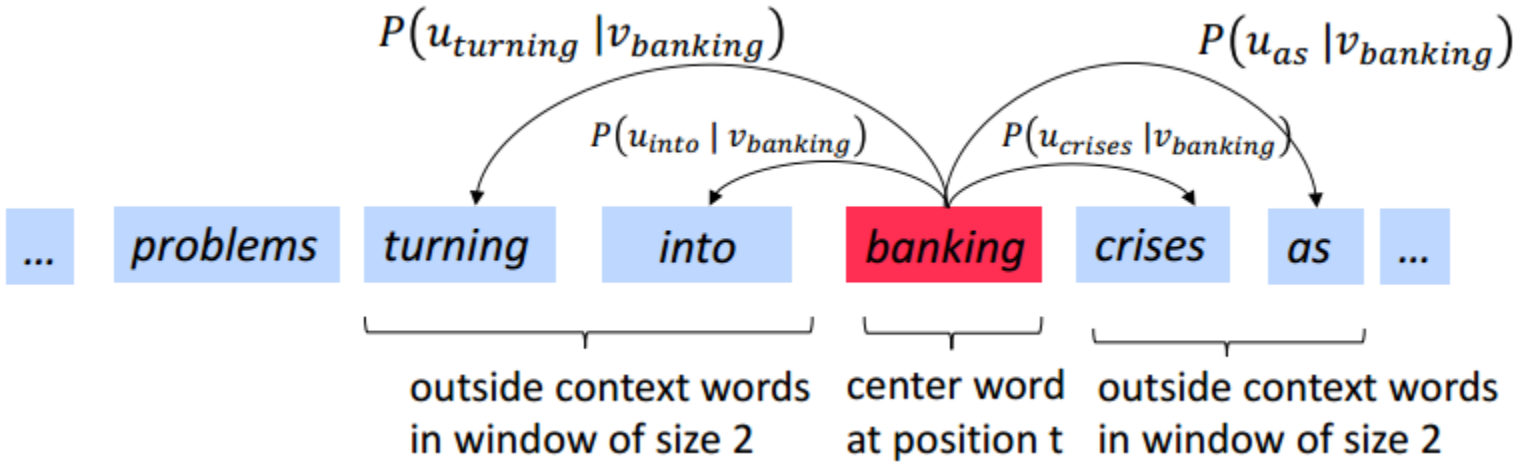
Idea：一个词的意思可以由该词的上下文信息表述，连接了词之间的语义关系。

…government debt problems turning into	***banking***	*crises as happened in 2009…*
…saying that Europe needs unified	***banking***	*regulation to replace the hodgepodge…*
…India has just given its	***banking***	*system a shot in the arm…*

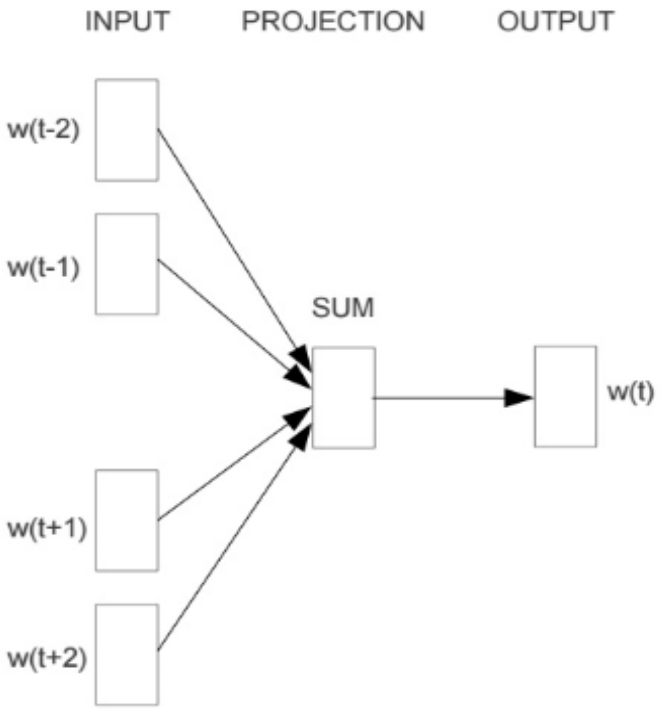
通过把每个词表示成一个密集向量，就可以方便的表示每个词之间的关系，那些含义相似（向量相近）的词会出现在相似的上下文中。

Word Embedding怎么做的？

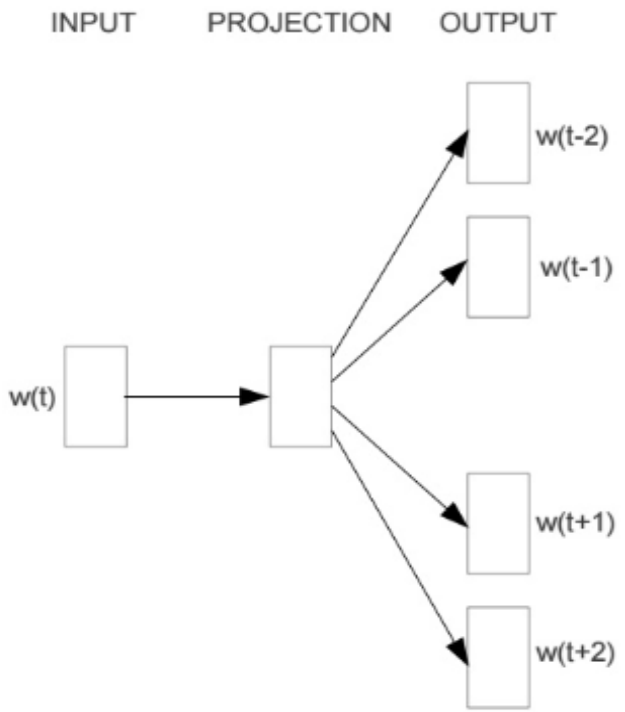
通过调整参数（词向量），使得给定中心词(center word)时，其上下文词出现的概率最大化。（或者，反过来，上下文预测中心词）



word2vec



CBOW



Skip-gram

词汇的相似度

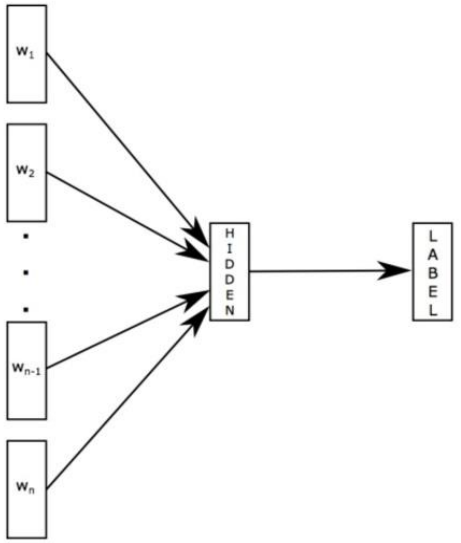
```
(EXIT to break): china

n vocabulary: 486

                      Word       Cosine distance
-------------------------------------------------
                    taiwan              0.768188
                     japan              0.652825
                     macau              0.614888
                     korea              0.614887
                       prc              0.613579
                   beijing              0.605946
                    taipei              0.592367
                  thailand              0.577905
                  cambodia              0.575681
                 singapore              0.569950
                  republic              0.567597
                  mongolia              0.554642
                   chinese              0.551576
```

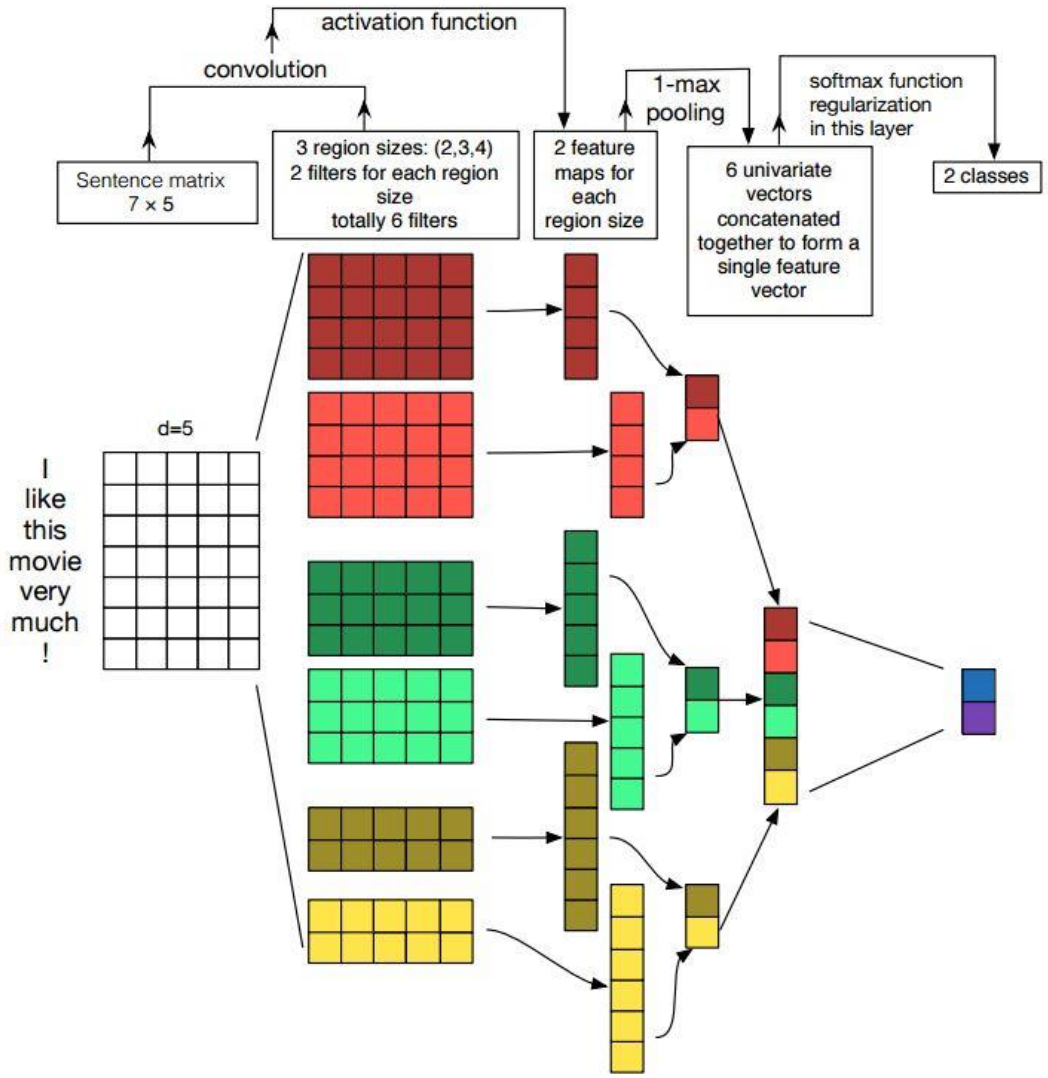
深度学习文本分类模型

FastText

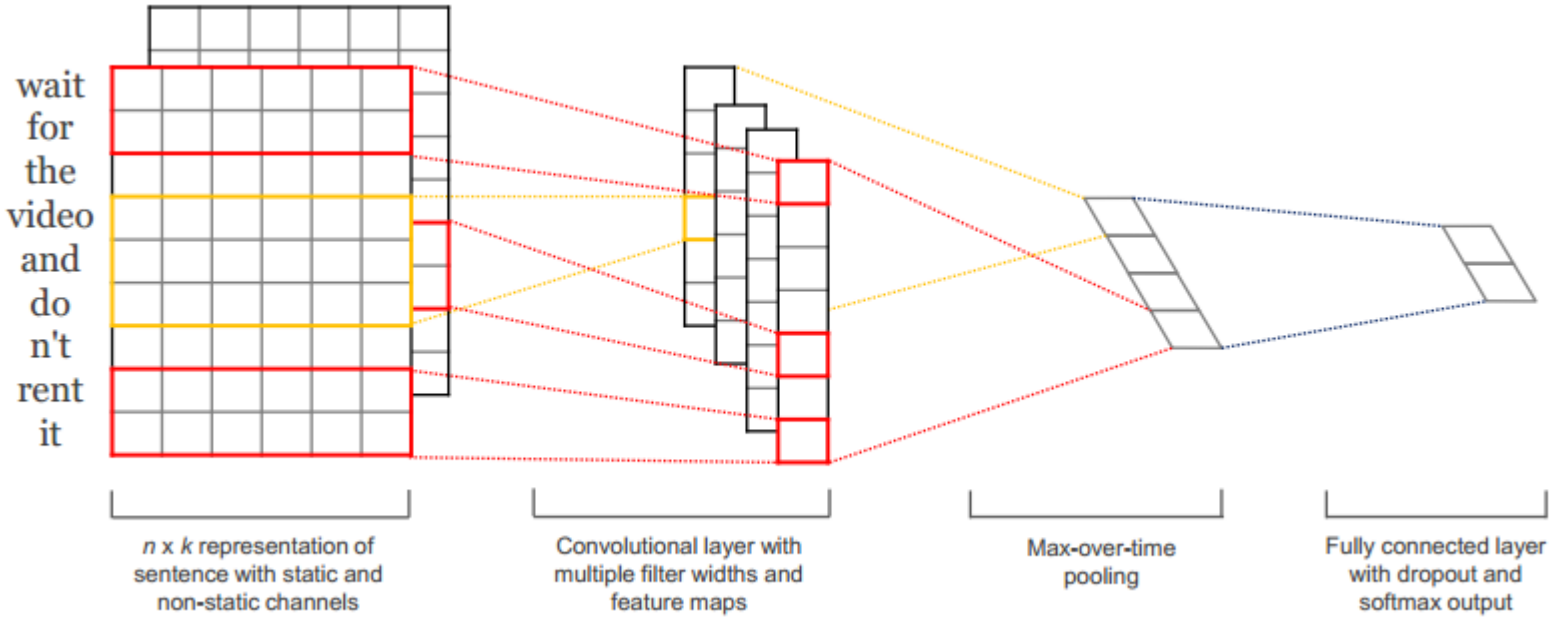


1. 将句子中所有词的词向量求平均，得到句子的向量表示
2. Hierarchical Softmax加速计算分类的概率
3. N-gram捕捉句子中（局部的）序列信息

TextCNN

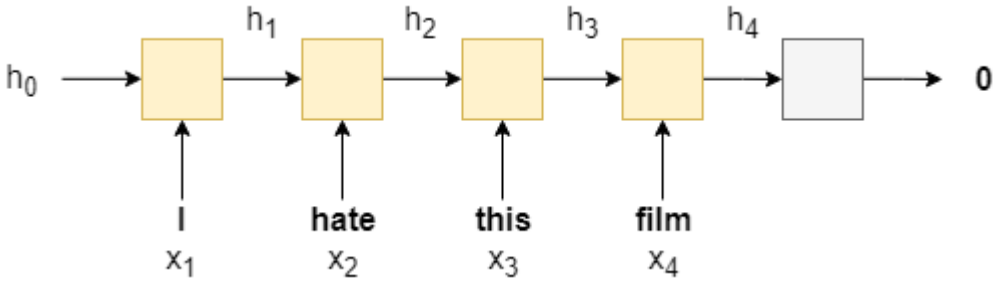

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

TextCNN

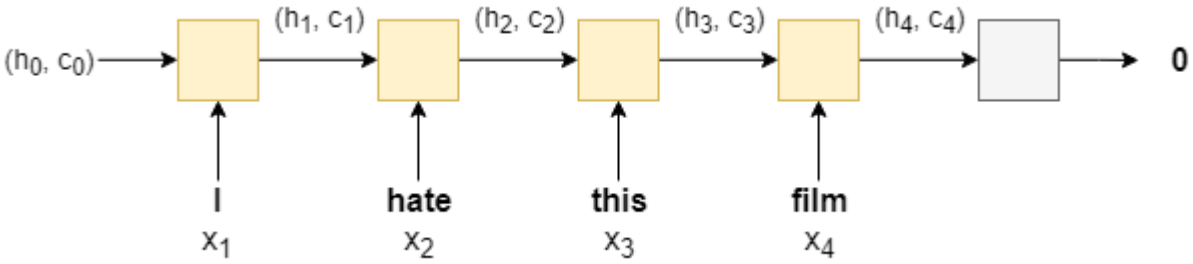


1. CNN-rand：词向量随机初始化
2. CNN-static：预训练的词向量进行初始化，并保持不变
3. CNN-non-static：预训练的词向量进行初始化，但可以被修改
4. CNN-multichannel：含有不变的词向量和可变的词向量channel

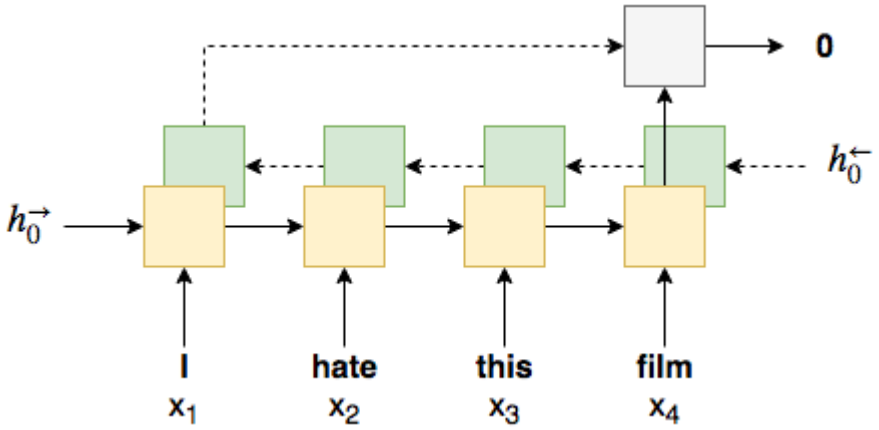
TextRNN



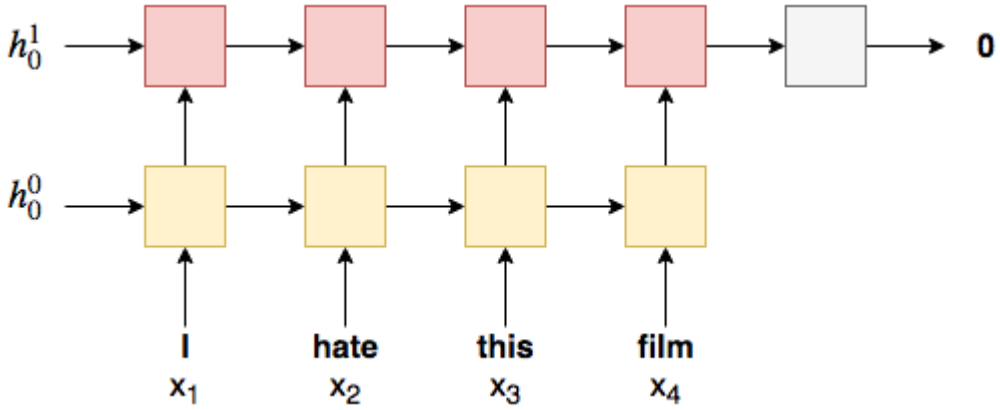
RNN



LSTM



Bidirectional RNN



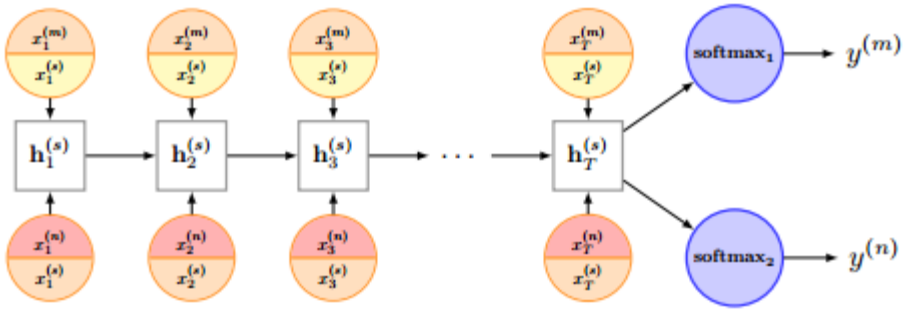
Multi-layer RNN

TextRNN

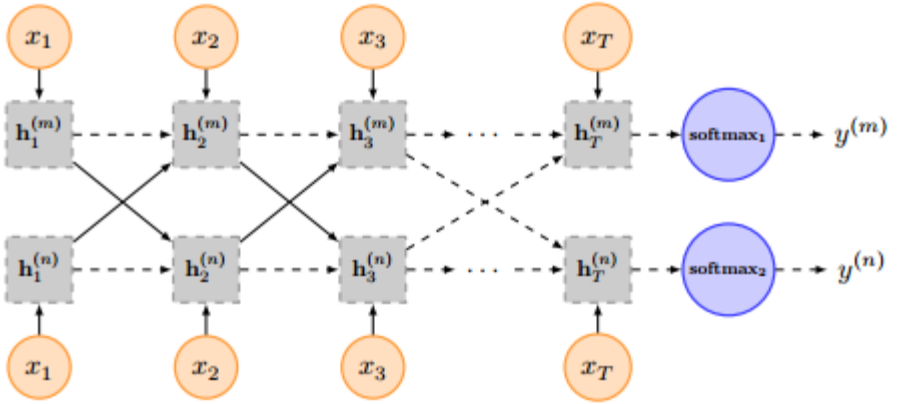
Multi-Task Learning

Dataset	Type
SST-1	Sentence
SST-2	Sentence
SUBJ	Sentence
IMDB	Document

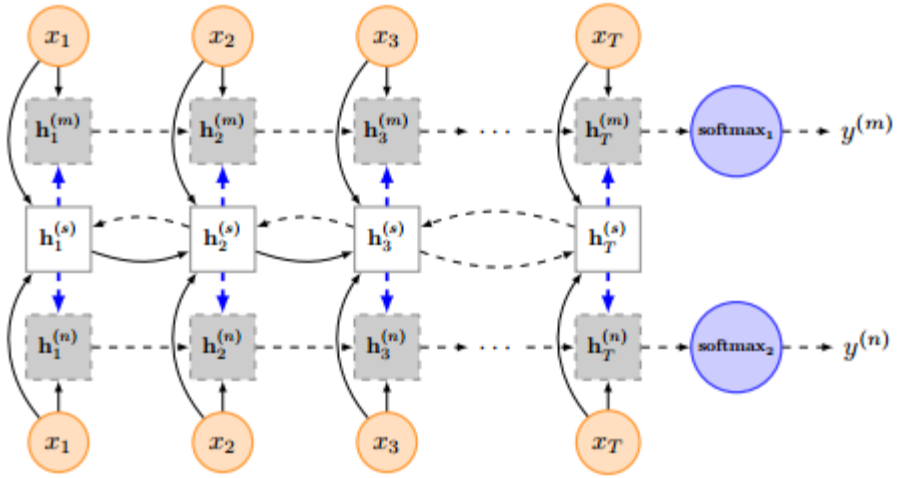
同时学习多个任务
通过特定的方式共享权重
让每一个任务都学习的很好



(a) Model-I: Uniform-Layer Architecture



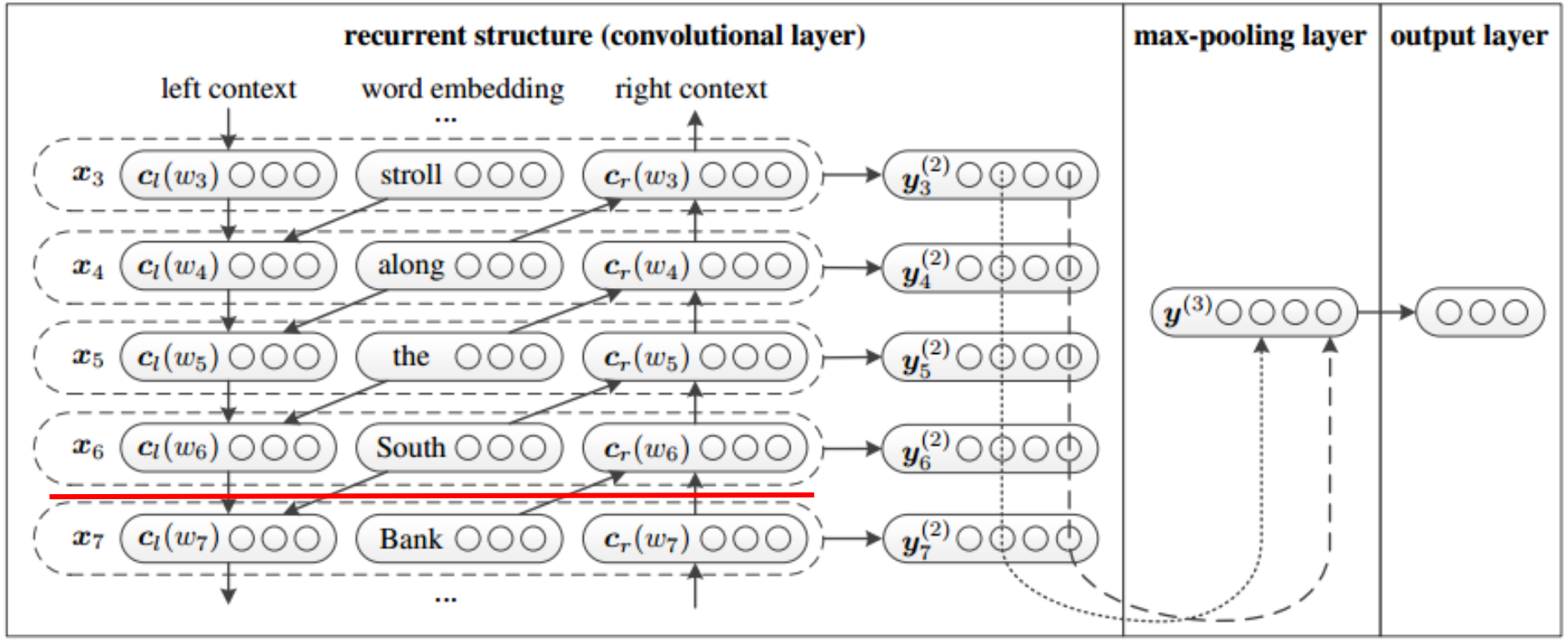
(b) Model-II: Coupled-Layer Architecture



(c) Model-III: Shared-Layer Architecture

TextRCNN

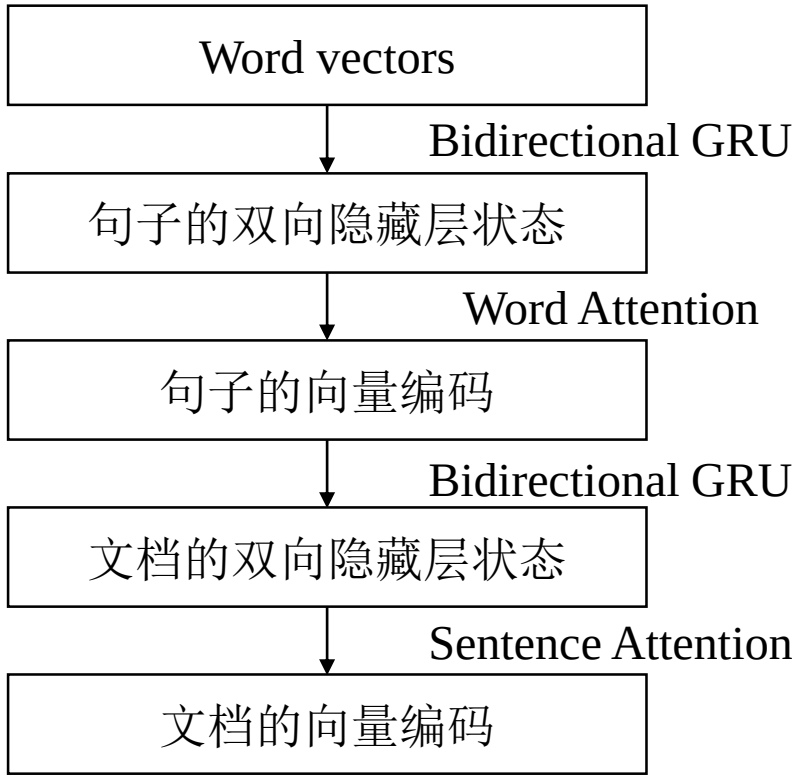
“A sunset stroll along the South Bank affords an array of stunning vantage points”

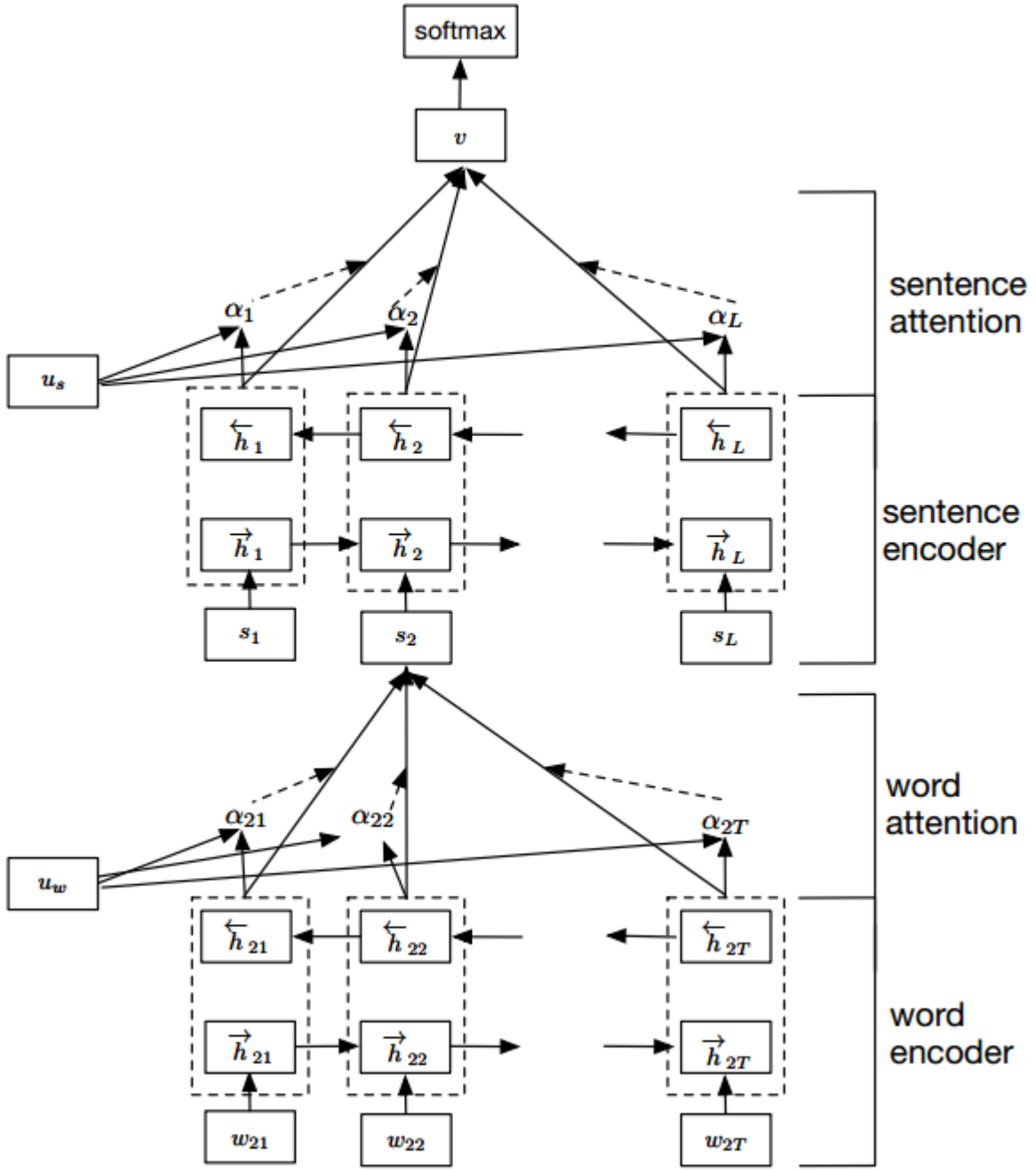


上文　　词向量　　下文　　词的表示　文本表示

HAN(GRU+Attention)

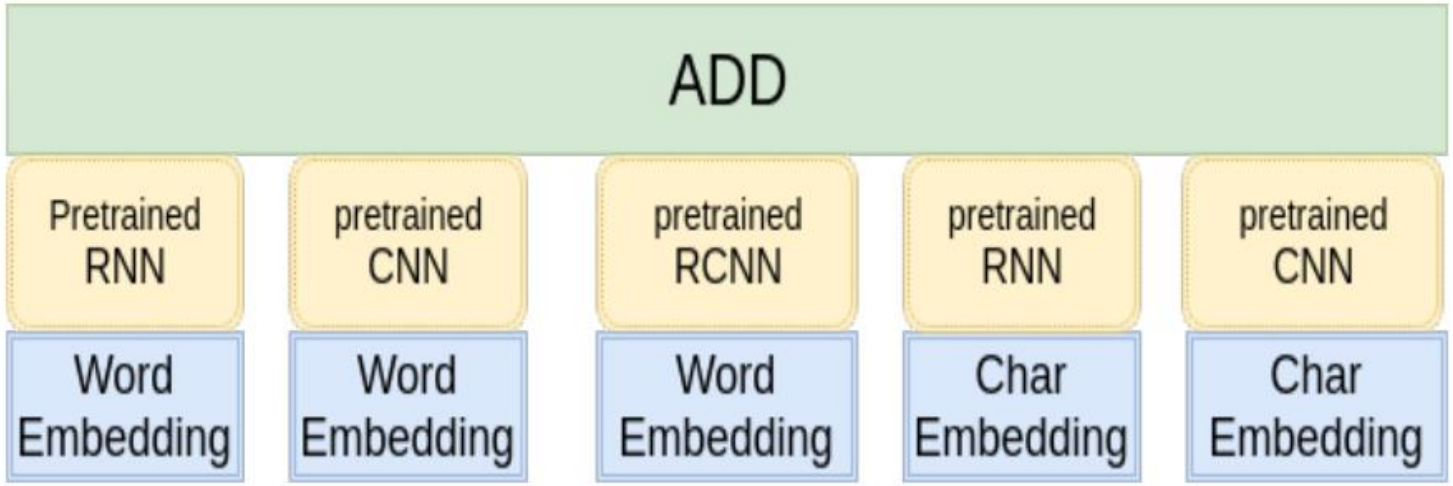
文章由句子组成，每个句子的重要性不同
句子由词语组成，每个词语的重要性不同
相同的词语在不同的上下文中重要性也不同





总结

- Char-CNN
- MemNN
- EntNet
- DMN
- …



使用深度学习方法进行文本分类，需要能够将文本表示成稠密连续向量的方法。
无论是使用预训练的词向量，还是通过模型去学习相应的参数。
所有的工作，都是在寻找一种表征文本的向量，能够更好的实现分类。

Thank You