# 文本表示简介


## 1 引言

文本分类是自然语言处理中研究最为广泛的任务之一，通过构建模型实现对文本内容进行自动分类，有很多应用场景，比如新闻文章主题分类，产品评论情感分类，检索中用户查询的意图分类等等。文本分类的大致流程：文本预处理，抽取文本特征，构造分类器。其中研究最多的就是文本特征抽取，更广义上说是**文本表示。**

关于文本表示，研究者从不同的角度出发，提出大量的文本表示模型。本文重点梳理现有模型，大致分为三类，即基于向量空间模型、基于主题模型和基于神经网络的方法，针对每类给出一些具有代表性的模型，阐述其基本思想，对于具体的细节，读者可以阅读给出的参考文献。

## 2 基于向量空间模型的方法

向量空间模型是将文本表示成实数值分量所构成的向量，一般而言，每个分量对应一个词项，相当于将文本表示成空间中的一个点。向量不仅可以用来训练分类器，而且计算向量之间的相似度可以度量文本之间的相似度。

最常用的是 TF-IDF 计算方式，即向量的维度对应词表的大小，对应维度使用 TF-IDF 计算。向量空间模型的优点是简单明了，向量维度意义明确，效果不错，但也存在明显的缺点，其一，维度随着词表增大而增大，且向量高度稀疏；其二，无法处理“一义多词”和“一词多义”问题。

在向量空间模型中，文档集合相当于表示成高维稀疏矩阵，如图 1 中所示，文档集合矩阵的维度是 N*V，其中 N 是文档数目，V 是词表的大小。为了更好的提升文本的语义表示能力，有人提出通过矩阵分解的方法，对高维稀疏矩阵进行分解，最为著名的便是潜在语义分析（Latent semantic analysis, LSA），具体而言，LSA 会构建一个文档与词项的共现矩阵，矩阵的元素一般通过 TFIDF 计算

得到，最终通过奇异值分解的方法对原始矩阵降维，可以得到文档向量和词项向量。如图 1 所示，分解后，每个文档可以用 k 维向量表示（k << V），相当于潜在语义分析实现对文档的低维语义表示。但是，以上过程通过矩阵分解得到，空间中维度的物理含义不明确，无法解释。

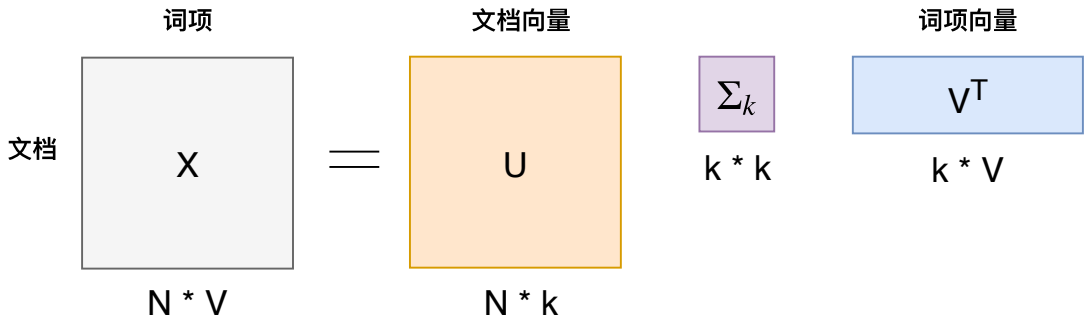


图 1　LSA

## 3 基于主题模型的方法

第 2 节中提到 LSA 算法通过线性代数中奇异值分解实现文档映射到低维语义空间里的向量，但是空间中每一个维度是没有明确物理意义的，主题模型尝试从概率生成模型的角度实现文本的表示，每一个维度是一个“主题（topic）”，这个主题通常是一组词的聚类，因此可以通过主题大概猜测每个维度所代表的语义，具有一定的解释性。

最早的主题模型 pLSA （probabilistic LSA），假设文档具有主题分布，文档中的词从主题对应的词分布中抽取。如图 2 所示，以 d 表示文档，w 表示词，z 表示主题（隐变量），则文档和词的联合概率 p(d, w)的生成过程可表示为：

$$\mathrm{p}(d,w) = p(d)\sum_{z} p(w|z)p(z|d)$$

其中 p(z|d)和 p(w|z)作为参数可以用 EM 算法进行学习。然而，pLSA 没有假设主题的的先验分布，导致参数随训练文档的数目呈线性增长，参数空间很大。

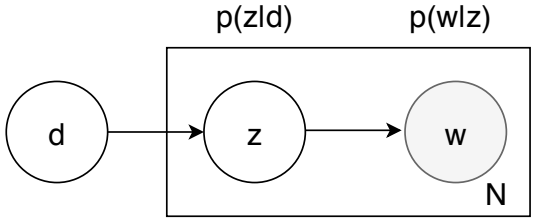


图 2　pLSA

于是，有人提出更加完善的主题的模型 LDA（Latent Dirichlet allocation），可以认为 pLSA 体现频率学派的思想，而 LDA 是贝叶斯学派的思想，LDA 在已

有的模型上中的 2 个多项式分布引入了狄利克雷先验分布，从而解决 pLSA 中存在的问题。如图 3 所示，每个文档的主题分布为多项式分布Mult(θ)，其中θ从狄利克雷先验分布Dir($\alpha$)抽取，同理，对于主题的词分布为多项式分布Mult(ϕ)，参数ϕ也是从狄利克雷先验Dir($\beta$)抽取得到。

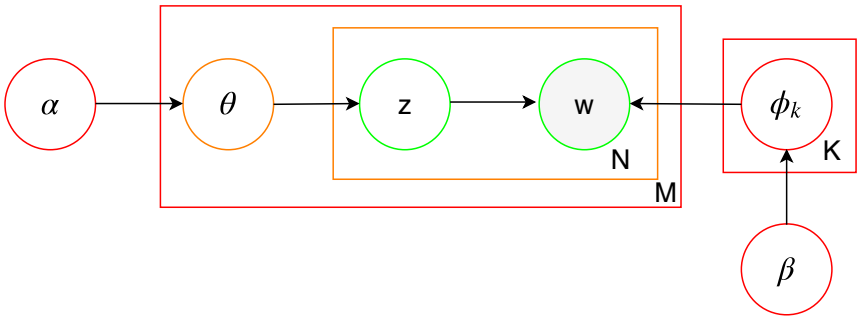


图 3　LDA

基于主题模型的方法，引入“主题”概念，具有一定的物理意义，从而得到文档的主题分布表示。当然，主题模型很存在一些问题，比如训练参数多导致训练时间长，对短文本的建模效果不好，主题数目的设置需要人工设定可能导致不合理。后来，也有很多人提出改进的方法，在这就不一一赘述了。

## 4 基于神经网络的方法

现今，基于神经网络的方法受到广泛关注，各种各样的模型被相继提出，本节总结其中最具有代表性的模型，将其分为三类：

第一类，基于词向量合成的模型，该类方法仅是在词向量基础上简单合成；

第二类，基于 RNN/CNN 的模型，该类方法利用更复杂的深度学习模型对文本进行建模；

第三类，基于注意力机制的模型，在已有神经网络模型基础上，引入注意力机制，提升文本建模效果。

### 4.1 基于词向量合成的模型

2003 年 Bengio 等人开始使用神经网络来做语言模型的工作，尝试得到词的低维、稠密的向量表示，2013 年 Mikolov 等人提出简化的模型，即著名的 Word2Vec，包含两个模型 CBOW 和 Skip-gram，前者通过窗口语境预测目标词出现的概率，后者使用目标词预测窗口中的每个语境词出现的概率。语义上相似或相关的词，得到的表示向量也相近，这样的特性使得 Word2Vec 获得巨大成功。

后来，Mikolov 等人又延续 Word2Vec 的思想，提出 Doc2Vec，核心思想是将文档向量当作“语境”，用来预测文档中的词。Doc2Vec 算法可以得到词向量和

文档向量。如图 4 所示，两个算法的思想基本一致。

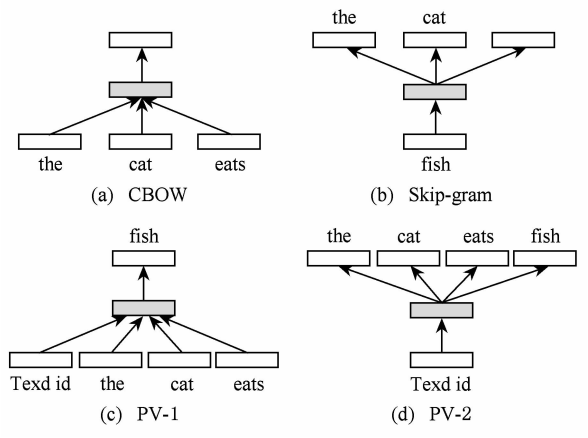


图 4　Word2Vec 和 Doc2Vec 比较

其实，也可以通过最简单的合成方式实现从词向量到句子向量的表示，fastText 就是这样简单有效的模型，如图 5 所示，输入层是词向量，然后通过把句子里的词向量平均就得到句子的表示，最后送到分类器中。不过，输入端会另外补充一些 n-gram 信息来捕捉局部序列信息。fastText 是线性分类模型，实验表明在诸多“简单”文本分类任务中表现出色，且具备训练速度非常快的优点，所以可以成为很好的 Baseline。

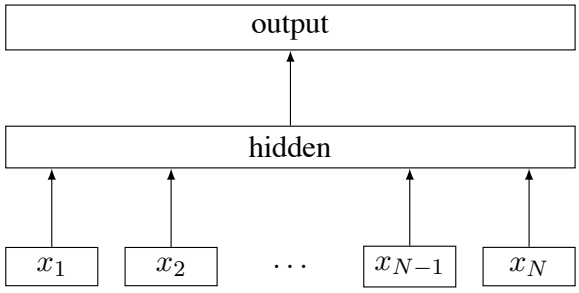


图 5　fastText 模型

## 4.2 基于 RNN/CNN 的模型

自然语言中，词构成句子，句子构成文档，有很多工作尝试合理表示词向量同时，也有很多模型被提出来建模句子和文档，其中最常见的网络结构便是 LSTM 和 CNN。

2014 年 Kim 提出基于卷积神经网络的文本分类，如图 6 所示，输入是句子对应的词向量矩阵，经过一层卷积层和一层 Max Pooling 层，得到句子的表示，送入到全连接层，最后 softmax 输出。卷积神经网络擅长提取重要的局部特征，在文本分类中，可以理解为不同大小的卷积核在提取不同 n-gram 特征。一般认为，卷积神经网络无法考虑长距离的依赖信息，且没有考虑词序信息，在有限的窗口下提取句子特征，会损失一些语义信息。

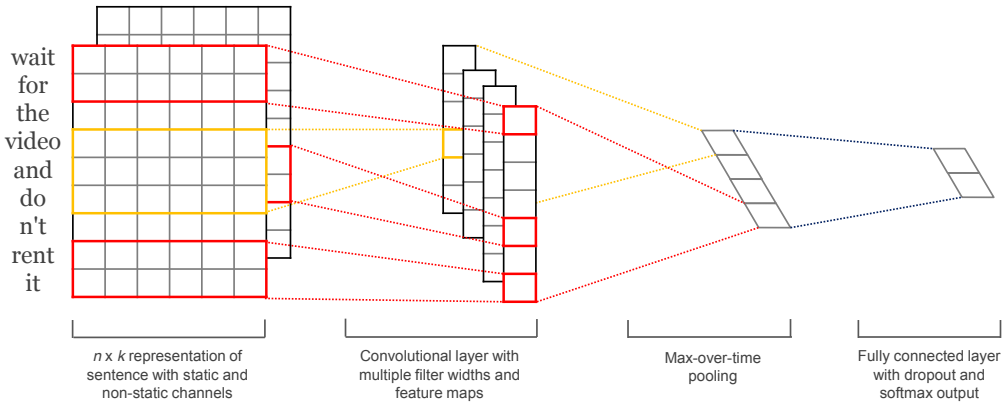


图 6　CNN 网络用于文本分类

针对 CNN 的不足之处，LSTM 和 GRU 等循环神经网络因为擅长捕捉长距离信息，所以也被大家尝试用来文本表示。如图 7 所示，图中利用双向 LSTM 来建模输入句子，输入是句子的词向量，输入至 BiLSTM 中进行序列建模。最后句子表示，可以尝试两种方法，其一，选择最后的 hidden state 用来表示句子；其二，所有 hidden state 的平均用于表示句子。

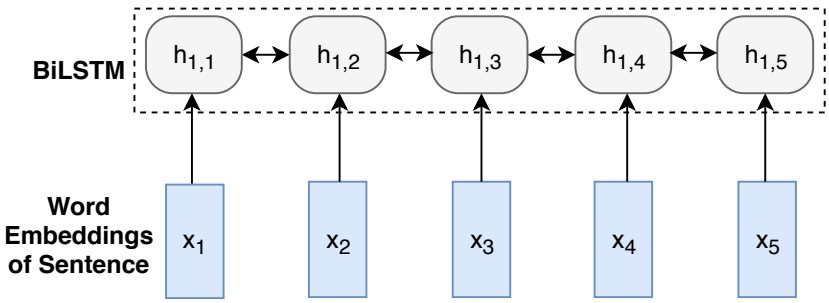


图 7　BiLSTM 用于文本表示

刚才分析到，CNN 擅长提取局部特征，而 LSTM 擅长捕捉长距离信息，不难想到，有人尝试结合两种网络的优点，提出 RCNN 用于文本建模。如图 8 所示，整个网络结构主要有两部分组成，循环结构和 Max Pooling。循环结构，可以理解为，在已有词向量为输入的基础上，通过双向 RNN 网络学习每一个词的左、右上下文信息，接着将三部分(left context, word embedding, right context)表示向量拼接起来，作为句子中每一个词的表示，然后使用变换得到中间语义表示；Max Pooling 层，采用 element-wise 的 max pooling 方式，可以从变长的输入中得到固定的句子表示。

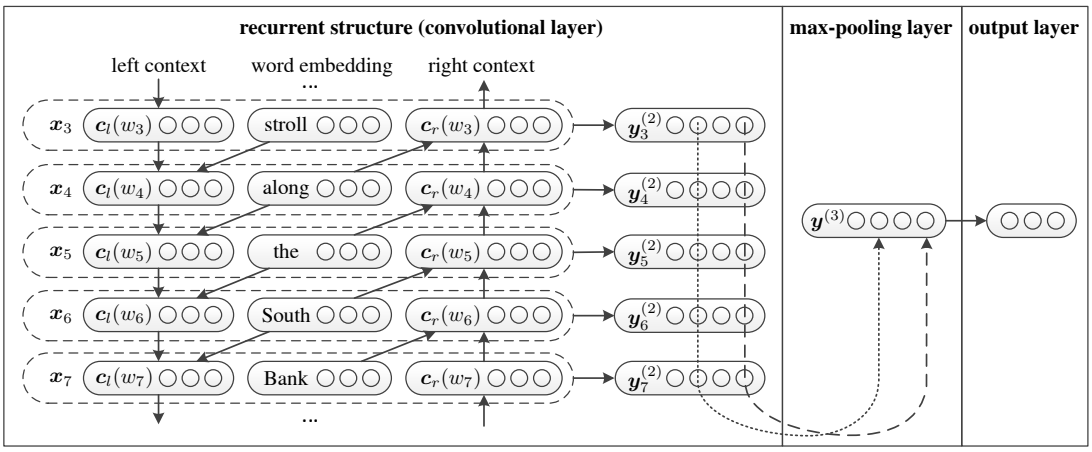


图 8　RCNN 用于文本表示

## 4.3 基于注意力机制的模型

注意力被认为是一种有效选择信息的方式，可以过滤掉大量与任务无关的信息，最先在机器翻译任务中被提出，解决 seq2seq 中 encoder 过程把源序列映射成固定大小的向量存在“损失”信息的情况。紧接着，Attention 被推广到各种 NLP 任务中，文本表示任务当然不例外。这里，主要介绍两种 Attention 的形式，Hierarchical Attention 和 Self-Attention。

Hierarchical Attention 网络结构，如图 9 所示，该模型基于两个基本假设，其一，文档是分层结构的，词构成句子，句子构成文档；其二，文档中不同词或句子提供的信息量不一样的，该模型适合用来表示包含多个句子的文档的表示问题。模型主要由 word encoder 和 sentence encoder，以及相应的 attention 组成，word encoder 部分用于得到句子的表示，该层的输入为句子的词向量，经过双向 GRU 后得到中间表示，word attention 部分对中间表示按 attention 值进行加权得到此句最终的句子表示；sentence encoder 部分用于得到文档的表示，使用 word encoder 部分得到文档中所有句子的表示后，将此作为 sentence encoder 部分的输入，类比 word encoder 的计算，可以最终得到文档的表示。简言之，利用 Hierarchical Attention 结构，一层词输入得到句子表示，一层句子输入得到文档表示。即使文档长度较长，但是在注意力的作用下，依然可以较好的捕捉到有效的特征信息，忽略无意义的输入。

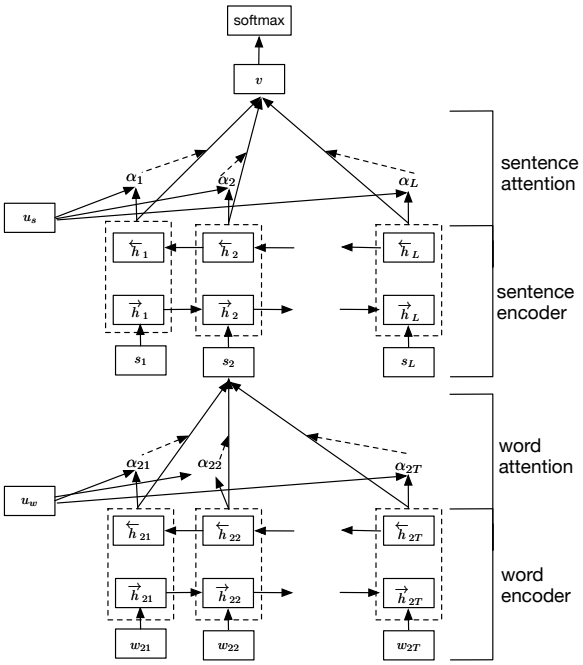


图 9　Hierarchical Attention

Self-Attention 网络结构，如图 10 所示，大多数神经网络模型将文本表示成一维的向量，但是此模型通过二维矩阵来表示句子，包括两部分，一部分是双向的 LSTM，另一部分是自注意力机制，自注意力机制实现对双向 LSTM 中所有的隐藏状态以不同权重的方式线形组合，每一次组合获得句子的一部分表示，多次组合便得到矩阵表示（图中矩阵 M）。

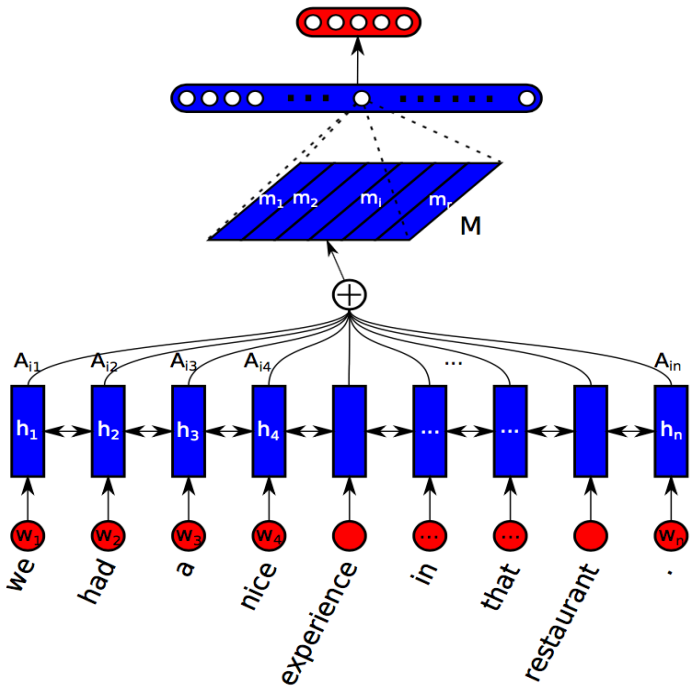


图 10　Self-Attention

## 5 总结

本文简述了具有代表性的文本表示模型，将现有模型分为三类进行介绍，包括基于向量空间模型、基于主题模型和基于神经网络的方法。不过，本文中提及的神经网络方法大部分都是有监督方法，通常都是结合具体的应用根据有监督的信息进行训练，其实也有大量的方法是通过无监督的方法获得普适性的文本表示，感兴趣的读者可以自行去翻阅相关文献。

## 参考文献

1. 信息检索导论，第 14、18 章
2. 短文本理解，计算机研究与发展 2016
3. LDA 数学八卦
4. Distributed Representations of Words and Phrases and their Compositionality, NIPS2013
5. Distributed Representations of Sentences and Documents, ICML2014
6. Bag of Tricks for Efficient Text Classification, EACL2017
7. Convolutional Neural Networks for Sentence Classification, EMNLP2014
8. Recurrent Convolutional Neural Networks for Text Classification, AAAI2015
9. Hierarchical Attention Networks for Document Classification, NAACL2016
10. A Structured Self-attentive Sentence Embedding, ICLR2017