传统文本分类方法 人工特征工程+浅层分类模型

TF-IDF方法

缺点：

1、传统做法主要问题的文本表示是高纬度高稀疏的，特征表达能力很弱，而且神经网络很不擅长对此类数据的处理；

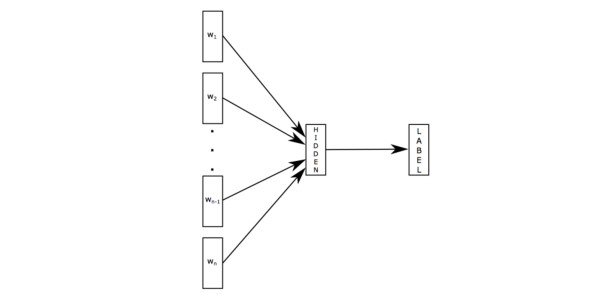
2、需要人工进行特征工程，成本很高。

基于深度学习的文本分类

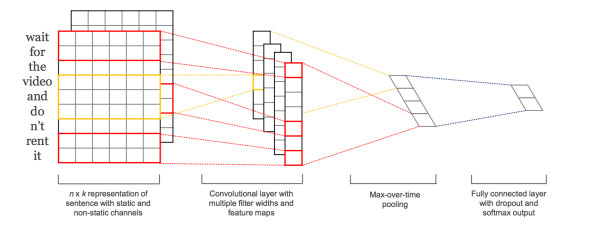
FastText文本分类

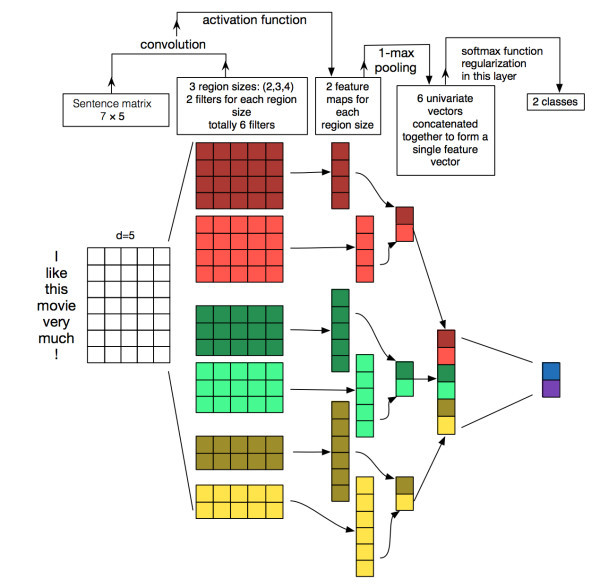
原理：把句子中所有的词向量进行平均（某种意义上可以理解为只有一个avg pooling特殊CNN），然后直接接 softmax 层，加入了一些 n-gram 特征的 trick 来捕获局部序列信息

缺点：模型过于简单，完全没有考虑词序信息



Text-CNN 文本分类



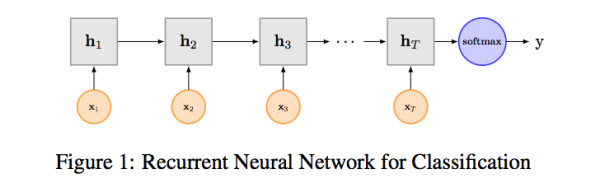


缺点：

固定 filter\_size 的视野，一方面无法建模更长的序列信息，另一方面 filter\_size 的超参调节也很繁琐

Text-RNN 文本分类

自然语言处理中更常用的是递归神经网络（RNN, Recurrent Neural Network），能够更好的表达上下文信息。具体在文本分类任务中，Bi-directional RNN（实际使用的是双向LSTM）从某种意义上可以理解为可以捕获变长且双向的的 "n-gram" 信息



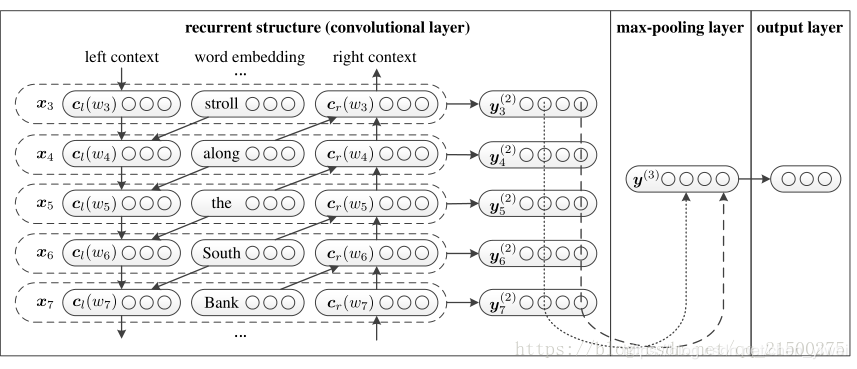
缺点：

RNN能够更好地捕捉上下文信息。这可能有利于捕获长文本的语义。然而，RNN是一个有偏倚的模型，在这个模型中，后面的单词比先前的单词更具优势。因此，当它被用于捕获整个文档的语义时，它可能会降低效率，因为关键组件可能出现在文档中的任何地方，而不是最后。

**TextRNN + Attention**

NN和RNN用在文本分类任务中尽管效果显著，但都有一个不足的地方就是不够直观，可解释性不好，特别是在分析badcase时候感受尤其深刻。而注意力（Attention）机制是自然语言处理领域一个常用的建模长时间记忆机制，能够直观的解释各个句子和词对分类类别的重要性

**TextRCNN（TextRNN + CNN）**



应用一个双向的循环结构，与传统的基于窗口的神经网络相比，它可以大大减少噪声，从而最大程度地捕捉上下文信息。此外，该模型在学习文本表示时可以保留更大范围的词序。其次，我们使用了一个可以自动判断哪些特性在文本分类中扮演关键角色的池化层（max-pooling），以捕获文本中的关键组件。我们的模型结合了RNN的结构和最大池化层，利用了循环神经模型和卷积神经模型的优点。此外，我们的模型显示了O(n)O(n)的时间复杂度，它与文本长度的长度是线性相关的。