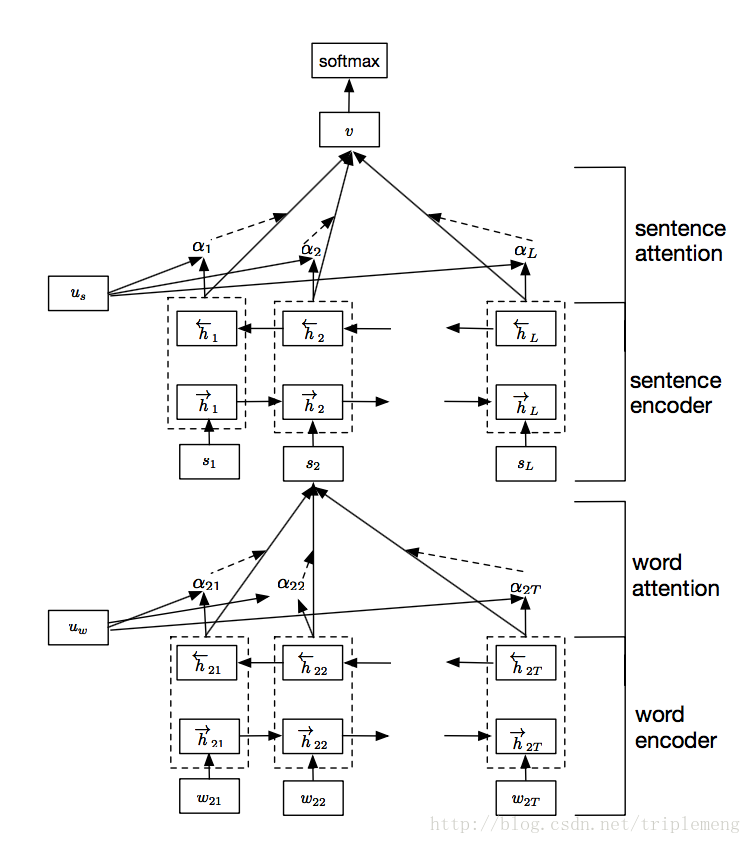
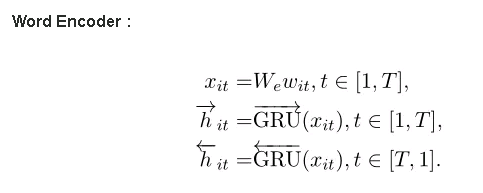
**链接：**Hierarchical Attention Networks for Document Classification

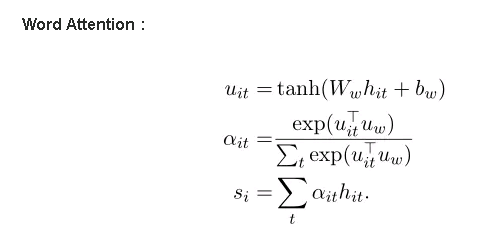
<http://blog.csdn.net/liuchonge/article/details/73610734>

**前言：**本文针对文本分类任务提出了一个层次化attention机制模型(**HAN**)，有两个显著的特点：(1)采用“词-句子-文章”的层次化结构来表示一篇文本。(2)该模型有两个层次的attention机制，分别存在于词层次(word level)和句子层次(sentence level)。从而使该模型具有对文本中重要性不同的句子和词的能力给予不同“注意力”的能力。作者在6个数据集合上进行了测试，并且相较以往的模型效果提升显著。最后，通过可视化说明该模型可以选择出含有丰富信息的词语和句子。  
**一 、写作动机**  
文本分类是一项基础的NLP任务，在主题分类，情感分析，垃圾邮件检测等应用上有广泛地应用。其目标是給每篇文本分配一个类别标签。本文中模型的直觉是，不同的词和句子对文本信息的表达有不同的影响，词和句子的重要性是严重依赖于上下文的，即使是相同的词和句子，在不同的上下文中重要性也不一样。就像人在阅读一篇文本时，对文本不同的内容是有着不同的注意度的。而本文在attention机制的基础上，联想到文本是一个层次化的结构，提出用词向量来表示句子向量，再由句子向量表示文档向量，并且在词层次和句子层次分别引入attention操作的模型。  
**二、模型构建**  
HAN的模型结构如Fig.1所示，它包含一个词序列编码器，一个词层面的attention层，一个句子序列编码器，一个句子层级的attention层。  
   
  
**1. 基于GRU的词序列编码器**  
GRU是RNN的一个变种，使用门机制来记录序列当前的状态。隐藏层有两个门(gate)，重置门(reset gate)和更新门(update gate)。这两个门一起来控制当前状态有多少信息要更新。更新门(update gate)是用来决定有多少过去的信息被保留，以及有多少新信息被加进来：  
重置门决定有多少过去的信息作用于候选状态，如果是0，即忘记之前的所有状态：

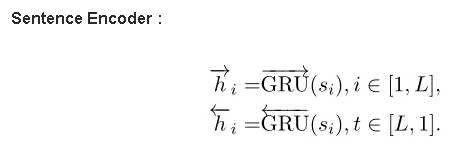
**2.词编码器（Word Encoder）**  
由词序列组成,组成的句子,首先把词转化成词向量，然后用双向的GRU网络,可以将正向和反向的上下文信息结合起来，获得隐藏层输出。对于一个给定的词语,经过GRU网络后，我们获得了一种新的表示,含了周围两个方向的信息。

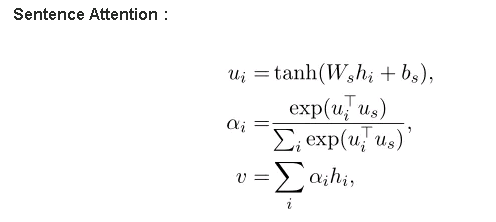


**3.词级别的attention机制**  
attention机制的目的是要把一个句子中，对句子的含义最重要，贡献最大的词语找出来。我们通过将输入到一个单层的感知机(MLP)中得到的结果作为的隐含表示。  
  
为了衡量单词的重要性,我们用和一个随机初始化的上下文向量的相似度来表示，然后经过softmax操作获得了一个归一化的attention权重矩阵。  
  
有了attention权重矩阵以后，我们可以将句子向量看作组成这些句子的词向量的加权求和。这里的上下文向量是在训练网络的过程中学习获得的。我们可以把当作一种询问的高级表示，比如“哪些词含有比较重要的信息？”



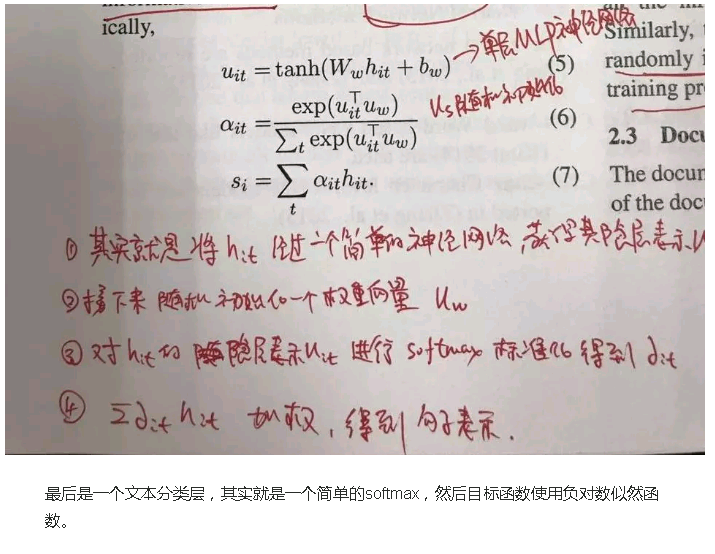
**4. 语句编码器(Sentence Encoder)**  
得到了句子向量表示以后，我们用类似的办法获得文档向量：  
对于给定的句子我们得到了相应的句子表示。这样获得的表示可以包含两个方向的上下文信息。

  
句子级别的attention：和词级别的attention类似，我们也提出了一个句子级别的上下文向量,来衡量一个句子在整篇文本的重要性。

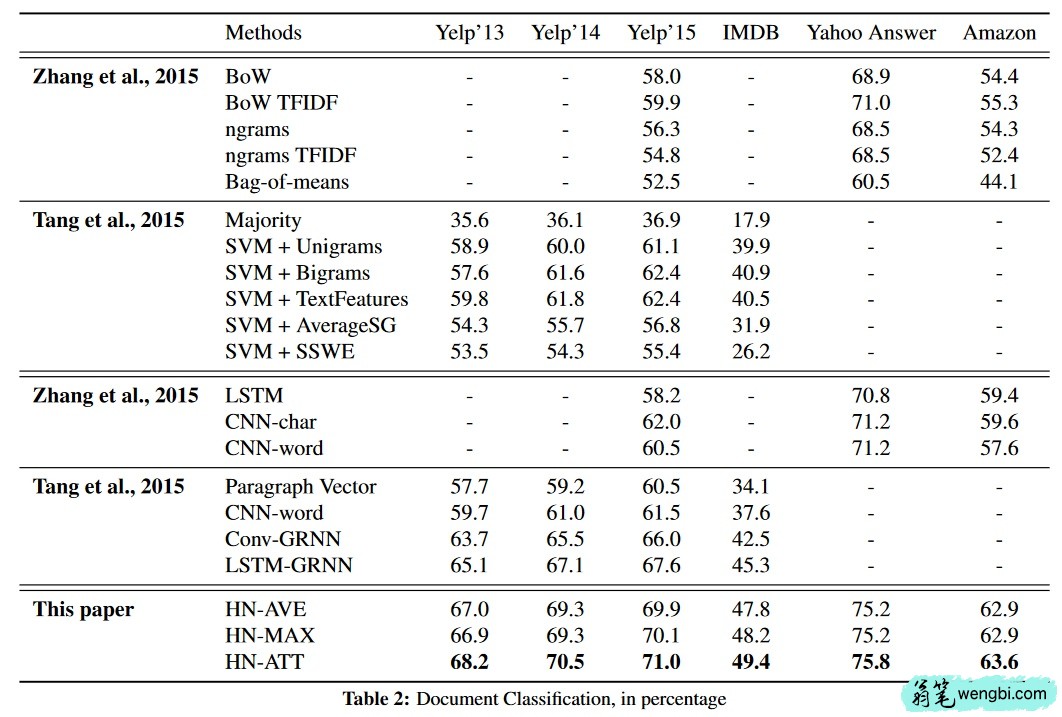
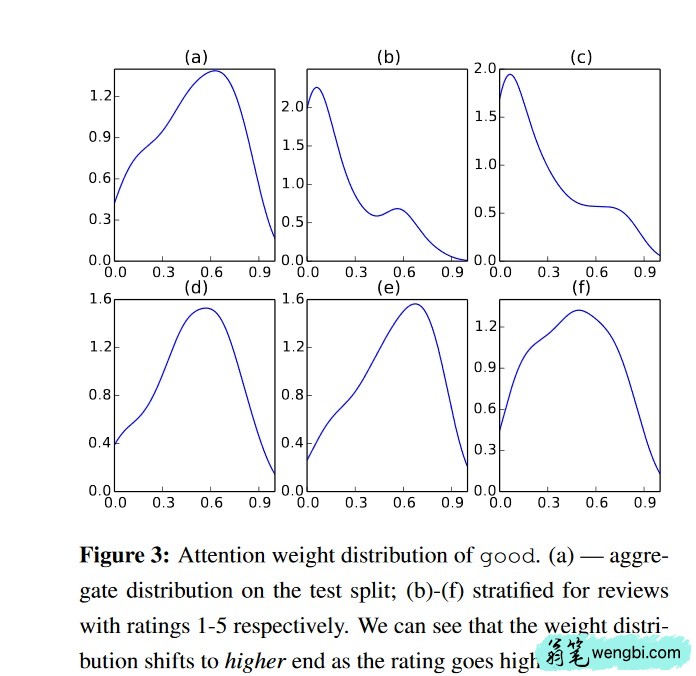
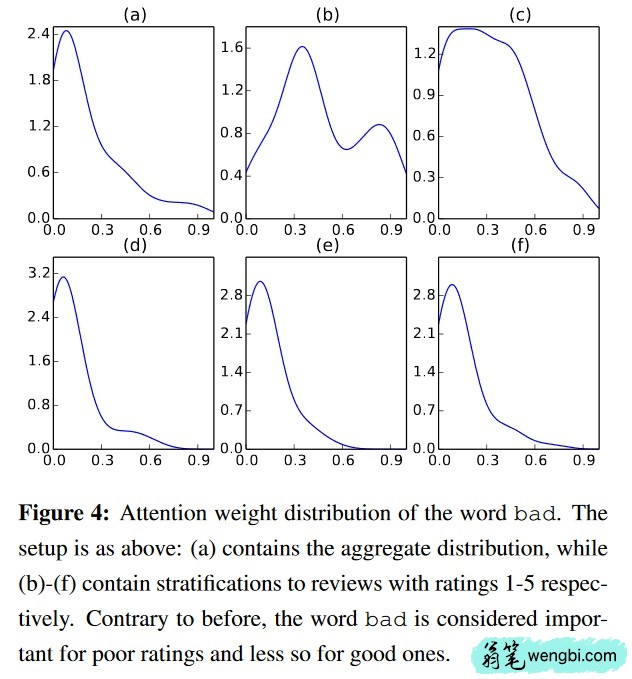
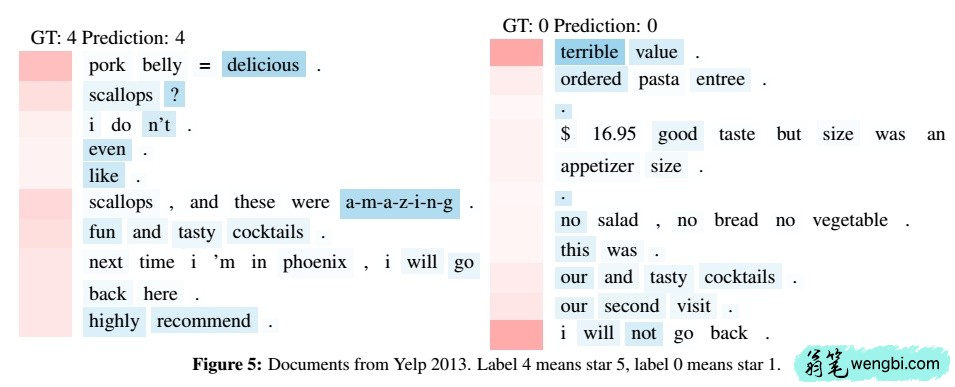
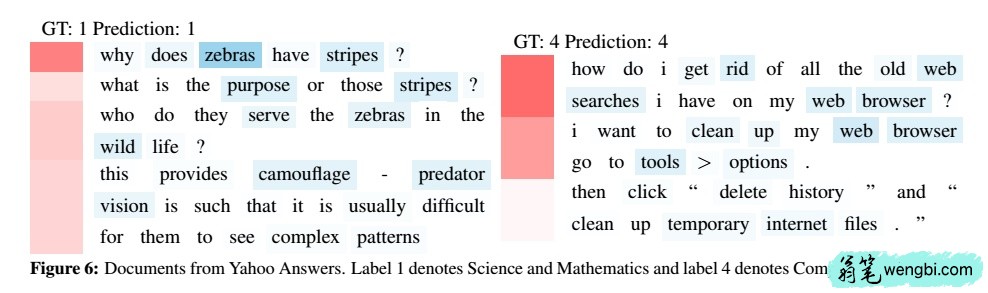


 我们获得了整篇文章的向量表示,最后可以使用全链接的softmax层进行分类。

结合上图，我们会发现ui是使用一个简单的神经网络对hi的隐层表示，然后α是其softmax之后的结果，然后使用α对hi进行加权计算得到最终的输出向量即可。总结如下图：

  
**三、实验**  
**1 数据集**  
论文中使用了六个数据集，Yelp reviews2013,2014,2015;IMDB reviews;Yahoo Answers;Amazon reviews。每个数据集合中80%的数据用作训练集合，10%的数据用作验证集合，剩余10%的集合用作测试集合。  
**2 实验指标与对比模型**

**线性模型：**使用手工构建的统计数据作为特征，多项式logistic回归作为分类器**SVM：**支持向量机+unigr，bigrams**word-based CNN:**基于词的卷积神经网络**Character-based CNN：**字符级别的卷积神经网络**Conv/LSTM-GRNN：**使用“词-句子-文本”的层次化结构表示文档，然后用带有门限制的RNN做分类。

**3 实验结果**  
**   
4 结果分析**  
  
根据上表的实验结果可以看出来，层次化attention模型在所有的六个数据集合上均取得了最好的效果。这种效果的提升是不受数据集大小限制的。在相对较小的数据集比如Yelp2013和IMDB上，我们的HAN模型超过基准模型的最好表现的比率分别为3.1%和4.1%。相同的在，大数据集上，我们的模型优于之前的最好模型的比例分别为3.2%，3.4%，4.6%，6.0%，分别在Yelp2014，Yelp2015，Yahoo Answer,Amazon Answer。  
一些没有使用层次结构来表示文本的神经网络分类算法比如CNN-word，CNN-char，LSTM在一些大数据集合上并没有超过基准模型太多。单从文本的结构化表示来看，HN-MAX,HN-AVG都可以显著提升CNN-word，LSTM等模型的性能。我们将层次话结构和Attention机制联合起来的模型，更是超过了单纯的层次化模型LSTM-GRNN。  
**5 可视化分析**  
   
上图是词“good”在IMDB数据集合种的的attention权值分布，可以看出，当评价等级为1时Fig3（b）,good的权值集中在比较低的一端，当评论的变得向好评方向时，good的权重也开始在高的一端分布。  
  
   
相反的，词“bad”在评价等级较低的评论里权值较大，在评价高的评论里权值小。  
   
情感分析任务种，从上图我们可以看到，在好评里，delicious,fun,amazing,recommend 等词所在的句子颜色深，权值大。在差评里，terrible，not等词占的比重大。  
   
在主题分类任务中，在科学自然类别里，权值比较大的词zebra,strips,camouflage,predator。 在计算机网络类别里web,searches, browsers的权值较大。  
四 总结  
本文提出了一种基于层次化attention的文本分类模型，可以利用attention机制识别出一句话中比较重要的词语，利用重要的词语形成句子的表示，同样识别出重要的句子，利用重要句子表示来形成整篇文本的表示。实验证明，该模型确实比基准模型获得了更好的效果，可视化分析也表明，该模型能很好地识别出重要的句子和词语。