北京郵電大學

本 科 毕 业 设 计（论文）

题目：基于循环神经网络的影评情感分析系统的设计与实现

姓　　名　王　超

学　　院　计算机学院

专　　业　计算机科学与技术

班　　级　2014211307

学　　号　2014211310

班内序号　07

指导教师　刘晓鸿

2018 年 6 月

北 京 邮 电 大 学

本科毕业设计（论文）任务书

学院	计算机学院	专业	计算机科学与技术	班级	2014211307
学生姓名	王超	学号	2014211310	班内序号	07
指导教师姓名	刘晓鸿	所在单位	计算机学院	职称	副教授
设计(论文)题目	（中文）基于循环神经网络的影评情感分析系统的设计与实现				
	（英文）Design and implementation of the movie-review sentiment classification system based on RNN				
题目分类	工程实践类☐ 研究设计类☐ 理论分析类☑				
题目来源	题目是否来源于科研项目 是☐ 否☑				
	科研项目名称：				
	科研项目负责人：				

主要任务及目标：

越来越多的人喜欢通过社交媒体的电影评论来了解电影并做出观影的决策；制片商也能够根据相关的影评来调整营销策略。项目需要完成的工作：获得社交媒体上的电影评论数据；搭建循环神经网络模型，对电影评论进行情感分析训练和建模；处理数据中可能存在的不均衡数据问题，改进得到的模型，提升模型分类准确率和召回率。将模型应用到其他可能场景，证明模型有效性及其泛化能力。

主要内容：

1.搭建基于 Tensorflow 的深度学习网络平台，并获取影评数据集；

2.使用不同结构的深度网络模型，对数据集进行训练，得到相应的深度神经网络；

3.通过训练时间和计算时间及网络规模之类特性对比，得到合适的神经网络，并对实验中的结果进行分析解释；

4. 基于上述得到的神经网络实现影评情感分析软件。

主要参考文献：

[1] Lei Zhang,Shuai Wang, Bing Liu, Deep Learning for Sentiment Analysis: A Survey, https://arxiv.org/abs/1801.07883

[2]Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP-14), pages 1746-1751.

<table>
<tr><td colspan="4">[3] Jin Wang,, Liang-Chih Yu,K. Robert Lai and Xuejie Zhang,Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics , pages 225–230,Berlin, Germany, August 7-12, 2016.</td></tr>
<tr><td colspan="4">进度安排：
3 月 5 日---4 月 1 日：搭建环境，读相关文献；
4 月 2 日---4 月 29 日：进行算法设计，比较不同方法的结果，并进行编码和调试；
4 月 30 日---6 月 3 日： 通过不同结果对比，优化设计实现；
6 月 5 日---6 月 22 日：撰写论文及答辩。</td></tr>
<tr><td>指导教师签字</td><td></td><td>日期</td><td>年 月 日</td></tr>
</table>

北 京 邮 电 大 学

本科毕业设计（论文）答辩决议书

<table>
<tr><td>学院</td><td>计算机学院</td><td>专业</td><td>计算机科学与技术</td><td>班级</td><td>2014211307</td></tr>
<tr><td>学生姓名</td><td>王超</td><td>学号</td><td>2014211310</td><td>班内序号</td><td>07</td></tr>
<tr><td rowspan="2">毕业设计
（论文）题目</td><td colspan="3">（中文）基于循环神经网络的影评情感分析系统的设计与实现</td><td>百分制成绩</td><td>89</td></tr>
<tr><td colspan="3">（英文）Design and implementation of the movie-review sentiment classification system based on RNN</td><td>五级分制成绩</td><td>良</td></tr>
<tr><td>指导教师姓名</td><td>刘晓鸿</td><td>所在单位</td><td>计算机学院</td><td>职称</td><td>副教授</td></tr>
<tr><td colspan="6">指导教师评语：（主要包含选题背景、意义；设计（论文）质量；设计（论文）成果、价值、创见性；论文撰写水平、文本规范程度；学生能力体现、工作量、工作态度；不足和希望等方面）
论文选题结合科研，具有应用价值。
论文研究循环神经网络对电影评论进行情感分析，设计了基于注意力的 IndRNN 模型，以处理后得到的含有输入序列注意力概率分布的语义编码为输入，可有效地突出关键词的作用。相关数据集的测评表明,该模型取得了较好的分类效果。
论文条理清楚，写作规范，按期完成了相关的任务。</td></tr>
<tr><td>指导教师评分
(满分 40 分)</td><td>36</td><td>签字</td><td></td><td>日期</td><td>2018 年 6 月 12 日</td></tr>
<tr><td colspan="6">答辩小组评语：（主要包含选题背景、意义；设计（论文）质量；设计（论文）成果、价值、创见性；论文撰写水平、文本规范程度；答辩准备、陈述、回答问题情况；不足和希望等方面）
论文选题结合科研，具有应用价值。论文按期完成任务书规定内容。论文条理清楚，开发验证工作正确。论文工作量大。专业英文翻译清楚流畅。答辩过程中回答问题正确。

答辩小组评分：53 (满分 60 分)
组长职称：教授 签字：
成员职称：副教授 签字：
成员职称：副教授 签字：
成员职称： 签字：
成员职称： 签字：
2018 年 6 月 12 日</td></tr>
</table>

注：毕业设计（论文）成绩由指导教师评分（满分 40 分）和答辩小组评分（满分 60 分）相加，得出百分制成绩，再按 100-90 分为“优”、89-80 分为“良”、79-70 分为“中”、69-60 分为“及格”、60 分以下为“不及格”的标准折合成五级分制成绩。

北 京 邮 电 大 学

本科毕业设计（论文）诚信声明

本人声明所呈交的毕业设计（论文），题目《基于循环神经网络的影评情感分析系统的设计与实现》是本人在指导教师的指导下，独立进行研究工作所取得的成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：_____

基于循环神经网络的影评情感分析系统的设计与实现

摘　要

在自然语言处理领域里，情感分析已经被广泛应用，一般划分为以下几个部分: 文本的预处理、文本特征的提取、机器学习分类器训练等方面。随着深层神经网络模型在图像识别，机器翻译等领域内的巨大进展，深层神经网络模型被证明在文本特征向量提取方面有着巨大的优势。

本文的研究内容是循环神经网络在情感分析领域对电影评论的研究。主要结合时下性能较好的循环神经网络方法，进行分析比较并结合。本文要基于 word2vec 使用模型, 采用了 Word Embedding 机制训练包含语义特征的词向量，避免了高维度输入易产生维度灾难的问题。同时，基于 Word Embedding 训练出的词向量具有相似度的特征，具备了一定的语义信息，作为 Attention-Based IndRNN 模型的输入，对二值分类器的性能具有一定的提高效果。

在文本特征提取问题上，基于时下较为优异的模型，本文设计了 Attention-Based IndRNN 和 Attention-Based Bi-IndRNN 模型，其中 IndRNN 相对传统 RNN 和 LSTM 来说，解决了梯度消失和梯度爆炸的问题。同时本文通过 Attention-Based 的方法，得到含有输入序列注意力概率分布的语义编码，并将其作为二值分类器的输入，减少了特征向量提取过程中的信息丢失和信息冗余，可以有效地突出关键词的作用。该方法可以得到包含注意力概率分布的语义特征，并对特征提取部分进行优化。基于豆瓣影评及 IMDB 影评数据集进行测评，结果表明: 这两个模型在多个情感分析任务中都取得了相对较好的效果。

关键词　循环神经网络　情感分析　长短时记忆　独立循环神经网络　注意力模型

Design and implementation of the movie-review sentiment classification system based on RNN

ABSTRACT

In the field of natural language processing, sentiment analysis has been widely used and is generally divided into the following sections: text preprocessing, text feature extraction, machine learning classifier training and so on. As deep learning technology has made great progress in image recognition, machine translation and other fields, the deep learning model has been proved to have great advantages in data preprocessing and feature extraction.

The research content of this paper is the research of film review in the field of sentiment analysis by cyclic neural network. Mainly combined with the current method of better performance of the circulatory neural network, analysis and comparison. This article is based on the word2vec use model, using the Word Embedding mechanism to train the word vectors containing semantic features, to avoid the problem of dimensional disaster caused by high-dimensional input. At the same time, the word vectors trained based on Word Embedding have similarity features and possess certain semantic information as input to the Attention-Based IndRNN model, and the performance of the binary classifier.

On the issue of text feature extraction, based on the current excellent model, this paper designs Attention-Based IndRNN and Attention-Based Bi-IndRNN models. IndRNN solves the problem of gradient disappearance and gradient explosion compared to traditional RNN and LSTM. At the same time, this paper adopts Attention-Based method to obtain the semantic code containing the attention probability distribution of the input sequence, and uses it as the input of the binary classifier to reduce the information loss and information redundancy in the feature vector extraction process, which can effectively Highlight the role of keywords. This method can get the semantic features containing attention probability distribution, and optimize the feature extraction part. Based on the evaluation of Douban film criticism and IMDB film review dataset, the results show that these two models have achieved relatively good results in multiple sentiment analysis tasks.

KEY WORDS Recurrent neural network sentiment analysis long short time memory independent recurrent neural network Attention model

目　录

第一章　绪论

1.1　研究背景及意义

信息传播的方式经历了语言符号、文字、电子传播等演变过程。21 世纪起，互联网逐渐成为人与人之间的沟通媒介。随着其发展，许多现代化的通讯交流方式建立了起来，越来越多的用户开始藉由互联网发表着自己的观点。由于这种现象，越来越多的观点和想法正在互联网上传播和发布。根据第 41 次《中国互联网络发展状况统计报告》，我国约有 55.8% 的人已经使用了互联网。任意一项商品或服务就会产生成千上万的评价信息，这就使得用户评论成为了一种越发重要的信息载体。例如，微信朋友圈、QQ 空间用户占比分别为 87.3% 和 64.4%，微博作为社交媒体，使用率也达到了 40.9%。知乎、豆瓣的用户使用率也分别有 14.6% 和 12.8%。从微博到豆瓣，社交网站和内容评论网站中蕴含了大量关于个人观点、评论的数据。

互联网包含海量的信息，人们可以利用它来帮助自己对特定问题做出决定。绝大多数人在决定某件事前会试图在网上找到相关产品，比如他们会考虑想去旅行的国家或考虑看的电影的评价。人们可以从这些网站中收集和分析关于特定主题的许多意见和观点。随着电子商务的发展，在电影市场的下游环节，用户可以通过在线票务系统选择影片、影院并在观影结束后进行点评，这一系列的行为所产生的大量数据，能够为电影产业的上游环节提供重要思考。在传统电影行业中是“我生产什么观众就看什么”，用户只能被动接受，可选择性小。而在互联网环境下，观众由产生观影意向到走进影院、完成观影，会经过“搜索关注、购票观看、评分评论”三个阶段，不管在哪一个阶段，用户的主体地位都已经被默许[1]。

2012 年，国内学者殷国鹏等采用信息采纳理论作为研究基础，讨论了消费者在购买决策中受评论的两类影响因素，即评论特征和评论者要素，并融合社会网络视角构建了评论有用性模型[2]。社交媒体的电影评论对于购票者做出决产生了一定的影响。制片商也能够根据相关的影评来调整营销策略。

但是，面对如此庞大的评论文本信息，传统的通过人工获取评论的情感倾向是一件特别费时费力的事情。因此，现在已经出现了对文本的情感分析过程自动化的需要。这将有助于人们能够以合理便捷的方式获得关于特定主题的意见和观点，而不是让他们搜索和阅读评论以获得最终意见。对于电影的上游，企业市场中，时间对企业来说比对普通用户更有价值，更需要一个自动化系统来帮助他们确定其产品和服务的用户的意见和看法。一个可以获取和分析用户评论以了解最终情绪的工具对企业更有价值。

情感分析已被证明对于几种自然语言处理（NLP）任务（如应答系统和信息提取）是有益的[3]。信息提取（IE）旨在提取与特定主题或需求相关的信息。例如，现在人们倾向于使用互联网，通过使用论坛或其他社交网络来传播他们对主题或问题的想法和想法。其中一些想法是积极的，而另一些想法和内容则有些负面。

这种在线传播情感的概念在文本分析中创造了一个新的领域，将传统的以事实和信息为中心的文本观点扩展到研究主题，以启用情感感知型应用。在过去的十年中，从文

本中提取情感在工业界和学术界都引起了高度关注。形式上，情感分析试图从人们的写作中提取人们的观点。这个主题包括许多领域，例如自然语言处理，机器学习和计算语言学。

目前为止，情感分析的主要方法分别为基于情感词典的情感分析方法和基于机器学习的情感分析方法。由于，基于情感词典的方法的性能相对机器学习方法来说，十分容易受到情感词典的大小与水平影响，因此本文将主要研究基于机器学习的方法来进行情感分析。

1.2 国内外研究现状

2003 年，J. Yi 等人提出通过提取情感特征基于情感词典的情感分析器[4]。2007 年，徐琳宏提出了根据分析句子结构提取情感特征，基于情感词典分析情感倾向[5]。基于规则、情感词典的情感分析方法长期有学者在研究，不过随着 Pang 等人提出了基于机器学习的方法如朴素贝叶斯 (Naive Bayes)、最大熵分类 (Max Entropy) 和支持向量机 (SVM) 等来对文本通过整体情感进行正负面分类[6]。由于基于情感词典的分析方法的进步空间相对机器学习方法更易受到语料库限制，在这之后，基于机器学习方法的情感分类模型被越来越多的学者研究。在这个过程中学者们发现，从文本中提取的特征被严重依赖，传统统计方法如 N-gram 模型存在维度灾难问题。根据 Bengio 等人在 2003 年的研究，神经网络能够有效解决维度灾难问题，初步使用到了词向量[7]。在 2010 年，Mikolov 等人提出了基于循环神经网络的语言模型来提高神经网络语言模型的性能，同样使用到了词向量[8]。在 2013 年，Mikolov 等人使用 CBOW 和 Skip-Gram 模型训练了具有语义特征的词向量[9]。之后，他们提出了基于 Skip-Gram 模型的改进，提出了词向量蕴含大量语法、语义的观点[10]。2014 年，Quoc 等人进行了词向量与 BOW 等模型的比较，发现其性能的优异性[11]。Mikolov 等人后来发布了开源模型 Word2vec，本文中使用的词向量就是 word2vec 训练得到的。

国外，2013 年，Vivek 等人研究了提高朴素贝叶斯在分类方面的性能的一些方法[12]。2014 年，Kim 提出了使用卷积神经网络训练模型进行情感分类，性能较佳[13]。2012 年，Socher 提出了 Matrix-Vector RNN 模型进行情感分类[14]。2013 年，Socher 对影评使用 RNTN(Recursive Neural Tensor Network) 进行情感分类，相对传统 SVN、Matrix-Vector RNN 等方法取得了较好的结果[15]。2014 年，dos Santos 提出一个新的深度卷积神经网络 (CharsSCNN) 对短文本进行情感分类取得了较好高的准确度[16]。Irsoy 等人提出了一个深度递归神经网络进行情感分析，性能相比以往 RNTN 等较为优异[17]。Chung 等人对 LSTM、GRU 模型进行了一些改进[18]。2015 年，Kai 将 LSTM 推广到树状网络拓扑结构 (Tree-LSTM) 进行情感分析[19]。2016 年，Lee 提出了一个基于递归神经网络和卷积神经网络的模型进行短文本序列分类[20]。Ghosh 等人将 LSTM 与传统机器学习方法相结合，有效提高了情感分类模型的性能[21]。

国内方面，2014 年，Tang 等人学习了 SSWE(sentiment-specfic 词向量) 模型用于

twitter 情感分类中[22]。Li 等人提出了自适应递归神经网络对 twitter 进行情感分类[23]。朱少杰提出了融合深度学习特征的半监督 RAE 方法进行情感分类[24]。2015 年，Tang 等人引入神经网络模型（ConvGRNN 和 LSTM-GRNN）用于文档级别的情感分类[25]。梁军等人提出了基于极性转移和 LSTM 递归神经网络的情感分析[26]。2016 年，刘艳梅提出了基于 SVM/RNN 的情感分类器[27]。2017 年，李丹将 LSTM 模型、NBSVM 模型以及 doc2vec 模型相融合进行了情感分析[28]。李松如改进了 RNN-Attention 模型，取得了较好的性能[29]。成璐提出了基于注意力机制的双向 LSTM 模型的情感分类应用[30]。2018 年，李松如等人提出了基于循环神经网络的情感词注意力模型进行情感分析，相比传统模型，性能有所提升[31]。

1.3 研究内容介绍

本文的主要研究内容和研究成果总结如下: (1) 本文的研究内容是循环神经网络在情感分析领域对电影评论的研究。主要结合时下性能较好的循环神经网络方法，进行分析比较并结合。本文要基于 word2vec 使用模型, 采用了词向量机制训练包含语义特征的词向量，避免了高维度输入易产生维度灾难的问题。同时，基于词向量训练出的词向量具有相似度的特征，具备了一定的语义信息，作为 Attention-Based IndRNN 模型的输入，对二值分类器的性能具有一定的提高效果。(2) 在文本语义提取问题上，基于时下较为优异的模型，本文设计了 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型，其中 IndRNN 相对传统 RNN 和 LSTM 来说，解决了梯度消失和梯度爆炸的问题。同时本文通过 Attention-Based 的方法，得到含有输入序列节点注意力概率分布的语义编码，并将其作为分类器的输入，减少了特征向量提取过程中的信息丢失和信息冗余，令关键词的作用凸显了出来。该方法可以得到包含注意力概率分布的语义特征，并对特征提取部分进行优化。

1.4 文本的组织结构

本文一共分为五个章节，主要内容如下所示:

第一章为绪论部分，该章主要介绍有关本文课题的研究背景及意义，在情感分析领域中的国内外研究现状，以及本文对中文文本情感分析主要采用的方法。

第二章为情感分析的相关技术介绍，主要介绍了本文的课题研究中需要了解到的一些基础知识，算法和一些具体的应用。主要包括了中文分词技术的介绍，以及在文本情感分析研究中的一些深度学习算法的极少。

第三章为词向量表示部分，主要介绍了词向量表示技术以及本文主要使用的神经网络预测模型。

第四章为本文的重点部分，详细地介绍了本文研究课题所采用的中文文本情感分析算法 Attention-Based IndRNN 对中文文本情感分析的具体过程。其中包括了使用 IndRNN 对中文文本进行情感分析的原因，传统的循环神经网络的不足之处。

第五章为对比实验部分，介绍了在收集的评论数据集上，例如豆瓣的电影评论和IMDB 评论数据集，进行对比实验得到的结果，并针对该结果做一些分析。

第六章为本文的总结与展望部分，主要讲述了本文在实验中遇到的一些问题和进一步需要改进的方向。

第二章　情感分析的相关技术综述

文本情感分析是时下比较常见的应用。本章将介绍情感分析主要运用到的一些相关技术。

情感分析的一般流程图如下:

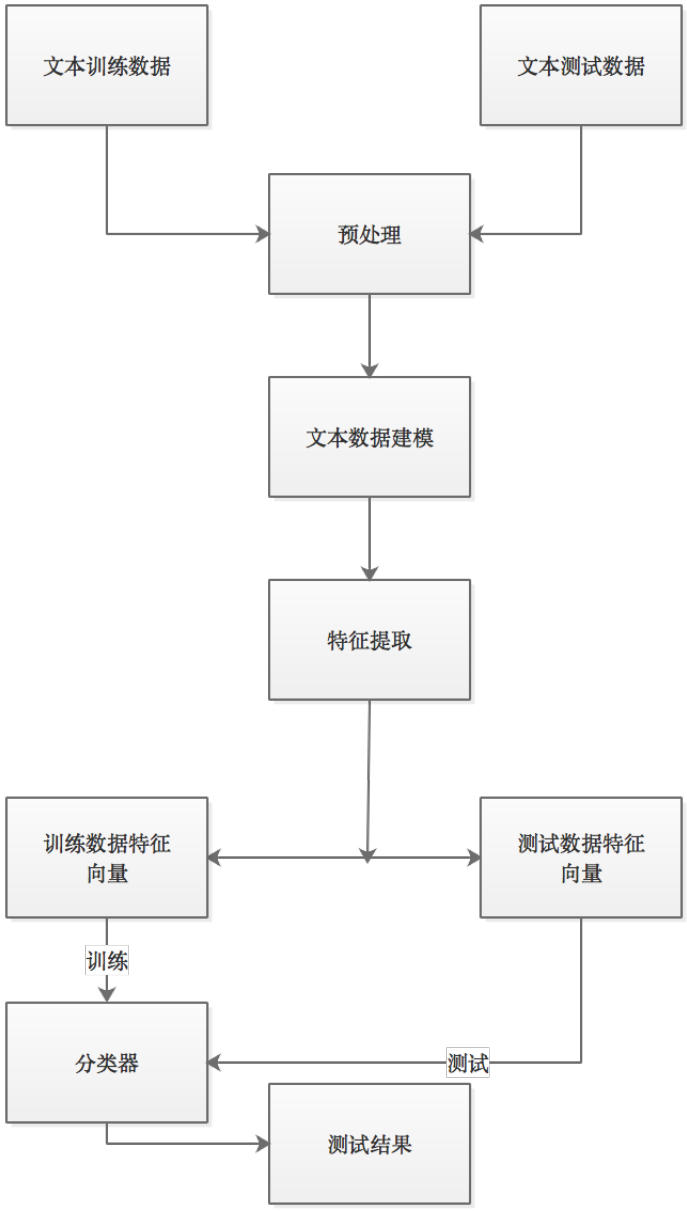


图 2-1　情感分析流程图

纵观中英文文本的情感分析的预处理过程，英文具有一种天然的分词结构，以空格为界可以很清晰的区分词义。而中文，天然以语句为单位，划分词义并不能直接从语句中得到。所以，相比英文的处理，中文文本在自然语言处理的过程中首先需要进行中文

分词。出于后续情感分析模型的训练中，语义的提取对词义的表示十分重视，因而中文分词的好坏对后续训练具有极大的影响。所以，在本章节中，将介绍中文文本的一些分词技术，为之后的情感分析打下一个好的基础。

主要有两种类型的情感分析方法，一类是基于情感词典的分析方法，一类是转化为文本分类问题来处理。

本文将中文文本情感分析转化为中文文本情感二值分类问题，之后得到情感倾向判断。

2.1 中文分词

对于英文语句而言，空格是天然的分隔符，单词之间非常自然的分词结构。而对于中文语句来说，词与词之间不存在这样的分隔符，只有能够比较明显的分隔的字、句、段。在中文文本中，有很多需要用多个字来表达情感的词与，无法直接分隔字来进行获取情感词。所以，首先需要对中文文本进行分词，从而才能用于训练模型。中文分词，是指将词按顺序逐个拆分，将一个完成连续的序列拆分成多个序列组合的过程。

中文分词技术主要有以下三类，基于词表的分词方法、基于统计语言模型的分词方法和基于序列标注的分词方法。

基于词表的分词方法主要由正向最大匹配法 (forward maximum matching method, FMM) 和逆向最大匹配法 (backward maximum matching method, BMM) 等组成。这种类型的技术也被称为字符串匹配分词算法。这个方法主要是将待匹配字符串按照某些策略与词表进行匹配，匹配到了就取出来。这种方法有效率高的优点，但对于歧义词、网络上的新词无法进行很好的划分。基于统计语言模型的分词方法主要有基于 N-gram 语言模型的分词方法。文本序列的划分存在多种结果，N-gram 模型是利用统计语言模型找出其中概率最大的一条。这种方法对于歧义词与网络新词的划分能得到比较好的结果。

基于序列标注的分词方法主要有基于隐马尔科夫模型 (Hidden Markov Model, HMM) 的分词方法、基于最大熵马尔科夫模型 (Maximum Entropy Markov Model, MEMM) 的分词方法、基于条件随机场 (Conditional Random Field, CRF) 模型的分词方法等。这类方法的基本思路是对汉字进行标注训练，结合词语出现频率与上下文，具有一定的学习能力。因此对歧义词与网络新词同样具有非常好的结果。

上述这些分词方法都有各自的优点，但也存在缺点。现在常用的分词器都是综合多种方法，使用机器学习方法与词典相结合进行分词。如 Jieba 分词器的词图扫描主要是藉由前缀词典来实现，生成一个包含句子中所有可能词组合而形成的的有向无环图 (DAG)，其最大概率路径主要使用动态规划算法来寻找，最后基于词频找出最佳划分组合。面对分词比较常见的新词问题，主要通过隐马尔科夫 (Hidden Markov Model，HMM) 模型，该模型由维特比算法 (Viterbi) 实现。该分词器由于性能优异，本文使用 Jieba 分词器来对文本进行分词处理。

2.2 使用机器学习算法的情感分析

在情感分析问题中，最重要的部分是通过某种机器学习算法来实现模型的训练和预测结果。一般是将情感分析作为分类问题，例如二值分类等等来处理。

分类问题，可以划分为三类方法，基于规则的分类模型、基于机器学习的分类模型和基于神经网络的方法。基于规则，可以使用例如决策树 (Decision Tree)、随机森林 (Random Forest)、RIPPER 算法等。基于机器学习，可以使用线性分类器 (逻辑回归)，支持向量机 (Support Vector Machine, SVM) 或者朴素贝叶斯分类器 (Naive bayes classfier, NB) 来处理。基于神经网络的方法，一般使用多层感知机 (Multiayer Perceptron, MLP), 卷积神经网络 (Convolutional Neural Network, CNN) 和循环神经网络 (Recurrent Neural Network, RNN) 等。

根据国内外现状，基于神经网络的方法在性能上较佳，因而本文主要采用了循环神经网络来进行训练。

2.2.1 RNN 模型

循环神经网络 (Recurrent Neural Networks, RNN) 是计算图包含有向周期的人工神经网络。与前馈神经网络不同，前向神经网络中信息严格按照层与层之间的一个方向流动，在循环神经网络（RNN）中，信息按照层与层之间的循环传播，以便模型的状态根据先前状态值的变化而变动。尽管前馈神经网络可以被认为是无状态的，但是 RNN 具有允许模型记忆关于其过去计算的信息的记忆器。这使得循环神经网络能够表现为动态时间行为和输入-输出对的模型序列。

因为它们可以构建出输入-输出对的时间序列模型，所以循环神经网络能够将信息持久化，能够提取序列数据中的特征，对于处理具有时间性的序列数据具有相当好的效果。详细地来说，RNN 会对过去的信息进行记忆并输入到当前单元的计算中，即隐藏层的输入包含上一时刻隐藏层的输出以及输入层的输出。以下是典型的 RNN：

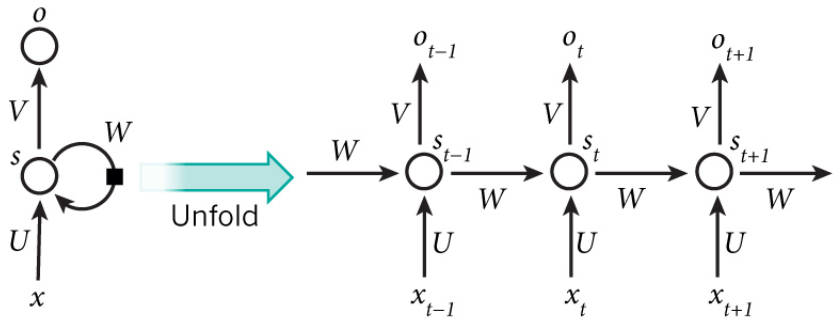


图 2-2 循环神经网络及其计算涉及的计算时间的展开

上图显示了一个正在展开（或展开）为完整网络的 RNN。通过展开，我们可以看到完整序列的网络。例如，如果我们关心的序列是 5 个单词的句子，则网络将展开成 5 层神经网络，每个单词一层。管理在 RNN 中发生的计算的公式如下：

首先，当前时间的输入由上一时间即 t-1 时间的隐藏层和当前时间即 t 时间的输入层一起计算而得。公式如下:

$$x(t) = w(t) + s(t-1)$$

其次，前面 t-1 时间内的所有词向量的时序信息都会被 t-1 时间的隐藏层记录下来，t 时间会读入当前时间的词向量 $w(t)$ 和 t-1 时间的隐藏层 $s(t-1)$，和转移矩阵 U 和 W 进行矩阵乘法，最后映射到 t 时间的隐藏层 s(t)。隐藏层的计算公式如下:

$$s_j(t) = f(\sum_t w_i(t) \times u_{ji} + \sum_j s_i(t-1) * w_{ji})$$

上式中，$f(z)$ 是隐藏层的激活函数。

最后，上述结果被传入输出层。输出层的大小与词表相同，词表中第 i 个词汇在语句中出现的概率由第 i 个节点输出的值表示。输出层计算公式如下:

$$y_k(t) = g(\sum_j s_j(t) \times v_{kj})$$

上式中，$g(z)$ 为输出层的激活函数。

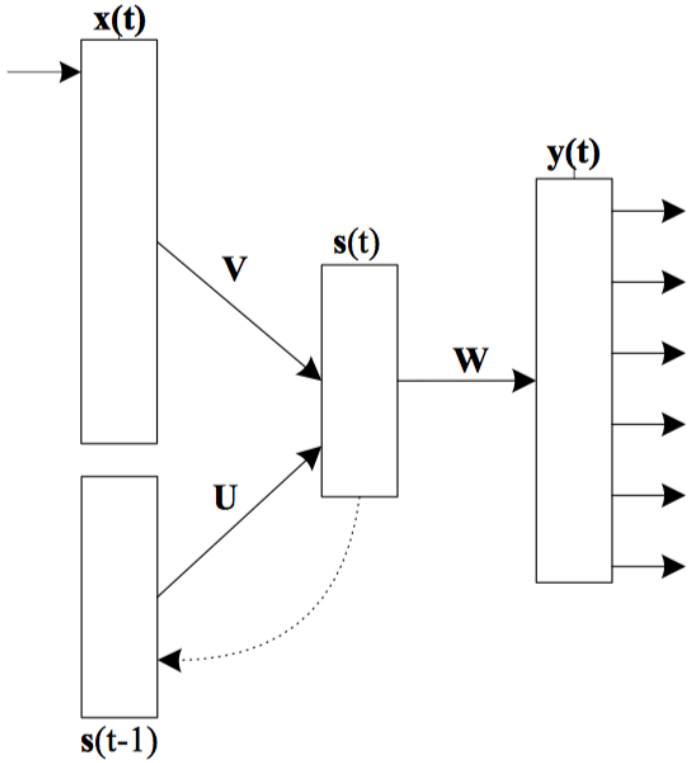


图 2-3 一个简单的循环神经网络

从上述循环神经网络的流程中，可以很清楚的看到可以优化的参数: 输入的词向量

$x(t)$, 隐藏层的向量 $s(t)$, 以及共享的转移矩阵 W, U 和 V。

RNN 模型的训练过程和传统神经网络模型训练过程是类似的，有一些基础名词需要了解。

2.2.2 LSTM 模型

RNN 能够运用历史信息来帮助当前的决策，但是也是存在局限性的。在实际实验中，训练模型时梯度会随着反向传播过程逐渐消失，只保持了较短范围内的序列信息。在某些需要较长的时间的信息来辅助角色时，RNN 的性能就显得有所不足了，在这样的背景下, 长短时记忆模型 (Long Short Term Memory, LSTM) 诞生了，引入了门 (Gate) 电路。

LSTM 通过输入门、遗忘门、输出门结构来控制循环神经网络中的各个时刻的状态。这里所说的三种门实际上是 sigmoid 神经网络和一个按位做乘法的操作，形象的来说，sigmoid 激活函数的全连接神经网络会输出 0 到 1 之间的数值，表示这个结构保留下来的信息。

LSTM 的隐藏层计算输出公式如下:

$$
\begin{aligned}
i_t &= \sigma\left(x_t U^i + h_{t-1} W^i\right) \\
f_t &= \sigma\left(x_t U^f + h_{t-1} W^f\right) \\
o_t &= \sigma\left(x_t U^o + h_{t-1} W^o\right) \\
\tilde{C}_t &= \tanh\left(x_t U^g + h_{t-1} W^g\right) \\
C_t &= \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right) \\
h_t &= \tanh(C_t) * o_t
\end{aligned}
\qquad \text{式（2-1）}
$$

这里，i, f, o 分别被称为输入，忘记和输出门。请注意，它们具有完全相同的方程，只是具有不同的参数矩阵（W, 是前一个隐藏层和当前隐藏层的循环连接，U 是将输入连接到当前隐藏层的权重矩阵）。他们关心所谓的门，因为 sigmoid 函数压缩这些向量在 0 和 1 之间的值，并且通过将它们与另一个向量相乘来确定“通过”的其他向量的数量。输入门定义了您想要通过的当前输入的新计算状态的多少。忘记门定义了你想要通过的前一个状态的多少。最后，输出门定义了你想暴露给外部网络的多少内部状态（更高层次和下一个时间步骤）。所有的门都有相同的尺寸 d_h，你的隐藏层状态的大小。

$\tilde{C}$ 是基于当前输入和先前隐藏状态计算的“候选”隐藏状态。C 是单位的内部记忆。它是前一个记忆乘以忘记门和新计算的隐藏状态乘以输入门的组合。因此，直观地说，它是我们想要如何组合先前的记忆和新的输入的组合。我们可以选择完全忽略旧的记忆（忘记全部 0 的门）或完全忽略新计算的状态（输入门全部为 0），但很可能我们需要在这两个极端之间进行一些操作。h_t 是输出隐藏状态，通过将记忆器与输出门相乘来计算。并非所有内部记忆器都可能与网络中其他设备使用的隐藏状态有关。

直觉上，简单的 RNN 可以被认为是 LSTM 的特例。如果将输入门全部固定为 1，将

所有 0 的忘记门（比如说，总是忘记以前的记忆）并将输出门设置为全 1（即暴露整个记忆），那么它基本就会获得传统的 RNN。

2.2.3 Bi-RNN 模型

单向的 RNN 结构能够通过前面时刻的输入来预测下一刻的输出，但也会有一些情况，预测的结果是由前面的输入与后面时刻的输入共同决定的。鉴于在这样的情况下，单向 RNN 的结构存在一些不足，所以提出了双向循环神经网络。单向 RNN 联系的是历史数据，那么双向 RNN 就成功联系了历史与未来。

从下图可以看到，双向 RNN 根据时间展开的结构，向前层和向后层共 2 个权值，即向前层和向后层的隐藏层与隐藏层之间的权值、向前层和向后层的隐藏层与输出层之间的权值，向前层和向后层一同连接输出层。

该模型的训练过程如下

- 前向传播
  1. 沿着时刻 1 到时刻 T 正向计算一遍，获取并保存每个时刻向前隐藏层的输出。
  2. 沿上述时间，反向计算一遍，获取并保存每个时刻向后隐藏层的输出。
  3. 正向和反向都计算完所有输入后，将根据每个时刻向前向后隐藏层获得最终输出。
- 反向传播
  1. 计算所有时刻输出层的 δ 项。
  2. 根据所有输出层的 δ 项，使用 BPTT 算法更新向前层。
  3. 根据所有输出层的 δ 项，使用 BPTT 算法更新向后层。

2.2.4 GRU 模型

门控循环单元 (Gated Recurrent Unit, GRU) 背后的原理与 LSTM 相类似，即用门控电路机制控制输入、记忆等信息从而实现预测, 它具有许多相同的属性，它将遗忘和输入门组合成一个“更新门”(Update Gate)。它还会合并单元格状态和隐藏状态，并进行一些其他更改。由此产生的模型比标准 LSTM 模型更简单，但其性能与序列建模的 LSTM 相当，但参数更少且更易于训练[18]。

4 个公式如下：

在时间 t，我们首先使用下述公式计算更新门 z_t

$$z = \sigma(x_t U^2 + s_{t-1} W^z)$$

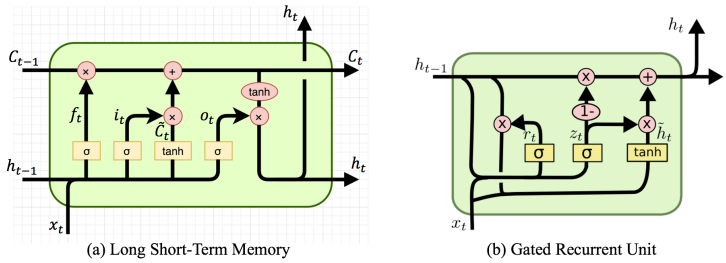


图 2-4 GRU 和 LSTM

其中 x_t 是第 t 个时间的输入向量，也就是，输入序列 X 的第 t 个分量，通过线性变换（乘以权重矩阵 $W(z)$）。h_{t-1} 保存前一个时间步 t-1 的信息，它同样也经过线性变换。更新门将这两条信息添加到 Sigmoid 激活函数中，并将激活结果压缩到 [0,1] 之间。

更新门帮助模型确定到底要将过去多少的信息传递给未来，或者需要传递前一时间步和当前时间步的信息。因为模型能决定复制过去的所有的信息以减少梯度消失的风险。

而复位门 (Reset Gate) 决定了过去的信息需要遗忘多少，公式见下:

$$r = \sigma(x_t U^r + s_{t-1} W^r)$$

该表达式与更新门的表达式相同，只是线性变换的参数和用处不同。

如前面更新门所述，h_{t-1} 和 x_t 先经过一个线性变换，再相加投入 Sigmoid 激活函数以输出激活值。

在复位门的使用中，新的记忆内容将使用复位门储存过去相关的信息，表达式见下:

$$h = tanh(x_t U^h + (s_{t-1} \circ r) W^h))$$

输入 x_t 与上一时间步 h_{t-1} 先经过一个线性变换，即分别右乘矩阵 W 和 U.

计算复位门 r_t 与 Uh_{t-1} 的 Hadamard 乘积，即 r_t 与 Uh_{t-1} 的对应元素乘积。由于之前计算的复位门是 0 到 1 的向量，它会控制门控大小。例如，如果某个元素的门控值为 0，那么它所包含的信息就完全被遗忘了。该 Hadamard 乘积将确定之前需要保留与遗忘的信息。两部分结果的和输入到双曲正切激活函数中。

在最后一步，网络需要计算 h_t，该向量将保留当前单元的信息并传递给下一个单元。在这个过程中，我们需要使用更新门，它决定了当前记忆内容 h'_t 和前一时间步 h_{t-1} 中需要收集的哪些信息。这一过程可以表示为:

$$s_t = (1 - z) \circ h + z \circ s_{t-1}$$

GRU 有两个门，一个复位门 (reset gate) 和一个更新门 (update gate)。直观地说，复位门会决定新输入如何与前面的记忆相合并，更新门定义了前面记忆保存到当前时间步的度量。如果我们将复位门设置为 1，更新门设置为 0，那样就会得到一个传统的 RNN

模型。

使用门控机制学习长期依赖关系的基本思路是和 LSTM 相似的，但仍然还存在一些关键的区别：

- 将输入门、遗忘门和输出门改变为两个门：更新门和复位门。
- 将单元状态 c_t 与输出合并为一个状态：h。
- LSTM 中的输入与遗忘门对应 GRU 的更新门，复位门直接作用于前面的隐藏状态。

2.3 本章小结

本章介绍了循环神经网络使用到的几个主要算法的原理及公式，梯度下降、随时间反向传播算法等。并介绍了循环神经网络的几个变体，LSTM 模型、结合了 LSTM 单元后的双向 RNN 即 Bi-LSTM 模型、GRU 模型。LSTM 相比传统的 RNN 模型优化了长期依赖问题，可以应对长度 500-1000 的文本。Bi-LSTM 在 LSTM 的基础上，保留了上下文信息。GRU 模型相比 LSTM 模型，结构上有所简化，运行时间上有所加速。

第三章　词向量技术

词向量技术是将文本转换为数学方式表达的一种模型。事实证明，许多机器学习算法和几乎所有深度学习体系结构都不能以原始形式处理字符串或纯文本。他们需要数字作为输入来执行任何类型的工作，无论是分类，回归等。在文本格式中存在大量数据的情况下，必须从中提取知识并构建应用。

文本表示主要分为基于独热 (one-hot Representation) 的表示方法与分布式表示 (Distributed Representation) 的表示方法。

词向量通常会将使用字典的单词映射到矢量，如“其/表现力/还算/不错”。如独热编码向量，其中 1 表示该单词存在的位置并且 0 代表其他地方。根据上述字典以这种格式，“不错”的矢量表示是 [0,0,0,1]，而“表现力”的是 [0,1,0,0]。

分布式表示方法现在主要分为基于矩阵 (Frequecny based Embedding) 的表示与基于神经网络（即 Prediction based Embedding) 的表示方法，词向量就属于基于神经网络的表示方法。另外，可以从上文的例子中看到，独热编码存在维度过大的问题，而词向量是将其压缩嵌入过后得到的，因而词向量也叫做词嵌入 (Word Embedding)。

基于矩阵的方法主要有 Count Vector、TF-IDF Vector、Co-Occurrence Vector 等方法。但是由于上述方法的性能上的局限性，在 Mikolov 提出 word2vec 后，基于神经网络的预测的方法展现了更佳的性能。不过，近来，Glove 也能提供语义等特征提取相差不多的词向量。

3.1　统计语言模型

词向量的上下文信息在神经网络模型诞生之前主要通过统计语言模型来得到。下文是基于统计的文本上下文表示的方法。

在数学术语中，如果文本中任意的词序列通过变量 W 来表示，则词序列通过按序排列的 n 个词语来表示 (即 $W = w_1 w_2 \cdots w_n$)，那么通过语言模型就可以得到该词序列 W 在这个文本中的概率 $P(W)$。由概率的乘积公式，可见如下所示:

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})$$

约束条件:

1. $P(w_1w_2\cdots w_n) \geq 0$
2. $\sum_{<x_1,x_2,\cdots x_n>\in V+} p(x_1, x_2, \cdots, w_n) = 1$

则 $P(w_1w_2\cdots w_n)$ 为 $(w_1w_2\cdots w_n)$ 的联合概率分布。常见的计算方法有 n-gram 模型方法、决策树 (Decision tree) 方法、最大熵模型 (Maximum Entropy Model, ME) 方法、最大熵马尔科夫模型 (Maximum Entropy Markov Model，MEMM) 方法、条件随机场 (Conditional random field，CRF) 方法等。

3.2 基于神经网络的分布式表示模型

基于神经网络的分布式表示模型现在主要有神经网络语言模型 (Neural Network Language Model ，NNLM)，基于语言模型的循环神经网络 (Recurrent Neural Network based Language Model，RNNLM),C&W 模型以及近来使用最多的 Mikolov 提出的 CBOW 和 Skip-gram 模型等。

word2vec 是 Google 发布的提供 CBOW 和 Skip-gram 模型的词向量训练工具。

word2vec 使用浅层神经网络来学习如何在特定的文本语料库中使用单词。word2vec 的输出是一个词向量，每个矢量（列）代表语料库中的一个单词，并提供该单词在上下文中的使用情况的数字描述。给定足够大的语料库，类似使用的两个词将具有相似的向量表示。例如，一个语料库可能会使用“学生”和“学生”这两个词来表示同一个事物，但从不在句子中一起使用它们。如果提供了足够的信息，word2vec 将能够学习这两个概念的上下文，并使它们的向量在数值上相似（即将它们放在语义空间中）。word2vec 的大多数应用使用余弦相似度来量化词/概念的接近度。word2vec 最适合于语料较大，局部一致且不含糊不清/隐喻的语料库。

3.2.1 CBOW 模型

连续词袋模型 (Continuous Bag of words，CBOW) 主要是给定了上下文预测某个单词出现的概率。

- W_i：单词表 V 中的第 i 个单词, one-hot 向量
- $v \in R^{n \square |V|}$: 输入词向量
- v_i:V 的第 i 列，n 维 W_i 的输入向量
- $U \in R^{|V| \square n}$: 输出词向量
- U_i: U 的第 i 行，n 维 W_i 的输出向量

目标函数使用交叉熵，用根据上文中提到的梯度下降法去更新每一个相关的词向量 U 和 V，根据下述算法就可以预测出中心词了。

算法 1 CBOW 模型

1: 生成输入的句子 $x^{(c)}$，采用 one-hot 编码 $(x^{(c-m)}, ..., x^{(c-1)}, x^{(c+1),...,x^{(c+m)}})$，上下文大小为 m。
2: 得到 embedded vectors,$v_{c-m} = Vx^{(c-m)}, ..., v_{c+m} = Vx^{(c+m)}$
3: 取平均 $\hat{v} = \frac{v_{c-m}+v_{c-m+1}+...+v_{c+m}}{2m} \in R^n$
4: 得到一个分数 vector，$z = U * \hat{v} \in R^{|V|}$
5: 将分数转化为概率 $\hat{y} = softmax(z)$

3.2.2 Skip-Gram 模型

Skip-gram 遵循与 CBOW 相同的拓扑结构。它只是翻转了 CBOW 的架构。Skip-gram 的目的是预测给定词的上下文。

目标函数使用交叉熵，用根据上文中提到的梯度下降法去更新每一个相关的词向量 U 和 V，根据下述算法就可以预测出上下文了。

算法 2 Skip-Gram 模型

1: 生成输入的词语 x，采用 one-hot 编码
2: 得到 embedded vectors,$v_c = V_x$
3: $\hat{v} = v_c$
4: 根据 $u = U * \hat{v} \in R^{|V|}$ 得到 2m 个分数 vector$u_{c-m}, ..., u_{c+m}$
5: $\hat{y} = softmax(u)$, 将分数转化为概率 $y^{c-m}, ..., y^{c+m}$

3.3 本章小结

生成词向量的方法中，分别有 CBOW 模型和 Skip-Gram 模型。在 CBOW 模型的结构中，模型从周围环境词的窗口中预测当前词。上下文词语的顺序不影响预测。在 Skip-gram 模型的结构中，模型使用当前词来预测环境词的周围窗口。skip-gram 体系结构权衡附近的上下文词语比更远的上下文词语更重。根据模型来看，CBOW 速度更快，而 skip-gram 生成速度较慢，但对于不常见的词语做得更好。均可以通过 Hierarchical Softmax 和 Negative sampling 策略来进行训练。

第四章　Attention-Based IndRNN 模型原理

本章主要讲述本文设计的情感分析模型:Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型。首先引入 Attention Model 思想，Attention Model 能够计算历史节点对当前节点的影响力权重，即注意力分配概率分布，有效地防止信息的丢失。基于 Attention Model 的 IndRNN 模型有效地解决了信息冗余、信息丢失等长期依赖问题，为文本分类提供优化之后的特征向量，提高分类器的性能。

4.1　IndRNN 模型

2018 年，独立循环神经网络由 Li 等人提出了[32]。在 IndRNN 中，循环输入用 Hadamard 乘积处理为

$$h_t = \sigma(W_{X_t} + u \odot h_{t-1} + b)$$

其中循环权重 u 是向量，$\odot$ 表示 Hadamard 乘积。每层中的神经元都是独立的，可以通过堆砌 2 层或更多层的 IndRNN 来实现神经元之间的连接。

对于第 n 个神经元，隐藏层的状态 h_n, t 可以由下式得出:

$$h_{n,t} = \sigma(W_n X_t + u_n h_{n,t-1} + b_n)$$

其中，w_n 和 u_n 分别是输入权重和循环权重的第 n 行。每个神经元仅在前一个时间步接收来自输入和它自己的隐藏状态的信息。也就是说，IndRNN 中的每个神经元独立地处理一种类型的时空模型。传统上，RNN 被视为时间上的、共享参数的多层感知器。与传统的 RNN 不同，IndRNN 模型为循环神经网络提供了新的视角，也就是说，随着时间的推移 (即通过 u) ，空间上逐渐集合 (即通过 w)。堆叠两层或多层可以加强不同神经元之间的相关性。在这种情况下，下一层的每个神经元处理上一层所有神经元的输出。

对于每层中随时间的梯度反向传播，可以为每个神经元独立计算 IndRNN 的梯度，因为在一个层中它们之间没有相互作用。对于忽略偏差 bias 的第 n 个神经元 $h_{n,t} = \sigma\square w_n x_t + u_n h_{n,t-1})$，假设在步骤 T 尝试最小化的目标是 J_n。然后梯度返回传播到时间步长 t

$$\begin{aligned}
\frac{\partial J_n}{\partial h_{n,t}} &= \frac{\partial J_n}{\partial h_{n,t}} \frac{\partial h_{n,T}}{\partial h_{n,t}} \\
&= \frac{\partial J_n}{\partial h_{n,t}} \prod_{k=t} T-1 \frac{\partial h_{n,k+1}}{\partial h_{n,k}} \\
&= \frac{\partial J_n}{\partial h_{n,t}} \prod_{k=t} T-1 \sigma\prime_{n,k+1} u_n \\
&= \frac{\partial J_n}{\partial h_{n,t}} u_n^{T-t} \prod_{k=t} T-1 \sigma\prime_{n,k+1}
\end{aligned} \qquad \text{式（4-1）}$$

其中 $\sigma\prime_{n,k+1}$ 元素激活函数的导数。可以看出，梯度只涉及标量值 u_n 的指数项，其

可以容易地调节，并且激活函数的梯度通常被限制在一定范围内。与 RNN 的梯度相比，IndRNN 的梯度直接取决于递归权重的值（其根据学习率以小幅度改变）而不是矩阵乘积（其主要由其特征值确定并且可以显着改变即使每个矩阵条目的变化很小）。因此，IndRNN 的训练比传统的 RNN 更强大。为了解决随时间变化的梯度爆炸和消失问题，我们只需要将指数项 $u_n^{T-t}\prod_{k=t}T-1\sigma\prime_{n,k+1}$ 调节到合适的范围。下面将进一步解释这一点，同时在 IndRNN 中保留长短记忆。

为了保持网络中的长期记忆，当前状态（在时间步骤 t）仍然能够在很长时间间隔后有效地影响未来状态（在时间步骤 T）。因此，时间步骤 T 处的梯度可以有效地传播到时间步长 t。通过假定最小有效梯度为 α，可以获得 IndRNN 神经元的重复性权重以保持长期记忆的范围。特别是，要保持记忆力 T-t 个时间步, $|u_n| \in [\sqrt[(T-t)]{\frac{\epsilon}{\prod_{k=t}T-1\sigma\prime_{n,k+1}}}, +\infty)$ 根据式（4-1 ）（忽略在时间步骤 T 从物 镜反向传播的梯度）。也就是说，为了避免神经元的梯度消失，应该满足上述约束条件。为了避免梯度爆炸问题，q 的范围需要进一步限制为

$$|u_n| \in [\sqrt[(T-t)]{\frac{\epsilon}{\prod_{k=t}T-1\sigma\prime_{n,k+1}}}, \sqrt[(T-t)]{\frac{\gamma}{\prod_{k=t}T-1\sigma\prime_{n,k+1}}}]$$

其中 γ 是最大的梯度值而不会爆炸。对于常用的激活函数，如 relu 和 tanh，它们的导数不大于 1, 即 $\sigma\prime_{n\square k+1}| \leq 1$。特别是对于 relu，其梯度为 0 或 1. 考虑到短期记忆对于网络性能也是重要的，特别是对于多层 RNN，对重复权重范围的约束随着 relu 活动功能可以放松到 $|u_n| \in [0, \sqrt[T-t]{\gamma}]$。当经常性权重为 0 时，神经元只使用当前输入的信息而不保留过去的记忆。这样，不同的神经元可以学习保持不同长度的记忆。请注意，关于轮回权重 u 的规定与梯度裁剪技术不同。对于梯度剪切或梯度范数剪裁，计算的梯度已经爆炸并被强制回到预定义的范围。以下步骤的渐变可能会继续爆炸。在这种情况下，依赖于该神经元的其他层的梯度可能不准确。相反，这里提出的规则基本上将梯度维持在适当的范围内，而不影响通过该神经元反向加速的梯度。

相比传统 RNN，这个模型有很多优点:

- 梯度消失和爆炸问题可以通过调节基于时间的梯度反向传播而解决
- 利用 IndRNN 可以保留长期记忆，处理长序列。根据文献中的实验表明，IndRNN 可以很好的处理的序列超过 5000 步，而 LSTM 处理 1000 步不到的序列。
- IndRNN 可以很好地利用 relu 等非饱和函数作为激活函数，并且训练结果具有非常好的鲁棒性。
- IndRNN 可以实现高效的多层堆叠以增加网络的深度，尤其是在层上具有残差连接的情况下。
- 由于各层神经元相互独立，很容易解释各层 IndRNN 神经元的行为。

4.2 Attention Model 思想

本节重点讲述自然语言处理领域的 Attention Model。在自然语言处理领域，Attention Model 经常结合 Encoder-Decoder 模型使用[33]。根据 Wang 等人在 2016 年的研究可以看到，引入了 Attention Model 后的 LSTM 情感分类模型性能相比对照基准有所提高[34]。本文通过介绍 Encoder-Decoder 的 Attention Model 机制，说明 Attention Model 的原理效果，并且引出 Attention Model 的其他应用方法。

以文本翻译为例，原始输入 $X = (x_1, x_2, \cdots, x_m)$，翻译到目标输出 $Y = (y_1, y_2, \cdots, y_n)$。

现在采用 Encoder-Decoder 架构模型，如下图

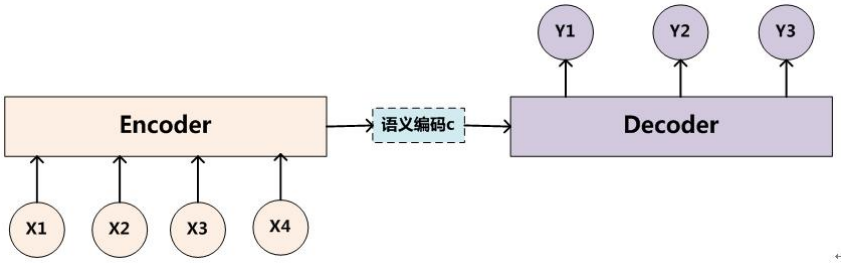


图 4-1 Encoder-Decoder 模型

Encoder 会利用整个原始句子生成一个语义向量，Decoder 再利用这个向量翻译成其它语言的句子。这样可以把握整个句子的意思、句法结构、性别信息等等。

Encoder 对 X 进行非线性变换得到中间语义向量 c:

$$c = G(x_1, x_2, \cdots, x_n)$$

Decoder 根据语义 c 和生成的历史单词 $(y_1, y_2, \cdots, y_{i-1})$ 来生成第 i 个单词 y_i:

$$y_i = f(c, y_1, y_2, \cdots, y_{i-1})$$

但是在生成目标句子 Y 的单词时，所有的单词 y_i 使用的语义编码 c 都是一样的。而语义编码 c 是由句子 X 的每个单词经过 Encoder 编码产生，也就是说每个 x_i 对所有 y_j 的影响力都是相同的，没有任何区别的。

为了克服普通的 Encoder-Decoder 结构的缺陷，一种新的模型被提出。其主要思想是选择编码过程中重要的部分来作为编码的输入。根据量化的思想，主要做法就是对编码阶段的每个输出增加一个权重（用 softmax 获得），而且这个权重是根据编码过程中的隐藏状态计算得到的，这样每一个序列都有一个不同的语义编码，由此决定序列中每一个部分的重要性。这样就可以很好的避免普通 Encoder-Decoder 模型的短板。

在注意机制之前，翻译依赖于阅读一个完整的句子并将所有信息压缩成一个固定长度的向量，就像你可以看到的那样，一个由数个单词表示的数百个单词的句子肯定会导致信息丢失，翻译不足等等。

然而，注意力部分解决了这个问题。它允许机器翻译器查看原始句子所保存的所有信息，然后根据当前处理的单词和上下文生成适当的单词。它甚至可以允许翻译器放大或缩小（侧重于本地或全局功能）。

结合了 Attention Model 之后，架构模型如下:

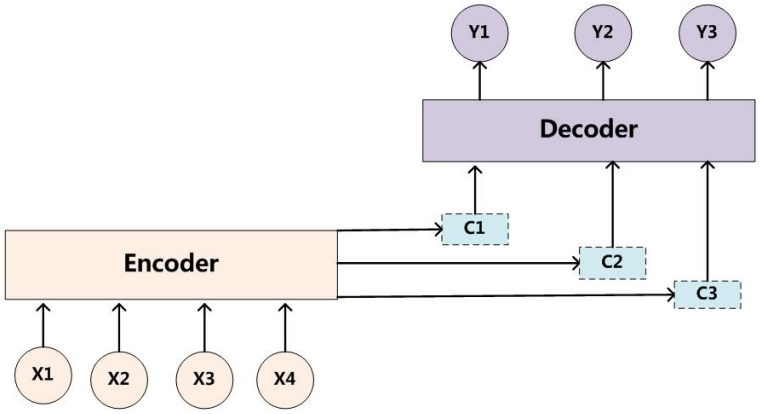


图 4-2 注意力模型

生成每个单词 y_i 时，都有各自的语义向量 C_i，不再是统一的 C。

$$y_i = f(C_i, y_1, \cdots, y_{i-1})$$

在编码阶段我们得到输入对应的状态值，然后计算得到状态值对于输出 y_i 的注意力概率分布，从而得到对应的语义向量 C_i，公式如下:

$$C_i = \sum_{j=1}^{T_x} a_{ij} \cdot h_j, \quad h_j = h(x_j)$$

其中 a_ij 是输入 x_i 对输出 y_i 的注意力概率。T 是输入的元素数量。

Attention Model 对于传统的模型具有相当明显的性能提升效果。Shang 等人对微博评论使用基于 Attention Model 的 Encoder-Decoder 模型进行文本翻译，混合使用 Global Attention 和 Local Attention，得到了不错的结果[35]。

4.3 Attention-Based IndRNN 模型

基于上文 Attention Model 的思想，本文提出了两个基于 Attention Model 的 IndRNN 改进模型。该模型主要引入了注意力机制，将注意力概率计算引入特征向量的权重计算中。

注意力概率分布的计算是 Attention Model 的关键，如下图:

上图中，atk 是节点 t 对于输出 k 的影响力权重，即注意力模型。不同的计算公式决定了 Attention Model 的不同作用，如 soft Attention、self Attention 等。

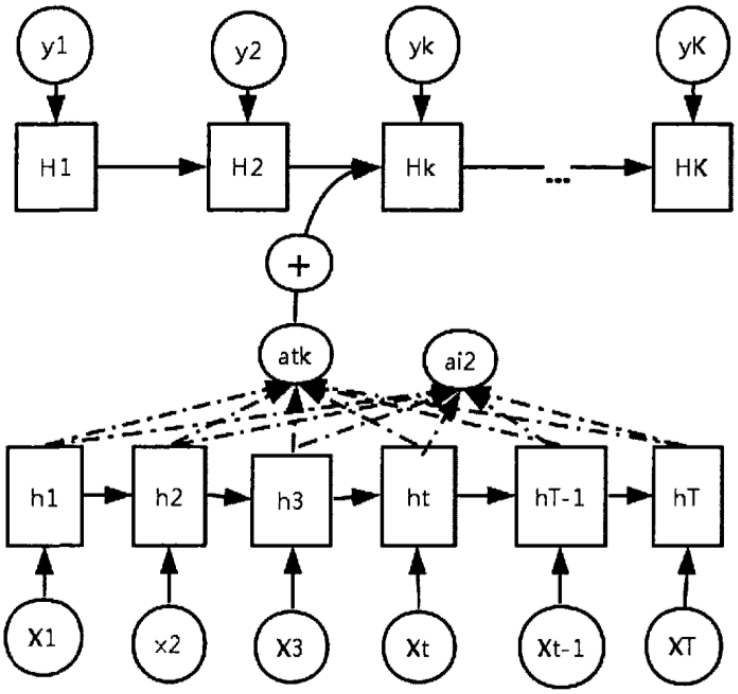


图 4-3 注意力概率分布计算方法

根据情感分析任务的要求，由序列输出为特征向量，因而在 Encoder 阶段引入 Attention 机制。

计算公式:

$$_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{N_x} exp(e_{ik})}$$

$$e_{ij} = V_a^T tanh(W_a \bullet s_{i-1} + U_a h_j)$$

其中，$W_a \in R^{n' \times n}$，$U_a \in R^{n' \times 2n}$

其中权重 α_{ij} 可理解为解码到第 i 个目标语言单词时，第 j 位置的源语言隐层状态对其的贡献值；亦可理解为目标语言中第 i 个目标语言单词的注意力或对齐到源语言第 j 位置的程度。若将整个解码过程中的 $_{ij}$ 输出，可以直观地看到上述的注意力信息。

4.4 情感分析器设计

本文的目标是情感分析问题，主要有文本的词向量表示、语义提取过程、二值分类器三个部分。本文根据文本表示的几种方法的比较择优，选择了近年使用较多性能较佳的词向量方法来通过词向量作为文本表示。语义提取阶段，使用本章节设计的 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型作为语义提取模型，在 Encoder 模型中使用 IndRNN 模型，并引入了 Attention Model，通过计算输入序列相对予以信息的

注意力权重，从而优化特征向量。二值分类器，这里主要使用 Softmax 回归模型作为分类器[36]。

情感分析器的结构如下图所示:

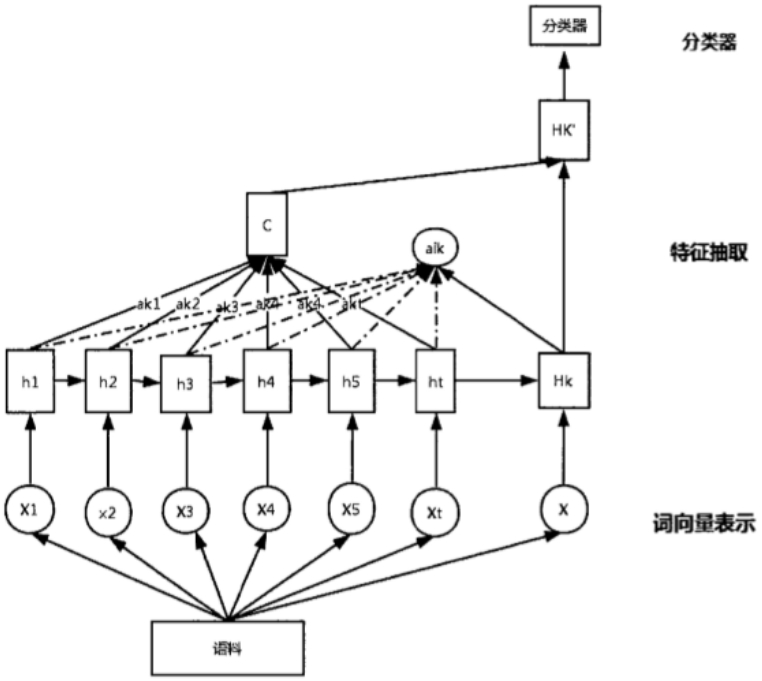


图 4-4 Attention-Based IndRNN 情感分析器的结构

(1) 词向量表示层: 基于 Word2vec 训练得到词向量模型，在词向量层由训练集数据集提取词向量。使用 Word2Vec 训练得到的词向量具有丰富的语义信息，如同义词的词向量具有相似性。实现的思路是，将数据集的单词通过模型训练为 K 维词向量，从而得到输入序列 $X = x_1, x_2, x_3, \cdots, x_T$ 其中 x_i 为 K 维词向量。整篇文本的 K 维向量输入 X' 通过对输入序列词向量进行累加求均值。

(2) 语义提取部分:Enocder 模型使用 Attention-Based IndRNN 模型。如上图所示，$h_1, h_2, h_3, \cdots, h_t$ 对应于 $x_1, x_2, x_3, \cdots, x_T$ 的隐藏层状态值。H_k 对应于输入 X' 的隐藏层状态值。

(3) 二值分类器部分: 采用 Softmax 回归模型构建二值分类器。输入向量是上述语义提取部分 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型输出的特征向量。

Attention-Based IndRNN 和 Attention-Based Bi-IndRNN 模型情感分析器的优点主要有下述几点:

(1) 通过词向量来实现词向量表示，可以有效减少向量维度，并且避免了特征稀疏问题。同时，词向量是基于了 Word2Vec 采用机器学习方法训练的语言模型，藉此得到的词向量包含丰富的语义信息，同义词与近义词的欧氏距离小，因而具有相似的词向量。这样得到的词向量输入到 Attention-Based IndRNN 模型，可以有效避免维度爆炸问

题，又由于丰富的语义信息，可以对 IndRNN 的输出结果进行优化。

(2) 在语义提取阶段，即 Encoder-Decoder 模型的 Enocder 阶段，使用 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型。IndRNN 有效地解决了 RNN 模型的长期依赖问题，避免了梯度爆炸和梯度消失问题。引入了 Attentin Model 思想后，得到了注意力分布的权重，对输出进行了一定的优化。本文提出了将 Attention Model 与 IndRNN 相融合的模型，可以有效提取出丰富的语义信息并有所突出。

(3)Attention-Based Bi-IndRNN 与 Attention-Based IndRNN 相比，多考虑了上文的信息，从中生成的语义向量具有更全面丰富的信息，所以理论上 Attention-Based Bi-IndRNN 的情感分析性能更好。

综上所述，Attention-Based IndRNN 情感分析模型和 Attention-Based Bi-IndRNN 模型在计算概率分布的同时加入了上下文信息，可以得到更加全面丰富的语义信息，。

4.5 本章小结

本章主要介绍了本文设计的情感分析器模型:Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型。Attention Model 的机制由于 IndRNN 生成的语义向量的局限性而被引入，并参考 Encoder-Decoder 模型的设计思想，通过得到输出关于历史信息的注意力概率分布，从而使性能得到提升。

第五章　对比实验与结果分析

5.1　实验设计

根据上述的理论基础，针对本文中用于对影评进行文本情感分析设计的 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型设计了对照实验。通过与 LSTM 模型，Bi-LSTM 模型，GRU 模型和 IndRNN 模型等对照的实验用于对比实验，来验证 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型对于情感分析的效果。情感分析的评价指标有多种，本实验通过准确率、F1 值来验证 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型在不同的语料库上的分析结果，并对结果进行分析。

5.1.1　实验环境

本文基于 Google 开源的机器学习框架 Tensorflow 上实现本文的实验设计。TensorFlow 是一个用于表达机器学习算法的接口，以及一个用于执行这些算法的实现，整合了目前比较流行的深度学习模型，诸如卷积神经网络 (CNN)，循环神经网络 (RNN)，以及 LSTM 模型等。张量 (Tensor)，解释如何创建，操作和访问张量，即 TensorFlow 中的基本对象。变量 (Variables)，详细说明如何在程序中表示共享的持久状态。图表和会话 (Graph and Sessions)，用来解释数据流图 (dataflow graphs)，TensorFlow 将计算表示为运算之间的依赖关系。会话 (sessions)，这是 TensorFlow 在一个或多个本地或远程设备上运行数据流图的机制。

本文具体的实验环境如下:

表 5-1　实验环境配置

操作系统	Ubuntu 14.04
开发语言	Python
开发平台	Tensorflow 深度学习框架
CPU	Intel 双核 2.5GHz
内存	6G
硬盘	100G SATA

5.1.2　实验数据集

本文训练词向量使用搜狗实验室新闻数据集，训练模型所用的主要数据集是豆瓣所爬取的电影短评以及一些公开的数据集。

5.1.2.1 词向量训练语料数据集

训练 word2vec 模型需要一个较好的提取语料特征的语料库，考虑到语料库的要求，使用了搜狗实验室提供的 2012 全网新闻数据 (1.54G) 和 Google News 数据集 (3.6G)。

(1) 中文语料库

由于 2012 全网新闻数据语料库为 XML 格式，并且其中文为 GBK 格式，而非通用的 UTF-8 格式，需要先进行转码及抓取指定元素内容的步骤，本步骤主要使用 iconv 及 grep 命令实现。

之后执行常规分词所要执行的预处理操作，首先过滤常见符号、数字、英文，之后使用 jieba 默认模型。其次，对训练集进行预处理操作。对训练集过滤非中文字符，使用 jieba 模型进行分词，再根据网上开放的一份较大的停用词表进行去除停用词。矩阵大小约为 572296*300。

(2) 英文语料库

由于 Google News 语料库过大，受实验环境限制，只使用 Glove 提取了部分词向量，矩阵大小约 400000*50。

5.1.2.2 豆瓣电影短评数据集

使用 python Requests 库对豆瓣电影进行爬虫，分析豆瓣网页请求，发现豆瓣分类网页中，后台发来 json response，前端从而进行解析显示。通过对该 url 发送请求来获取豆瓣电影的唯一标记 id 以及 url。

之后，根据电影 id 访问对应电影的短评首页，通过 BeautifulSoup 分析网页抓取短评所在元素以及该短评对应评分，从而根据评分分类 positive 和 negative 标签，将其按评分保存到各个 txt 中。将 10、20 评分作为 negative 数据集，40、50 评分作为 positive 数据集。

数据集获取完成后，进行清洗操作，对文本去重操作。去除重复文本，可以保障不会直接导致测试集与训练集的重复，而引起准确率过高。

抓取到的数据集为不平衡数据集，为了保障该数据集的准确率与后续实验的 IMDB Review 数据集具有一定的对照价值，因此对数据集进行了抽样，保障了 positive 与 negative 数据集的平衡性。

5.1.2.3 IMDB 电影评论数据集

互联网电影资料库 (Internet Movie Database, IMDB)，属于世界上最为详细的电影数据库。斯坦福处理过的该数据集在自然语言处理领域应用广泛，是进行文本相关实验的主要实验数据集[37]。

文本在实验中采用的 IMDB 数据集，训练集和测试集 positive 和 negative 各由 12500 条组成。

表 5-2 语料数据集信息

	训练样本	测试样本	平均文本长度
豆瓣短评数据集	8680 条/类	868 条/类	46.5
IMDB Review dataset	12500 条/类	1250 条/类	268

5.1.3 实验具体设计

为了检验本文设计的模型对于情感分析的效果，本文设计了几个实验:

- LSTM 模型
- Bi-LSTM 模型
- GRU 模型
- IndRNN 模型
- Attention-Based IndRNN
- Attention-Based Bi-IndRNN

上述模型参数设置如下: Embedding 层将训练集映射为 word2vec 模型维度的词向量。单层隐藏层设置为 64 个节点。损失使用 Adam 算法进行优化。情感分析器采用 softmax 分类器，接近每个 epoch 测试一次，batch 大小为 24。

5.2 实验结果和分析

本文根据上述实验模型在 2 个数据集上进行实验，经过多次训练调优，选取实验结果最好的数据。

表 5-3 实验结果

	豆瓣短评		IMDB Dataset	
模型	**准确率**	**F1 值**	**准确率**	**F1 值**
LSTM(3 layer)	88%	0.881	85.3%	0.845
GRU	86.5%	0.863	85.8%	0.858
Bi-LSTM	87.3%	0.875	85.1%	0.849
IndRNN(5 layer)	87.6%	0.877	84.5%	0.846
Attention-Based IndRNN	87.7%	0.876	85.8%	0.855
Attention-Based Bi-IndRNN	**88.2%**	**0.885**	**86.6%**	**0.863**

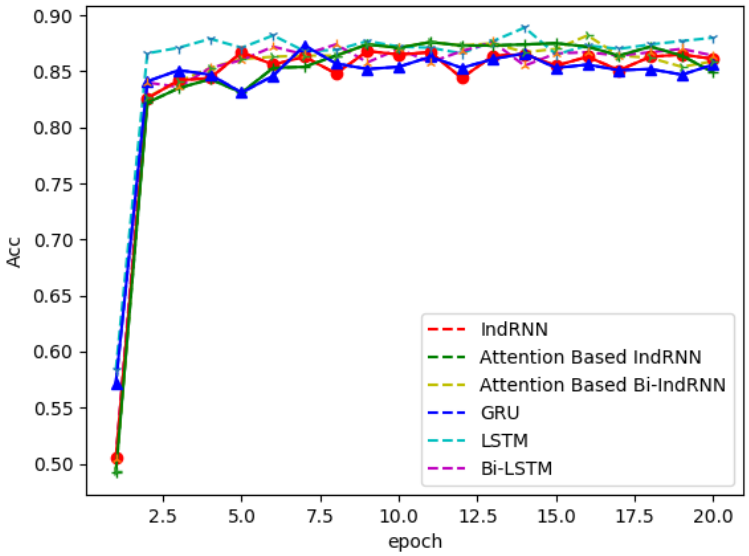


图 5-1 豆瓣影评数据集对照实验结果

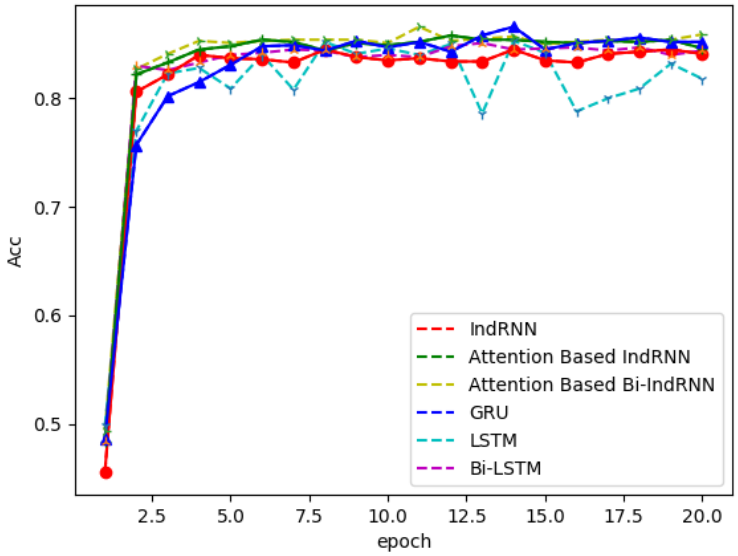


图 5-2 IMDB Review Dataset 对照实验结果

从表中的实验数据来看，实验的结果基本上可以说明本文设计的 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型能够有效地提高情感分析的效果。

下面对实验结果进行具体分析:

1. 在豆瓣数据集上，引入了注意力机制后的 IndRNN 相比对照组的实验准确率和 F1 值都较高，相比 IndRNN 准确率提高了 0.1%。在 IMDB 数据集上，引入了注意力机制的 IndRNN 相比对照组的实验准确率提高了 1.3%，F1 值提高了 0.009。从总体上来说，Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型在不同训练集上对于情感分析的准确率和 F1 值均有提高，证明了这两个模型的有效性。

2. 根据 Attention-Based IndRNN 和 Attention-Based Bi-IndRNN 模型在 IMDB 数据集和豆瓣短评数据集上效果来看，IMDB 准确率提高了 1.3%，而豆瓣数据集上仅提高了 0.1%，IMDB 上的提升效果稍好一些，这说明本文使用的注意力机制对于长文本情感分析的提升效果更好一些。

3. Attention-Based Bi-IndRNN 相比单向的 Attention-Based IndRNN 性能较好，结合了双向神经网络后，豆瓣数据集的准确率提高了 0.5%，F1 值也提高了 0.009，在 IMDB 数据集上准确率又提高了 0.8%，F1 值提高了 0.008。说明，双向神经网络对上下文信息的提取能力，能够有效提高情感分析的性能。

总的来说，本文设计的 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型对于情感分析任务的准确率具有一定的提高效果。Attention-Based Bi-IndRNN 对于情感分析任务性能最好。

5.3 本章小结

本章节主要介绍实验的设计以及组织过程，并对实验结果进行了分析。首先介绍了实验环境，对实验的平台 Tensorflow 进行了介绍。然后介绍了实验主要使用的 2 个语料库。随后给出了本文几个实验模型的设计方案和模型参数。最后根据实验结果从多个角度分析了本文设计的模型对于情感分析器的影响，从而验证了本文模型的有效性。

第六章　总结与展望

随着进入了大数据时代，曾经没有足够的数据来源来让我们分析，而现在如何从海量数据中获取到其中的价值点成为了重要的研究课题。本文针对中文文本情感二分类问题，在传统循环神经网络的基础上，实现了一种基于长短时记忆单元的循环神经网络语言模型。

6.1　论文主要工作

本文的主要工作是寻找性能较佳的循环神经网络的变体模型，使新的模型可以在影评文本情感分析上有更好的运行效果。

近年来，深度学习飞速发展，不少深度学习思想的实现使得在自然语言处理领域里的许多任务出现了更好的解决方法。

本文在情感分析器的三个部分，主要做了以下工作:

1. 在文本的词向量表示部分，本文采用了 word2vec 实现的 CBOW 模型进行词向量化，有效减少了向量维度从而避免了维度灾难的问题，并且避免了特征稀疏问题。这样得到的词向量包含丰富的语义信息，同义词与近义词具有相似的词向量，作为语义特征提取部分的输入，对情感分析的性能具有提高效果。

2. 在语义特征提取部分，即 Encoder-Decoder 模型的 Encoder 阶段，本文在原有 IndRNN 模型基础上，引入了一种注意力机制，设计了 Attention-Based IndRNN 模型。其中，IndRNN 模型相比 LSTM 模型更好地解决了长期依赖问题。并且，引入的注意力机制，得到了注意力分布的权重，对特征提取的输出进行了一定优化。该模型输出的编码，作为后续分类器的输入，有效地提取出了更丰富的语义信息并有所突出。

3. 为了进一步提高情感分析器的性能，本文中结合了双向 RNN 的结构，设计了 Attention-Based Bi-IndRNN 模型，多考虑了上文的信息，从而生成了更全面的语义向量，将其与单向 Attention-Based IndRNN 模型进行对照，从而得到上下文信息对于情感分析的影响。

综上所述，本文将 IndRNN 模型应用于情感分析，并引入了注意力机制，设计了 Attention-Based IndRNN 模型。考虑到文本上下文信息的输入对于输出的语义向量的效果，设计了 Attention-Based Bi-IndRNN。最后，对比了 Attention-Based IndRNN 模型和 Attention-Based Bi-IndRNN 模型与 LSTM 模型、GRU 模型、IndRNN 模型、Bi-LSTM 模型对在 2 个数据集上情感分析的性能，验证了提出的模型的有效性。

6.2 论文工作展望

- 在本文实验使用的近年提出的 IndRNN 模型之外，传统的 LSTM 模型还有很多变体，例如 Grid LSTM 等，同样也有很好的实验效果，下一步可以从这些变体中学习一些优点逐步引入。
- 本文的实验环境是单机，性能较为有限，下一步计划引入如 SRU 等提高传统 LSTM 模型运行时间的模型，将运行时间也纳入对照参数。
- 本文中主要使用的模型设计是 end to end 的深度学习模型，与传统的机器学习模型没有交集，考虑到近来不少论文中结合了传统模型后，在情感分析任务上性能有所提升，下一步可以研究传统的语义提取模型与本模型的融合对于情感分析性能的影响。
- 本文的情感分析仅处理了二分类问题，下一步可以考虑将中性纳入分类范围，进行多分类处理。

参考文献

[1] 沈尧．中国电影在线票务发展研究 [学位论文]. 北京：中国电影艺术研究中心，2016．

[2] 殷国鹏，刘雯雯，祝珊．网络社区在线评论有用性影响模型研究——基于信息采纳与社会网络视角 [J]. 图书情报工作. (16)．2012：140–147．

[3] Pang Bo，Lee Lillian．Opinion Mining and Sentiment Analysis [J]. Foundations and Trends® in Information Retrieval. 2 (1–2). 2008，July：1–135．

[4] Yi J, Nasukawa T, Bunescu R et al. Sentiment Analyzer: Extracting Sentiments about a given Topic Using Natural Language Processing Techniques [C]. In Third IEEE International Conference on Data Mining. 2003：427–434.

[5] 徐琳宏．基于语义资源的文本情感计算 [学位论文]. 大连：大连理工大学，2007．

[6] Pang Bo, Lee Lillian, Vaithyanathan Shivakumar. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques [C]. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. Stroudsburg, PA, USA. 2002：79–86.

[7] Bengio Yoshua, Ducharme Réjean, Vincent Pascal et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research. 3 (Feb). 2003: 1137–1155.

[8] Mikolov Tomas, Karafiat Martin, Burget Lukas et al. Recurrent Neural Network Based Language Model [J].: 4.

[9] Mikolov Tomas, Chen Kai, Corrado Greg et al. Efficient Estimation of Word Representations in Vector Space [J]. arXiv:1301.3781 [cs]. 2013, January.

[10] Mikolov Tomas, Sutskever Ilya, Chen Kai et al. Distributed Representations of Words and Phrases and Their Compositionality [M] // Burges C J C, Bottou L, Welling M et al. Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013: 2013: 3111–3119.

[11] Le Quoc，Mikolov Tomas．Distributed Representations of Sentences and Documents [J].： 9．

[12] Narayanan Vivek, Arora Ishan, Bhatia Arjun. Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model [C]. In Intelligent Data Engineering and Automated Learning – IDEAL 2013. 2013：194–201.

[13] Kim Yoon. Convolutional Neural Networks for Sentence Classification [J]. arXiv:1408.5882 [cs]. 2014, August.

[14] Socher Richard, Huval Brody, Manning Christopher D et al. Semantic Compositionality through Recursive Matrix-Vector Spaces [J].: 11.

[15] Socher Richard, Perelygin Alex, Wu Jean et al. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank [C]. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA. 2013：1631–1642.

[16] dos Santos Cicero, Gatti Maira. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts [C]. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland. 2014：69–78.

[17] Irsoy Ozan, Cardie Claire. Deep Recursive Neural Networks for Compositionality in Language [M] // Ghahramani Z, Welling M, Cortes C et al. Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2014: 2014: 2096–2104.

[18] Chung Junyoung, Gulcehre Caglar, Cho KyungHyun et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling [J]. arXiv:1412.3555 [cs]. 2014, December.

[19] Tai Kai Sheng, Socher Richard, Manning Christopher D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks [J]. arXiv:1503.00075 [cs]. 2015, February.

[20] Lee Ji Young, Dernoncourt Franck. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks [J]. arXiv:1603.03827 [cs, stat]. 2016, March.

[21] Ghosh Monalisa, Sanyal Goutam. Document Modeling with Hierarchical Deep Learning Approach for Sentiment Classification [C]. In Proceedings of the 2Nd International Conference on Digital Signal Processing. New York, NY, USA. 2018 : 181–185.

[22] Tang Duyu, Wei Furu, Yang Nan et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification [C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland. 2014 : 1555–1565.

[23] Dong Li, Wei Furu, Tan Chuanqi et al. Adaptive Recursive Neural Network for Target-Dependent Twitter Sentiment Classification [C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Baltimore, Maryland. 2014 : 49–54.

[24] 朱少杰．基于深度学习的文本情感分类研究 [学位论文]. 哈尔滨：哈尔滨工业大学，2014．

[25] Tang Duyu, Qin Bing, Liu Ting. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification [C]. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. 2015 : 1422–1432.

[26] 梁军，柴玉梅，原慧斌 等．基于极性转移和 LSTM 递归网络的情感分析 [J]. 中文信息学报. (05)．2015：152–159．

[27] 刘艳梅．深度学习技术下的中文微博情感的分析与研究 [J]. 软件. (05)．2016：22–24．

[28] 李丹．基于长短时记忆网络的中文文本情感分析 [学位论文]. 北京：北京邮电大学，2017．

[29] 李松如．基于循环神经网络的网络舆情文本情感分析技术研究 [学位论文]. 厦门：华侨大学，2017．

[30] 成璐．基于注意力机制的双向 LSTM 模型在中文商品评论情感分类中的研究 [J]. 软件工程. (11)．2017：4–6+3．

[31] 李松如，陈锻生．采用循环神经网络的情感分析注意力模型 [J]. 华侨大学学报 (自然科学版). (02)．2018：252–255．

[32] Li Shuai, Li Wanqing, Cook Chris et al. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN [J]. arXiv:1803.04831 [cs]. 2018, March.

[33] Vaswani Ashish, Shazeer Noam, Parmar Niki et al. Attention Is All You Need [J]. arXiv:1706.03762 [cs]. 2017, June.

[34] Wang Yequan, Huang Minlie, zhu xiaoyan et al. Attention-Based LSTM for Aspect-Level Sentiment Classification [C]. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas. 2016 : 606–615.

[35] Shang Lifeng, Lu Zhengdong, Li Hang. Neural Responding Machine for Short-Text Conversation [J]. ArXiv e-prints. 1503. 2015, March: arXiv:1503.02364.

[36] 张冲．基于 Attention-Based LSTM 模型的文本分类技术的研究 [学位论文]. 南京：南京大学，2016．

[37] Maas Andrew L, Daly Raymond E, Pham Peter T et al. Learning Word Vectors for Sentiment Analy-

sis [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. Stroudsburg, PA, USA. 2011 : 142–150.

致　　谢

在本篇本科学位论文完成之际，本人也将结束 4 年的本科生生活。在结束大学生活，走向工作岗位之际，衷心向帮助过我的老师、同学、朋友及家人表示感谢。

本论文使用基于 LaTeX 的本科生毕业设计模板书写，衷心感谢 Caspar Zhang 等人开源的这份模板。

外 文 原 文

Document Modeling with Hierarchical Deep Learning Approach for Sentiment Classification

Monalisa Ghosh
Research scholar: Dept. of CSE
National Institute of Technology, Durgapur, India
monalisa_05mca@yahoo.com

Goutam Sanyal
HOD: Dept. of CSE
National Institute of Technology, Durgapur, India
nitgsanyal@gmail.com

ABSTRACT
Sentiment analysis has recently been considered as most active research field in NLP domain. Deep learning is a growing trend of machine learning due to its automatic learning capability with impressive results across different NLP task. In this paper a model is proposed to analyze the deep sentiment representation based on CNN and LSTM (modified version of RNN) network. We aim to improve the performance of traditional machine learning method by merging them with deep learning techniques to tackle the challenge of sentiment prediction of massive amount of unsupervised product review dataset. We make our model first learn to sentence representation with CNN. Next, the semantics of sentences are encoded with LSTM network for document representation. We conduct experiments on two review datasets based on movie review with evaluation metric 'accuracy'. The result shows that proposed model outperformed traditional machine learning as well as baseline neural network model

CCS Concepts
• **Computing methodologies~Information extraction**

Keywords
Sentiment Analysis, Deep learning, traditional machine learning, Convolutional neural network (CNN), recurrent neural networks, LSTM, embedding algorithm etc.

1. INTRODUCTION
An opinion is a viewpoint or judgment about a specific thing and act as a key influence to an individual process of decision making. Opinion plays an important role in human being's life, because of People's beliefs and the choices they make are always depending on how others see and evaluate the world.

Sentiment analysis, also known as opinion mining is the process of determining the emotional tunes behind a series of words, in recent years, it has been receiving a lot of attention from the researchers. This field has many interrelated sub problems rather than a single problem to solve, which makes this field more challenging. Sentiment classification process can be done in three levels mainly: document level, sentence level and feature level.

In Document level, the entire document is classified based on the positive or negative opinion expressed by the authors.

Sentiment classification at the sentence level, considers individual sentence to identify whether the sentence is positive or negative. In feature level, classify the sentiment with respect to the specific aspects of entities. In this study, document level sentiment analysis has been taken into consideration.

Sentiment classification generally relies on two types of techniques, i.e., lexicon based and machine learning based techniques.

Machine learning approaches classify the sentiments based on training as well as test data sets. The lexicon based approach doesn't require any prior training data set for sentiment analysis. It uses large amount of linguistic resources with predefined list of words, where each word is associated with a specific sentiment. There are few researchers applied hybrid approaches by combining both approaches to machine learning and lexical to improve the sentiment classification performance. Deep learning has an extremity over the classical machine learning algorithms to perform the task of sentiment analysis, because of its capability to handle the challenges faced by sentiment analysis. In last few years, many researchers have tried to merge the concept of traditional machine learning with deep learning to develop more precise sentiment classifier to improve the classification task on any text content with minimal constraints.

In this paper, we approach a model based on two deep neural methods CNN and LSTM to learn continuous document representation for sentiment classification. The method is on the basis of the principle of compositionality. Specifically, the approach composes document modeling in two steps. In the first step, it uses convolutional neural network (CNN) to produce sentence representations from word representations. Next, LSTM applied for document representation from sentence level. Afterwards, max over pooling technique applied over the feature map and take the maximum value of features. Finally, the pooled features are used in a softmax layer for classification.

The remaining part of this paper is the following.

Section 2 presents the existing works which can relate to our approach. Section 3 discusses the basic Procedure of Sentiment Classification of review dataset. Section 4 addresses the background including LSTM networks and convolution operators. We then describe our architectures for sentence modeling and document modeling. In Section 5, and report experimental results in Section 6.

ICDSP 2018, February 25–27, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6402-7/18/02…$15.00
DOI: https://doi.org/10.1145/3193025.3193046

2. RELATED WORK

Neural network models based on deep learning have achieved great prosperity in many NLP related tasks. Document level Sentiment classification is a basic problem in sentiment analysis with the aims of identifying the sentiment label of a document. The first successful architecture of deep learning is based on word modeling with semantic vector space [1][2][3]. They introduced word embedding technique, where each word is represented as a real valued vector. The researcher [4] described a neural bag of words model that uses the dynamic k-max pooling operator. This model achieves good performance on sentiment classification task without feature engineering. Kim [5] applied simple convolutional neural networks with static vectors and got excellent results on different datasets.

Zhang et al.[6] in 2015 proposed a character-level CNN for text classification and achieve competitive results. Different types of LSTM models are applied in previous research. Specially, the model which encode the assumption and hypothesis separately with two LSTMs [8], a shared LSTM applied by Rockt ̈aschel et al[7].The tree structured LSTMs proposed by Tai et al. for sentiment classification. Few researchers combined the CNN and LSTM model for sentence classification [9]. The hierarchical structure of CNN and LSTM also applied by Tang et al.[10]. They first use of CNN or LSTM to produce sentence representations from word representations. Next, gated recurrent neural network is applied to encode semantics of sentences for document modeling.

Tai et al. (2015) [11] proposed to combine the standard LSTM to tree-structured topologies and got the superior results over a sequential model of LSTM. In [12], the researcher proposes three different models for multi-task learning with recurrent neural networks.

Vo et al.[13] introduced a Vietnamese corpus for sentiment classification, which collected from Vietnamese commercial web pages. They applied CNN and LSTM to generate information channels for Vietnamese sentiment analysis.

Wei et al. approached [14] a transfer learning framework based on a convolutional neural network and a long short-term memory model. This architecture automatically identify whether a post expresses confusion, determine the urgency of the post and finally polarity will be classified.

3. THE BASIC PROCEDURE

In our work we follow the basic procedure to perform sentiment classification by machine based learning method. The classification method summarized into several steps as described below.

A. The review dataset is preprocessed in such a way that supervised learning algorithm can be applied. Data processing is required to remove noisy, inconsistent and incomplete by considering tokenization, stop words removal, stemming method.

B. The features considered for this proposal are Unigram features and two-word (bi tagged) feature in POS based fix pattern. Composite feature set created by combining unigram and bi-tagged feature.

C. Tf-Idf method used to assign a particular score for each individual feature and top ranked features will be considered as input to supervised ML algorithm

D. Finally, train the supervised machine learning classifier SVM and NB with the different feature vector for classification the dataset. SVM and NB classifier considered as baseline method for evaluate our proposed model.

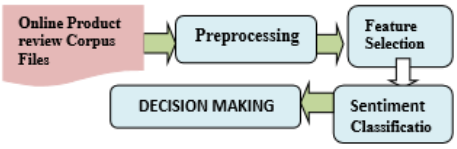


Figure 1. Architecture of the basic framework for sentiment classification using machine learning method.

4. PROPOSED MODEL

Long Short-Term Memory (LSTM) network performs surprisingly well in previous studies regarding sentiment classification problem. This is a modified version of the recurrent neural networks by handling the issue to model long range dependency. The key element of this deep learning network is memory cells, where the information can be stored. In our approach, to model the document semantic representations, we adopt Long Short-Term Memory (LSTM) network and convolutional neural network (CNN) through a hierarchical structure combined of word level and sentence level.

We consider CNN with multiple convolutional filters to catch local features [10] from every possible window of different size, where each filter considered as a feature detector. Max-pooling layer is used in our method to maintain the uniformity in the size of the sentence vector, because the output length of convolution layer based on the length of input sentence. The performance of this layer for classification task is better in comparison with simple averaging. We prefer the CNN and LSTM are very latest composition model for document level as well as sentence level classification [5][4][15]. In our approach, we don't require any external parser or parsing result to capture long distance dependencies for high quality sentence representation. In recent years, CNN based model [23] performs surprisingly well in wide range of NLP task such as sentence and document modeling for class prediction [4][5],object recognition [17],and other traditional NLP tasks [16] etc.

Let's discuss about how this combination of LSTM and CNN network use to learn document semantic representations. At first we define word vector representation. The generic word vector as a pre-trained can be extract [2] [1] with embedding learning algorithm (word2vec). Sometimes word vector captured in an unsupervised way with semantic and syntactic information. In our approach, we consider the word vectors that are not trained. In general, we assume a product p ε P contain a review from an user. This review can be represent as a document d with N number of sentences {s1,s2,…,.sN}.Then consider, the length of the sentence J-th is l_J. The J-th sentence sJ be formed of l_J words as $\{w_1^J, w_2^J, \ldots..w_{l_J}^J\}$, convolutional filters of CNN apply to extract local features from sentence sJ.

4.1 Word Level

Each word is represented as continuous and real valued vector and we embedded the word in a sentence into low dimensional space [21]. This process is known as word embedding [18]. In this way, each word w_i^J is mapped to its embedding representation $w_i^J \varepsilon R^d$ and the filter $F \varepsilon R^{d \times |L|}$, here d is the dimension of word vector and L is the size of the window.

LSTM network can produce hidden state representation [19]. For the given word w_1^J of sentence sJ, the current state and hidden state of the cell are respectively c_t^J and h_t^J can be updated with the previous cell state c_{t-1}^J and hidden h_{t-1}^J at time step ***t*** in the following manner.

$$i_t^J = \sigma\,(W_i\,.\,[\,h_{t-1}^J\,;\,w_t^J] + b_i) \quad (1)$$
$$f_t^J = \sigma\,(W_f\,.\,[\,h_{t-1}^J\,;\,w_t^J] + b_f) \quad (2)$$
$$o_t^J = \sigma\,(W_o\,.\,[\,h_{t-1}^J\,;\,w_t^J] + b_o) \quad (3)$$
$$\hat{c}_t^J = \tanh\,(W_c\,.\,[h_{t-1}^J\,;\,w_t^J] + b_c) \quad (4)$$
$$c_t^J = f_t^J\,\Theta\,c_{t-1}^J + i_i^J\,\Theta\,\hat{c}_t^J \quad (5)$$
$$h_t^J = o_t^J\,\Theta\,\tanh(c_t^J) \quad (6)$$

Where σ denotes the logistic sigmoid function and Θ represents the element-wise multiplication. Note that (i,f,o) are gate activation with the forget gate ft, the input gate it, and the output gate ot. These gates decide how to update the cell from one state to another state. $W_{\{i,f,o,c\}}$,$b_{\{i,f,o,c\}}$ are the set of parameters that we have to train.

4.2 Sentence Modeling

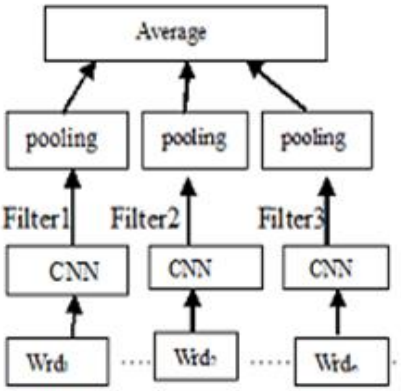


Figure 2. An example of sentence modeling

CNN composed of multiple layers with a number of shared parameters. Each layer performs a specific task of alternating its input into useful representation. We try CNN, which follows as the second layer to compose the sentence model. This convolution neural network consists of multiple convolutional filters of different width. In this work, we apply three convolutional filters of different width in order to fetch the local semantics of N-grams such as unigrams, bigrams and trigrams in a sentence. Suppose, a sentence as input of n words $[w_1,w_2,w_3\ldots,w_k,\ldots w_n]$. Let us consider p and b are shared parameters and l be the window size of a filter F. For CNN layer $w_i\ \varepsilon\ R^d$ be the embedding representation of word w_i with dimension d. In general, the input of a linear layer is the concatenation of word embeddings as $w_{i:i+l-1} = \{w_i, w_{i+1}, \ldots w_{i+l-1}\}\ \varepsilon\ R^{dl}$. Now we can determine the output of linear layer as follows

$$O = p.\,w_{i:i+l-1} + b \quad (7)$$

Where $p\ \varepsilon\ R^{l_o \times dl}$ and $b\ \varepsilon\ R^{l_o \times dl}$ are the shared parameters, l_o considered as output length of linear layer. Next, to extract the global features of a sentence, we feed the output of linear layer to the max-over-time pooling layer. A max-over-time pooling layer is added on top of the convolution neural network. Finally, the pooled features are used in a softmax layer for classification.

4.3. Document Modeling:

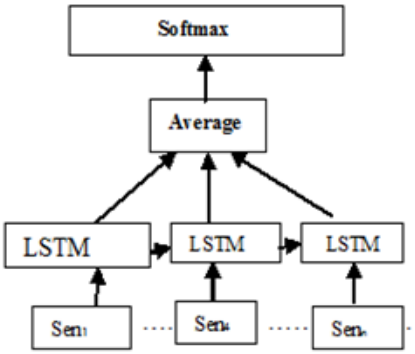


Figure 3. An example of document modeling

Our proposed architecture is not limited to sentence modeling, the obtained sentence vectors are provided to compose the document model. Let's consider the processed document as input to our model, where the document consists of n number of sentences $[s_1,s_2,s_3,\ldots,s_n]$. We apply LSTM for document composition in the same way as implemented for sentence modeling. The hidden state of LSTM network feed forward to the mean pooling layer. In this way, we get the mean value for the sentences.

CNN will be placed just top of LSTM like sentence modeling. Finally, the classification task can be completed with pool features by the max pooling and softmax layer.

5. EXPERIMENTS

5.1 Dataset: In this section, we conduct experiment to evaluate the effectiveness of our model on various benchmark data sets. Our main task is to perform sentiment classification on these different domain data set. Specifically, we use movie review and movie review datasets to evaluate the performance of our model and compare to existing base line model. The evaluation metric of these two datasets are *Accuracy.* We consider 80% of the data for training purpose, 10% for validation, and the remaining 10% for testing, unless stated otherwise.

SST2: Stanford Sentiment Treebank is an extension of movie review dataset. It is same as SST1, but this review dataset includes only binary labels.

IMDB The movie review data set contains 5,331 positive and 5,331 negative reviews, movie reviews [22]

5.2 Pre-processing:

Tokenization or segmentation: It can be done by splitting documents into a list of words. In our experiment we use the *Stanford Tokenizer* to obtain the tokens.

Removal of stop words: Some of the high frequency stop words are to be removed (prepositions, irrelevant words).

5.3 Experimental Settings:

We implement our model based on Python library. The hyper-parameter settings of the deep learning neural networks may depends on the data set being used for the experiment. For product review data set, we considered hyper parameters such as number of filters, filter length in CNN; memory dimension in LSTM; which layer to apply, etc. In our proposed architecture, we use one CNN layer and 1 LSTM layer. The filter size Or filter length 1,2,3 are used for single convolutional layer to captured the local features.

5.4 Baseline Methods

We compare our proposed model with the following baseline methods including traditional machine learning approaches such as SVM,NB etc. for document level sentiment classification.

5.4.1. Traditional paradigm:

- SVM+Ngrams: SVMs based methods are elaborate [10] with Unigram, Bigram and Unigram + Bigram as features.
- NB + Ngrams: Naïve Bayes classifier with Unigram features and Bigram features [4] also considered as baseline method.
- NBSVM-bi: SVM and Multinomial NB with bigram features are specified in [20] previous research.
- Text Features are defined based on [24], which includes word and character ngrams, sentiment lexicon features, cluster features etc.

5.4.2. Neural network Paradigm:

- CNN-rand, CNN-static, CNN-nonstatic and CNN-multichannel: These four variants of the CNN model proposed by [5] based on the different usage of word vectors for sentiment classification purpose.
- Standard-LSTM: Standard Long Short Term Memory Network by [11] considered as baseline to compare with our approach.

6. RESULT AND DISCUSSION

The experimental results of our proposed model on different datasets are shown in Table 1. This result focuses its effectiveness in comparison with other baseline method. We evaluate each dataset with the metric 'Accuracy' and the best method in each dataset will be marked in bold.

Table 1. Comparisons with baseline models on movie review dataset. Binary is a 2-classification task. The first block contains other baseline methods of traditional approach. The second blockare methods related to convolutional neural networks. The third block contains CNN method for word representation and Char representation. The fourth block contains methods using LSTM. The last block is our model.

Model	**SST-2**	IMDB
SVM+Ngram (Socher et al., 2013)	79.4	--
NB+Ngrams (Kalchbrenner et al., 2014)	80.5	--
NBSVM-bi (Wang and Manning, 2012)	---	91.2
SVM + text feature (Kiritchenko et al)	87.8	—
		—
CNN-rand (Kim, 2014)	82.7	—
CNN-static(Kim, 2014)	86.8	—
CNN-nonstatic (Kim, 2014)	87.2	—
CNN-multichannel(Kim, 2014)	85.1	—
		—
Standard-LSTM (Tai et al.,2015)	86.7	----
bi-LSTM (Tai et al., 2015)	86.8	---
SA-LSTM (Dai and Le, 2015)	88.1	92.8
LSTM+CNN (Our implementation)	86.7	88.9
CNN+Ngram (Our implementation)	89.7	93.2

From Table 1, we can see that the neural network approach (CNN, LSTM) outperform the traditional methods for different datasets. The neural network approach more effective in composing the semantic representation of text data. However, it must say that SVM classifier is extremely strong method with Ngram features than other baseline method in comparison.

CNN based approach achieve better accuracy when comparing the Convolutional Neural Networks and Recurrent neural network (RNN) to Recursive NNs using the movie review data set. This is because RNN suffers from the vanishing gradient and Gradient Explosion problems. The improved version of RNN such as Long Short Term Memory Models (LSTM) works surprisingly very well for sentence as well as document modeling.

7. CONCLUSION:

In this paper, we introduce neural network model () for document level sentiment classification. The approach is to first learns sentence representation with Long Short Term Memory network and after that the semantics of sentences are encoded with convolutional neural network for document representation. We conduct experiments on two review datasets based on movie review with evaluation metric 'accuracy'. The result shows that proposed model outperformed traditional machine learning as well as baseline neural network model with better accuracy on sentiment classification dataset.

8. REFERENCE:

[1] J. Pennington, R. Socher, and Christopher D Manning, "Glove: Global vectors for word rep-resentation" Proceedings of EMNLP, 12:1532–1543,2014.

[2] D. Bahdanau, K. Cho, and Y. Ben-gio. "Neural machine translation by jointly learning to align and translate". arXiv preprint arXiv:1409.0473, 2014.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S Cor-rado, and J. Dean. "Distributed representa-tions of words and phrases and their compositionality". In Proceedings of NIPS, pp. 3111–3119, 2013.

[4] N. Kalchbrenner, E. Grefenstette, and P. Blun-som. “A convolutional neural network for mod-elling sentences”. arXiv preprint arXiv:1404.2188,2014.

[5] Y. Kim. “Convolutional neural networks for sen-tence classification”. In Proceedings of EMNLP, pp. 1746–1751, Doha, Qatar, October,2014.

[6] X. Zhang, J. Zhao, and Y. LeCun. .”Character-level convolutional networks for text classification”. arXiv preprint arXiv:1509.01626,2015.

[7] T. Rockt ̈aschel, E. Grefenstette, K. Moritz Hermann, Tom´aˇs Koˇcisk`y, and P. Blunsom, ”Reasoning about entailment with neural attention”. In Proceedings of the 2016 ICLR, San Juan, Puerto Rico,2016.

[8] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning, “A large annotated corpus for learning natural language inference”. In Proceedings of the 2015 EMNLP, pp 22–32, Lisbon, 2015, Portugal.

[9] S Lai, L Xu, K Liu, and J Zhao, “Recurrent convolutional neural networks for text classification” In Twenty-Ninth AAAI Conference on Artificial Intelligence,2015.

[10] D. Tang, B. Qin, and T. Liu. “Document modeling with gated recurrent neural network for sentiment classification”. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432,2015

[11] K. Sheng Tai, R. Socher, and Christopher D. Manning. “Improved semantic representations from tree-structured long short-term memory networks”. In Proc. ACL.

[12] Liu et al., Pengfei Liu, Xipeng Qiu, Jifan Chen, and Xuanjing Huang. Deep Fusion LSTMs for Text Semantic Matching. In ACL, 2016.

[13] Quan Hoang Vo, Huy Tien Nguyen, Bac Le, Minh-Le Nguyen, “Multi-channel LSTM-CNN model for Vietnamese sentiment analysis” Proceedings of KSE October,2017.

[14] X. Wei, H. Lin, L. Yang and Y. Yu, “A Convolution-LSTM-Based Deep Neural Network for Cross-Domain MOOC Forum Post Classification”, Information **2017**, 8, 92; doi:10.3390/info8030092.

[15] Rie Johnson and Tong Zhang, “Effective use of word order for text categorization with convolutional neural networks”. NAACL,2015.

[16] Ronan Collobert and Jason Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In Proceedings of ICML, pp. 160–167. ACM,2008.

[17] Y. LeCun, L ́eon Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”. Proceedings of the IEEE, 86(11): pp.2278–2324.1998

[18] Yoshua Bengio, R ́ejean Ducharme, Pascal Vincent, and Christian Janvin. 2003. “A neural probabilistic language model. Journal of Machine Learning Research”,

[19] S. Hochreiter and J. Schmidhuber. “Long short-term memory. Neural computation”, 9(8):pp.1735–1780, 1997.

[20] Sida Wang and Christopher D Manning, “Baselines and bigrams: Simple, good sentiment and topic classification”. In Proceedings of ACL: Short Papers-Volume 2, pp. 90–94. Association for Computational Linguistics,2012.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Cor-rado, and Jeff Dean, “Distributed representations of words and phrases and their compositionality” In Proceedings of NIPS, pp. 3111–3119,2013.

[22] Peng, Long & Ding., “Feature selection based on mutual information: criteria of max-dependency”, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach.Intell.27:1226-1238,2005.

[23] Mingbo Ma, Liang Huang, Bing Xiang, and Bowen Zhou,. “Dependency-based convolutional neural networks for sentence embedding”. In Proceedings of ACL-IJCNLP, volume 2, pp. 692-700, 2015.

[24] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. “Detecting aspects and sentiment in customer reviews”. In Proceedings of the 8th Internation- Workshop on Semantic Evaluation (SemEval 2014) Nrccanada 2014, pp. 437–442, 2014.

外 文 译 文

基于分层深度学习方法的情感分类的文本建模

Monalisa Ghosh, Goutam Sanyal
National Institute of Technology, Durgapur, India

摘要

情感分析最近在自然语言处理领域，已经被认为是最活跃的研究领域。由于深度学习在多个不同的自然语言处理任务上具有非常可观的成果，因此已经被认为是机器学习的一个热门成长趋势。在这篇文章中，一个基于卷积神经网络 (CNN) 和长短时记忆 (LSTM) 网络 (循环神经网络 (RNN) 的修订版本) 的模型被推荐来分析深层的情感表达。我们计划期望依靠使用深度学习技巧，来合并这些网络，从而实现处理海量无监督产品评论数据集的情感预测的挑战，从而提高传统机器学习方法的表现能力。首先，我们使用 CNN 来让我们的模型学习句意表示。其次，使用 LSTM 网络对句子语义进行编码来实现文本表示。我们在两个电影数据集上进行实现，使用准确度来作为评估单位。结果表示推荐的模型相比传统机器学习模型和基本神经网络模型有所提高。

关键字

情感分析, 深度学习, 传统机器学习, 卷积神经网络 (CNN), 循环神经网络, LSTM, 嵌入算法等

引言

意见是关于具体事物的观点或判断，对个人决策过程具有十分重要的影响。意见在人类的生活中发挥着极为重要的作用，因为人们的信仰和选择取决于他人如何看待和评估世界。

情感分析也被称为意见挖掘，是确定一系列词语背后的情感曲调的过程，近年来它一直受到研究人员的大量关注。这个领域有许多相互关联的子问题，而不是一个要解决的问题，这使得这个领域更具挑战性。情绪分类过程主要可以分为三个层次：文档层次，句子层次和功能层次。

在文档层面，整个文档根据作者所表达的正面或负面意见进行分类。

在句子级别的情绪分类，考虑单独的句子来确定句子是积极的还是消极的。在特征层面，将关于实体特定方面的情绪分类。在这项研究中，文档级的情绪分析已经被考虑在内。

情感分类通常依赖于两种类型的技术，即基于词典的技术和基于机器学习的技术。

机器学习方法基于训练和测试数据集对情绪进行分类。基于词典的方法不需要任何提前训练的数据集来进行情感分析。它使用大量的语言资源和预定义的单词列表，每个单词都与特定的情感有所关联。很少有研究人员将机器学习的方法和基于词典的方法两种方法结合起来应用混合方法来提高情感分析的性能。深度学习相对于用于执行情感分

析任务的经典机器学习算法的性能有些优势，因为它能够处理情感分析所面临的挑战。在过去几年中，许多研究人员试图将传统机器学习的概念与深度学习相结合，从而开发更精确的情感分类器，通过最小的约束来改善对任何文本内容的分类任务。

在本文中，我们基于两种深度神经方法卷积神经网络 (Convolutional Neural Network,CNN) 和长短时程记忆网络 (Long Short Time Memory, LSTM) 来处理模型，从而学习情感分析的连续文档表示。该方法是基于组合性原则。具体来说，该方法组成文档建模主要分为两步。在第一步中，它使用卷积神经网络（CNN）从单词表示中产生句子表示。接下来，LSTM 从句子级别申请文档表示。之后，将 max over pooling 技术应用于特征映射并获取特征的最大值。最后，汇集的特征用于 softmax 层进行分类。

本文的其余部分如下。

第 2 部分介绍了与我们的方法相关的现有工作。第 3 节讨论审查数据集的情绪分析的基本程序。第 4 节介绍了包括 LSTM 网络和卷积算子的知识背景。然后我们描述我们的语句建模和文档建模架构。在第 5 节中根据实验设计进行实验，并在第 6 节中根据实验得到的结果进行分析并得到结论。

相关工作

基于深度学习的神经网络模型在许多 NLP 相关任务中取得了巨大的成功。文档级别情感分类是情感分析中的一个基本问题，其目的是识别文档的情感标签。深度学习的第一个成功的体系结构是基于带有语义向量空间的单词建模 [1] [2] [3]。他们引入了词嵌入技术，每个词都被表示为一个实值向量。研究人员 [4] 描述了一个使用动态 k-max 池操作符的词语模型神经袋。该模型在没有特征工程的情感分类任务上取得了良好的表现。Kim [5] 将静态向量应用于简单的卷积神经网络，并在不同的数据集上得到了很好的结果。

张等人 [6] 在 2015 年提出了用于文本分类的字符级 CNN 并取得竞争结果。在以前的研究中应用了不同类型的 LSTM 模型。特别是用两个 LSTM [8] 分别对假设和假设进行编码的模型，这是一个由 Rocktaschel 等人应用的共享 LSTM [7] .Tai 等人提出了树形结构的 LSTM。用于情感分类。很少有研究者将 CNN 和 LSTM 模型结合起来进行句子分类 [9]。Tang 等 [10] 也应用了 CNN 和 LSTM 的层次结构。他们首先使用 CNN 或 LSTM 从单词表示中产生句子表示。接下来，应用门控递归神经网络对文档建模语句进行语义编码。

Tai 等人（2015）[11] 提出将标准 LSTM 与树形拓扑相结合，并在 LSTM 的顺序模型上获得了优越的结果。在 [12] 中，研究人员提出了三种不同的模型，用于循环神经网络的多任务学习。

Vo 等人 [13] 为情绪引入了一个越南语料库分类，这是从越南商业网站收集的页面。他们申请 CNN 和 LSTM 来生成信息越南情绪分析渠道。魏等人。接近 [14] 基于卷积神经网络和长期短期记忆模型的转移学习框架。该体系结构自动识别帖子是否表达混乱，确定帖子的紧迫性并最终将极性分类。

基本理论

在我们的工作中，我们遵循基于机器学习方法进行情感分类的基本过程。分类方法总结为几个步骤，如下所述。

A. 审查数据集进行预处理，以便可以应用监督学习算法。数据处理需要通过考虑标记化，停用词删除，词干方法来消除噪音，不一致和不完整。

B. 本建议考虑的功能是基于 POS 的固定模式中的 Unigram 功能和双字（双标记）功能。通过结合 unigram 和 bitagged 功能创建的复合功能集。

C. 用于为每个单独特征和最高排序特征分配特定分数的 Tf-Idf 方法将被视为监督 ML 算法的输入。

D. 最后，训练具有不同特征向量的监督机器学习分类器 SVM 和 NB 以对数据集进行分类。SVM 和 NB 分类器被认为是评估我们提出的模型的基准方法。

推荐模型

长期短期记忆（LSTM）网络在以前关于情感分类问题的研究中表现出色。这是通过处理这个问题来模拟远距离依赖的递归神经网络的修改版本。这种深度学习网络的关键要素是记忆单元，信息可以记忆在其中。在我们的方法中，为了对文档语义表示进行建模，我们采用长短期记忆（LSTM）网络和卷积神经网络（CNN），通过结合词级和句级的分层结构。

卷积神经网络由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和池化层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构。与其他深度学习结构相比，卷积神经网络在图像和语音识别方面能够给出更好的结果。这一模型也可以使用反向传播算法进行训练。相比较其他深度、前馈神经网络，卷积神经网络需要考量的参数更少，使之成为一种颇具吸引力的深度学习结构。

卷积层是构建卷积神经网络的核心层，它产生了网络中大部分的计算量。卷积层的参数是有一些可学习的滤波器集合构成的。每个滤波器在空间上（宽度和高度）都比较小，但是深度和输入数据一致。举例来说，卷积神经网络第一层的一个典型的滤波器的尺寸可以是 5x5x3（宽高都是 5 像素，深度是 3 是因为图像应为颜色通道，所以有 3 的深度）。在前向传播的时候，让每个滤波器都在输入数据的宽度和高度上滑动（更精确地说是卷积），然后计算整个滤波器和输入数据任一处的内积。当滤波器沿着输入数据的宽度和高度滑过后，会生成一个 2 维的激活图（activation map），激活图给出了在每个空间位置处滤波器的反应。

LSTM 通过输入门、遗忘门、输出门结构来控制循环神经网络中的各个时刻的状态。这里所说的三种门实际上是 sigmoid 神经网络和一个按位做乘法的操作，形象的来说，sigmoid 激活函数的全连接神经网络会输出 0 到 1 之间的数值，表示这个结构保留下来的信息。

我们考虑有多个卷积滤波器的 CNN 从每个不同大小的窗口捕捉局部特征 [10]，其中每个滤波器被视为特征检测器。由于卷积层的输出长度基于输入句子的长度，因此我们的方法中使用最大共享层来保持句子向量大小的一致性。与简单平均相比，该层用于

分类任务的性能更好。我们更喜欢 CNN 和 LSTM 是文档级和句子级分类的最新组合模型 [5] [4] [15]。在我们的方法中，我们不需要任何外部解析器或解析结果来捕获高质量句子表示的长距离依赖关系。近年来，基于 CNN 的模型 [23] 在各种 NLP 任务中表现出色，例如用于类别预测 [4] [5]，对象识别 [17] 和其他传统 NLP 任务 [16] 的句子和文档建模等。

我们来讨论一下 LSTM 和 CNN 网络如何使用它来学习文档语义表示。首先我们定义词向量表示。作为预训练的通用词向量可以通过词嵌入学习算法（word2vec）来提取 [2] [1]。有时词汇向量以无监督的方式捕捉到语义和句法信息。在我们的方法中，我们考虑未经训练的词向量。一般来说，我们假设产品 p□P 包含来自用户的评论。该评论可以表示为具有 N 个句子 s1，s2，...，.sN 的文档 d。然后考虑，句子 J-th 的长度是 1J。第 J 个句子由 j 个词组成，如 J J J w 1，w 2，...，w l J，CNN 的卷积滤波器用于从句子 sJ 中提取局部特征。

词语等级

每个单词表示为连续的和实值的矢量，我们将这个单词嵌入到一个低维空间中 [21]。这个过程被称为词嵌入 [18]。以这种方式，每个词 w i 映射到其嵌入表示 J w iεR d 和滤波器 FεR d×| L | ，这里 d 是单词向量的维数，L 是窗口的大小。LSTM 网络可以产生隐藏状态表示 [19]。对于句子 sJ 的给定单词 □1，单元格的当前状态和隐藏状态分别为 □□ 和 □□ 可以用 □□ 前一个单元状态 □1 -1 和隐藏 □□-1 更新为时间步 t 按以下方式。其中 σ 表示逻辑斯蒂尔德函数，Θ 表示单元乘法。请注意，（i，f，o）是使用忘记门 ft，输入门和输出门 ot 激活门。这些门决定如何将单元从一个状态更新到另一个状态。W i，f，o，c，b i，f，o，c 是我们必须训练的一组参数。句子模型

根据 CNN 由多层共享参数组成。每个图层执行将其输入交替为有用表示的特定任务。我们尝试使用 CNN 作为第二层来组成句子模型。这个卷积神经网络由多个不同宽度的卷积滤波器组成。在这项工作中，我们应用三个不同宽度的卷积滤波器来获取 Ngrams 的本地语义，例如句子中的 unigrams，bigrams 和 trigrams。假设一个句子作为 n 个词 [w 1，w 2，w 3 ...，w k，... w n] 的输入。让我们考虑 p 和 b 是共享参数，l 是滤波器 F 的窗口大小。对于 CNN 层 w iεR d 是维数为 d 的单词 w i 的嵌入表示。通常，线性层的输入是字嵌入的连接，如 w i：i + l-1 = w i，w i + 1，... .w i + l-1εR dl。现在我们可以如下确定线性层的输出 O= p. w+b

其中，pε□□□× 和 bε□□□× 是共享参数，l o 被认为是线性层的输出长度。接下来，为了提取句子的全局特征，我们将线性图层的输出提供给最大时间池图层。在卷积神经网络的顶部添加最大随时间汇聚层。最后，汇集的特征用于 softmax 层进行分类。文本模型

我们提出的体系结构不限于句子建模，所提供的句子向量被提供来组成文档模型。让我们考虑处理后的文档作为我们模型的输入，其中文档由 n 个句子 [s 1，s 2，s 3，...，s n] 组成。我们将 LSTM 应用于文档组合，其方式与语句建模相同。LSTM 网络的隐藏状态反馈给平均共享层。这样，我们得到了句子的平均值。CNN 将被置于 LSTM 的顶

部，如句子建模。最后，分类任务可以通过最大池和 softmax 层的泳池特征来完成。

实验

数据集

在本节中，我们进行实验来评估我们模型在各种基准数据集上的有效性。我们的主要任务是对这些不同的域数据集进行情感分类。具体来说，我们使用电影评论和电影评论数据集来评估我们模型的性能，并与现有的基线模型进行比较。这两个数据集的评估指标是准确度。除非另有说明，我们认为 80％的数据用于培训目的，10％用于验证，其余 10％用于测试。SST2：Stanford Sentiment Treebank 是电影评论数据集的延伸。它与 SST1 相同，但该评论数据集仅包含二进制标签。

IMDB 电影评论数据集包含 5,331 个正面评论和 5,331 个负面评论，电影评论 [22]

预处理

标记或分段：可以通过将文档分成单词列表来完成。在我们的实验中，我们使用 Stanford Tokenizer 来获取令牌。

停用词的去除：某些高频停用词将被删除（介词，无关词）。实验设置

我们基于 Python 库实现我们的模型。深度学习神经网络的超参数设置可能取决于用于实验的数据集。对于产品评论数据集，我们考虑了超参数，例如滤波器数量，CNN 中的滤波器长度; 记忆维度在 LSTM; 要应用哪一层等等。在我们提出的架构中，我们使用一个 CNN 层和一个 LSTM 层。滤波器大小或滤波器长度 1,2,3 用于单个卷积层捕获局部特征。

对照方法

我们将我们提出的模型与以下基准方法进行比较，包括传统的机器学习方法，如 SVM，NB 等，用于文档级别情感分类。传统范式

- SVM + Ngrams：基于 SVM 的方法与 Unigram，Bigram 和 Unigram + Bigram 一样精心制作 [10]。
- NB + Ngrams：具有 Unigram 特征和 Bigram 特征的朴素贝叶斯分类器 [4] 也被认为是基线方法。
- NBSVM-bi：[20] 以前的研究指定了 SVM 和具有两个特征的多项式 NB。
- 文本特征是基于 [24] 定义的，其中包括单词和字符 ngrams，情感词汇特征，群集特征等。

神经网络范式

- CNN-rand，CNN-static，CNN-nonstatic 和 CNNmultichannel：文 [5] 提出的 CNN 模型的这四种变体基于用于情感分类的词向量的不同用法。
- 标准 LSTM：标准长期短期记忆网络 [11] 被认为是比较我们的方法的基线。

结果讨论

我们提出的模型在不同数据集上的实验结果如表 1 所示。这个结果集中了与其他基线方法相比的有效性。我们使用度量“精度”评估每个数据集，每个数据集中最好的方法将用粗体标出。

表 1. 与电影评论数据集上的基准模型的比较。二进制是一个 2 分类任务。第一个块包含传统方法的其他基线方法。与卷积神经网络有关的第二种阻塞方法。第三个块包含用于字表示和 Char 表示的 CNN 方法。第四个块包含使用 LSTM 的方法。最后一块是我们的模型。

从表 1 可以看出，对于不同的数据集，神经网络方法（CNN，LSTM）优于传统方法。神经网络方法在组成文本数据的语义表示方面更有效。但是，必须指出，SVM 分类器与 Ngram 特征相比是非常强大的方法，与其他基准方法相比较。

当使用电影评论数据集将卷积神经网络和递归神经网络（RNN）与递归 NN 进行比较时，基于 CNN 的方法实现更好的准确性。这是因为 RNN 患有梯度消失和梯度爆炸问题。长期短期记忆模型（LSTM）等改进版本的 RNN 在句子和文档建模方面的效果非常好。

结论

在本文中，我们引入神经网络模型（Neural Network）进行文档级情感分类。该方法是首先用长短期记忆网络学习句子表示，然后用卷积神经网络对文本表示语义进行编码。我们基于评估度量“准确度”的电影评论对两个评论数据集进行实验。结果表明，该模型比传统的机器学习和基准神经网络模型在情感分类数据集上具有更高的准确率。

北 京 邮 电 大 学

本科毕业设计（论文）开题报告

学院	计算机学院	专业	计算机科学与技术	班级	2014211307
学生姓名	王超	学号	2014211310	班内序号	07
指导教师姓名	刘晓鸿	所在单位	计算机学院	职称	副教授
设计（论文）题目	（中文）基于循环神经网络的影评情感分析系统的设计与实现				
	（英文） Design and implementation of the movie-review sentiment classification system based on RNN				

毕业设计（论文）开题报告内容：（主要包含选题的背景和意义；研究的基本内容和拟解决的主要问题；研究方法及措施；研究工作的步骤与进度；主要参考文献等项目）

1. 选题的背景和意义

信息传播的方式经历了语言符号、文字、电子传播等演变过程。21 世纪起，进入了互联网时代，越来越多的用户开始藉由互联网发表着自己的观点，传播个人的信息。任意一项商品或服务就会产生成千上万的评价信息，这就使得用户评论成为了一种越发重要的信息载体。

伴随着互联网的发展，电子商务的领域也逐渐扩散，电影与互联网产业也渐渐跨界融合。电影市场的下游环节，用户通过在线票务系统选择影片、影院并在观影结束后进行点评，这一系列的行为所产生的大量数据，能够为电影产业的上游环节提供重要思考。

在传统电影行业中是“我生产什么观众就看什么”，用户只能被动接受，可选择性小。而在互联网环境下，观众由产生观影意向到走进影院、完成观影，会经过“搜索关注、购票观看、评分评论”三个阶段，不管在哪一个阶段，用户的主体地位都已经被默许。[1]

2012 年，国内学者殷国鹏等[2]以信息采纳理论为研究框架，探讨了消费者在购买决策中采纳与接受在线评论信息的两类影响因素，即评论本身特征和评论者要素，并结合社会网络视角构建了在线评论有用性模型。社交媒体的电影评论对于购票者做出决产生了一定的影响。制片商也能够根据相关的影评来调整营销策略。

但是，面对如此庞大的评论文本信息，传统的通过人工获取评论的情感倾向是一件特别费时费力的事情，因此，如何利用自然语言处理领域的相关技术针对性地对评论文本的情感倾向进行自动化地挖掘与分析成为当今热门且很有意义的研究课题。

2. 研究的基本内容和拟解决的主要问题

2.1 研究的基本内容

主要内容：

1.搭建基于 Tensorflow 的深度学习网络平台，并获取影评数据集；

2.使用不同结构的深度网络模型，对数据集进行训练，得到相应的深度神经网络；

3.通过训练时间和计算时间及网络规模之类特性对比，得到合适的神经网络，并对实验中的结果进行分析解释；

4. 基于上述得到的神经网络实现影评情感分析软件。

2.2 拟解决的主要问题

1) 学习 word2vec 原理及应用，学习 CBOW 与 Skip-Gram 神经网络语言模型，以及 TF-IDF 模型，根据训练性能调整 word2vec 模型。
2) 学习 CNN、RNN、LSTM、GRU 等网络模型， 比对不同结构的深度网络模型，最终选择合适的神经网络
3) 学习过抽样、欠抽样、集成学习等方法，解决数据不均衡问题

3. 研究方法及措施

3.1 获得社交媒体上的电影评论数据；

在具体的研究过程中，通过网络爬虫分别采集了所需的评论数据，数据采集来源为豆瓣网（www.douban.com）。 豆瓣网作为国内最主要的以电影、图书及音乐作为纽带的社交网站，它不仅有关于国内外各种电影的用户评论，并且在用户群体中既有专业影评人，也有普通观影用户。 基于以上原因，本文选择豆瓣网影评作为研究对象，并对其评价数据进行实例分析。

3.2 搭建循环神经网络模型，对电影评论进行情感分析训练和建模；

寻找合适的语料库（如 jieba 语料库等）对数据集进行分词, 去除停用词，完成对数据集的预处理操作。

基于 word2vec 使用 CBOW 或 Skip-Gram 模型得到词向量。[3]考虑到可能 word2vec 无法区分文本中词汇的重要程度，可能进一步引入 TF-IDF 模型计算 word2vec 词向量的权重，提出加权 word2vec 模型。[4]或者通过构建情感要素词典（如 HowNet 的“情感分析用词语集”）捕捉含情感要素的词，通过构建词的情感特征向量来表示这些词的情感要素，然后与 word2vec 词向量进行特征融合，构成多元特征词向量。[5]

对 CNN 模型、传统 RNN 模型、LSTM 等模型进行性能比对，选择合适的神经网络。[6–8]

3.3 处理数据中可能存在的不均衡数据问题，改进得到的模型，提升模型分类准确率和召回率；

对文本分类的数据集中的不均衡问题处理的最终目标是在确保总体的分类准确度不降低或降低不大的前提下，提高少数类的分类精度。[9]

在数据不均衡情况下，模型性能评价指标只采用准确度(accuracy)是不恰当的。所以，使用如混淆矩阵、 F1 得分、ROC 曲线等更有效的评价作为评价标准。

对数据不均衡有下面几类处理方法:

3.3.1 数据集层面的方法

可以采用过抽样如 SMOTE 过采样算法、欠抽样策略等策略进行调整。[10]

3.3.2 算法层面的方法

可以通过改变概率密度、集成算法（Boosting、抽样和集成算法的融合）等方法进行调整。[10]

3.4 将模型应用到其他可能场景，证明模型有效性及其泛化能力。

4. 研究工作的步骤与进度

3 月 5 日---4 月 1 日：搭建环境，读相关文献；
4 月 2 日---4 月 29 日：进行算法设计，比较不同方法的结果，并进行编码和调试；
4 月 30 日---6 月 3 日： 通过不同结果对比，优化设计实现；
6 月 5 日---6 月 22 日：撰写论文及答辩。

5. 主要参考文献

[1] 沈尧. 中国电影在线票务发展研究[D]. 中国电影艺术研究中心, 2016.
[2] 殷国鹏, 刘雯雯, 祝珊. 网络社区在线评论有用性影响模型研究——基于信息采纳与社会网络视角[J]. 图书情报工作, 2012(16): 140–147.
[3] ZHANG L, WANG S, LIU B. Deep Learning for Sentiment Analysis : A Survey[J]. arXiv:1801.07883 [cs, stat], 2018.
[4] 张谦, 高章敏, 刘嘉勇. 基于 Word2vec 的微博短文本分类研究[J]. 信息网络安全, 2017(01): 57–62.
[5] 李科. 基于多元特征融合和 LSTM 神经网络的中文评论情感分析[D]. 太原理工大学, 2017.
[6] KIM Y. Convolutional Neural Networks for Sentence Classification[J]. arXiv:1408.5882 [cs], 2014.
[7] WANG J, YU L-C, LAI K 等. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model[C]//2016: 225–230.
[8] 李丹. 基于长短时记忆网络的中文文本情感分析[D]. 北京邮电大学, 2017.
[9] 谢娜娜. 基于不均衡数据集的文本分类算法研究[D]. 重庆大学, 2013.
[10] 陶新民, 郝思媛, 张冬雪等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报(自然科学版), 2013(01): 101–110+121.

指导教师签字		日期	年 月 日

注：可根据开题报告的长度加页。

北 京 邮 电 大 学

本科毕业设计（论文）中期进展情况检查表

学院	计算机学院	专业	计算机科学与技术	班级	2014211307
学生姓名	王超	学号	2014211310	班内序号	07
指导教师姓名	刘晓鸿	所在单位	计算机学院	职称	副教授
设计（论文）题目	（中文）基于循环神经网络的影评情感分析系统的设计与实现				
	（英文）Design and implementation of the movie-review sentiment classification system based on RNN				

主要内容: (毕业设计（论文）进展情况，字数一般不少于 1000 字)

一、豆瓣短评爬虫

使用 python Requests 库对豆瓣电影进行爬虫，分析豆瓣网页请求，发现豆瓣分类网页中，后台发来 json response，前端从而进行解析显示。通过对该 url 发送请求来获取豆瓣电影的唯一标记 id 以及 url。

之后，根据电影 id 访问对应电影的短评首页，通过 BeautifulSoup 分析网页抓取短评所在元素以及该短评对应评分，从而根据评分分类 positive 和 negative 标签，将其按评分保存到各个 txt 中。

二、预处理

训练 word2vec 模型需要一个较好的提取语料特征的语料库，考虑到语料库的要求，使用了搜狗实验室提供的 2012 全网新闻数据(1.54G)。

由于该语料库为 XML 格式，并且其中文为 GBK 格式，而非通用的 UTF-8 格式，需要先进行转码及抓取指定元素内容的步骤，本步骤主要使用 iconv 及 grep 命令实现。

之后执行常规分词所要执行的预处理操作，首先过滤常见符号、数字、英文，之后使用 jieba 默认模型进行分词，之后使用该分词后语料库训练 Word2Vec 模型，并将其 vocabulary 及其对应关系模型保存为 numpy array，便于后续处理。得到的 Word2Vec 模型形状为 572297*300。

其次，对训练集进行预处理操作。对训练集过滤非中文字符，使用 jieba 模型进行分词，再根据网上开放的一份较大的停用词表进行去除停用词。

三、LSTM 模型对照实验

RNN 模型能够将信息持久化，能够处理序列数据。但对于长序列数据，会出现梯度下降、梯度爆炸等问题，导致无法得到优良的结果。而 LSTM 的提出解决了 RNN 在长期依赖方面的不足。本次实验主要设置了 LSTM 与传统 RNN 进行比照。

目前已完成任务

在设置循环神经网络模型参数前，为了确定神经网络步长，读取训练集每篇语料词长，以及频次，使用 matplotlib 可视化得知语料内容长度满足正态分布，绝大部分语料长度在 50 以内，因而确定步长为 50。超过 50 的则丢弃，不足则以 0 补齐。

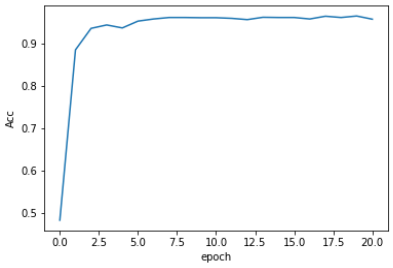
实验中设置的 LSTM 主要为 4 层，输入层、隐藏层、dropout 层和输出层。输入层，即 Embedding 层，设定输入参数句长为 50、向量维度（400）等。第二层，隐藏层，设定 LSTM 单元数为 64。第三层为 dropout 层，减少过拟合。第四层为输出层，即 softmax 层。

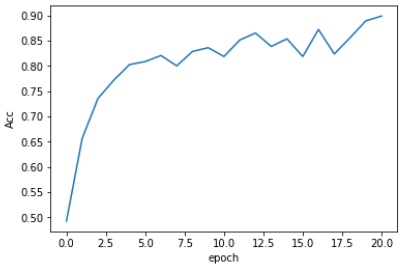
输入，通过 word2vec 模型提取特征矩阵 idx，将语料数值化，每条语料得到一个 50*400 的矩阵。

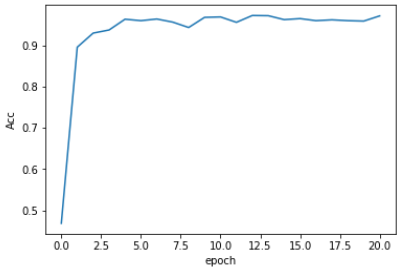
训练集经过 shuffle 打乱，防止模型不稳定，损失函数跳跃严重。

输入向量经过 embedding，馈送到 dynamic_rnn。之后主要采用 tensorflow 框架的 lstm 单元来完成循环神经网络模型的训练。

传统 RNN 模型和 GRU 模型参数设置与上面 LSTM 模型基本相同，将使用的 tensorflow 的 lstm 单元更换成了 rnn 单元，稍稍根据输出矩阵形状调整矩阵，主体与 LSTM 相同，仅设置了一层隐藏层。







四、结果分析

使用 matplotlib 可视化测试集的 accuracy 指标后可以看到，上左图是 LSTM 模型，右图是传统 RNN 模型，下左图是 GRU 模型。在设置了均衡训练集的情况下，三者训练时间相近，在经过 20 轮训练后已经趋于稳定。单层 RNN 的准确率在 81%，2 层的 RNN 准确率反而下降到了 76%，而单层 LSTM 的准确度有 96%，单层 GRU 的准确度高达 97%。可以看到，LSTM 和 GRU 相对传统 RNN 具有极大优势。根据文献得知，GRU 和 LSTM 性能在很多任务上不分伯仲。因此，在长序列文本的情感分析的问题上，将主要基于 LSTM 或 GRU 模型进行优化训练。

五、应用

搭建了一个根据输入评论内容可以返回情感分析结果的网页。前端提交输入的内容后，

<table>
<tr><td></td><td colspan="3">后端调用神经网络模型得出预测值，从而返回情感分析结果。技术上主要使用 python Flask Web 框架构建一个 wsgi 服务器，渲染引擎主要使用 Jinja2 后端渲染。不过由于模型较大，加载模型时间约 5 分钟，单次提交评论测试，响应时间略久，达到了近 10 秒之多。</td></tr>
<tr><td></td><td colspan="3">是否符合任务书要求进度</td></tr>
<tr><td rowspan="2">尚需完成的任务</td><td colspan="3">添加 ROC、F1 等评估标准对不平衡数据测试进行评估。
增加对照模型如 GRU、IndRNN，从而进行优化
完成毕业论文，并准备答辩</td></tr>
<tr><td colspan="3">能否按期完成设计（论文）</td></tr>
<tr><td rowspan="2">存在问题和解决办法</td><td>存在问题</td><td colspan="2">损失函数过拟合问题</td></tr>
<tr><td>拟采取的办法</td><td colspan="2">1. 调整模型参数
2. 增加训练数据
3. 添加 dropout 层</td></tr>
<tr><td>指导教师签字</td><td></td><td>日期</td><td>年 月 日</td></tr>
<tr><td>检查小组意见</td><td colspan="3">负责人签字： 年 月 日</td></tr>
</table>

注：可根据长度加页。