北京林业大学

大学生创新创业训练计划

创新训练项目任务书

项目名称：	基于 CNN 与 LSTM 的文本情感分析算法研究		
主 持 人：	陈飞阳	**专业年级**	计算机科学与技术（创新实验班）2016级
联系电话：	17801114068		
电子邮箱：	chenfeiyang@bjfu.edu.cn		
指导教师：	**崔晓晖**	**学　　院**	**信息学院**
申请日期：	2018 年 03 月 27 日		

北京林业大学

一、项目基本情况

项目名称	基于 CNN 与 LSTM 的文本情感分析算法研究			
项目级别	国家级		项目类型	创新训练
学校资助经费	2.0（万元）		项目编号	201810022064
主 持 人	姓名	联系方式	学号	专业班级
	陈飞阳	17801114068	161002109	计创 16 班
参 加 人	王超群	15652630259	161002101	计算机科学与技术 16-1
	陈楠	17610602157	161002107	计创 16 班
	史金明	13161303799	161002112	计创 16 班
指导教师	姓名	联系方式	职称	所在学院
	崔晓晖	17801043500	讲师	信息学院
项目阶段检查时间	一年半期项目：2018 年 9 月、2019 年 3 月			
项目结题时间	□ 2019 年 9 月（一年半期项目）			

二、预期成果

本项目预期将有以下成果：

1. 发表 SCI 或 EI 检索论文一篇；

2. 嵌入神经网络模型新算法的文本情感分析网站 demo。

三、项目经费预算

支出科目	预算金额（元）	具体支出内容	预算编制说明
合计	20000		
1.实验材料费	3300	数据集获取及软件服务费	数据集获取及软件服务费
2.设备费 （原则上不许购置通用办公设备，如电脑、相机、打印机、复印机、移动硬盘等，购置的设备在项目结题后须交还学校）			
3.图书资料购置费 （购置的图书资料在项目结题后须交还学校）	1500	购置 NLP 领域图书及参考资料等	购置 NLP 领域图书及参考资料等
4.项目办公费 （记录本、笔、文件夹、档案袋、电池等物品购置费，不许购置办公耗材，如硒鼓、墨盒、复印纸、优盘、接线板等，原则上不超过 200 元）	200	记录本、笔、文件夹、档案袋、电池等物品购置费	记录本、笔、文件夹、档案袋、电池等物品购置费
5.打印制作费 （原则上不超过 500 元）	500	打印、复印费等	打印、复印费等
6.测试化验加工费			
7.论文发表费	8000	论文版面费等	预计发表 SCI 或 EI 检索论文 1 篇，版面费 8000 元
8.知识产权事务费 （如申请专利等）			
9.文献及信息检索费			

10.京外差旅费 （原则上不超过项目总经费30%）	6000	外地研讨费	往返北京、上海、深圳等地，开展NLP领域调研
11.市内调研公共交通费 （原则上不超过500元）	500	市内研讨费	往返中科院研究所等地
12.项目研究成果参赛费			
13.其他支出（具体列明）			
说明：经费预算应符合项目申请书内容。项目经费批准后，无客观原因预算不得调整			

四、研究内容及解决的关键问题

为实现文本情感分类，需要重点研究两方面科学问题的解决方案，分别是词向量的情感信息表示方法以及在此基础上文本分类方法。前者是后者的研究基础，后者通过前者进行情感分类。具体研究内容为：

1. 研究具备情感倾向信息的词向量表示方法。在现有 word2vec 词向量表示方法的基础上，面向情感分析的词向量表达需要，研究具有情感倾向信息的词向量表示方法，解决 word2vec 难以用于情感分类场景的问题。

2. 研究用于情感分类的神经网络结构。在现有文本分类方法的基础上，面向情感分类任务，通过深度学习的方法，研究用于情感分类的深度学习网络结构，解决现有神经网络在情感分类任务过程中存在的准确率低以及泛化能力弱等问题。

除上述研究内容外，本项目会将模型以网站的形式进行实际应用。网站可以处理访问者上传的文本内容，并以可视化的方式实时展示文本的处理结果，直观地展现情感分析结果的准确性。拟解决的关键问题为：

1.词向量在表达上缺乏情感信息的问题。由于词向量着重体现的是语义信息，导致对于一些语义相近而情感相反的词没有区分度，如 bad 与 good。所以直接使用通用的预训练词向量可能会降低模型的效果。因此，我们需要使用带有情感信息的词向量。本项目将寻找一种方法来优化预训练的词向量，使得在保证一定语义关系的情况下，语义相近，情感不同的两个词，距离更远；语义相近，情感相同的两个词，距离更近，从而得到既有语义信息还带有情感信息的词向量。

2.LSTM 神经网络模型在进行文本分类过程中存在以下问题。首先它侧重于保存句子的历史信息，然而有时候只看前面的词是不够的。其次 LSTM 应用在大数据集时计算量大，训练时间长。第三，LSTM 在短文本和训练语料相对有限的情况下优势较弱，对局部信息的捕获能力不强。因此本项目将在 LSTM 的基础上寻找一种方法来有效解决上述问题。

五、项目实施阶段计划

2018.5——2018.6　补充自然语言处理（Natural Language Processing，NLP）基础知识，阅读大量相关论文，把握学界动态；

2018.7——2018.8　实现基准方法，多种最先进算法结构，在SemEval、SST、IMDB等数据集上训练模型并测试，记录实验结果。提出新方法并实现，在实践中改进；

2018.9——2019.6　对比不同模型产生的结果，整理算法与实验结果，投稿论文（SCI/EI），并针对论文审稿意见，进行稿件的修改和完善；

2019.7——2019.9　准备演示系统并实际应用。

六、研究方法、技术路线、实验方案

一．研究方法

第一步，搜集并下载目前国际认可度较高的文本分类与情感分析数据集。

第二步，把训练集和测试集中的文本进行预处理。（包括删除非字母数字字符，删除停用词，小写转换，词性还原等，考虑词性还原（将「am」「are」「is」等词语统一为常见形式「be」））

第三步，词向量的表示。目前主流深度学习模型采用连续词袋模型（CBOW）产生 word embeddings，来表示词语。

在 word embedding 的基础上，我们将提出如下新的情感词向量学习框架（如图 1 所示）。这种新框架将情感信息和语义信息分开训练,得到两个维度空间的向量,即情感空间(sentimental space)和语义空间(semantical space)。从图 1 中可以看出,每个词向量都被分为两个部分(情感和语义)。其中语义空间部分使用 skip-gram 模型中的训练方法,情感空间部分将滑动窗口的所有词加和后映射到一起，作为 softmax 层的输入。训练情感空间时,使用滑动窗口中的所有词的映射作为 softmax 层的输入,并通过反向传播更新词向量的情感信息。（前三步对应流程图如图 2 所示。）

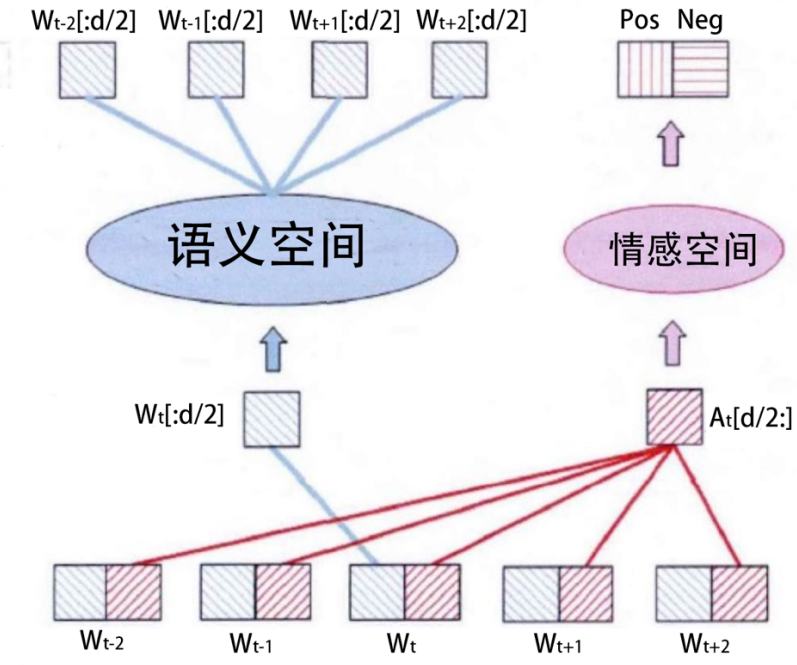

图 1.情感词向量示意图

图 2.神经网络训练前流程图

第四步，神经网络的训练。

当前 state-of-the-art 的方法有 TextCNN，TextRNN 等。下面介绍 TextCNN 的结构与原理（如图 3 所示）。

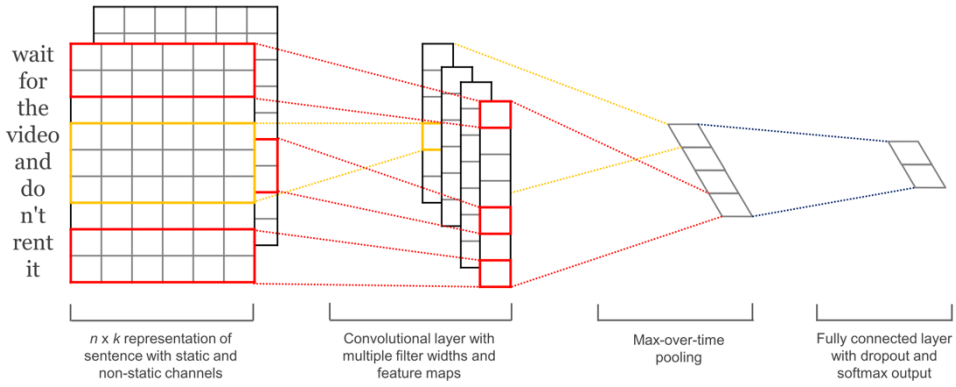

图 3.TextCNN 结构

1.输入层

可以把输入层理解成把一句话转化成了一个二维的图像：每一排是一个词的 word2vec 向量，纵向是这句话的每个词按序排列。

所谓的 static 和 non-static 的 channel 解释如下：

CNN-static: 所有的 word vector 直接使用无监督学习即 Google 的 Word2Vector 工具(COW 模型)得到的结果，并且是固定不变的；

- CNN-non-static: 所有的 word vector 直接使用无监督学习即 Google 的 Word2Vector 工具(COW 模型)得到的结果，但是会在训练过程中被 Fine tuned；

从输入层还可以看出 kernel 的 size。很明显 kernel 的高(h)会有不同的值，因为我们需要获得的是纵向的差异信息，也就是不同范围的词出现会带来什么信息。

2.卷积层

由于 kernel 的特殊形状，因此卷积后的 feature map 是一个宽度是 1 的长条。

3.池化层

使用 MaxPooling，并且一个 feature map 只选一个最大值留下。这被认为是按照这个 kernel 卷积后的最重要的特征。

4.全连接层

全连接层使用带 dropout 的全连接层和 softmax。

在 CNN 和 LSTM 的基础上，我们提出了两种新的网络模型。

第一种模型借鉴图像识别中性能优越的残差网络的思想，进行跨层连接（如图 4 所示）。这种结构可以有效解决纯粹加深 CNN 网络层数的几种缺点：

1. 参数太多，容易过拟合，若训练数据集有限；

2. 网络越大计算复杂度越大，难以应用；

3. 网络越深，梯度越往后穿越容易消失（梯度弥散），难以优化模型

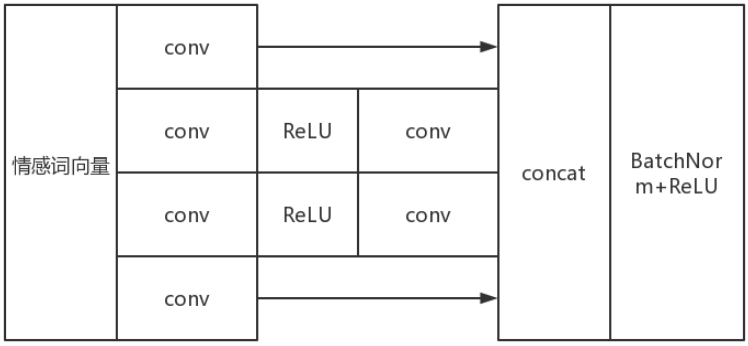

图 4. 网络模型一

第二种模型融合了LSTM和CNN两种网络结构，并且加深了CNN的卷积层数（如图5所示）。由于网络深度的增加，有效提高了它对文本特征的提取能力 ，结合LSTM后，可以更好地保留历史信息，弥补了CNN的不足。

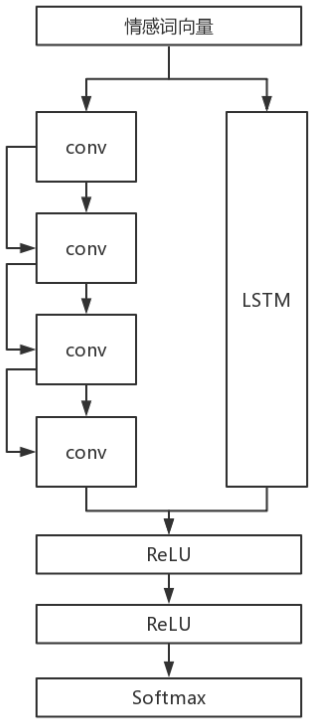

图5. 网络模型二

二． 技术路线

在国际主流情感分析数据集的基础上，我们首先进行情感词向量（senti-embedding）的学习框架研究，这是第一个研究内容。随后，我们将训练好的情感词向量送入以神经网络模型为基础的文本建模框架，该框架将在 CNN 和 LSTM 的基础上进行改进，这是我们的第二个研究内容。具体框架图如图 6 所示。

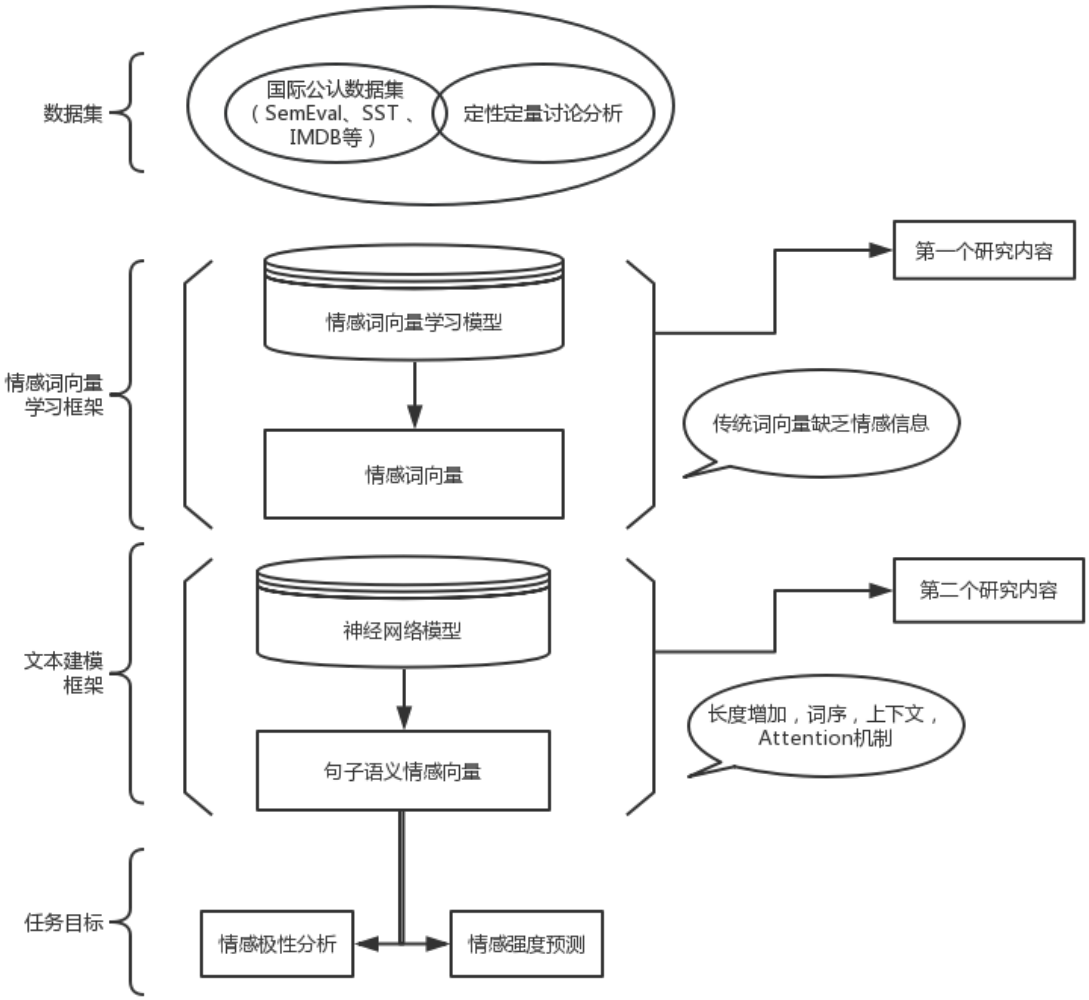

图 6.本项目主要研究框架图

三．实验方案

1. 数据集及实验设置

为比较基于神经网络的文本建模方法及其与情感词向量结合后在情感分析任务中的表现，我们设计了基于 IMDB、SST、SemEval 等数据集的情感分类实验。

2. 基准系统

以词袋模型+SVM 作为基准系统。

3. 前沿模型当前主流模型

将传统语言学特征、特定文本领域内特征结合起来，训练 word embeddings，之后送入 TextCNN 或 TextRNN 进行训练。

4. 设计 CNN 和 LSTM 相结合的新算法

根据实验技术路线，通过训练不同模型，并对比产生的结果，改进网络结构，从而提出自己的新算法，并在实践中改进和优化。

5. 实验结果及分析

整理算法与实验结果，对数据进行可视化分析，并通过网站的形式实际应用。

七、其他说明

为确保北京林业大学大学生创新创业训练项目（以下简称项目）顺利进行，明确学校与项目组学生的责任与权利，现进行以下说明：

1、学校提供政策和经费支持，项目所在学院负责组织开展项目阶段检查、中期检查和结题验收工作。

2、项目组保证在项目实施过程中严格遵守学校有关规定，恪守学术规范、项目按计划开展、主动配合阶段检查、中期检查和结题验收工作。由于客观原因，项目需要进行变更，项目组须按要求及时提出书面申请，待学校批准后方可做出项目变更。

3、项目组确保经费专款专用，所有票据必须为正式发票，发票抬头写“北京林业大学”，发票背面需要有项目主持人签字、指导教师签字、学院经费负责人签字，方可凭票报销。

4、项目无故终止、逾期不能完成、弄虚作假或结题验收不合格者将追究项目组责任，并追回项目前期投入，且项目组所有成员大学期间不能再参加项目。

5、项目主持人有权利剔除对项目不负责的成员，但需按要求及时向学校提出书面申请。被主持人剔除的成员以及无故退出项目组的成员大学期间不能再参加项目。

八、项目组执行承诺

项目组成员均已清楚项目内容和工作分工，并将严格遵守项目管理办法的规定，恪守学术规范，保证项目研究时间，合理有效使用经费，实事求是按计划开展研究工作，按期按要求完成项目各项工作，及时报送项目有关材料。 项目组所有成员签字： 日　期：

九、项目审批意见

（一）指导教师意见及承诺
签字：　　日期：
（二）学院意见

主管领导签字（公章）：　　日期：

（三）学校意见

同意立项。

单位（公章）：　　日期：2018 年　月　日