分类号：_____ 密级：_____

UDC ：_____ 编号：_____

专业硕士学位论文

（工程硕士）

# 基于深度学习股票价格波动预测的研究


硕士研究生 ：李洪强

指 导 教 师 ：王科俊　教授

企 业 导 师 ：张博实　教授

工 程 领 域 ：控制工程

论文主审人 ：冯伟兴　教授



哈尔滨工程大学

2018 年 5 月

分类号：_____ 密级：_____

UDC ：_____ 编号：_____

专业硕士学位论文

（工程硕士）

# 基于深度学习股票价格波动预测的研究


**硕士研究生**：李洪强

**指 导 教 师**：王科俊　教授

**学 位 级 别**：工程硕士

**工 程 领 域**：控制工程

**所 在 单 位**：自动化学院



**论文提交日期**：2018 年 5 月

**论文答辩日期**：2018 年 6 月

**学位授予单位**：哈尔滨工程大学

Classified Index:

U.D.C:

A Dissertation for the Professional Degree of Master

（Master of Engineering）

# Research on Prediction of Stock Price Fluctuations Based on Deep Learning


**Candidate:** Li Hongqiang

**Supervisor:** Prof. Wang Kejun



**Academic Degree Applied for:** Master of Engineering

**Engineering Field:** Control Engineering

**Date of Submission:** May., 2018

**Date of Oral Examination:** Jun., 2018



**University:** Harbin Engineering University

# 哈尔滨工程大学

# 学位论文原创性声明

本人郑重声明：本论文的所有工作，是在导师的指导下，由作者本人独立完成的。有关观点、方法、数据和文献的引用已在文中指出，并与参考文献相对应。除文中已注明引用的内容外，本论文不包含任何其他个人或集体已经公开发表的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者（签字）：

日期：　　年　月　日

# 哈尔滨工程大学

# 学位论文授权使用声明

本人完全了解学校保护知识产权的有关规定，即研究生在校攻读学位期间论文工作的知识产权属于哈尔滨工程大学。哈尔滨工程大学有权保留并向国家有关部门或机构送交论文的复印件。本人允许哈尔滨工程大学将论文的部分或全部内容编入有关数据库进行检索，可采用影印、缩印或扫描等复制手段保存和汇编本学位论文，可以公布论文的全部内容。同时本人保证毕业后结合学位论文研究课题再撰写的论文一律注明作者第一署名单位为哈尔滨工程大学。涉密学位论文待解密后适用本声明。

本论文（□在授予学位后即可　□在授予学位 12 个月后　□解密后）由哈尔滨工程大学送交有关部门进行保存、汇编等。

作者（签字）：　　导师（签字）：

日期：　　年　月　日　　年　月　日

# 摘　　要

股票为市场经济的产物，随着国家经济的不断发展，股票交易在金融市场中有着重要的意义。由于股票具有高风险和高收益的特点，股票的涨幅将直接影响投资者的利益，因此对股票价格未来走势的预测具有可观的应用价值。股票市场具有较强的随机性，各种因素错综复杂、主次关系变化不定，股票价格不仅受买卖双方、投资者的主观意识等因素影响，同时也与国内外重要的政治事件、国家重大经济政策、国际形势变化、战争影响有直接关系，而这些重要信息一般会以新闻的形势发布，因此在股票预测中加入新闻因素的影响，可以提高模型对股票价格未来走势预测的准确性。本文将股票历史价格波动数据和股票公司相关的新闻信息融合并进行分析。由于股市具有高阶非线性特点，而传统时间序列模型很难解决高阶非线性问题，通过对预测方法的大量分析，深度学习可以解决高阶非线性问题而且在训练过程中避免了传统人工神经网络在训练时出现的过拟合问题。因此本文将采用深度学习网络模型对股票市场未来的走势进行预测。本文的主要研究工作如下：

1、数据库的建立及预处理：实验的数据库分为两部分，一部分为有关苹果公司及相关科技公司的新闻信息，另一部分为苹果公司股票价格的历史数据。两部分数据集将通过爬虫的方法在网络上进行爬取。由于模型的输入只能是数字特征向量，所以需要将新闻信息使用 NLP 进行特征向量的转换，并对数据库进行预处理。

2、应用 CNN 对股票价格波动进行预测：由于 CNN 具有权值共享和局部感知的特性，有效的减少了参数量，可以避免过拟合现象的发生。实验中将采用 CNN 对股票价格波动进行预测。验证卷积神经网络模型对股票价格波动预测的准确度优于传统时间序列 ARIMA 模型对股票价格波动预测的准确度；并分别对滑动窗口长度选择及新闻的文本处理方式的不同都会对股票价格波动预测的精度有所影响。

3、应用 LSTM 对股票价格波动进行预测：实验中应用 LSTM 网络对股票价格波动进行预测。验证了 LSTM 网络对股票价格波动的预测优于 CNN 神经网络对股票价格波动预测的精度；并且验证了新闻信息对股票价格波动预测效果有促进作用，同时相关公司新闻的种类越多、数据量越充分，股票价格预测的效果越好。

本文研究发现，以新闻数据和股票历史价格数据为实验的数据集作为长短时间记忆网络模型的输入，有利于提高股票价格波动的准确性，对未来股票价格波动预测的进一步研究有重要意义。

**关键词**：股票预测；深度学习；新闻；CNN；LSTM

# ABSTRACT

The stock is the product of a market economy. With the continuous development of the national economy, stock trading has an important significance in the financial market. Due to the characteristics of high risk and high returns of stocks, the rise of stocks will directly affect the interests of investors. Therefore, the prediction of the future trend of stock prices has considerable application value. The stock market has strong randomness, various factors are complex and the relationship between the primary and secondary changes is uncertain. The stock price is not only affected by factors such as the buyer and the seller, the subjective awareness of the investor, but also important political events at home and abroad, and major national economic policies. There is a direct relationship between changes in the international situation and the impact of war. These important information will generally be released in the news situation. Therefore, adding news factors to the stock forecast can increase the accuracy of the model's prediction of the future trend of the stock price.This article combines the stock historical price volatility data and the news information related to the stock company and analyzes it.Because of the high-order nonlinear characteristics of the stock market, traditional time series models can hardly solve high-order nonlinear problems. Through extensive analysis of prediction methods, deep learning can solve high-order nonlinear problems and avoid traditional artificial neural networks in train when over fitting problems.Therefore, this paper will use the deep learning network model to predict the future trend of the stock market. The main research work of this paper is as follows:

1. Database establishment and preprocessing: The experimental database is divided into two parts, one part is the news information about Apple and related technology companies, and the other part is the historical data of Apple's stock price. The two-part dataset will be crawled on the network by crawlers. Since the input of the model can only be a digital feature vector, it is necessary to use NLP to transform the feature vector and preprocess the database.

2. Using CNN to predict the fluctuation of stock price: As CNN has the characteristics of weight sharing and local perception, it effectively reduces the parameter amount and avoids the occurrence of over fitting. In the experiment, CNN will be used to predict stock price fluctuations. Verifying that the convolutional neural network model is more accurate than the traditional time series ARIMA model in predicting stock price fluctuations, and the differences in the sliding window length selection and the news text processing method will affect the stock price fluctuation. The accuracy of the forecast affects.

3. Use LSTM to forecast stock price fluctuations: In the experiment, LSTM network is used to forecast stock price fluctuations. It verifies that the prediction of stock price fluctuations by LSTM network is better than the accuracy of prediction of stock price fluctuations by CNN neural network; And it verifies that news information has a positive effect on the prediction of stock price fluctuations,and the more variety and the more data of relevant company news, the better the effect of stock price prediction.

This study finds that using datasets with news data and stock historical price data as input to the LSTM will help improve the accuracy of stock price volatility, which is of great significance for further research on the prediction of stock price volatility in the future.

**Keywords:** Stock Forecasting; Deep Learning; News; CNN; LSTM

# 目　录

# 第 1 章 绪论

## 1.1 课题的背景及研究意义

随着人工智能领域[1]的快速发展，越来越多的相关技术被应用到金融预测领域，如：信用卡的评级、风险投资、风险评估评估和股票价格波动预测等等诸多行业。因此很多金融领域研究员便向人工智能抛出橄榄枝，希望能在人工智能领域占有一席之地。基于人工神经网络的金融预测模型在股市上具有很强的现实意义和应用性[2-4]。股票为市场经济的产物，股票交易在金融市场中有着重要的意义[5]。股票市场与国家政治经济等紧密相连，是金融市场中的重要组成部分之一，经济学家不仅将它称为是国家经济的“晴雨表”和“报警器”，而且在一定程度上反映着一个国家的经济状况和经济实力。在当下的金融领域，股市对整个金融圈有着重要影响，股票投资俨然已成为大部分人日常生活的一部分，股民时刻关注股票的跌涨。如果能够准确地预测股票的波动趋势，同时对股民进行有效的引导，给出抛出或者买进的意见，这对股民投资有很大的帮助[6]。

预测指的是在掌握现有的信息基础上，按照一定的规律和方法来对未来的事情进行推算，以预先了解未来事情发展的方向和结果。股票预测是金融预测的一个分支，指以股市信息和准确地调查资料为依据，从股票市场的历史、现状和规律性出发，运用科学恰当的方法和统计对未来股票市场的发展做出预测[7]。按预测时间的长短可分为：短期预测（一般指三个月以内）、中期预测（指三个月到一年）和长期预测（一年以上的时间）。由于股票市场受多种不确定性因素的影响，价格波动的规律常常表现出较强的非线性，而且股票市场处理的信息非常庞大，对算法的要求非常高。因此，股票市场的预测往往不尽人意。股票价格波动有诸多影响因素，除了股票市场自身因素的影响以外，还受限于各种政治、经济，以及交易技术方面和投资者心理的等因素。概括起来主要分为两大类：一类是基本因素，另一类是技术因素。

基本因素是指除股票市场以外的政治与经济因素及其他的因素的波动和变化对股票市场价格的波动趋势产生决定性的影响。一般来说，基本因素可分为人为操纵性因素、经济性因素、政治性因素和其他因素等。

政治因素是指影响股票价格波动的政治措施、国内外的重大政治事件、经济政策、政府发展计划、刚刚实行的法令等都会影响到股票价格波动。政治形势的更改也会对其产生重要的影响，主要表现在：

（1） 国内外重要的政治事件。譬如国内外政治风波会影响投资者的心理状态，从而影响股民的投资水平，间接的影响股票价格波动状况。

（2） 国内外经济政策的改变。譬如国家扶持项目的股票价格则会呈上升状态，而

违背国家制度的公司股票价格则会持续走低。如：目前国家对环保十分重视，对于以清洁能源为主要业务的公司股票价格大部分会持续走高，而以煤炭、铁矿等公司的股票价格将会持续走低，由于环境保护国家鼓励清洁能源对煤炭等需求将逐渐减少从而影响公司的经营状况，导致了该公司的股票价格下跌。这些消息大部分都会通过新闻文本进行传达，因此需要着重掌握这些重要信息来对股票的波动进行更准确的预测。

（3） 国际形势变化。如投资者抓住外交关系改进的机会并及时的购进相关跨国公司的股票，便会获得很大的利润。

（4） 战争影响。战争造成国家经济发展延缓，民心动荡不安，从而影响股民的投资状态。但战争会使军工产业繁荣兴盛，但凡与军工相关的公司其股票价格必然上涨。此时投资者应该乘胜加注购买军工相关的股票，售空因战争而亏损的股票。

对未来股票价格的预测不仅通过股票历史价格进行预测，还需要了解股票价格波动的影响因素。而大多数影响因素都会以新闻信息的形式发布，比如国家的政策、公司的经济状况等等都是在新闻中有所体现，而这些信息将会直接影响该公司未来股票价格的上涨下跌。因此加入新闻信息对股票价格波动的影响会比单一利用历史数据对未来股票价格预测的准确性高。本文主要介绍通过新闻信息和历史数据的融合对未来股票价格波动进行预测。

## 1.2 国内外研究现状

由于股票具有高风险高收入的特性，成为人们追捧的投资方式之一。因此很多人对股票价格波动预测的研究产生强烈的兴趣，众多研究方法使得股票价格预测的准确度逐渐提高，为投资者提供了很大的帮助。

传统的股票预测方法有 K 线图法、柱状图分析法、形态分析法、道氏分析法、点数图法、移动平均线法、黄金分割比螺旋历法、趋势分析法等等。上述方法可以实现预测短期的股票价格波动的趋势。一般传统的方法分析需要预知各种影响参数，并需要了解这些参数的修改情况。传统时间序列预测方法主要是通过建立股价及综合指数之间的时间序列相关辨识模型来预测股市未来变化。其中包括 ARCH（Auto Regrssive Conditional Heteroskedasticity Model,自回归条件异方差模型）、ARMA（Auto Regrssive Moving Average Model,自回归移动平均模型）以及指数平均预测法、指数平滑法、季节性变化法等[8-10]。但一般的时间序列模型很难处理高度非线性函数，所以预测的结果往往不尽如人意。随着计算机技术和人工智能的不断发展，神经网络成为一种高效的处理高度非线性函数的模型，并能够依据数据本身内在分布建立函数关系，具有良好的非线性逼近能力和抗噪能力。因此人工网络便逐步地应用在股票价格预测上，也提高了预测的准确性。

目前国内外对股票价格波动的预测按时间发展分为以下几类：

（1） 基于传统时间序列的股票价格预测研究

为了可以更好地描述股市共有的波动特性，Engle 等人首次提出 ARCH（自回归条件异方差）模型，在处理具有波动特性的时间序列数据问题上预测效果较为理想[11]。随着自回归条件异方差模型倍受广泛关注，研究者便将自回归条件异方差模型应用于分析股票等具有时序性波动特征中，大量实验数据表明：自回归条件异方差模型非常适用于具有时间序列特性的数据。但是，自回归条件异方差模型也存在一定的缺陷，譬如自回归条件异方差模型在处理低阶数据过程中拟合能力较强，但拟合高阶时序性数据时，预测效果并不理想。为弥补 ARCH 模型在处理高阶时序性数据拟合程度较差的问题，Bollerslev 对 ARCH 模型算法加以改进，并提出了广义自回归条件异方差模型（GARCH），提高模型对多样数据的拟合能力[12]。GARCH 模型相对于 ARCH 模型，加入了历史数据的条件方差，能够较好地描述指数日收益率序列中存在的条件方差的时变性和簇集性等特征。王博等则应用 ARMA－GARCH 模型对上海证券综合指数的日收益率为研究对象，同时分别讨论了序列残差项服从 GED 分布、正态分布等不同条件下模型的预测能力[13]。在大量实验数据验证下，ARMA－GARCH 模型的准确率远高于 ARMA 模型，表明 ARMA－GARCH 模型能够较为准确的预测上海证券综合指数日收益率的时间序列条件均值。梁恒运用了 GARCH 模型分别对深证成份指数和上海证券综合指数的非对称波动特征进行了详细叙述，并考虑到两种指数在不同时期的波动特性存在差异，因此将数据样本划分为牛市、熊市和牛熊市混合的三种不同情况。在对沪深指数收益率涨幅分析过程中，发现其时间序列均显现出较强的非对称特性以及相似的波动趋势，但从宏观角度分析，上深证成指日收益率比上证综指的日收益率震荡激烈[14]。廖敏辉选用了 GARCH 时间序列模型中的 TGARCH 模型对股票市场的价格波动性情况进行了详细的描述，文中分析了股票市场中存在非对称波动的主要原因，同时对沪深股票市场提出了卖空机制的建议，以此建议来弱化非对称波动对股票价格的影响[15]。刘璐等用 GARCH 模型对亚洲四个国家（中国、印度、日本、韩国）股票日收率的波动进行研究，结果发现各国的指数日收益率波动情况比较相似，但是日韩两国日收益率的波动状况比中印两国的略大[16]。

（2） 基于神经网络的股票预测研究

由于股票价格是高度非线性函数，而传统时间序列模型很难解决这种高度非线性函数，所以预测的结果往往差强人意、不利于给予投资者建议。随着人工神经网络的不断发展，很多研究员使用人工神经网络对股票价格进行预测，并取得了很好的研究效果。

目前普遍认为最早将神经网络引入到股票预测是 IBM 公司的工程师 White，在 1988

年期间他利用神经网络模型对 IBM 公司的普通股的日收益率进行预测[17]。但是经过对数据的学习后，预测的结果准确率并不高，他认为其中的主要问题是人工神经网络陷入了局部最小值而却不能解决。虽然效果并不理想，但是开创了神经网络对于股票预测应用的先河。Pesaran 和 Timmermann 在 1999 年以伦敦证券指数过去 25 年的数据为样本进行试验预测，试验证明基于神经网络模型对股票价格波动的预测的效果比较好，其指数变化的准确性可以达到 60%[18]。文献[19-22]根据不同的预测目标和研究对象来选择与其对应的变量，使用误差反向传播算法进行网络的训练，对金融市场价格进行预测[19-22]。文献[23-24]对误差反向传播算法进行改进，将其应用于股票价格跌涨趋势的预测，取得了很好的效果，在一定程度上证明了股票市场不完全满足随机游走理论，从而说明股票市场是可以进行预测的[23-24]。文献[25]使用四层回归神经网络，对五个不同的股票市场的历史价格交易量来预测下一年的股票交易量，给个人投资提供参考，是一个成功的探索实验[25]。在国内也有很多研究员对使用神经网络对股票价格波动能够进行预测。武振、郑丕谔将小波神经网络用遗传算法进行改进，并对股票代码为 600019 的宝钢股份、股票代码为 600104 的沪市股票进行研究预测，得出结果证明：此模型具有预测精度高收敛速度快的特点[26]。牛东晓、王建军、李莉于 2009 年提出基于相似度与神经网络的短期协同预测模型，该模型与单独的反向传播算法预测模型具有较高的预测精度[27]。江弋等人提出基于径向基神经网络的预测模型，采用 K-meas 聚类算法来动态计算径向基网络的中心，对径向基网络的权值根据梯度下降算法进行自适应调整，该实验对某公司 11 年股票的收盘价格为样本进行实验预测，取得的效果比较满意[28]。蔡红等人提出了基于主成分分析（PCA）的 BP 神经网络的股票价格波动预测，该方法首次使用主成分分析方法提取几个影响股票价格波动的主要因素，从而降低影响股票价格波动因素的维度，与仅仅使用 BP 神经网络进行预测相比运算效率与精度都有所提高[29]。尹璐等人使用遗传算法对 BP 神经网络的初始值和阈值进行优化，在数据挖掘的启发下实现对股票价格波动进行发掘和预测[30]。从预测的误差中分析，随着时间的推移，预测值和真实值之间的绝对误差成上升趋势变化，因而该模型不适用股票中长期价格波动的预测，仅仅适用于短期股票价格波动预测。在文献[31]中，受限制玻尔兹曼机（RBM）被训练编码股票的月收盘价格，然后微调预测每只股票的价格是否会超过中值变动或低于它[31]。该策略与简单的动量策略进行比较，并证实所提出的方法在年化收益方面取得显着改善。文章中使用卷积神经网络模型对股票价格波动进行预测，文章中对不同网络进行对比，证明卷积神经网络比 SVM、MLP 效果好，预测的准确度更高[32]。文献[33]同样使用卷积神经网络对股票价格波动进行预测，性能评估结果表明，该模型具有较高的精度，召回率和准确度；当作为交易模式的一部分使用时，提供投资者买或者买的选择机会[33]。在深度学习网络模型中，长短时间记忆网络（LSTM）对股票价格波动的预测效果也有了突

出的表现，孙瑞奇使用 LSTM 网络对股票价格波动进行预测研究，文章主要讨论了 BP 神经网络、RNN 神经网络和 LSTM 神经网络对股票价格进行短期预测的可行性并作出相应对比，验证了 LSTM 网络对股票预测的准确性较好，但文章中忽略了影响股票价格的其他因素。

（3） 基于文本信息的股票价格预测研究

虽然股票的影响因素比较多，但是多数因素都会在新闻中体现出来，研究人员为了更加准确的对股票价格进行预测，便通过新闻方面的因素来对股票的价格进行预测。由于股票起初在国外发展起来，所以国外对股票的研究时间长于国内的研究人员，对技术的掌握能力起初也比国内成熟。

2010 年,Feng Li 对国外财经网上的 150 万条财经评论对 45 家企业道琼斯指数的影响进行分析，结果显示公司股票的评论信息内容会对股票市场的走势产生较大影响[34]。2013 年，Geva T 等人通过将股票历史价格变化数据和股评的信息进行融合并进行分析，实验结果表明该模型的预测效果更加准确。但仅仅根据文本内容进行情感分析，将其加入到预测模型中仍有较大的局限性。文本情感分析不够准确，新闻数据量不充足，都会造成预测结果不够精准。但是将文本信息与股票历史数据融合可以进一步提高预测的准确率[35]。2014 年，EJD Fortuny 等人在文本信息处理问题上加以改进，采用支持向量机将新闻评论信息的情感类别进行分类处理，并建立预测模型。实验结果表明准确的文本情感分类对预测模型的拟合能力有一定的提高[36]。在 2015 年，Skuza 等人爬取了 Twitter 上股票信息的相关文本，并进行情感分析，发现社交网络中文本所展现的情感与股票价格变化存在一定的联系[37]。

Vaishali Ingle 等人通过对互联网上的新闻进行收集，使用 TF-IDF 统计方法找到文本或语料库的中对股票市场有价值的文本。并应用 HMM 模型去提取股票市场的和指示的趋势应用算法进行预测。对于每周收集的股票数据，特定公司的实际收盘价和预期收盘价显示的误差率大约在 0.2 至 3.9％的范围内。在文献[38]中，将金融时间序列数据和自然语言处理的组合，使用卷积神经网络对未来股票价格波动进行预测，实验结果证明此种方法相对于其他算法准确度有一定的提高[38]。在文献[39]中，使用技术分析变量的神经网络模型已经被应用于上海股票市场的预测[39]。该实验比较了两种学习算法和两种权重初始化方法的性能。结果表明反向传播效率可以通过多重线性回归权重初始化的共轭梯度学习来提高实验结果的准确度。国内在此方面也有相对研究，张世军等人于 2014 年将网络文本评论和股票历史数据信息结合，并用 SVM 的模型进行股票波动预测，使得传统的股票分析方法和数据挖掘分析方法优势互补，能够达到更好的预测方法，但是该篇文章并没有考虑时间序列等其他影响[40]。Xiao Ding、Yue Zhang 等人通过在新闻中提取关键词并表示成密集向量，然后将卷积神经网络应用在长期和短期时间的价格波动

上，通过这个方法对 S&P500 指标的预测和个人股票的预测，准确率提高了将近 6%[41]。朱梦君、许伟等人在建立预测模型中，将金融相关的微博内容加以语义分析。分析过程综合了文本的感情表述、博主的影响力等因素。虽然分析全面，但是内容获取比较单一，而且在微博中的部分内容都是包含个人感情色彩的，终会有一些偏激对后期的预测有所影响[42]。

## 1.3 目前股票预测存在的问题

本文对股票预测的相关文献进行归纳和总结，阐述各种方法的优缺点。传统股票价格波动预测方法需要知道建立模型中的各种参数，且需要知道建立模型时参数修改的时机和情况，操作比较复杂，需要了解更多股票预测的知识和影响股票价格波动的因素，这些传统股票预测的缺点对于普通炒股者来说难度很大。既要知理论又需懂技术，不易操作，不适合模型的快速建立。随着技术的发展，虽然部分传统时间序列模型解决了多种参数的修改技术问题，由于股市具有高阶非线性特点，而传统时间序列模型很难解决高阶非线性问题，所以预测的结果往往差强人意、不利于给予投资者建议。而随着计算机技术的不断发展，神经网络的出现有效地解决了处理高度非线性技术问题，具有很好的非线性逼近能力和抗噪能力。但是由于神经网络层数少，对大量数据处理时间长，模型训练容易出现过拟合。而深度学习网络模型却解决了以上的技术难题，很好的完成大量数据的处理问题，对于过拟合问题也做了很好的优化，方便了模型的建立。

自然语言处理方法对于文本信息的处理也很好的解决了文本信息处理技术问题。便有很多研究员通过对新闻因素的分析来预测未来股票价格波动的趋势。股票价格波动的影响因素虽然很多，但大部分影响因素都将直接或间接通过新闻信息来进行公布，所以新闻信息是股票价格波动影响因素的重要部分。虽然新闻信息直接会影响未来股票价格的波动趋势，但是新闻只是影响因素的一部分且只能作为辅助信息来对未来股票价格趋势进行预测，仅通过单一的新闻信息作为数据集进行建模预测，所预测的效果不会很理想。

## 1.4 本文主要内容及章节安排

目前金融预测领域常用的方法为通过股票的历史数据集来对未来公司的股票价格进行预测，忽略了其他影响股票价格波动的因素，而这些影响因素很大程度决定了未来股票价格的跌涨程度，例如：国家重视环境保护，则清洁能源公司、新型环保材料公司的股票价格则会涨停成为牛市，而像煤矿、钢厂等重污染企业的股票价格则会暴跌成为熊市。通常这些影响因素都会以文本信息的形式通过新闻发布出来。本文将利用新闻作为辅助信息来对公司未来的股票价格波动情况进行预测研究。

本文主要研究苹果公司股票价格波动与苹果公司的新闻信息及与该公司相关联的五家科技公司（其中包括 ACLS、CAMP、CSLT、CYOU、RPD）的新闻信息间的关系，并运用深度学习网络模型对苹果公司股票四个指数（开盘价、收盘价、最高价、最低价）的变化趋势进行研究，选取实验效果比较好的模型作为最终的训练模型。

实验中选用苹果科技公司的历史股票价格和该公司及相关公司的新闻信息来作为深度学习网络模型的数据库，并选取多个网络模型进行实验对比，选取实验结果比较好的模型应用于实际股票预测中。

论文分为 4 章，具体内容如下：

第 1 章 绪论。阐述本文课题的背景及其研究意义，并对国内外历史研究现状进行了简洁阐述，给出了目前股票预测方法的不足与改进，对于本课题的主要研究内容进行简单的介绍。

第 2 章 数据库的建立及预处理：本章节主要介绍数据库的构成、建立及预处理过程。数据库主要有苹果公司股票价格的历史数据、苹果公司新闻信息及其他五家科技公司（ACLS、CAMP、CSLT、CYOU、RPD）的新闻信息构成。实验中通过两种方式对股票历史价格和新闻信息进行爬取，并对数据进行预处理工作。由于模型的输入为特征向量，因此需要对新闻文本进行处理，本章节简单介绍了四种自然语言的处理方式。

第 3 章 基于 CNN 股票价格波动预测的研究：本章节详细阐述了卷积神经网络对股票价格波动预测的模型，并对模型的网络架构进行详细的分析进而说明该模型应用的合理性。同时简要介绍了卷积神经网络的基本原理、特点及运算规则。实验部分介绍了实验的操作环境及对预测的评价标准；验证卷积神经网络模型对股票预测的效果好于传统时间序列网络 ARIMA 模型；对几种不同文本处理方式进行对比，选取效果较好的处理方式并分析主要原因；由于股票价格是时间序列的数据，实验中验证滑动窗口长度对股票波动预测有所影响。

第 4 章 基于 LSTM 股票价格波动预测的研究：本章节详细介绍了长短时间记忆网络对股票价格波动预测的模型，并详细说明了模型的网络结构。对长短时间记忆网络的发展、算法原理及长时间记忆网络的优点进行阐述。实验部分分析了不同网络结构对股票价格预测的影响；验证了新闻数据集对于股票价格的波动有积极地影响，同时新闻信息的类别越多对于未来股票价格波动的预测结果越好。

# 第 2 章 数据库的建立及预处理

对一个需要建模的问题进行分析研究过程中，模型都是以数据为前提进行构建，所以数据库的建立是进行模型构建的基础也是模型构建最重要的一步，因此数据库是整个过程中一个重要的部分。实验中可以利用数据库对模型构建的性能进行评估。一个算法的性能主要在于算法的唯一性和可行性，这些性质可以在数据库上进行实现，且由数据库的大小和数据质量进行体现。优质的数据库才能建立鲁棒性能好、泛化性能强的模型。对于股票价格波动预测系统而言，这些影响因素可以为新闻的信息种类，文本的格式，新闻编辑者对公司的个人感情色彩等一系列因素[43]。这些因素直接会影响模型建立模型的鲁棒性能。获取公司的新闻信息有多种方法，如通过新浪财经、谷歌 API、雅虎 API、Intrinio 的 API、爬虫等都可以检索新闻的文本特征。本章节主要介绍两种实验方法获取苹果公司及相关五个科技公司的新闻信息。由于股票价格波动影响因素诸多，需要在不同影响环境下有一个良好的实现效果，需要获得大量的信息，而本实验的获取新闻信息的渠道有限，因此不能保证在该数据库下进行实现的算法具有普遍性，且样本量较少，不能很好地表现出样本间的差异性。本文的实验效果在本实验的数据库中有较好的表现，具有一定的特殊性。所以在应用到实际中时需要有足够大的样本集，在新闻信息中需要通过多种渠道进行获取，以便获取更加精准的信息保证算法具有很好的鲁棒性和泛化性能。

## 2.1 新闻数据库的建立

### 2.1.1 爬虫获取新闻

本文实验主要研究苹果科技公司的股票价格波动规律。由于新浪财经网站对金融企业的信息公布相对于其他网站比较多，新闻消息比较客观公正，可以及时更新各公司的相关新闻消息，因此本实验将从新浪财经网站上爬取公司的新闻信息。在财经网站中将选择 Python 爬虫来爬取新闻的主标题、新闻的发布时间、公司的名称等信息标签。

#### 2.1.1.1 爬虫工具简介

随着网络的快速发展，万维网络已成为大量信息的载体，对于如何有效提取、利用网络上的大量信息成为一个挑战，网络爬虫应运而生[44]。网络爬虫又被称作网页机器人、网页蜘蛛，是一种按照设定的规则而自动去抓万维网络中信息的脚本或程序。网络爬虫按照实现技术和网络结构大致分为以下几类：聚焦网络爬虫、通用网络爬虫、深层网络爬虫、增量式网络爬虫。实际的网络爬虫系统通常由几种爬虫技术相结合实现[45-46]。

网络爬虫的基本工作流程分为以下几个步骤，流程图如图 2.1 所示：

（1） 首先选取一部分挑选的种子 URL。

（2） 将选好的 URL 放到将要抓取的 URL 队列中。

（3） 从将要抓取的 URL 队列中读取队列中的 URL，进行 DNS 解析，并且获得主机的 IP 地址，将 URL 所对应的网页进行下载，然后存储到已经下载的网页库中。同时，将这些 URL 放入到已获取的 URL 队列中。

（4） 进行分析已获取队列里的 URL，从已下载网页数据中分析其他的 URL，并和已经抓取的 URL 进行比较重复比较去掉重复的 URL，最后将去重的后的 URL 放入待抓取的 URL 队列中，从而进入下一个循环。

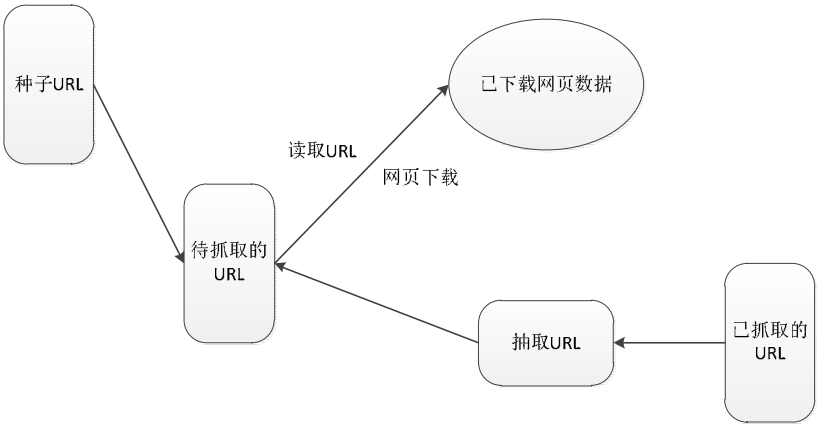



图 2.1 网络爬虫生成结构

#### 2.1.1.2 新闻数据的抓取

本文主要利用爬虫做特定网页中内容的爬取，爬虫选取的主要原则是开源、结构简单、可结构化抓取网页中特定内容的题目，通过按着自己定义的规则处理爬取新闻时遇到的字符串问题。处理字符串的时候通常使用正则表达（也称为模块表达式）。正则表达式就是描述一系列字符串排列的一套规则。如，需要寻找新浪财经网站中有关苹果公司的新闻标题，其他的无用信息则需要过滤，此时需要查看新闻标题的格式，然后建立一个正则表达式来表示网站中所有新闻的标题，这样便可以利用该正则表达式从网站的网页提取出满足该规则的一系列字符串出来，而这些满足此规则的字符串就是该网站中新闻标题的格式。这样便能将该网站中所有的新闻标题提取出来，并过滤掉无用的信息。实验中使用的爬虫语言为 Python 语言，因此会使用 re 模型实现正则表达式的功能[47-48]。

爬虫爬取新闻标题测试如下：

（1） 使用 Python 爬虫，首先安装两个 Python 的第三方库，一个为 request，另一个为 BeautifulSoup，需要在终端（cmd）进行 pip 安装。

（2） 打开新浪财经新闻首页，搜索苹果科技公司的相关新闻信息（网址：http://biz.finance.sina.com.cn/usstock/usstock_news.php?symbol=AAPL，如图 2.2 所示：


图 2.2 苹果公司新闻信息

（3） 获取网页的 html 后，需要使用 BeautifulSoup 库进行解析，然后在分析新浪财经新闻的 html 结构。进入新闻后，找到第一个新闻的题目鼠标右键就可以定位到题目所在的 html 代码，如图 2.3 所示：然后可以看到<h1>标签，找到该新闻标题的 class 元素，便可以获取该新闻的标题、时间。


图 2.3 财经新闻 html 结构图

（4） 运行后可以得出新闻的信息，将其保存在 CSV 文件中，如图 2.4 所示：

| | | |
|---|---|---|
| AAPL | 2016/5/24 | More challenges than cheer for Apple chief on Asia tour |
| AAPL | 2016/5/25 | India says Apple must sell locally-sourced goods to set up stores - source |
| AAPL | 2016/5/30 | India discussing Apple's request for FDI rules waiver, Nirmala Sitharaman says |
| AAPL | 2016/6/1 | Apple plans to sell $1 bln of 30-yr bonds in Taiwan -sources |
| AAPL | 2016/6/2 | S&P 500 closes at seven-month high on data boost |
| AAPL | 2016/6/20 | India opens the door for Apple retail with new FDI rules |
| AAPL | 2016/6/22 | Samsung takes fight to Apple with mobile wallet strategy |
| AAPL | 2016/6/23 | High-end smartphone market set to grow in India this year: Report |
| AAPL | 2016/7/15 | S&P 500's record highs held back by Apple's falling stock price |
| AAPL | 2016/7/18 | Japan's Son chased $32 billion ARM deal by the sea in Turkey |
| AAPL | 2016/7/21 | Pokemon Go seen making billions for Apple |
| AAPL | 2016/7/22 | Silicon Valley leads avalanche of quarterly reports |

图 2.4 新闻数据

### 2.1.2 Intrinio 的 API 获取新闻数据

通过爬虫获取的新闻数据比较繁琐，爬虫过程可能会丢失部分信息，而且在新浪或其他网络金融网站的信息保存时间比较短，由于 Intrinio 包含多家金融信息且含有 API

端口，可以通过自身的网络端口获取大量公司的新闻数据而且可以获取三年内的新闻数据，但是新闻获取比较单一，因此本实验将通过此方法进行新闻数据集的扩充。

#### 2.1.2.1 Intrinio 简介

Intrinio 使用机器学习和基于规则的算法收集财务数据、整理及标准化数据组，以便投资者进行数据分析。Intrinio 的数据价格合理、易于使用。主要通过数据民主化推动普及金融技术的新时代。在 Intrinio 的财务数据中可以获得每只股票的历史股价和及时的历史股价，在有的平台上获取的数据可能存在 15 分钟的延迟，而在该网站能及时的获取股票价格历史数据及其他相关金融信息的功能，对模型预测未来股票价格的波动提供了很好的帮助，使得预测的准确性得到了提高。Intrinio 公司的 API 包括超过 45 种类型的公司业务，用于超过 130000 种国际证券。这些公司的变动数据字段可用于美国所有主要交易所和许多著名的国际交易所。每种类型的公司业务都有许多与该公司相关的数据点。例如，股息是一种公司行为，并且有多个与该行为有关的数据点，包括该日期和支付日期，这使得企业行为 API 是多维的，意味着一个合理的公司事件可以具有与该事件相关的多个数据点。而这些数据点将通过 Intrinio 的 API 以嵌套的 JSON 格式返回。

#### 2.1.2.2 Intrinio 获取公司新闻

Intrinio 有多种 API，其中包含：数据点、历史数据、筛选、价格、交换价格、公司、证券、指数、拥有者、证券交易所、最新的 SEC 文件、公司新闻、财务报告等等 AIP，并可以保存成多种文件形式，JSON、CSV、Excel、Sheets 等。本文将使用新闻 API 来获取苹果公司及其他五家科技公司的相关新闻，并将提取的文本数据保存为 CSV 文件格式。

打开 Intrinio（网址：https://intrinio.com）在工具中选择 API 浏览器，找到公司新闻，这样便可以下载苹果公司的新闻信息，如图 2.5 所示：同样获取其他五个公司的新闻。

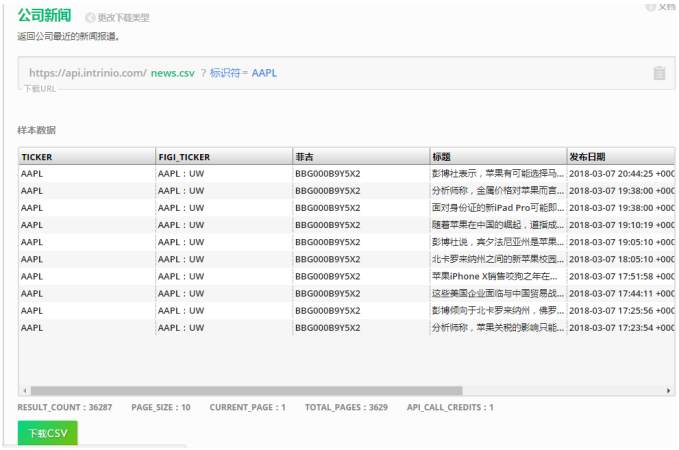


图 2.5 AAPL 新闻数据

## 2.2 新闻数据的处理

### 2.2.1 新闻数据的初步处理

本实验将使用爬虫技术来爬取从 2015 年 1 月 2 号到 2017 年 12 月 29 日之间共 755 个工作交易日的苹果公司及五家科技公司的新闻信息，并保存成为 CSV 文件形式，这些在网上发布的新闻信息文本不具有规范性，一般存在着重复啰嗦、成分残缺、用词不当、前后矛盾、含有噪声、不文明用语等问题，这些语句问题的存在不利于数据的分析。而在获取新闻时是按照一定的规则处理新闻文本而且这些新闻文本数据的结构化相对整齐，同时获得的文本数据大多数是通过金融网页获取的，这样只需处理掉重复啰嗦、无意义符号的文本信息从而使数据质量得到提高。

通过 Excel 打开新闻文本，可以看到需要的公司名称、时间列表、新闻内容等信息。每天包含几十条新闻信息，这样便会有上万条信息信息等待处理。虽然 Excel 具有数据处理了功能，可以十分方便的对表格中大量数据进行运算、统计（筛选、分类汇总等），而且还可将表中的数据按需求转换成直观的图表形式；可以将多个单元格进行合并，同时也可以将一个表格按不同要求进行分离。但是 Excel 的缺点在于容量小、并发性差、速度慢。访问 Excel 时，系统需要把整个文件载入内存再处理，如果数据量很大，内存消耗会很大甚至无法打开；这样在处理文本信息时会在打开 Excel 中浪费大量时间，造成电脑的卡顿。因此在处理大量数据时本实验选用 Python 进行数据的处理，在 Python 语言中有一个 Pandas 模块。Pandas 对非空值数据计算速度很快，据不完全统计 9800 万数据也只需要 28.7 秒。对于控制的数据处理也只需要 350 秒左右，这样便可以快速的进行数据处理。

Pandas 提供 DataFrame.describe 方法查看数据的基本信息（DataFrame 为 pandas.read 读取数据后保存的数据格式，为了方便常使用 df 表示），包括数据查看（通常使用 DataFrame.info（）可以查看数据的前五行信息）和行列统计。由原数据一般都会包含一些空值（Null），这样对数据分析的效率和时间产生影响，通过预览了数据后，需要对这些无效数据进行处理。使用 DataFrame.isnull()函数判断数据中的空值，与其相反的是 DataFrame.notnull()函数，Pandas 会对表中的所有数据进行 null 计算，并以 True/False 作为结果进行填充。这样在得知那些数据为空值后便进行数据空值的填充。其次对空值进行数据的填充。一般情况下使用数字零对空值进行填充，这样不会破坏原始数据之间存在的规律，保证了建模后结果的准确性。根据数据的结构和实验要求等也会使用平均数。最大值等其他描述性统计量来代替空值。用数字零对空值进行填充所用到的函数为 DataFrame.fillna(0)。如果只想在全部数据的列中进行数据数据填充，需要加上 axis 和

how 两个参数描述。

### 2.2.2 新闻特征转化为特征向量

经过预处理后的文本特征需要处理成计算机能认识的语言，也就是说将汉字信息转化为特征向量信息，而自然语言处理便可以将文字转化为特征向量[49-50]。因此该实验中主要应用自然语言处理（Natural Language Processing，简称 NLP）对新闻信息进行特征处理。

自然语言处理是语言学领域、人工智能领域和计算机科学领域的分支学科，是语言学、计算机科学、人工智能关注人类（自然）语言和计算机之间的相互作用的领域。主要研究的目的是如何让计算机运用和处理自然语言[51]（即人们日常使用的语言）。自然语言处理并不是像一般地语言学专家去研究自然语言，而是专注于研制能够有效地实现自然语言通信的计算机系统，尤其是软件系统。因而它是计算机科学的一部分。自然语言处理在广义上分为两部分，其一是自然语言理解，是指能让电脑读“懂”人类的语言；另一部分为自然语言生成，是指把计算机数据转化为自然语言[52]。由于自然语言具有其形式与其意义之间是一种多对多关系的特点。但从计算机处理数据的角度来看，必须消除这种歧义，便被人们认为这是自然语言理解的中心问题，即要把具有潜在歧义的人类语言输入转换成某种无歧义的计算机语言的内部表示。

近年来，随着计算机和人工智能领域的迅速发展，自然语言处理也有了重大的突破，出现了各类的智能机器人。如，在 2014 年 9 月 16 日江苏卫视的《芝麻开门》节目中迎来了首位智能机器人的参加，该机器人由百度研发并命名为叫做小度。小度在节目中不仅与主持人互动调侃，而且还以准确地回答和迅速的反应答对 40 道包含影视、历史、文学和音乐类型的题目闯过四关，其表现全场观众惊叹不已。智能机器人的优越表现可以看出自然语言处理的飞速发展和取得的优异成绩。

文本分析是自然语言处理算法的主要应用领域。但文本分析的原始数据却无法直接输入到算法中，因为这些原始数据是一些列符号，而大多数算法的期望输入是固定长度的数值特征向量而不是长度不一的文本特征。为了解决这些问题出现了多种工具进行文本分析，本文主要介绍几种常用的处理方法：

（1）词袋模型（Bag of Words）

Bag of Words 顾名思义就是将某些 Word 打成包，就像人们经常在收拾物品时把同一类的物品装到一个盒子里一样，或是即使是随意打包一些物品时也是为了方便寻找或携带，在对大数据处理时为了能够携带数据中的大量信息，与其一个个处理数据倒不如对数据进行打包容易一些[53-54]。在信息检索的应用中，Bag of Words 模型对于一个文本来说忽略其语法、语序和句法，将其看做一个词的集合或者说是一个词的组合，每一个

文本的词都是相互独立的，不依赖与其他词的出现与否，或者将一篇文章转化为任意一个词都不会受前后句子的影响而是独立存在的。例如有下面两个文本：

1：Lee likes to play basketball,Chen likes too.

2：Lee also likes to play football games.

基于文本内容，构造一个词典：

Dictionary={1:“Lee” ,2.“like”,3.“to”,4.“play”,5.“basketball”,

6.“also”,7.“football”,8.“games”,9.“Jim”,10.“too”}。

这个词典一共有 10 个不同的单词，需要根据字典的索引序号可以得到一个 10 维向量表示但某个单词出现的次数

1：[1,2,1,1,1,0,0,0,1,1]

2：[1,1,1,1,0,1,1,1,0,0]

Bag of Words 实现步骤：

步骤 1：将给定的大数据进行聚类处理，找到这组数据适当的聚类中心点（Vocabulary）。

步骤 2：寻找每一个训练集在改聚类中心的一个低维表示。

步骤 3：在得到每一个数据集的低维表示后选择合适的模型进行训练。

步骤 4：对新来的样本映射到聚类中心，然后利用训练好的模型就行训练。

（2）Word2vec（Word to vector）

Word2vec 也被称为 word embeddings，中文译为“词向量”， 其主要作用就是将自然语言中的字词转为计算机可以理解的稠密向量（Dense Vector）[55]。在 Word2vec 出现之前，自然语言处理经常把字词转为单独的离散的符号，也就是 One-Hot Encoder。Word2vec 主要分为 CBOW（Continuous Bag of Words）和 Skip-Gram 两种模式。CBOW 是从原始语句推测目标字词；而 Skip-Gram 正好相反，是从目标字词推测出原始语句。CBOW 对小型数据库比较合适，而 Skip-Gram 在大型语料中表现更好。Word2vec 算法是一个逻辑回归（分类）问题，使用最大似然估计。在已知某个单词 h，要最大化下一个单词$w_t$出现的概率，使用 softmax 函数做分类，则问题的数学描述如下：

$$\mathrm{P}\left(w_t/_h\right) = softmax\left(score(w_t,h)\right) = \frac{expscore(w_t,h)}{\sum_{w'\in\vartheta} expscore(w_t',h)} \qquad (2\text{-}1)$$

这里$\mathrm{score}(\mathrm{w_t},\mathrm{h})$计算上下文是 h 时单词是$\mathrm{w_t}$的概率，显然上面的除法和指数计算是比较耗时，将上述公式取对数进行简化，目标函转换如下：

$$J_{ML} = logP\left(w_t/_h\right) = score(w_t,h) - logexpscore(w_t',h) \qquad (2\text{-}2)$$

其基本模型就是简单的神经网络，如图 2.6 所示：

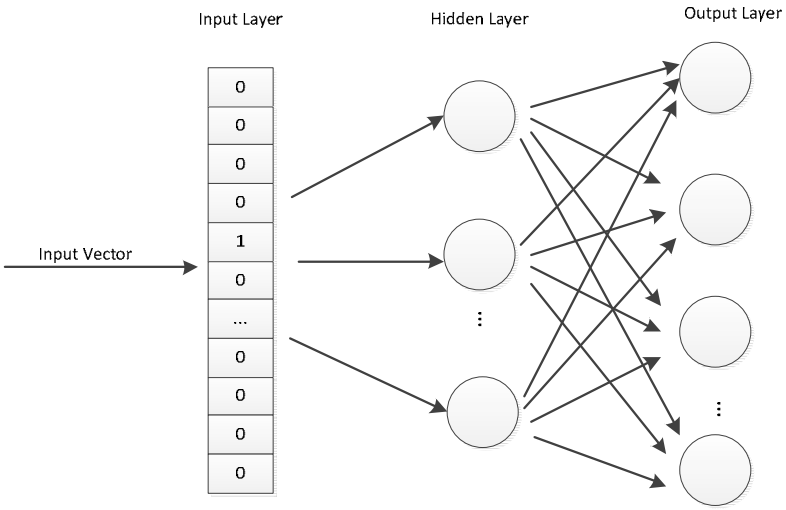



图 2.6 神经网络结构图

输入是 One-Hot Vector，Hidden Layer 没有激活函数，也就是线性的单元。Output Layer 维度跟 Input Layer 的维度一样，用的激活函数为 Softmax。最后获取的 dense vector 就是 Hidden Layer 的输出单元。也就是 Input Layer 和 Hidden Layer 之间的权重。

（3）TF-IDF（term frequency–inverse document frequency）

TF-IDF（term frequency–inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF(term frequency, 词频)指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化（分子一般小于分母区别于 IDF），以防止它偏向长的文件[56-57]。IDF(inverse document frequency,逆向文件频率)是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语文件的数目，再将得到的商取对数得到：

$$\mathrm{idf}_i = log\frac{|D|}{|\{j:t_i \in d_j\}|} \tag{2-3}$$

式中： $|D|$ ----- 语料库中的文件总数

$|\{j: t_i \in d_j\}|$ ----- 包含词语$t_i$的文件数目（即$n_{i,j} \neq 0$的文件数目）

其中，如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用$1 + |\{\mathrm{j}: \mathrm{t_i} \in \mathrm{d_j}\}|$然后$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$某一特定文件内的高词语频率，以及该词语在整个文件集合中的低频率词语，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

某一特定文件内的高频率词语，以及该词语在整个文件集合中的低频率文件，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语。

（4）pysentiment

该模块为 Python 情绪分析实用程序，主要用于文本数据的情绪分析，根据文本的

内容进行数据的评分，这样就将文本信息转化为数字特征向量，这也是本文应用的模块，对股票新闻数据进行情绪分析并基于一定的评分。pysentiment 模块是一个在字典框架中进行情感分析的库。库中提供了两种词典，即哈佛 IV-4、Loughran 以及 McDonald Financial Sentiment Dictionaries，这是用于一般和金融情绪分析的情感词典。如：快乐，痛苦，美德和恶习，这些词通常也被分类为正面或负面，快乐、美德表示正数，痛苦、恶习表示负数。

pysentiment 是 Python 中自带模块，使用前只需要通过 pip install pysentiment 命令便可以安装，安装后在程序前导入此模块便可应用。图 2.7 便是将新闻信息根据情绪分析转化的向量特征。其中 POLARITY 便是根据新闻信息的情感色彩转化的数据特征，而 SUMMARY_SCORES 则是对文本情感分析的结果。

In [4]: aapl_news

Out[4]:

| | TICKER | DATES | DATE | SUMMARY | SUMMARY_SCORES | POLARITY |
|---|---|---|---|---|---|---|
| 0 | AAPL | 42514 | 2016/5/24 | More challenges than cheer for Apple chief on ... | {'Subjectivity': 0.199999960000008, 'Negative'... | -0.999999 |
| 1 | AAPL | 42515 | 2016/5/25 | India says Apple must sell locally-sourced goo... | {'Subjectivity': 0.0, 'Negative': 0, 'Polarity... | 0.000000 |
| 2 | AAPL | 42520 | 2016/5/30 | India discussing Apple's request for FDI rules... | {'Subjectivity': 0.0, 'Negative': 0, 'Polarity... | 0.000000 |
| 3 | AAPL | 42522 | 2016/6/1 | Apple plans to sell $1 bln of 30-yr bonds in T... | {'Subjectivity': 0.0, 'Negative': 0, 'Polarity... | 0.000000 |

图 2.7 新闻向量特征转化

ACLS、CAMP、CSLT、CYOU、RPD 五家公司的新闻数据也可以通过爬虫或 intrinio 爬取。

## 2.3 股票历史数据库的建立及预处理

### 2.3.1 历史数据库的获取

数据集的选择及表达形式的有效性都将会对模型的设计和训练后的结果有着重要的影响。本实验选取的为苹果公司（AAPL）从 2015 年 1 月 2 号到 2017 年 12 月 29 日 755 个工作日的历史数据，包含开盘价、收盘价、最高价、最低价四个股票基本指数[]。

（1）开盘价

开盘价为当天股市开市后股票的第一笔成交价格。如果当天开盘后 30 分钟内没有股票的交易，则会将前一天股票的收盘价作为当日的开盘价。

（2）收盘价

指每一天股市在截止收盘时股票最后一次成交的价格。如果当天无交易则以前一天的收盘价作为今日的收盘价

（3）最高价

指股票在当天交易时最高的一次价格作为股票的最高价。

（4）最低价

指股票在当天交易中最低的一次交易价格作为股票的最低价。

由于大部分文章仅仅对股票价格的单一指数进行预测，如预测股票价格的开盘价或是最高价等属性。但是股票的属性之间存在这一定的联系：将收盘价和同一日中的最高价、最低价和开盘价进行比对，根据情况的不同来判断未来股票价格的走势特点：若收盘价比开盘价低，则说明该支股票正面临着市场压力，需要有调整的要求。若收盘价比开盘价高，则说明该支票股具有一定的抵抗性。其中，若在下跌中低开高走，则表明有较低的资金流入；但若是高开高走，则表明该支股票正处于强势上涨过程；

本实验将对这四个股票指数作为历史数据集，增加了数据间的联系性，提高预测结果的准确度。

### 2.3.2 历史数据的处理

噪声和数据的非平稳性是股票数据的主要特征，同时由于历史数据比较多可能在爬取过程中出现丢失或数据异常的情况。这些因素会严重影响训练的时间和预测的准确精度，因此在进行构建模型之前，需要对历史数据进行初步的预处理。数据预处理有多中方法，实验中主要应用数据的清理、数据变换两种方法。数据清理主要的目的是去除原始数据中的噪声、数据异常和数据缺失等[58]；数据变换包括数据的归一化、变换、降维等操作。本实验的预处理主要包括数据缺失值填充、数据异常处理、数据平滑处理、归一化四种方式。

数据异常处理主要针对的是历史数据中小于数字零的值。对于缺失值和数据异常的处理方法主要是将历史数据附近的数据进行替换或填充如果数据串中只有一个数据缺失就用缺失数据前面的历史数据进行填充，若存在两个以上的数据缺失则选择缺失数据前后的历史数据进行填充替换。处理的主要原因是：对于金融时间序列的历史数据，特别是股票价格的历史数据，数据的整体波动次数多、波动比较剧烈，一个缺失数据或是异常数据的影响比较大，预测的数据往往会与真实数据相差比较大，因此本实验主要利用该种方法进行数据的填充和替换，避免造成影响。

数据平滑处理：为了提高预测的准确性，需要对时间序列的历史数据进行去噪处理也就是对历史数据进行平滑处理[59]。主要方法有小波变换、kernel 平滑、傅里叶变换、滑动平均等。kernel 平滑是一经典的平滑方法，在时间序列的预测中取得了良好的去噪效果。本实验将核平滑应用在数据的处理上。基本思想是利用 kernel 函数对数据的中心点附近的数据分配权重，距离中心点越小，分配的权重越大。公式如下：

$$m_h(x) = \frac{\sum_{t=1}^{T} K_h(x-X_t)Y_t}{\sum_{t=1}^{T} K_h(x-X_t)} \tag{2-4}$$

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}} \frac{n!}{r!(n-r)!} \tag{2-5}$$

式中：$K_h(x)$ ----- 核函数
$\{K_h(x)\}$ ----- 历史数据
$X_t$ ----- 样本中心点$Y_t$的坐标
$m_h(x)$ ----- 历史数据$\{Y_t\}$经过处理后所对应的序列值
X ----- 中心点与附近样本点的距离
H ----- 核平滑的宽度

## 2.4 数据库的处理

### 2.4.1 历史数据和新闻数据融合

本文数据集内包含股票公司的新闻信息和股票的历史数据，因此需将两种数据集融合成为一个完整的数据集，然后在对完整的数据集做进一步的数据处理。由于数据集包含新闻数据和股票历史价格数据两部分内容，因此在输入到网络模型前需要将两种数据集进行融合。由于股票的交易日为正常的工作日，也就是说股票历史数据一周只有 5 天的数据信息，一年只有 270 天左右的历史数据信息。而新闻信息每天都有，这样便会造成两种数据集的长度不一致。而数据融合的基本条件为：多种数据的长度必须保证一致。本文将历史数据进行扩充到达新闻信息的长度，为了扩充后的数据尽量与原历史数据有一致的波动性，将历史数据中周六、周日等节假日的数据用最近的历史数据进行填充，比如周六的股票价格使用周五的股票价格进行填充，而周日的历史价格便使用周一的股票价格进行填充。这样既保证了填充的数据与原始数据有一致的波动性，也起到了数据扩充的作用。

数据融合按抽象的层次主要分为三大类别：数据级融合、决策融合、特征级融合。数据级融合又被称为传感器融合，指利用计算机技术对按时序方式获取的若干信息在一定准则下进行自动化分析，以完成所需要的评估和决策而进行的数据信息处理技术[60]。它直接让计算机从传感器中采集未经处理的原始数据进行融合，数据融合时造成的信息缺失比较少，但是受环境影响比较大，而且需要各传感器间需要具有绝对的量级匹配度才能进行计算。由于股票需要保证数据的准确性和完成性，噪声信息相对较少，因此该实验应用数据级融合技术。

在数据融合时选择股票历史价格的数据（包含开盘价、收盘价、最高价、最低价）和经过自然语言处理得到新闻数据的特征向量集进行数据融合，在融合时需要注意必须要保证数据集的 shape 形状必须一样，否则再融合时会出现报错。如图 2.8 所示是融合后的数据集中的一部分信息。

Out[9]:

| | OPEN_PRICE | CLOSE_PRICE | HIGH_PRICE | LOW_PRICE | POLARITY |
|---|---|---|---|---|---|
| 0 | 109.07 | 109.58 | 109.6200 | 107.310 | 0.000000 |
| 1 | 108.01 | 110.38 | 111.0136 | 107.550 | 0.000000 |

图 2.8 数据融合

图中可以看到融合后的数据集，POLARITY 中为零的是由于在此时此刻没有新闻信息的公布，则实验中将使用数字零对缺失值进行数据补充。

### 2.4.2 数据库的预处理

数据集的预处理好坏直接会影响到实验结果的好与坏，为了提高结果准确性的精度和训练时的收敛速度需要对数据进行归一化处理，使得预测后的数据变化率不会过大。本试验使用的归一化方法为 MinMaxScaler（最小最大标准化），该标准化的方法是将全部数据集缩放到给定数据的最大值和最小值之间，通常是 0 与 1 之间的数据。这种方法的好处是当数据集的标准差非常小时，数据中存在稀疏数据（零元素）需要保住零元素的存在；也就是说当数据间数据差距特别大时，在模型学习时往往会忽略掉数值很小数据的存在，导致模型训练的不完整从而导致模型训练失败。

MinMaxScaler 公式为：

$$\mathrm{X_{std}} = \frac{X - X.\min(axis=0)}{(X.\max(axis=0) - X.\min(axis=0))} \tag{2-6}$$

$$\mathrm{X_{scaled}} = \mathrm{X_{std}} * (\max - \min) + \min \tag{2-7}$$

式中： $X.\min$ ----- 数据中最小值

$X.\max$ ----- 数据中最大值

## 2.5 网络模型参数初始化

深度网络模型参数初始化的目的是减弱非凸优化目标函数对初始值的依赖性，尽可能的避免所求的解（即模型的参数）过早的陷入局部最优。模型参数的初始化主要有四种：零化（初始参数设置为零）、完全随机（服从于高斯分布）、带尺度约束的随机（尺度因子在-1 与 1 之间）和 Xavair-glorot（不同分布下的半随机初始化）。研究表明，带有尺度约束的 Xavair-glorot 初始化参数的方式是最好的。另外在卷积网络中，超参数的选择也非常重要，例如一般倾向于使用小滤波器（如$3 \times 3$的尺寸）和小步长（Stride），这样就不会减少参数的数量，从而提升整个网络的准确率。此外，常用的池化尺寸是$2 \times 2$，一维卷积池化尺寸为$2 \times 1$，可以保持平移不变性的同时，有效的降低参数量。本实验中为了保证数据的时序性，需要对模型赋予不同的权重，距离预测值的近的初始权重要比距离预测值远的初始权重赋值要大，体现出近距离的股票价格和新闻数据对当天的股票价格影响较大，距离较远的股票价格和新闻数据对当天的股票价格影响较远的特性。

## 2.6 本章小结

数据库一直都是深度学习的一个重要部分，数据的质量和数据量的大小决定着训练后模型的好与坏。本章节介绍了四个主要的内容：第一部分介绍了新闻数据库的建立，将通过两种方式进行新闻数据的获取，Python 爬虫和通过 intrinio 的 API 获取数据，也简单的叙述了两种获取方法的原理和实验步骤；第二部分主要介绍了通过自然语言处理将文本数据集转化为数字特征向量，并简单介绍了几种常用的方法和本实验主要应用的方法。由于本文处理后的数据集并不是非常大，毕竟新闻的数据集在获取上存在困难，任何网站也不会大量保存某个公司三年之前的新闻，因此本篇文章主要通过实验来验证新闻对股票价格波动有促进作用。如果有大量的数据库本实验的效果会更好，预测的准确性会更精确。第三部分内容主要简单介绍了苹果公司历史数据的获取及预处理过程；第四部分介绍了数据融合，将历史数据与新闻信息进行融合，然后对融合后的数据进行简单的数据处理即归一化处理，便于以后模型的建。

# 第 3 章 基于 CNN 股票价格波动预测的研究

卷积神经网络目前是深度学习中应用最为广泛的模型，卷积神经网络大部分都应用于图像处理方面、实体识别、文本分类方面，但是卷积神经网络还可以应用于语音、金融等一维时间序列的处理方面。本章实验主要应用于卷积神将网络对苹果公司的股票价格波动进行预测。

本章实验主要通过以下三方面部分进行：

1 通过与传统时间序列模型（自回归积分滑动平均模型）进行对比，验证卷积神经网络对股票价格波动的预测效果比传统时间序列模型好。

2 对不同滑动窗口长度对股票价格的影响进行实验对比，选取效果好的滑动窗口长度。

3 通过对多种新闻信息的处理方法进行比较并选择出最优的处理方法作为下次实验文本的处理方式。

图 3.1 为本实验的基本流程图。


图 3.1 卷积神经网络预测流程图

## 3.1 国内外相关研究

在金融行业中所产生的数据是巨大的并且是非线性。为了模拟这种动态数据，需要能够分析隐藏模式和潜在动态的模型。深度学习算法能够通过自学习过程来识别和利用数据中存在的交互和模式。与其他算法不同，深度学习模型可以有效地对这些类型的数据建模，并且可以通过分析数据中的交互和隐藏模式来提供良好的预测。文献[61]可以看到各种深度学习模型在多元时间序列分析中的应用[61]。文献[62]介绍了使用神经网络模型对金融时间序列进行第一次建模的尝试[62]。在这次实验中试图模拟一个神经网络模型解码 IBM 的资产价格变动的非线性规律。虽然这项实验有很大的局限性，但它有助于建立针对 EMH 的证据[63]。文献[64]提到使用卷积神经网络对三家公司（两家 IT 公司，一家制药公司）的股票价格进行预测，文中基于滑动窗口方法对短期的未来股票价格进行预测[64]。窗口大小被固定为 100 分钟，重叠了 90 分钟的信息，并且预测 10 分钟后的股票价格。通过与自回归积分滑动平均模型(Autoregressive Integrated Moving Average Model,简记 ARIMA)进行对比，实验效果证明卷积神经网络模型预测的误差百分比远远低于 ARIMA 模型。该文章使用在特定时刻提供的信息进行预测。这是由于 CNN 不依

赖任何以前的信息进行预测的原因。它仅使用当前窗口进行预测，这使模型能够理解当前窗口中发生的动态变化和模式。上述文章详细证明了卷积神经网络对股票价格波动预测效果比传统时间序列效果好。但是上述文章中并没有考虑到公司的新闻因素对股票价格波动的影响因素。文章开始提到影响股票价格波动因素很多，诸如：国际形势、战争、国内重要的政治事件、国家重大经济政策等等。但是这些因素的影响一般都会以新闻的信息表现出来让人们看到。本实验将考虑这些因素的影响来对未来股票价格波动进行预测。

本章节实验利用卷积神经网络具有时间和空间上的平移不变性，将卷积神经网络的思想应用在股票价格波动的预测（时间序列的预测）上。通过将历史数据和新闻因素双方面的影响来对股票价格波动进行预测。由于卷积神经网络由多层卷积层和池化层（下采样层）顺序连接构成的神经网络，能够准确地从原始数据中获取有效地特征描述，同时由于权值共享可以节省大量的训练时间。根据以往的实验表明卷积神经网络不能很好的预测时序信息的数据，因此需要通过加入滑动窗口技术来让数据变为时序信息。

## 3.2 基于股票价格预测的卷积神经网络模型

卷积神经网络是由卷积层和池化层构成的特征抽取器。在卷积神经网络的卷积层中，一个神经元仅仅与部分邻层神经元相连接，并通过权值共享和局部感知两种手段使其能够实现对未来图像的仿射不变性。权值共享是指整张图像使用同一套局部权值参数处理，图像的局部特征在整张图像上具有位置无关的特性这也意味着通过相同的权值模板可以对整张图片某一个特征进行特征提取。也就是在同一个卷积层中，一般都包括多个特征平面(featureMap)，而每个特征平面都由一些神经元构成，在同一特征平面上的神经元权值共享，而共享的权值便是卷积核。卷积核的大小一般都以随机小数的形式进行初始化。通过网络的不断训练，卷积核将学习得到一系列合理的权值。卷积核（共享权值）的优点是减少网络各层神经元间的连接，而且同时又降低了训练过程中的过拟合风险。局部感知的特性主要受生物视觉系统局部敏感的启发而来，一般认为人对外界环境的认知都是从局部到全局，而图像的空间联系也是局部的像素联系较为紧密，而距离较远的像素相关性则较弱。因而，每个神经元其实没有必要对全局图像进行感知，只需要对局部进行感知，然后在更高层将局部的信息综合起来就得到了全局的信息。如下图 3.2 所示：左图为全连接，右图为局部连接。

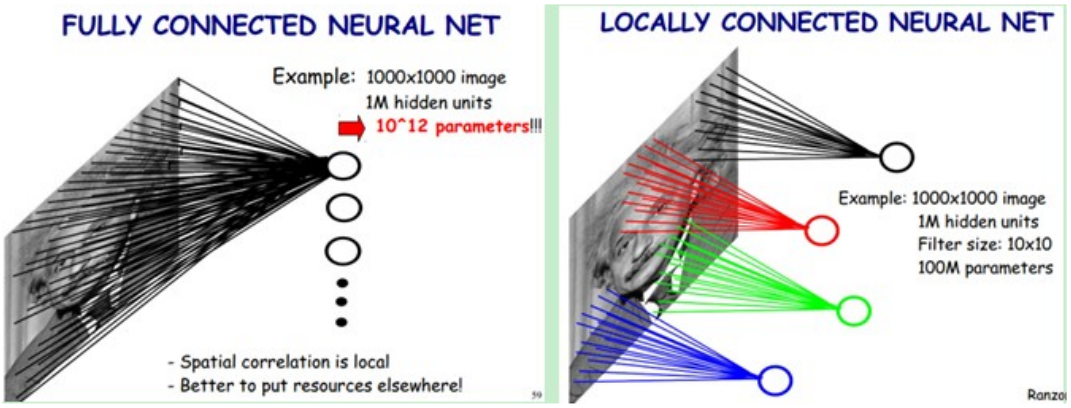



图 3.2 卷积神经网络的局部感知

池化层也被称为下采样层，是将图像按窗口的打小划分为不重叠区域，然后对每一个区域中的元素进行聚合。本质上池化操作执行特征类型或空间的聚合，降低空间的维度。其主要的思想是：减少数据的计算量，刻画系统的平移不变性；减少下一层的输入维度（核心是有效的降低下一层的参数变量），可以有效的控制过拟合风险。池化的操作有多种形式：最大池化、范数池化、平均池化和对数池化等，在实验中经常使用的是最大池化，就是该区域范围内的最大值。对窗口为 2x2 的池化进行操作，处理完的图像大小其长、宽各位原来的一半，也就是输出的尺寸为原图片的 1/4 大小。如图 3.3 所示：

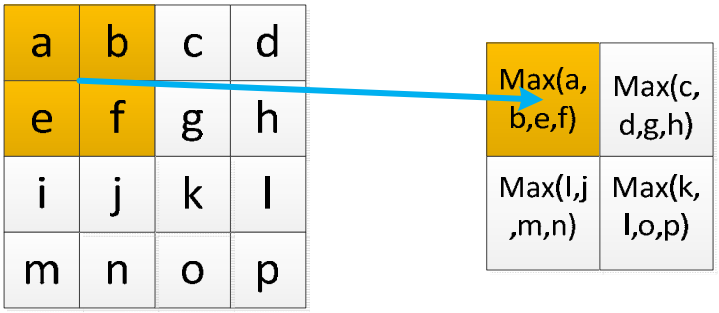



图 3.3 最大池化示意图

根据卷积可以约减不必要的权值连接，引入局部或稀疏连接，带来的权值共享策略大大地减少参数量相对地提高了数据量，从而可以避免过拟合现象的发生，而且由于卷积操作的不变性使模型学到的特征具有拓扑对应性、鲁棒性。

根据卷积神经网络的这些特性，本实验将采用卷积神经网络对股票价格波动进行预测。

### 3.2.1 输入样本选择

由于卷积神经网络不具有记忆功能，因此需要通过对输入数据进行处理来保证卷积神经网络具有记忆功能。本实验中选择使用滑动窗口技术来保证卷积神经网络具有记忆

功能。滑动窗口即通过规定窗口长度截取一定范围内的历史数据，作为卷积神经网络模型的输入，输出的值为预测的数据也就是模型预测的该输入下的股票价格值，窗口的长度被称为时间步长。滑动窗口就是时间步长固定，随着时间的移动来截取数据。窗口的截取方式主要包含三种方式：倾斜窗口、界标窗口、滑动窗口。由于股票价格受前一时刻影响比较大，而随时间的延伸对此时刻股票价格的影响也将随之衰弱。因此实验将主要应用滑动窗口技术从历史数据中截取数据条作为研究对象，这样弥补了卷积神经网络不能准确的预测时间序列的缺点。滑动窗口只有一个参数就是时间步长 N,也就是滑动窗口的宽度。该实验主要通过前 N 个历史数据来预测第 N+1 个股票的价格，也就是将前 N 个数据作为输入样本送到模型中，预测出来的结果为第 N+1 天股票的价格。根据这种方式划分建立了其大量的数据集，然后划分出训练集和测试集。训练集和测试集的标签为第 N+1 天的股票开盘价、收盘价、最高价、最低价四个指数的历史数据。本实验中滑动窗口大小将以固定为 10 天历史数据为例，重叠了 9 天的信息，并且预测将来 10 天后的股票价格。滑动窗口原理图如图 3.4 所示：

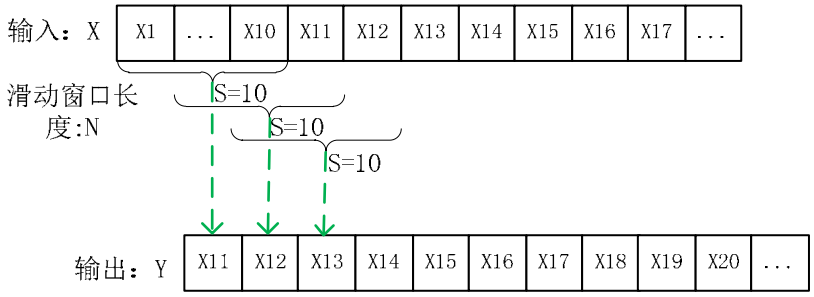



图 3.4 滑动窗口原理图

若将数据集作为卷积神经网络的输入还需要将数据集转化为满足卷积神经网络要求的形式：输入 shape（samples，steps，input_dim）的 3D 张量。

### 3.2.2 构建卷积神经网络的股票价格预测模型模型

实验的数据集为一维数据，因此本实验采用一维卷积神经网络，一维卷积神经网络与二维卷积神经网络最大的不同主要表现二维卷积神经网络选取的卷积核为 N*N 的卷积核，而一维卷积神经网络选取的为 N*1 的卷积核；一维卷积过程如图 3.5 所示，显示的为 Valid 卷积，滤波器在卷积神经网络中也被称为卷积核[65]，图中卷积核的长度为$2 \times 1$。

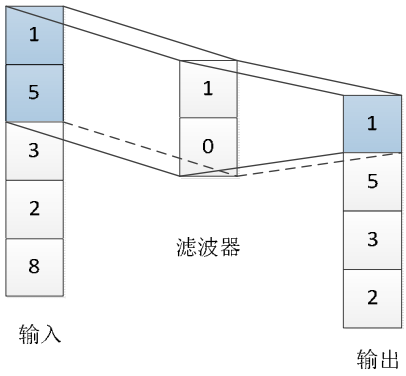



图 3.5 一维卷积图示

而二维卷积过程如图 3.6 所示，其中卷积核大小为$3 \times 3$。

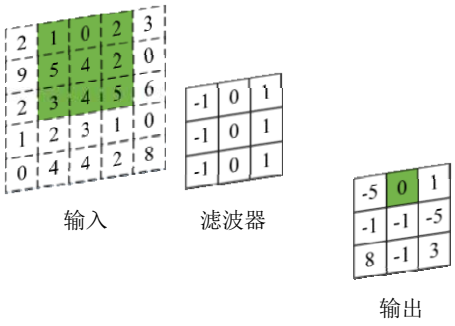



图 3.6 二维卷积过程

二维卷积神经网络处理的为图片，而一维卷积神经网络处理的为一维数据组，如：金融数据、语音等。由于影响卷积神经网络训练结果的因素主要包括卷积层数目、每层卷积核的大小和数量、池化层数目及过滤器的类型选择。对于时间序列的数据，某一时刻的数据受前一时刻的影响比较大，但是受很长时间之前因素的影响却很小，因此这也决定了输入数据滑动窗口长度不会太长，卷积神经网络的卷积层数和池化层数也不会很多，根据这些特点本实验将采用两层卷积网络和两层池化网络进行实验。

由于本实验的数据维度比较小，所以简化了卷积神经网络的结构，尤其是卷积核的大小，本实验中选取卷积核大小为 2x1 的卷积核，运算过程如图 3.7 所示，降低了网络计算的复杂度。这也是与处理 2 维图像卷积网络结构的最大不同之一。如果每次卷积、池化过程中层数不同，这样会增加网络结构的复杂度，不利用网络结构的训练，因此本实验都选用相同大小的卷积核和过滤器。如图 3.7 所示是实验中其中一个网络训练完成后打印出来的卷积网络的结构图，通过结构图可以清楚地看清每一层的输出 shape 和每一层参与训练的神经元个数和一共参与训练的神经元总数。为了更加方便直观的查看卷积神经网络的输入、输出的 shape 和网络运行的规律，将通过流程图将网络结构图表现出来，如图 3.8 所示：

```
print(model.summary())

_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_28 (Conv1D)           (None, 9, 64)             704
_________________________________________________________________
max_pooling1d_23 (MaxPooling (None, 4, 64)             0
_________________________________________________________________
conv1d_29 (Conv1D)           (None, 3, 128)            16512
_________________________________________________________________
max_pooling1d_24 (MaxPooling (None, 1, 128)            0
_________________________________________________________________
flatten_2 (Flatten)          (None, 128)               0
_________________________________________________________________
dropout_10 (Dropout)         (None, 128)               0
_________________________________________________________________
dense_10 (Dense)             (None, 4)                 516
=================================================================
Total params: 17,732
Trainable params: 17,732
Non-trainable params: 0
_________________________________________________________________
None
```

图 3.7 卷积网络框架

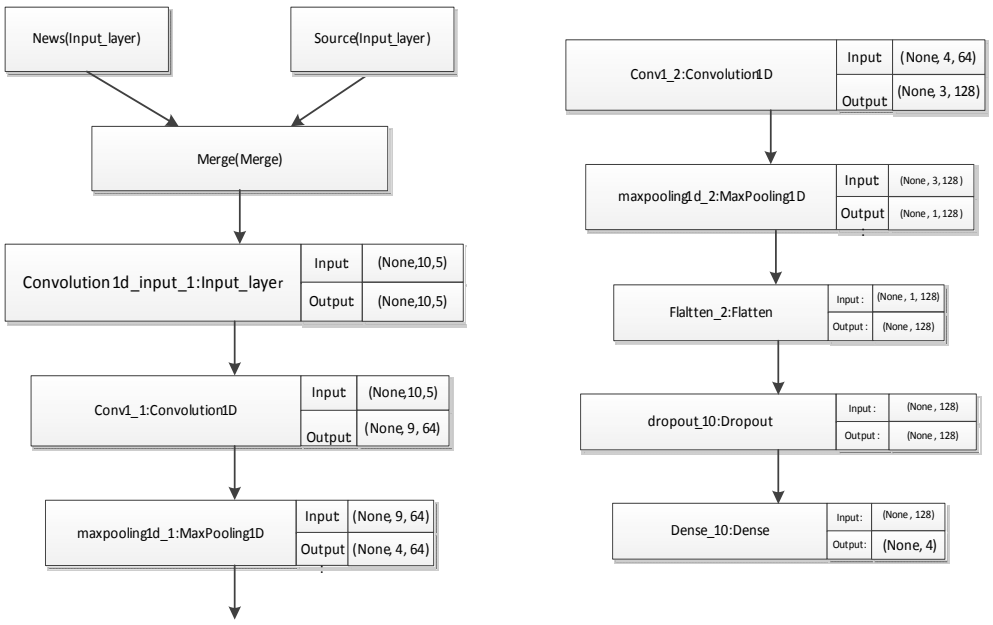



图 3.8 卷积网络实验流程图

通过图 3.8 可以直接看到在卷积神经网络模型之前需要将经过处理自然语言处理后的新闻特征和股票历史价格进行融合作为一个整体的数据集。然后再作为卷积神经网络的输入。卷积神经网络的输入为（None,10,5）,其中数字 10 表示时间窗的长度，也就是通过前 10 的数据集来预测下一天的股票的开盘价、收盘价、最高价、最低价这四个股

票指数。数字 5 表示数据集的维度，有开盘价、收盘价、最高价、最低价以及通过第二章中将新闻根据感情分析处理的特征。经过一次卷积后输出的 shape 为（None,9,64）。在卷积过程中使用的步长为 1 的卷积核，卷积核的个数为 64，这也说明了卷积神经网络的另一个特点：多卷积核。这样使模型在卷积过程中学到不同的数据特征。经过卷积层的输出将会作为池化层的输入，池化中选的为 2*1 的核，经过池化后的输出为（None,4,64），经过两次卷积、池化后，卷积的输出为（None,1,128），最后通过一个由四个神经元组成的全连接层作为网络模型的最后输出。

### 3.2.3 卷积神经网络基本原理概述

20 世纪 60 年代，Wiesel 和 Hubel 在研究猫视觉的皮层中用于方向选择和局部敏感神经元时发现了其独特的网络结构可以有效地降低反馈神经网络的复杂性，纽约大学的 Yann LeCun 教授与 1989 年根据这种特点提出来卷积网络（Convolutional Network）也被人们称作为卷积神经网络（Convolutional Neural Network,CNN）[66]，它是一种专门处理具有高维网格型数据（即张量 Tensor）的前馈神经网络[67]。如处理时间序列数据（可看为是在时间轴上有规律采样而形成的一维网络结构）及图像数据（可被看作二维的像素）有着良好的处理效果，因为它的神经元可以响应部分周围范围内的邻层神经。卷积神经网络最擅长处理大型图像数据，例如由二维数据表达的灰度图像、三维图像（高、宽、RGB 通道）表示的彩色图像等，也在该领域取得了举世瞩目的成绩。2012 年在 ILSVRC 图像识别竞赛中，Alex Krizhevsky 运用卷积神经网络设计的 AlexNet 而成为人们的焦点，以压倒性优势取得了该场比赛的冠军，体现了卷积神经网络在图像分类识别问题上的强大优势，瞬间使卷积神经网络成为深度学习中人们追捧的网络模型之一。卷积神经网络在诸多应用领域都表现出优异的成绩，如行人检测、信号处理、人脸识别、表情识别、图像分类、文本识别等均有新的发展与成果，而且成绩也是比较突出的。

“卷积神经网络”说明该网络使用的卷积这种数学运算，是一种特殊的线性运算。卷积网络指的是那些在网络层中使用卷积运算替代一般的矩阵乘法运算的神经网络。CNN 与传统的人工神经网络算法的主要区别在于共享和非全连接[68]。通过拓补结构建立层与层间的非全连接关系来降低训练参数的数目，这也是 CNN 的主要思想。CNN 在经过多层网络的反馈训练学习之后能自动提取出隐藏在数据里的有效特征的卷积表示，然后通过逐层的卷积、池化将能够提取出隐藏在数据里的拓扑结构特性。当网络的层数变深，提取的特征也将逐渐变得抽象，最终将达到获得数据的平移、旋转以及缩放的不变形的特征表示[69]。

随着研究者的不断深入探索，越来越多的优秀网络结构[70]得以提出，简单对卷积神经网络进展以及经典网络结构进行介绍，如图 3.9 所示。

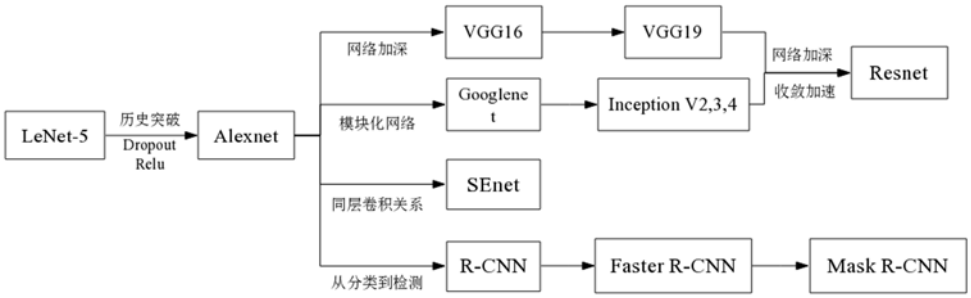



图 3.9 卷积神经网络发展趋势

(1) 卷积运算

卷积是分析数学中的一种运算。在通常的形式中，卷积运算就是对两个实变函数进行的运算方式。在数学中卷积的定义为：假设f(x)和g(x)分别为 R 上的两个可积函数，则积分为：

$$s(t) = \int_{-\infty}^{+\infty} f(\tau)g(x-\tau)d\tau \quad (3\text{-}1)$$

称为f(x)和g(x)的卷积，记为：

$$s(t) = (f * g)(x) \quad (3\text{-}2)$$

容易验证，

$$(f * g)(x) = (g * f)(x) \quad (3\text{-}3)$$

并且$(f * g)(x)$认为可积函数。卷积运算通常用星号（*）表示。

在卷积网络中，卷积运算参数的输入被称之为输入和核函数，输入称为特征映射。在机器学习应用中，输入一般都是多维数组的数据，然而核一般都是由学习算法优化后得到的多维数组参数。由于输入与核中的每一个元素都要明确的分开存储，通常假设在存储了数值的有限点集外，函数的值都将为零。这将意味着可以通过对有限数组元素的求和来实现无限求和的过程。但是经常在多个维度上对数据进行卷积运算，如将一个二维的图像 I 作为数据的输入，若需要一个二维的核 K：

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (3\text{-}4)$$

等价转换后得到：

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \quad (3\text{-}5)$$

在机器学习库中为了实验的简单方便通常选取第二个公式，因为 n 和 m 的有效取值范围更小。

卷积运算主要通过三个重要的思想来改进机器学习系统：稀疏交换（spare interactions）、等变表示（equivariant representations）、参数共享（parameter sharing）。传统的神经网络将会使用矩阵乘法来建立输入与输出关系。其中参数矩阵中的任何一参数都描述了一个输入单元与输出间的交互，即每一个输出单元都将会与输入产生交互[71]。

然而，卷积神经网络具有稀疏权重（也叫稀疏交互或者系数连接）的特征。这将会使核的大小远远小于输入数据的大小。举一个示例，对于一张具有上万个像素点图片的处理，只需通过使用几十到上百个像素点的核来检测小的有意思的特征点，像图像的边缘信息。这样数据的存储参数变少，不仅减少了模型的存储需求而且还提高论文统计效率，同时输端的计算量也将变小。如图 3.10 所示，对于一个输入单元$x_3$，当 s 是由矩阵乘法算出时，s 中所有的单元都受到$x_3$的影响，连接不是稀疏的；当 s 是有一个卷积核宽度为 3 的卷积产生时，s 中只有三个参数$x_2x_3x_4$受到$x_3$的影响。这样输入与输出之间的参数连接就极大程度上的减少了。

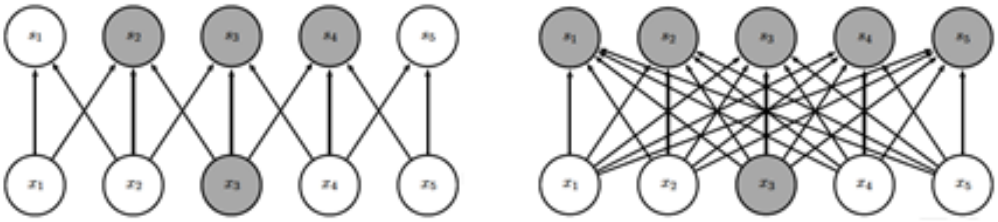


图 3.10 稀疏连接示意图

参数共享指的是同一个模型中多个函数使用同一个参数。传统神经网络计算一层输出时权重矩阵中的每一个元素只使用一次，当他与输入的一个元素相乘后就不会再使用。而卷积运算中的参数共享使得我们只需要对一个参数的集合进行改变就能够使得所有的模型都统一的进行更新，这样能够显著的减小模型的存储需求。卷积操作是一种滑动窗口滤波形式，而卷积核就是滤波器。卷积核对图像实现卷积过程示意图如图 3.8 所示：

(2) 权值更新

卷积神经网络训练的过程可简化成前向神经网络传播的计算和反向传播神经网络参数更新两个步骤。在此之前该网络需要对权值、阈值等参数进行随机的初始化或给定数值的初始化。前向传播训练主要对网路东阿输入数据进行求值保存，后向传播过程根据网络训练的输出和实际的输出计算出每层的误差值，从而根据反向传播算法从深层到浅层依次对偏置和权重进行更新，使得误差值不断降低。卷积神经网络通过对训练集不对迭代更新减少损失函数数值使其达到局部最小值或全局收敛，达到设定的指标要求后停止迭代训练结束。

卷积神经网络本质上就是神经网络的一种特殊形式，只是在前几层的网络处理上有所不同，因此可以把卷积核看做神经网络的权值W，而采样层也是卷积运算的一种形式。因此我们通过人工神经网络来了解卷积神经网络的权重更新情况。

具体权值更新的传播算法入戏：

（1） 神经网络第 l 层的输出函数为：

$$x^l = f(u^l) \tag{3-6}$$

$$u^l = W^l x^{l-1} + b^l \tag{3-7}$$

对于卷积神经网络特殊的结构和权值共享的特点，可以通过简单的误差反向传播来实现该网络全参数的更新，定义该网络的平方误差代价函数为：

$$E^N = \frac{1}{2}\sum_{n=1}^{N}\|t_k^n - y_k^n\|_2^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{n_0}(t_k^n - y_k^n)^2 \quad (3\text{-}8)$$

式中： $t_k^n$ ----- 第 n 个样本的标签$t^n$的第 k 维

$y_k^n$ ----- 第 n 个样本的输出（即预测）样本的第 k 维

训练网络的主要目标是通过不断地更新权值来使预测的数值不断地接近真实数据。

（2） 反向传播算法

反向传播算法的网络权重的更新主要是将预测值和真实值的误差传给每一层，让前面的每一层不断地更改他们的权重来是卷积神经网络的预测更加准确。具体的说就是，用神经元得$\sigma$值不断地缩放其输入数据，然后将达到的偏导数与一个负的学习率进行相乘便是该层神经元的权值更新：

$$\Delta w^l = -\eta\frac{\partial E}{\partial w^l} = -\eta x^{l-1}(\delta^l)^T \quad (3\text{-}9)$$

第 L 层的权值为：

$$\frac{\partial E}{\partial w^l} = \frac{\partial E}{\partial \mu^l}\frac{\partial \mu^l}{\partial w^l} = \frac{\partial E}{\partial \mu^l}x_i \quad (3\text{-}10)$$

$$\delta^l = \frac{\partial E}{\partial \mu^l} = \frac{\partial E}{\partial f(\mu^l)}\frac{\partial f(\mu^l)}{\partial f(\mu^l)} = \sum_k -\left(t_k - f(\mu^l)\right)f'(\mu^l) \quad (3\text{-}11)$$

第 I 层权值为：

$$\frac{\partial E}{\partial w^l} = \frac{\partial E}{\partial \mu^l}\frac{\partial \mu^l}{\partial \mathrm{w}} = \frac{\partial E}{\partial \mu^l}x^{l-1} = x^{l-1}(\delta^l)^T \quad (3\text{-}12)$$

对于卷积神经网络卷积层的每一种输出的特征图$x_j$为：

$$x_j^l = f\left(\beta_j^l down\left(x_j^l\right) + b_j^l\right) \quad (3\text{-}13)$$

down 表示为下采样，β为乘性偏置，b 为加性偏置，一般卷积网络中没有β。

第 i 层的误差为：

$$\delta_j^l = f'\left(\mu_j^l\right)\circ\left(\sum_{j=1}^{M}\delta_j^{l+1}\circ k_{ij}\right) \quad (3\text{-}14)$$

式中：°表示为矩阵每个元素相乘

则所得到的偏置导数为：

$$\frac{\partial E}{\partial b_j} = \sum_{\mu,\nu}\left(\delta_j^{l+1}\right)_{\mu,\nu} \quad (3\text{-}15)$$

## 3.3 实验准备

### 3.3.1 实验环境搭建

深度学习中大数据的训练对计算机要求较高，基于现有环境对深度学习框架的选择

也十分重要，本小节以硬件设施、软件依赖和深度学习框架顺序介绍股票预测系统实验环境的搭建。

（1） 硬件设施：

本文实验所用硬件设施主要参数为：CPU 型号：Inter Xeon E7-8890，主频 22GHz；内存型号：Kingston Fury DDR4，容量 8*16G；硬盘型号：WD，容量 2TB;显卡型号：Tesla GTX 1080 GPU；显存型号：NVIDIA GDDR5X ，频率为 10Gbps。本文在上述设备的基础上进行深度学习环境的搭建。

（2） 软件依赖：

本文所采用的操作系统型号与版本为 Ubuntu 14.04 LTS。采用的开源框架是 tensorflow，tensorflow 是谷歌公司推出的基于张量流计算的开源深度学习平台，由于 tensorflow 的可用性高、灵活性强，并且谷歌公司投入了大量研发精力，一经推出，便成为了全世界最流行的深度学习框架之一。TensorFlow 支持 CNN、RNN 和 LSTM 算法，可被用于金融预测、语音处理或图像识别等多项深度学习领域。

由于深度学习涉及到极大的矩阵计算，这使得随着 GPU 性能的要求增高，对运算平台有效性和高速性的要求也随之增加。NVIDIA 显卡公司上市一种名为 CUDA（Compute Unified Device Architecture）的并行运算结构，这种平台使得 GPU 可以快速求解包括矩阵计算在内的各种计算问题，它含有 CUDA 指令集架构（ISA）和 GPU 并行计算引擎两部分。GPU 通过 CUDA 架构可以使得图形显示和数据并行计算获得速度上的极大提升。GPU 的作用不仅仅限于图形显示，运用 GPU 对数据进行并行加速处理成为当前研究的热点。Cudnn8.0 是 NVIDIA 为深度学习专门研发的一类 GPU 加速库，目前仅支持本品牌 GPU。Cudnn 支持大多数深度学习相关算法如卷积，池化，Softmax，各种激活函数，Tesnor 四维向量转换等。Cudnn8.0 强调性能，易于上手，使用代价也很低，同时因为其插入式设计，还易于将其集成到更高级的深度学习框架中去，使那些框架能够基于 Cudnn8.0 在 GPU 上实现高性能的并行计算。

基于以上事实基础，本文采用 CUDA8.0 搭配 Cudnn8.0+tensorflow1.40-GPU +keras+Python3.5 进行深度学习的相关实验。

### 3.3.2 评价标准

评价一个模型训练的好坏程度需要通过一些指标来进行衡量，通过指标数据的分析来说明模型的训练效果，本文主要通过均方根误差、确定系数两个评价指标来说明模型的优势程度。

均方根误差（Root Mean Square Error，RMSE)：是均方误差（Mean Squared Error，MSE）的算术平方根，表示预测值和真实值差的平方和后与预测值数量 N 比值的平方

根，也就是预测值和真实值之差平方的期望值平方根。公式如下：

均方误差：

$$\mathrm{MSE} = \frac{1}{N}\sum_{t=1}^{N}(y\_pre_t - y\_real_t)^2 \tag{3-16}$$

均方误差根：

$$\mathrm{RMES} = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y\_pre_t - y\_real_t)^2} \tag{3-17}$$

式中： N ----- 数据的数量
$y\_real_t$ ----- 数据的真实值

RMSE 值越小说明预测模型对实验数据的描述更加准确。

确定系数（R-square）：通过给定的数据的变化来表示数据拟合的好坏。公式如下：

$$R^2 = 1 - \frac{\sum(\mathrm{y\_pre} - \mathrm{y\_real})^2}{\sum(\mathrm{y\_real} - \mathrm{y\_mean})^2} \tag{3-18}$$

式中： y_pre ----- 预测值
y_real ----- 数据的真实值
y_mean ----- 数据的均值
R ----- 相关系数

相关系数的平方值便是决定系数 R-square。R-square 的取值范围是$(-\infty, 1]$，通常为[0,1]。当值为 1 时，则表示预测的结果与真实数据完全匹配的情况；当值为 0 时，表示预测结果和没进行预测一样；负数没有任何意义。数据值越接近 1，说明模型对数据真实值的理解能力越强，模型对数据拟合程度的也较好。

## 3.4 实验结果及分析

本章节实验主要运用一维卷积神经网络对股票价格波动进行预测并对新闻信息采用不同的处理方法展开实验对比，根据评价标准（均方误差根、确定系数）选择出效果最好的一种方式作为后续新闻处理方式的首选方法。为了保证实验的公平性在对比实验中需要保证网络训练参数的不变，表格如下 3.1 所示：

表 3.1 网络训练参数

| 参数 | 参数值 |
| --- | --- |
| 激活函数（activation function） | Relu |
| 学习率（learning rate） | 0.0001 |
| 损失函数（loss function） | mse |
| 优化函数（Optimizer） | Adam |
| 迭代次数（epoch） | 200 |
| 样本批次大小（batch） | 10 |
| 卷积层 | 2 |
| 池化层 | 2 |

其中，迭代次数 epoch 是个基准值，本章节主要以训练网络中 loss 不再下降或达到设置迭代次数为参考，此时停止网络训练。因为深度学习网络模型经过训练后，往往会在训练集上有较好的实验效果，但是这对网络模型的鲁棒性和繁华性能的评判并不准确，此时应该用网络并没有参加过训练的数据来进行模型的测试，也就是测试集，这也是建立测试集的目的，本文中给出的实验结果均为网络模型在测试集上的效果。由于本文对股票价格进行预测属于回归问题而不是分类问题，所以采用得损失函数为 mse（mean_squared_error），在评价指标中不使用准确度来进行衡量模型的好坏。

优化器的选择：keras 中有多种优化器可以供使用者进行选择，如：BGD（Batch gradient descent），SGD（Stochastic gradient descent），MBGD（Mini-batch gradient descent），Adagrad，RMSprop，Adam。本文选择了 Adam 自适应优化器。主要有以下原因：

（1） 由于 Batch gradient descen 在一次更新中，便会对整个数据集计算梯度，这样计算量比较大，计算起来非常慢，不能投入新数据实时更新模型。

（2） Stochastic gradient descent 由于更新比较频繁，造成 cost function 产生严重的震荡。

（3） Mini-batch gradient descent 不能保证很好的收敛性，学习率选择比较困难，选择值太小，收敛速度将会很慢，如果值选择太大，loss function 将会在极小值处不断地震荡甚至造成偏离。

（4）Adam 不仅具有 Adadelta 和 RMSprop 可以存储过去梯度平方$v_t$的指数衰减平均值，而且还像 momentum 一样保持了过去梯度$m_t$的指数衰减平均值，公式如下：

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (3\text{-}19)$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (3\text{-}20)$$

如果$m_t$和$v_t$被初始化为零特征向量，那将就会向零偏置，因此做了偏差校正，通过计算偏差校正后的$m_t$和$v_t$来抵消这些偏差：

$$\widehat{m_t} = \frac{m_t}{1-\beta_1^t} \quad (3\text{-}21)$$

$$\widehat{v_t} = \frac{v_t}{1-\beta_2^t} \quad (3\text{-}22)$$

而且经实践证明，Adam 比 Adagrad，Adadelta，RMSprop 等自适应性学习方法效果要好。

建议超参数的设定值为:$\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10e-08$。

### 3.4.1 卷积网络与传统时间序列网络对股票价格预测的对比

本实验选取典型的时间序列模型对股票价格波动预测与卷积神经网络模型对股票价格波动预测进行对比。传统时间序列模型对股票价格波动的预测中没有考虑新闻因素的影响，为了实验的准确性、公平性，卷积神经网络对股票价格波动的预测中也将不加入新闻因素的影响。卷积神经网络依然采用表 3.1 中的网络训练参数，但是输入有所改变，去掉了新闻信息的输入层和融合层，只剩下历史数据的输入，而窗口长度本实验中将采用 N=10 进行实验测试，训练网络结构如图 3.11 所示，为了更加直观清晰的观察将其化成流程图形式，如图 3.12 所示：

```
print(model.summary())
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 9, 64)             576
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 4, 64)             0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 3, 128)            16512
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 1, 128)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 128)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 4)                 516
=================================================================
Total params: 17,604
Trainable params: 17,604
Non-trainable params: 0
_________________________________________________________________
None
```

图 3.11 历史数据卷积网络结构图

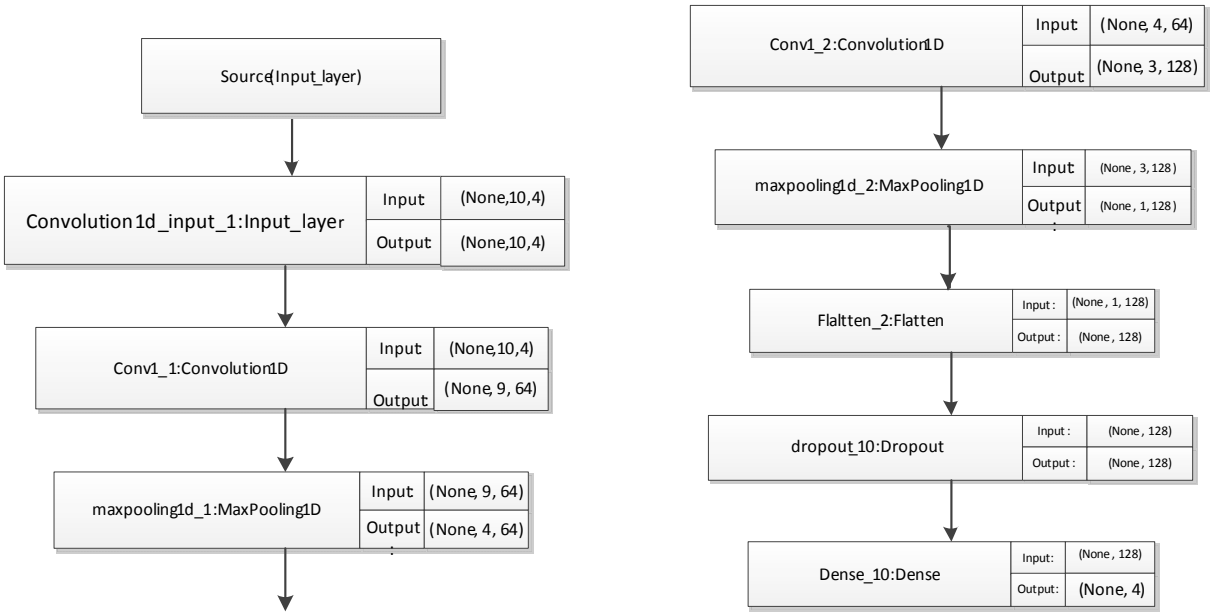



图 3.12 卷积网络框架

通过卷积网络结构图和框架流程图更加直观的可以看出，在整个训练过程中不仅仅输入层发生了变化，而且在每层的神经元个数也比图 3.8 的卷积网络结果神经元个数要少，总的神经元个数也就相对减少，因此网络在训练过程中也就会更加迅速一些，时间相对较少一些。

经过训练得出训练过程的损失图，如图 3.13、3.14 所示：

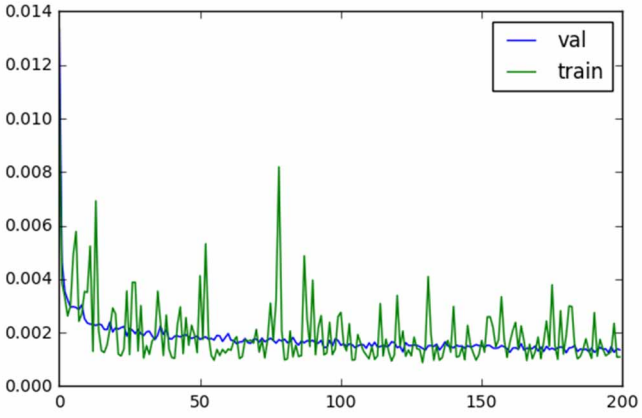



图 3.13 卷积神经网络训练的预测损失

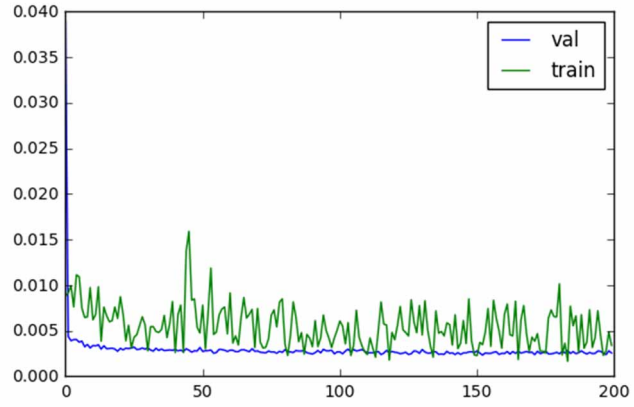



图 3.14 ARIMA 网络训练损失

其中横坐标表示 epoch 次数，纵坐标表示损失大小。通过两张图可以看出虽然在训练过程中出现诸多的波动，但是从总体上观看训练集合、验证集的损失都是在逐渐下降，网络模型是在不断地在探索数据间的关联不断进行学习的。可以说两个模型在学习的效果还是比较好的。

通过表 3.2 可知，ARIMA 网络模型对股票价格波动的预测均方根误差比卷积神经网络训练后得出的均方根误差数值较大，R2 值比卷积神经网络相对较小。通过两个评价标准的对比可知道卷积神经网络在对股票价格波动上的预测会比经典时间序列模型对股票价格波动的预测效果相对较好，拟合能力较强。卷积神经网络能取得这样好的效果与卷积神经网络的网络独特的网络结构有很大联系，可以深度学习数据间存在的相互联系。

表 3.2 评价标准对比效果对比

| 模型类型 | Train Score（RMSE） | Test Score（RMSE） | Train R2 | Test R2 |
|---|---|---|---|---|
| 卷积神经网络模型 | 2.62 | 9.64 | 0.38432 | 0.33662 |
| ARIMA 模型 | 6.32 | 10.56 | 0.36582 | 0.31685 |

### 3.4.2 基于滑动窗口长度对股票价格预测的影响

由于股票价格为时间序列数据，所以模型训练的效果受滑动窗口长度的影响，实验中将采取滑动窗口长度分别为 1、5、10、15 天历史数据进行实验验证，网络模型将采用图 3.11 中网络模型进行模型训练，通过评价指标查看滑动窗口长度与预测准确度的关系，表格如下：

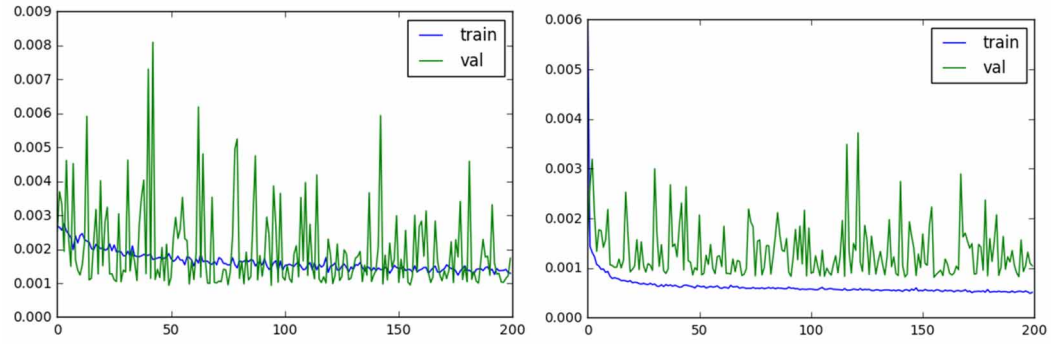
表 3.3 滑动窗口长度评价

| 滑动窗口长度 | Train Score（RMSE） | Test Score（RMSE） | Train R2 | Test R2 |
|---|---|---|---|---|
| 1 | 5.68 | 11.65 | 0.33638 | 0.31609 |
| 5 | 4.68 | 10.13 | 0.35671 | 0.32962 |
| 10 | 2.62 | 9.64 | 0.38432 | 0.33662 |
| 15 | 4.89 | 10.38 | 0.36358 | 0.33326 |

通过表格可以看到滑动窗口长度为 10 时预测的效果最好，因为股票数据为时间序列数据，受时间的影响很大，时间长度过短所包含的时间信息不够充分，在训练过程中很难把握重要的数据信息，根据较短的数据信息很难预测出未来股票的价格波动情况；如果滑动窗口长度过长，造成时间跨度比较大，数据量变小，而股票价格波动受临近时间段的影响比较大而受较远时间段的影响相对比较小。时间跨度大，数据量变少则会使模型很难抓到重要的信息，造成预测模型不准确。所以只有滑动窗口长度适当才会很好的预测出股票价格波动的情况。

### 3.4.3 基于不同新闻信息处理方式对股票价格预测的影响

本实验主要采用 Bag of Words、Word2vec、TF-IDF、pysentiment 四种方式对新闻信息进行处理来提取新闻信息中重要特征。通过评价标准来选取处理方式中比较好的文本处理方式。为了验证该实验的效果，在整个实验中选取的股票价格历史数据不变、新闻信息内容不改变，卷积层和池化层不进行改变、卷积的时间窗口依然为 10 不进行更改。模型网络结构在本章节图 3.7 已经写明。下图为四中文本处理方式的损失图，其中左上为 Bag of Words 对新闻处理的损失图、右上为 Word2vec 对新闻处理的损失图、左下为 TF-IDF 对新闻处理的损失图、右下为 pysentiment 对新闻处理的损失图：



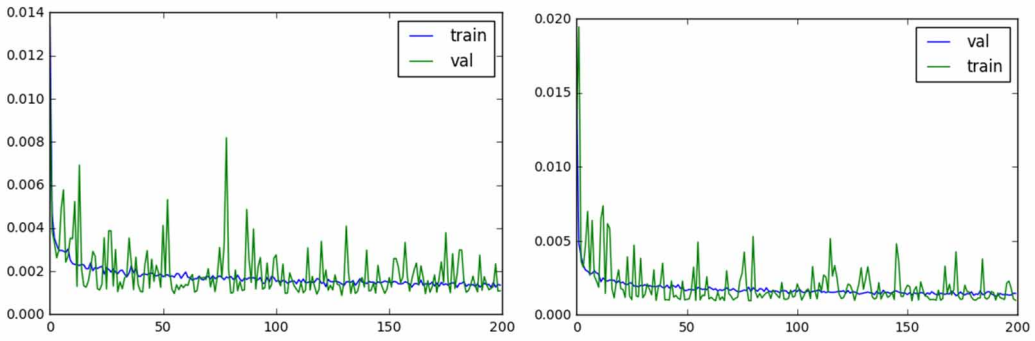



图 3.15 新闻损失图

图 3.15 表示模型在训练集和验证集上的表现。其中，横坐标表示迭代次数，纵坐标表示损失。由图可知，前三张图的训练集效果下降趋势相对明显，但是在验证集的损失波动性很大，而且图一的波动性非常大，主要原因可能有数据中有一些噪声的存在，新闻数据的处理方式相对比较建议不能合理的分析出新闻的感情色彩，使得模型在训练中泛化能力相对较弱从而造成在验证时损失出现波动，但是最后的大趋势还是不断地下降，说明模型还是在不断地学习，只是训练后的模型鲁棒性能比较差。而第四张图在训练过程中损失出现波动，随着迭代次数的增加波动逐渐下降，说明模型训练会比前三个模型相对较好，通过验证集验证，看到损失逐渐下降，证明模型的鲁棒性能比较好，模型训练比较成功。

根据表 3.4 中信息可观察到，pysentiment 根据新闻感情色彩将新闻信息进行向量特征的转化得到的效果比其他三种效果相对较好，主要原因在于公司的新闻存在很多的感情色彩的词语，表述了最近公司的经营状况，要是只是通过单词出现的频率或是对将单词进行分割使任何单词与语境没有任何关系等很难对新闻所要表述的重要信息进行剖析，因此融合后的数据集也不能很好地预测股票价格的波动情况，而 pysentiment 根据感情色彩对新闻进行特征向量转化，能很好地把握新闻中所要突出的感情，比如：在新闻词语中表达了正面情感的词语，“good”、“wonderful”、“amazing”或对公司称赞的词语：“经营状况比较好”、“产品畅销”等词都具有感情色彩，而且这些词直接影响股票价格的波动。通过对新闻信息进行感情色彩评级，增加新闻的信息在股票历史数据的权重比可以很好的提高模型训练后的预测效果。

表 3.4 数据处理的评价值

| 新闻处理方式 | Train Score（RMSE） | Test Score（RMSE） | Train R2 | Test R2 |
|---|---|---|---|---|
| Bag of Words | 5.68 | 10.52 | 0.36807 | 0.34684 |
| Word2vec | 4.85 | 10.29 | 0.39872 | 0.34827 |
| TF-IDF | 4.68 | 9.89 | 0.43501 | 0.38205 |
| pysentiment | 2.62 | 9.64 | 0.43562 | 0.38592 |

## 3.5 本章小结

本章提出了利用深度卷积神经网络对苹果公司股票价格波动进行预测。首先进行验证卷积神经网络对股票价格波动的预测会比传统时间序列模型对股票价格波动的预测效果较好。由于传统时间序列模型大多数对历史数据股票价格进行预测，为了保持公正性本实验使用卷积神经网络通过对股票历史价格波动进行预测。通过训练损失和训练后的评价标准得出的实验结果，证明卷积神经网络比传统时间序列模型对股票价格波动预测效果较好。实验中又验证影响股票价格波动情况与滑动窗口的长度有关，由于股票受此及附近时刻的影响比较大，受较久远时刻的影响比较小，尤其当加入新闻信息后，对于时间长度的选取更为重要，因为此刻的新闻对后两天的影响比较大，而当随着时间的延续，此刻新闻的影响也将随之变小。因此滑动窗口长度选取过长或过短都不利于股票价格的预测。最后通过验证新闻信息处理方式的不同对实验结果有所影响，选择 Bag of Words、Word2vec、TF-IDF、pysentiment 四种处理方式进行实验对比验证，通过评价标准的对比，最终得出 pysentiment 对新闻信息的处理方式比其他三种效果会好，主要原因是由于 pysentiment 对信息信息进行感情色彩评分，这样会很好的把握新闻中重要信息对股票价格波动的影响。

# 第 4 章 基于 LSTM 股票价格波动预测的研究

在前面章节介绍和分析中得知，深度神经网络架构能够很好捕捉隐藏的动态信息并能够进行准确预测，CNN 架构能够识别趋势的变化，它使用在特定时刻提供的信息进行预测。但由于股市经常发生突然变化，发生的变化可能并不总是规律的，也可能并不总是遵循相同的周期。卷积神经网络对于这些由于时间变化而产生的波动无法精确预测而长短时间记忆网络（LSTM）具有记忆功能，可以根据自身学习明确对过去价格进行记忆还是遗忘，解决了由于卷积神经网络不具有时间记忆功能而不能很好对股市的突发变化进行处理的缺点。由于公司和部门的不同，趋势的存在和存在的时期将有所不同，对这些趋势和周期的分析将为投资者带来更多利润。为了分析这些信息，该实验通过加入新闻作为辅助信息使用长短时间记忆网络来对股票价格波动进行预测。

## 4.1 国内外相关研究

长短时间记忆网络（Long-Short term memory，LSTM）是一种很常用的深度学习网络模型，在许多问题上被证明是非常成功的，因为它能够区分最近和早期的数据问题，通过给予不同的权重同时忘记它认为与预测下一个输出无关的记忆信息。这样，与其他仅能记忆短序列的传统神经网络相比，它能够很好的处理较长的时间序列。

本实验中使用的长短时间记忆网络与传统的前馈网络有所不同，它们不仅在单一方向上具有神经连接，换句话说，神经元可以将数据传递到先前的或同一数据层。在这种情况下，数据不会“并且其实际效果是短期记忆的存在，此外还有神经网络已经因训练而产生的长期记忆。LSTM 是由引入的，它旨在通过解决在处理长数据序列时经常性网络将遭受的消失梯度问题来获得更好的性能。通过“门”单元来实现误差流恒定，当“门”的信息不需要时，可以进行权重调整以及梯度截断。

鉴于 LSTM 网络在解决 NLP 等时序问题上有着卓越的表现，有些研究使用新闻文本数据作为输入来预测价格趋势，但是也有一些研究使用历史价格数据来预测股票价格在一天内的波动趋势。文献[72]中应用 LSTM 网络模型对股票历史价格进行预测，并与 MLP 和 Random Forest 模型进行比较，通过准确率、召回率、F 值作为评价指标实验结果表明 LSTM 网络模型在各项指标中都会比其他两种模型效果好，文中提到根据新闻信息对股票价格波动预测效果也很好，但是文中并没有加入新闻因素的影响[72]。国内使用 LSTM 网络模型对股票预测的研究员很少，孙瑞奇使用 LSTM 网络对股票价格波动进行预测研究，文章主要讨论了 BP 神经网络、RNN 神经网络和 LSTM 神经网络对股票价格进行短期预测的可行性并作出相应对比，验证了 LSTM 网络对股票预测的准确性较好，但文章中也忽略了影响股票价格的其他因素[73]。本实验通过加入新闻信息的影

响来对股票价格波动进行预测。

实验基本流程图如图 4.1 所示：

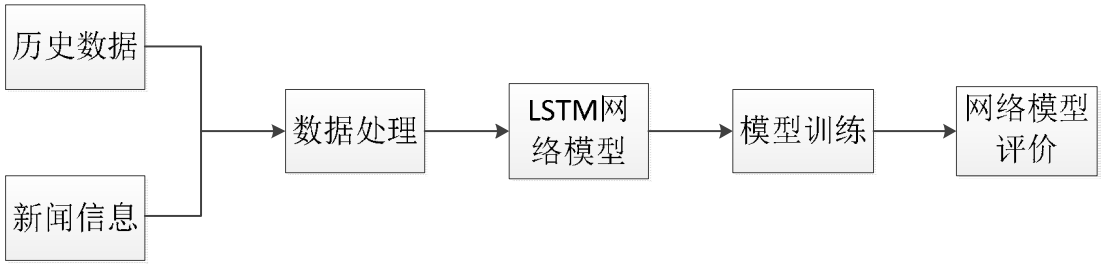



图 4.1 LSTM 股票预测基本流程图

## 4.2 基于股票价格预测的长短时间记忆网络模型

股票价格受前一时刻影响比较大，而随时间的延伸对此时刻股票价格的影响也将随之衰弱。虽然卷积神经网络通过在输入数据处理中加入滑动窗口技术，使输入的数据可以联系到前一时刻的数据特征具备了基本的时序性。但是长短时间记忆网络便是为“记忆”而生的网络，该网络可通过网络中的“门”限功能实现对数据的记忆功能，并通过模型的训练对输入的历史数据进行取舍。如果模型需要此时刻前的历史数据则将需要的进行保留并作为下一时刻输入的一部分，若模型不需要该时刻前的历史数据则将历史数据遗忘。

由于股票历史数据和新闻信息都具有时序性，而新闻对股票价格波动的一般都会有一段时间的延续性，当天的新闻可能会对后几天的股票价格产生影响，因此本章将利用长短时间记忆网络对股票价格进行预测。根据实验中数据集的大小，实验中将采用一层和两层长短时间记忆网络对股票价格波动进行预测，通过对比实验选取最适合的网络层数。

### 4.2.1 输入样本选择

由于长短时间记忆网络具有记忆功能，便不需要使用滑动窗口技术对输入数据进行初步处理。模型的数据输入是经过融合后的数据集，不需要做其他的处理。模型每一次输入数据的长度为 N，这与卷积神经网络设置的滑动窗口长度起到一个作用，训练集的标签将采用数据中第 N+1 天的股票历史价格作为前 N 天训练集的标签，这样便实现了通过前 N 天的股票价格来预测第 N+1 天的股票价格。对数据集进行分割，其中将数据集的 80%作为训练集，剩余 20%其作为测试集。若将数据集作为 LSTM 网络的输入还需要将数据集转化为满足 LSTM 网络要求的形式：输入 shape 为（samples，timesteps，input_dim）的 3D 张量。

### 4.2.2 构建长短时间记忆网络的股票价格预测模型

本实验仍在文章 3.3 的实验环境中进行实验验证。网络模型主要选择 LSTM 模型对股票价格波动进行预测。模型将以一层网络为例，该层网络以技术指标、定价数据和新闻辅助信息作为输入。输入层具有 5 个特征维度，其由价格回报数据（开盘价，收盘价，最高价，最低价）和新闻数据构成。输出层激活函数为 tanh 函数，最后由一个以 4 个神经元组成的全连接层网络作为最后预测的输出，并通过 RMSE、R2 评价指标来评价网络模型的稳定性。由于此时的股票价格受短时间的股票价格影响很大，受长时间段的股票价格影响很小，尤其加入新闻信息的影响。因为新闻只对短时间的影响大，当时间延续新闻的影响效应也就随之减少。因此在输入样本训练时需要将距离标签近的股票价格和新闻的初始权重设置要比距离标签时间远的初始权重要大，这样保证了新闻对近距离的影响大的影响效果。

网络模型如下图 4.2 所示，通过网络模型可以查看 LSTM 网络每层神经元得个数、输入层和输出层的维度。为了方便直观的观看 LSTM 网络的框架，将通过流程图形式进行描述，如图 4.3 所示：

```
print(model.summary())
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 64)                17664
_________________________________________________________________
dense_1 (Dense)              (None, 4)                 260
=================================================================
Total params: 17,924
Trainable params: 17,924
Non-trainable params: 0
_________________________________________________________________
None
```

图 4.2 LSTM 网络模型

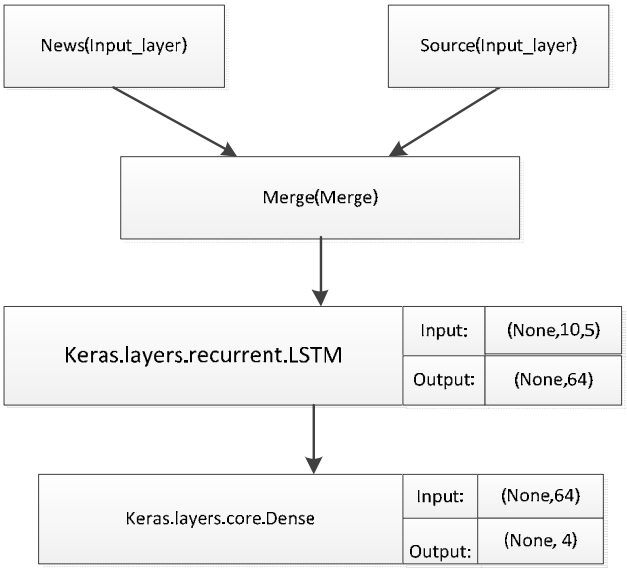



图 4.3 网络流程图

### 4.2.3 循环神经网络基本概述

长短时间记忆网络是循环神经网络的一种变体结构，因此在介绍长短时间记忆网络前，将对循环神经网络进行简单概述。

1986 年 Rumelhart 等人首次提出循环神经网络（Recurrent Neural Network, RNN），循环神经网络主要为处理时间序列数据而产生。时间序列数据也就是序列数据前后之间存在着很强的关联性、数据之间不互相独立，前面的数据会对后面的数据产生重要的影响，甚至后面的数据会对前面的数据同时产生重要的影响（双向循环神经网络）[74]。循环神经网络的主要特点为：它同时具有前馈神经网络和反馈神经网络的人工神经网络，其反馈神经网络是将该网络的输出经过一个或多个时间序列点后送给自身或其他后续的神经元，即一个时间序列的输出不仅仅与当前的输入有关系而且还与该神经元本身的输入和前期的输入共同作用产生的结果，网络的权值也表现在对所有的信息进行存储并应用在损失函数中[75-77]。循环神经网络的结构图如图 4.4 所示：

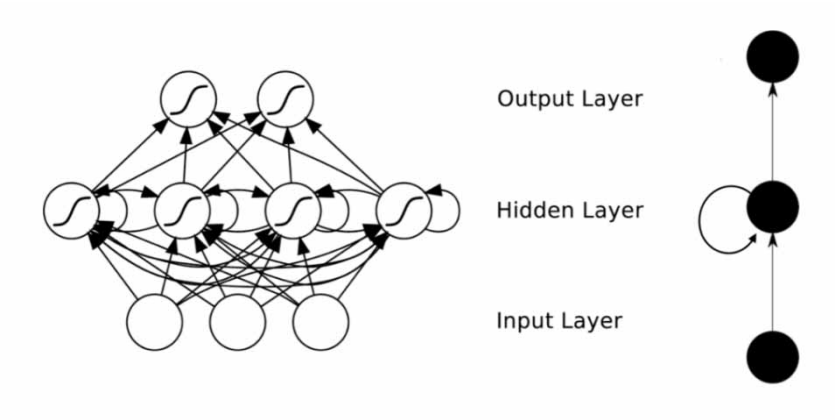



图 4.4 循环神经网络结构

通过上图可以看到隐层节点之间存在连接，在误差反向插播时，隐层自身和节点之间存在权值与偏置的更新。

RNN 循环神经网络基本形式如图 4.5 所示。x是输入时间序列的向量特征，h表示网络隐层状态，o表示网络的输出层。输入层与隐层间通过参数矩阵$U$ 互相连接，不同时刻的隐层间以矩阵W互相连接，一层 u 输入层之间以矩阵V互相连接。

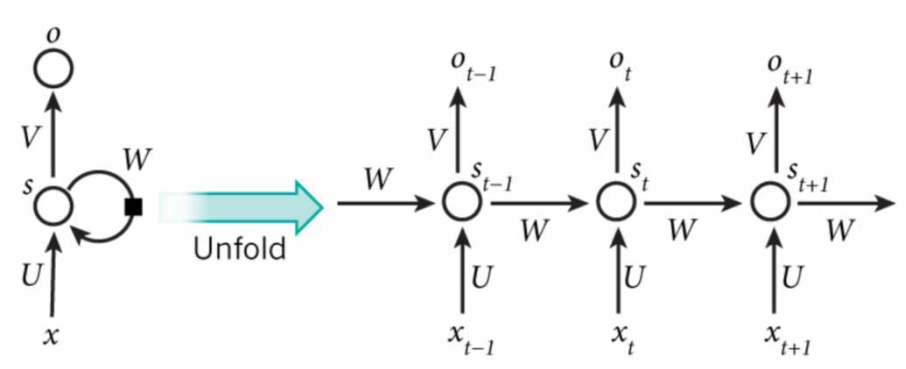



图 4.5 RNN 的折叠形式和展开形式

通过公式表示 RNN 的前向传播为：

$$a_t = b + Wh_{t-1} + Ux_t \quad (4\text{-}1)$$

$$h_t = tanh a_t \quad (4\text{-}2)$$

$$o_t = c + Vh_t \quad (4\text{-}3)$$

$$\hat{y_t} = softmax(o_t) \quad (4\text{-}4)$$

式中：$x_t$ ----- t 时刻输入

$h_t$ ----- t 时刻的隐层状态

$o_t$ ----- t 时刻输出

$\hat{y}$ ----- 经过归一化后的预测概率

参数共享也存在 RNN 中，同一时刻中矩阵参数 U、W、V 没有发生变化，使用同一组数据。

1 RNN 的权值更新和目前的弊端

RNN 网络的权值更新和使用的是 BPTT（Back-Propagation Through Time）算法，其是 BP 算法在 RNN 结构上的一个变体形式，在 RNN 网络模型上进行梯度计算。

其公式为：

$$s_t = tanh(U_{xt} + W_{st-1}) \quad (4\text{-}5)$$

$$o_t = softmax(V_{st}) \quad (4\text{-}6)$$

式中： $s_t$ ----- t时刻的隐藏层状态值

$o_t$ ----- t时刻输出值

如上图可知，对于每个隐层神经元来讲，每次输入除上层神经源的信息流入外还有本层来自上一时间点的信息流入，这个结构就使得 RNN 能够对时域信息进行处理。

循环神经网络也存在弊端，由于 RNN 的参数更新需遵循误差方向传播，在 k 时刻如果选用 Tanh、Sigmoid 等激活函数时便会出现因为输入而接近两端的结果，函数导数将会趋于零从而引起“梯度消失”问题;若选用 Tanh 或 Relu 函数当输入值很小时，则会因导数的链式法则而带来“梯度爆炸”的问题。由于这两个问题的存在 RNN 在实际应用过程中很难处理超过时间序列长度为 10 的数据序列如图 4.6 所示：

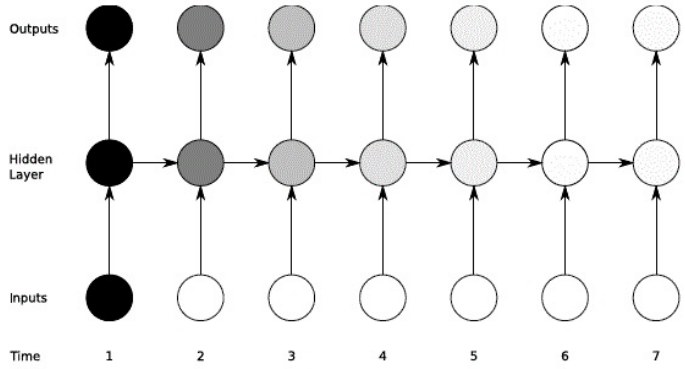



图 4.6 RNN 展开图

对于以时间序列信息输入为优势的 RNN 而言，这将是其最大的劣势。Bengio 等人也通过实验验证了这个结论。此外由于 RNN 在权值更新时仅仅会对上一时间点的输出和此刻时间点求梯度，这将意味着 RNN 会对上一时间点的全部信息进行学习而不会舍去任何信息，这显示对复杂而快速变化的输入信息是不可取得。对于“梯度消失”问题，目前 RNN 的结构除了慎重选取参数的初始值、层间使用正则化外，并没有其他的有效方法解决此问题;“梯度爆炸”可以通过梯度裁剪进行优化，或者更换一种算法——LSTM,下一小节进行详细介绍。

2 长短时间记忆网络

由于循环神经网络网络存储信息比较差，在训练网络时很难达到理想的目标，而且

基于原生的循环神经网络难以保持长时间的信息记忆，而对股票历时间序列预测对过去时间信息的把握相当重要，这样循环神经网络便不能很好的达到理想的目标。根据 RNN 不能把握过去时间数据的信息这一缺点，Sepp Hochreiter 与 Jürgen Schmidhuber 在 1997 年提出的最早提出了长短时间记忆网络模型（Long Short Term Memory,LSTM）。他们通过在循环神经网络中引入门限机制，对时间点的信息就行更新和整合有效地解决了循环神经网络容易出现梯度爆炸或梯度弥散的问题。LSTM 模型已成为对深度时间信息学习的最有效手段，它在各个领域都得到了充分应用。

LSTM 主要由输入门、输出门、遗忘门和记忆单元组成，其中输入门、忘记门、输出门都控制记忆模型实现读写和丢失的特性操作。LSTM 与 RNN 主要的不同就是 LSTM 网络具有一个状态参量来保存时域信息，随着模型的输入数据将逐个被送入长短时间记忆网络里，其中有用的信息经过筛选将被保存下来存入状态参量矩阵中，无用的信息将会被抛弃。也就是说在整个网络的训练过程中，状态变量一直随着时间不断地更新。其方式如下图 4.7 所示:

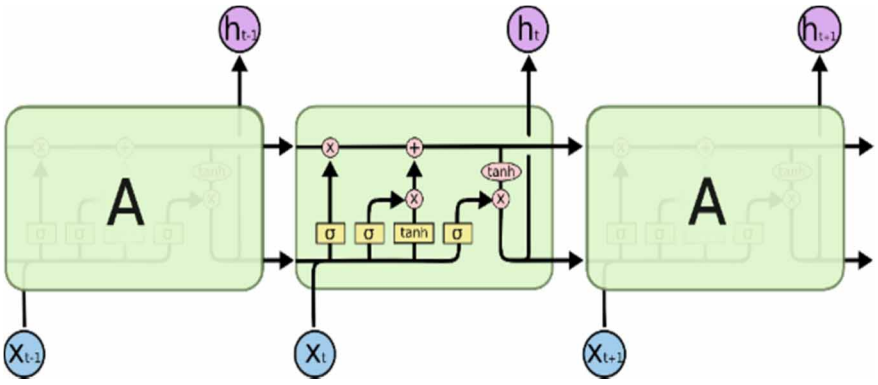



图 4.7 长短记忆网络整体结构

长短时间记忆网络是以门的形式来实现对状态的取舍，门的结构如图 4.8 所示：

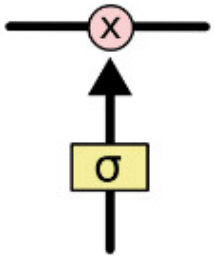



图 4.8 长短记忆网络的门结构

由于门结构是一个非线性 Sigmoid 函数与和点积操作构成。通过下图 4.9 可以看出函数范围在 0 到 1 之间，因此门控制器描述了数据信息通过的比例，Sigmoid 取值为零时表示信息没有通过，或者理解为将会忘记输入输入或之前信息。相反，Sigmoid 取值为 1 时则表示输入信息完全保存在网络。这样使门结构具有了取舍功能，而且去饱和性也可以避免对输入信息的过度学习。长短时间记忆网络有三个 Sigmoid 函数对状态进行

操作，遗忘门、输出门、输入门。

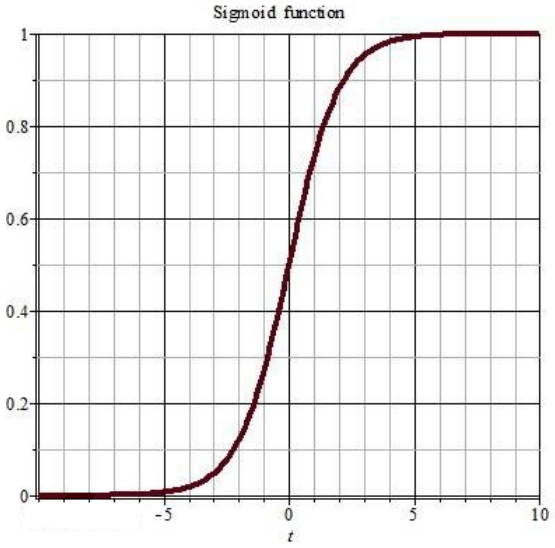



图 4.9 Sigmoid 激活函数示意图

（1）.遗忘门

在 LSTM 模型中，只有上一时刻的记忆单元的输出才会传到此时单元中，其他数据只能在本单元内可见，遗忘门对网络信息的输入进行选择，摒弃无用信息单元。上文介绍遗忘门由门单元来实现该功能，公式如下：

$$f_t = Sofmiod\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (4\text{-}7)$$

其中$W_f$表示忘记门的权矩阵，$h_{t-1}$表示上时刻的状态量，$b_f$代表该层网络的偏置，由公式可知，遗忘门由此时刻的时间点输入和上一时刻时间点状态量$h_{t-1}$和$x_t$作为输入并经过非线性变化，输出一个[0,1]之间的数值，而该值则代表了忘记网络信息的比例，若数值为零则代表该门遗忘掉上单元的所有信息，反之则说明该门将上单元信息就行保留。

（2）.输入门

输入门对当前时刻记忆单元的状态参量进行存储，公式如下：

$$i_t = Sofmiod\left(W_f[h_{t-1}, x_t] + b_i\right) \quad (4\text{-}8)$$

$$\widehat{c_t} = tanh\left(W_f[h_{t-1}, x_t] + b_c\right) \quad (4\text{-}9)$$

首先，输入门同样通过 Sigmoid 函数生成一个对新增状态量取舍比例的量。然后，本时刻的新增状态则用 tanh 函数计算而得，之所以在这里使用 tanh 非线性函数是因为其函数输出在-1 到 1 之间，起到将输入的均值调整为 0 的作用，而 Sigmoid 输出在 0 到 1 之间正好可以起到放缩的作用，但其均值为 0.5，不利于后续激活函数的处理。

然后，神经元开始对状态进行更新，公式如下：

$$c_t = f_t * c_{t-1} + i_t * \widehat{c_t} h\left(W_f[h_{t-1}, x_t] + b_c\right) \quad (4\text{-}10)$$

将上一时刻的状态与忘记门获得的比率相乘，求出取舍后的信息；再将其与缩放后

的新增状态相加就可实现状态的更新。

（3）.输出门

控制记忆单元的输出状态，公式如下：

$$o_t = Sifmoid(W_o[h_{t-1}, x_t] + b_o) \quad (4\text{-}11)$$

$$h_t = o_t * tanh(c_t) \quad (4\text{-}12)$$

最后，我们将确定输出层的内容。首先门函数将通过当前的状态放缩并决定输出部分的内容；然后 tanh 函数将其保持在[-1,1]之间；最后使他们相乘作为最后的输出。

## 4.3 实验结果及分析

本章节主要运用 LSTM 网络对股票价格波动进行预测，实验中主要验证新闻信息的存在对股票预测有积极影响。由于存在对比实验必须保证网络参数一致性，网络参数如下表所示：

表 4.1 网络训练参数

| 参数 | 参数值 |
|---|---|
| 激活函数（activation function） | Tanh |
| 学习率（learning rate） | 1e-04 |
| 损失函数（loss function） | MSE |
| 优化函数（Optimizer） | Adam |
| 迭代次数（epoch） | 200 |
| 样本批次大小（batch） | 10 |

### 4.3.1 基于不同网络层数对股票价格预测的影响

本节实验主要采用不同网络层数对股票价格影响进行实验对比。实验中使用苹果公司股票价格的历史数据进行预测，网络模型仍然采用图 4.3 的网络模型，只是在模型的输入上进行改进，仅仅将苹果公司的历史数据作为模型的输入，忽略其新闻因素的影响。由于本文实验的数据集并不是很大，所以只对两层网络和一层网络进行实验对比，并通过评价标准选取较好的网络模型。

表 4.2 不同层数实验对比

| | Train Score (RMSE) | Test Score (RMSE) | Train R2 | Test R2 | 运行时间 |
|---|---|---|---|---|---|
| 一层 LSTM | 1.98 | 4.17 | 0.48987 | 0.43057 | 10min |
| 二层 LSTM | 2.36 | 5.63 | 0.44652 | 0.41526 | 13min |
| 卷积神经网 | 2.62 | 9.64 | 0.43562 | 0.38592 | 11min |

| | | | | | |
|---|---|---|---|---|---|
| 络 | | | | | |

表格中加入卷积神经网络对股票价格波动预测的最好结果进行对比分析。通过上述表格可以看出，LSTM 网络对股票价格波动的预测结果都会比卷积神经网络的预测结果要好，也验证了长短时间记忆网络对时间序列的预测结果好于卷积神经网络。

通过表格可知，一层 LSTM 网络中 RMSE 的值比两层网络的小，R2 值一层网络比两层网络的大，而 RMSE 得值越小说明预测值和真实值的误差比较小；R2 值越大说明预测值越拟合真实的值。因此一层网络对股票价格预测的准确度比较高。由于实验数据量比较小，网络层数过多可能会丢失部分有用的特征点，而且层数越多训练时间越长，会造成不必要的浪费；对股票价格预测的 LSTM 网络使用一层网络和一个全连接层便可，这样不仅预测的效果好，而且还会节省能源，节约训练时间。

### 4.3.2 基于新闻因素对苹果股票价格预测的影响

该实验将有关苹果公司的新闻作为输入的一部分，由于 3.3.2 实验验证 pysentiment 对新闻信息的处理效果比其他三种文本处理方式好，在该实验中将采用 pysentiment 对新闻信息进行向量特征转化。

本实验采用的网络模型为两层网络，其中包含一个 LSTM 层和一个含有四个神经单元的全连接层，网络结构图如 4.2 所示。训练过程实验损失图如图 4.10 所示：

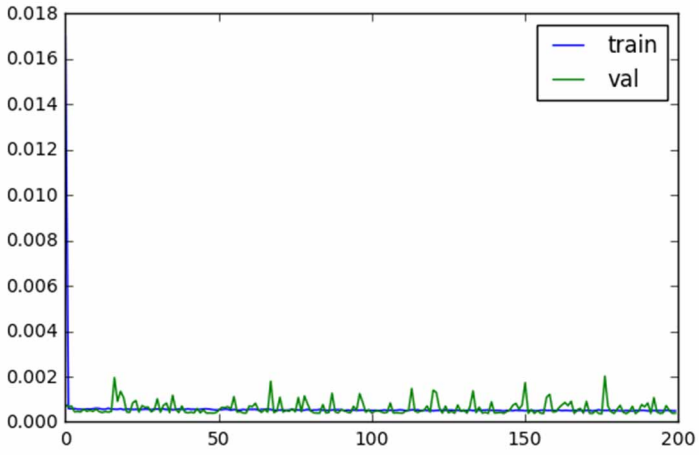



图 4.10 加入新闻损失图

为了说明新闻对股票价格预测具有积极作用，将加入单一历史数据对股票价格波动预测的对比实验。保证实验的公平性仍需使用上述实验所用的 LSTM 网络模型。训练过程实验损失图如图 4.11 所示：

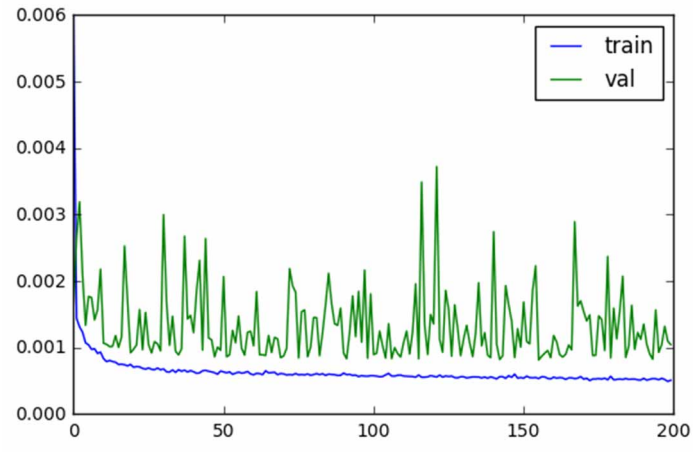



图 4.11 单一历史数据的损失图

根据以上两种损失图可以观察到，蓝色曲线表示训练集损失，绿色曲线表示验证集损失，横轴表示迭代次数，纵轴表示损失。通过图 4.10 可以看出训练集损失和验证集损失都从 0.016 瞬间降到 0.001 附近，然后在逐渐下降；而图 4.11 中训练集损失和验证集损失都在 0.005 下降到 0.001 左右附近，但是验证集的损失不断剧烈波动，说明模型的泛化性能不好，由于数据集较小，训练时可能出现过拟合，但大趋势是在逐渐下降。通过观察两幅图，虽然验证集都有所波动，但是图 4.10 的波动比较小。说明新闻信息的加入可以是模型的训练效果更好，使模型的鲁棒性、泛化性更好。

通过评价指标判断两种模型的优劣，实验结果如下 4.3 表所示：

表 4.3 评价指标

| 数据类型 | Train Score (RMSE) | Test Score (RMSE) | Train R2 | Test R2 |
|---|---|---|---|---|
| 历史数据 | 1.98 | 4.17 | 0.48987 | 0.43057 |
| 历史数据和新闻信息 | 1.82 | 3.17 | 0.53862 | 0.45964 |

根据表 4.3 中信息可以观察到，历史数据和新闻信息融合后的 RMSE 比单纯历史数据的 RMSE 小，历史数据和新闻信息融合后的确定系数 R2 比单纯历史数据的确定系数高，说明历史数据和新闻信息对网络模型训练的拟合能力更强，也就是模型可以很好地将预测数据逼近真实数据。主要原因是由于股票的影响因素很多，除了根据历史数据波动规律分析未来股市的发展规律外，还有其他的影响因素。诸如：国家的政策调控、利息调整、公司内部的调整等等这些都是影响未来股市的重要因素。仅仅通过历史数据来推测未来股市的发展规律，不能够很好地寻找其内部的规律，因此本实验中加入新闻的影响因素来更加准确地预测未来股票的发展规率。通过表格也验证了此观点，新闻信息对股票价格预测是有积极促进作用的，可以提高股票预测的准确性。

为了表述模型的预测效果，以苹果公司股票价格中开盘价与预测后的开盘价进行比较，如图 4.12。

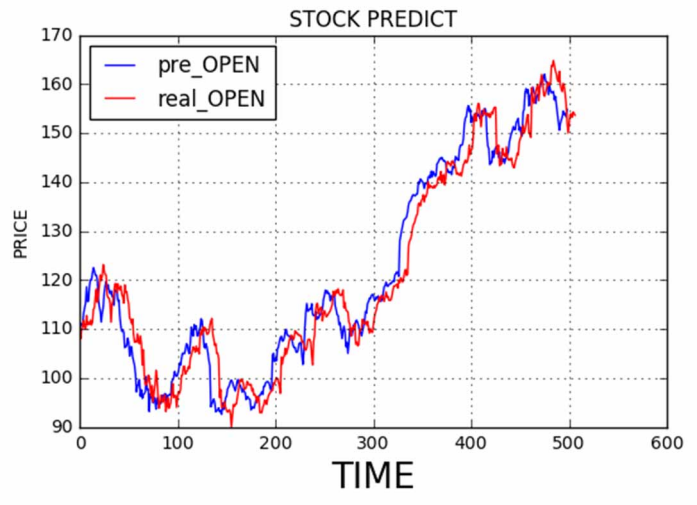



图 4.12 股票开盘价对比

图中红色曲线表示股票价格的真实数据，蓝色曲线表示预测股票价格的波动趋势。横坐标表示股票价格的时间天数，纵坐标表示股票的价格，单位为美元。从图中可以看到股票预测价格在股票真实价格附近波动，波动趋势相差不多。通过预测的股票价格该模型几乎能给出未来股票价格的波动趋势，但准确率可能不会很高，只能给股民一个参考意见，出现上述的主要原因是新闻的数据量比较小，新闻种类比较单一不能很好地提供苹果公司的相关信息，使模型不能很好地学习股票价格与新闻间的相关性。因此该模型还需要进行优化调整。

### 4.3.3 基于多种科技公司新闻因素对苹果股票价格预测的影响

本实验通过选取与苹果公司相关的多家科技公司新闻信息作为新闻数据集，其中有 ACLS、CAMP、CSLT、CYOU、RPD 五家科技公司，这五家科技公司在某一方面都会与苹果公司存在竞争或合作关系，而新闻信息内容可以体现出公司的经济状况，五家公司的经营状况的好与坏直接会影响苹果公司的经营状况。本实验依然选取实验 1 中的模型，模型层数、参数不做调整，下图是模型训练的损失图，如图 4.13 所示。

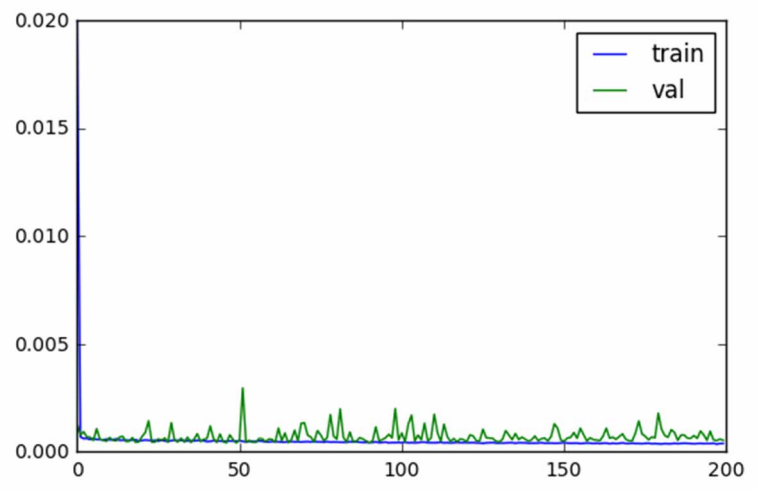



图 4.13 多新闻信息的损失图

通过损失图可以观察到，验证集的波动状况比 4.11 图中验证集的波动状况较小，说明多新闻因素对模型的鲁棒性有所改变。通过损失图可以看出网络模型是在不断的学习，损失是持续下降的，在开始网络模型损失下降速度比较快，迭代次数增多损失下降梯度逐渐变缓。

通过评价指标与实验 2 结果进行对比，表格如下：

表 4.4 新闻信息量对比实验

| 数据类型 | Train Score (RMSE) | Test Score (RMSE) | Train R2 | Test R2 |
|---|---|---|---|---|
| 历史数据和新闻信息 | 1.82 | 3.17 | 0.53862 | 0.45964 |
| 历史数据和多家新闻信息 | 1.61 | 3.05 | 0.58416 | 0.46336 |

通过表格可以看出取多家新闻信息可以更加准确预测苹果公司股票价格未来的波动趋势。由于多家新闻信息可以更好地获取有关苹果公司的信息，与苹果公司存在竞争关系的公司业绩比较好那么苹果公司的经营业绩可能会出现滑坡；恰恰相反如果竞争对手公司经营比较惨淡那么苹果公司经营状况可能会上升。而大部分信息都会通过新闻信息表示出来，而大量的新闻信息对苹果公司股票价格预测的准确度有促进作用，因此模型的训练也会比其他模型效果好，预测的结果会更加准确。

将采用训练好的模型对苹果公司股票价格的波动情况进行预测，如图 4.14 所示：

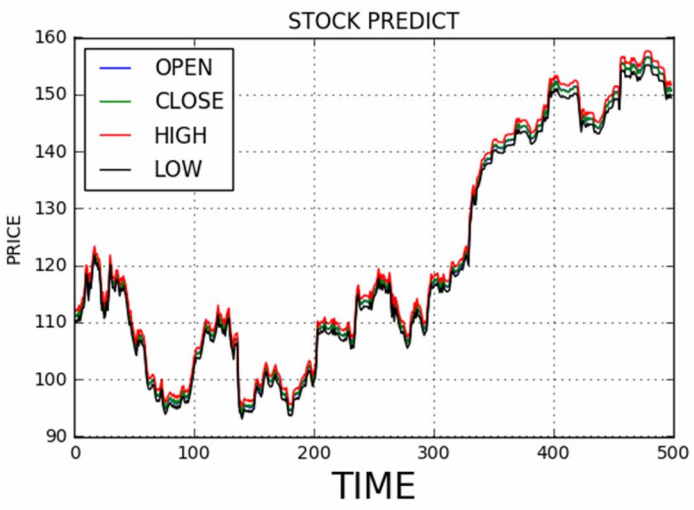



图 4.14 价格趋势波动

本图呈现出模型对苹果公司的开盘价、收盘价、最高价、最低价四种股票价格预测的曲线图。横坐标表示股票价格的时间天数，纵坐标表示股票的价格，单位为美元。为了更加直观明了的对比真实价格和预测价格的波动趋势，以苹果公司的真实的开盘价格与模型预测的开盘价格进行对比，如图 4.15 所示：

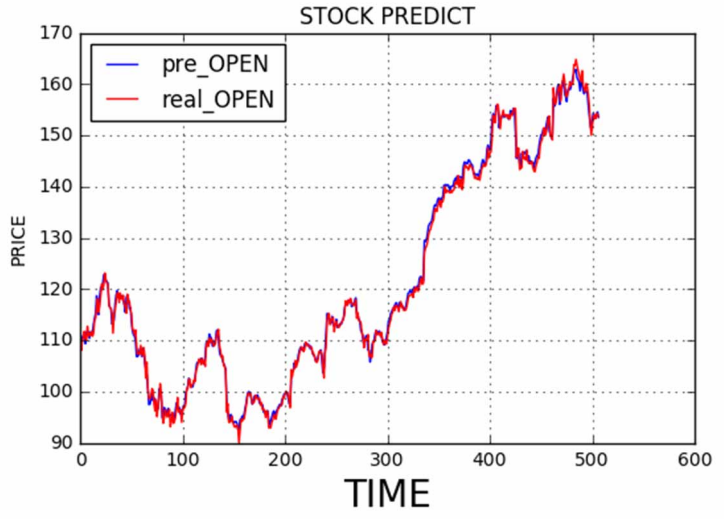



图 4.15 开盘价格对比

图中红色曲线表示股票价格的真实数据，蓝色曲线表示预测股票价格的波动趋势。从图中可以看到股票预测价格很好的拟合了股票的真实价格，相对实验 2 中预测结果，该模型可以准确的给出未来股票的价格波动情况，模型的预测效果提高很明显。通过给定未来股票的价格波动情况，股民可以决定自己手中的股票是抛还是继续购买，给股民提供了有力的参考意见，保证了股民的收益，达到稳赚不赔的目的。由于数据库的样本量比较小，因此模型泛化性能并不理想，只能在该数据集中表现出良好的效果，若应用到实际的预测中还需要扩大数据库，使新闻的样本集种类增多。

## 4.4 本章小结

本章实验通过验证长短时间记忆网络模型对股票预测的效果比卷积神经网络模型对股票价格波动预测的效果好，及验证了新闻信息对股票价格波动的预测有促进作用，且新闻种类越多对股票价格波动的预测效果会越好。章节前主要对长短时间记忆网络模型的构建进行了详细了介绍。实验主要分为三个实验进行验证，首先对不同网络层数对股票价格波动预测的影响进行分析验证，通过结果看出一层 LSTM 网络的预测结果会比两层网络的预测结果好，而且每一种结果都会比卷积神经网络的最好结果要好；其次加入苹果公司的新闻信息与实验一进行对比，结果证明新闻对股票价格波动预测有促进作用；最后引入其他五家科技公司的新闻信息进行实验的对比，预测结果证明新闻的种类预测对股票预测的结果会越好，提高了预测的精度。但是由于本文实验的数据库比较小，而新闻获取的途径相对单一，获取的种类也是比较少，所以训练后模型的泛化能力不强，不能很好的对苹果公司的价格波动进行预测。若应用到实际模型预测，需要扩大实验的数据库，增加新闻的获取途径及新闻种类，而且股票价格的影响因素诸多，新闻只是影响因素的一部分，需要精确的预测股票价格还需要考虑其他的影响。

# 结　论

当今全球经济处于快速发展状态，人民生产总值不断提升，但是金融市场发展却不是很成熟，研究金融时间序列预测问题有着非常重要的理论意义和应用价值，正是由于这个原因金融市场是一个非常热门的研究方向尤其是股票市场预测研究。每一位投资都想付出较少的成本去追求最大的收益，然而股票市场具有高风险与高收益并存的特性，所以对股票价格预测的研究可以在很大程度上帮助股民和投资者提高收益、避免股票市场的暴跌而带来的巨大损害。目前已经有很多国内外研究员和学者对该问题进行研究，希望可以找到高效准确的预测模型提供给股民帮助。本文借鉴和总结国内外的一些研究结论和经验，系统的研究了股票价格波动预测模型，对股票的影响因素做了比较全面的了解，并着重研究了新闻因素对股票市场特征及股票价格波动影响，由于深度学习模型具有良好的自适应性、自组织性、具有很强的学习能力、抗干扰能力，可以从历史数据中提取影响股票波动的各种因素，挖掘新闻信息中的内在关联，依据数据本身的联系进行建模，同时可以克服传统方法中的许多局限和不能很好解决非线性问题的困难，也避免了很多的人为因素影响。因此本文使用卷积神经网络和长短时间记忆网络通过利用新闻信息和历史数据对股票价格波动进行预测。实验中将选用苹果公司的股票价格、新闻信息及其他五家科技公司的新闻信息作为数据库，由于新闻数据获取比较困难，一般金融网站只会保留两年的新闻信息，所以新闻量比较小而且获取渠道比较单一，因此本实验算法的泛化性能不是很好，只适合本实验的数据库，若应用到实际中需要扩大实验得数据库，增加新闻信息的类别及获取渠道。

本文详细分析了目前对股票价格波动预测方法存在的缺陷和技术难点，并结合深度学习，找到了一些解决这些问题的方法。

1、数据集选取。由于股票价格受多种因素的影响，目前大部分研究员只是通过历史数据（开盘价、收盘价、最高价、最低价）对股票价格波动进行预测却忽略了其他的影响因素，而大部分影响因素都会通过新闻表现出来，将新闻信息作为数据集的一部分可以很好的提高预测结果。本文新闻获取方式通过爬虫方式来获取苹果公司及其相关公司的新闻信息，验证新闻信息对股票价格波动预测具有促进作用，而且新闻信息量越多，预测效果越好。

2、选用卷积神经网络对股票价格波动进行预测研究。现有的股票价格波动预测一般都是用传统时间序列模型或人工神经网络模型对股票价格波动预测，人工神经网络在处理非线性数据比传统时间序列模型好，但是在训练中容易出现过拟合、训练时间过长，而深度卷积网络由于具有权值共享这样可以节省训练时间避免过拟合。实验方面本文对卷积神经的概念进行简介，常用单元及主要误差传播算法进行了基本描述。然后本文对

优秀卷积网络特点和其发展趋势进行总结，且对训练误差反向传播原理进行了介绍。同时，第三章对本文实验所采用的设备与配置进行说明。由于股票价格历史数据为一维数据，因此本章实验采用一维卷积网络对股票价格波动进行预测。将卷积神经网络对股票价格波动能够预测的结果与 ARIMA 模型进行对比，验证卷积神经网络对于股票价格波动预测的优越性；通过实验验证滑动窗口的长度会影响股票价格预测的准确性，不同的新闻处理方式对模型的拟合能力也有所影响。

3、选用长短时间记忆网络对股票价格波动进行预测研究。由于一维卷积神经网络不具有时序性，不能很好地预测时间序列的数据，因此第四章将使用长短时间记忆网络对股票价格波动进行预测。实验方面本文对长短时间记忆网络的概念进行简介，介绍长短时间记忆网络的对于时间记忆的优势。本章实验主要验证时间序列网络对股票价格波动预测的效果比卷积神经网络效果好，同时验证新闻信息有助于提高预测的效果而且新闻信息量越多则预测的效果越好。

虽然本文结合深度学习在股票价格波动预测领域取得了一定得成果，但是由于自身对深度学习和股票价格波动影响因素的理论研究不够充分以及一些外在条件约束，仍有许多问题有待进一步解决，主要集中在一下几个方面：

1、由于时间方面因素，对于深度学习模型掌握不是很充分，不能很好的进行网络调参，因素不能保证所选的网络层数和参数为最佳。

2、在数据库建立方面，由于新闻数据集的搜集比较困难，短期时间的搜集还是比较容易的，但是长时间的新闻获取比较困难，任何网站不会保存某公司的三年以上的新闻信息，所以对于长时间的信息获取是比较困难。对于时间比较长的新闻每年可能有几十条，如果这样新闻空缺太多从而造成数据空缺太多，造成数据量比较小，只能做小部分实验验证，不利于模型的训练，因此需要长时间的搜寻新闻信息，扩充数据集实验结果还会有好的提升。

3、在对数据进行缺失值处理方面，缺失值填充和选取的是填零补充和最近距离补充的选择会对实验结果产生影响，需要进一步实验测试不同处理方法和阈值对实验结果的影响。

4、网络结构的调整。本文在实现的过程中只是针对单一的任务进行了简单的调整，考虑到影响股票价格波动的外在因素很多，考虑诸多因素的影响后可以更好地提取数据中的内在的关联信息。而且随着新的网络结构的不断出现，考虑使用不同的网络结构或者其组合可能会取得更好的结果。

# 参考文献


[1] 蔡自兴，徐光祐.人工智能及其应用[M].北京：清华大学出版社，2004.

[2] Refenes A N，Zapranis A,Francis G.Stock performance modeling using neural networks: a comparative study with regression models[J].Neural Networks,1994,7(2)：375～388.

[3] Prof Dr Matthias Schurmann , DipL-Kfm ,Thomads Lohrbach.Comparing artificial neural networks with statistical methods within the field of stock market prediction[J].Neural Computing and Applications,2000(3):596-609.

[4] 肖冬荣，柳亚婷，高健.基于改进 BP 网络的经济发展预测模型及应用[J].科技信息，2007（26）：55-56.

[5] 涂宇.零起点股票投资[M].北京：清华大学出版社，2009.

[6] 朱元.证券投资学原理.北京：立信会计图书出版社，2003:7478.

[7] 沈冰.股票投资分析.重庆：重庆出版，2002:94.

[8] 刘长虎，陶建格，崔衍秋.股票价格指数的投资功能.市场论坛，2004，（3）：71～72.

[9] 姚睁.我国推出股价指数期货的意义与政策设计.当代财经，1995，（12）：22.

[10] 麻卫华，李玉红.股指期货与我国股票市场发展.金融教学与研究，2004，（5）：50～54.

[11] Engle,R.F,Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation[J].Econometrics,1982,987-1007.

[12] Bollerslev,T.Generalized autoregressive conditional heteroskedasticity,Journal of Econemetrics,31,307-327.

[13] 王博．基于 ARMA－GARCH 模型的上证指数实证分析邮科学技术与工程，2012,(05):1219-1221

[14]梁恒.基于 GARCH 族模型的我国沪深股市波动非对称性研巧[ D ].安徽大学,2014.

[15] 廖敏辉．我国股市波动非对称性研，2007,03:55-58.

[16] 刘巧，张倩.亚洲地区股票指数收益率的波动性研巧——基于 GARCH 族模型[J]．2011，01：34-39.

[17] White,H. Economic Prediction using networks:the case of IBM daily stock returns[C].In: Proceeding of the IEEE International Conference on Neural Networks,California,1988: 451-458.

[18] M.H.Pesaran and A.Timmermann,A Recirsive Modeling Approach to Predicting UK Stock Returns[J].Economic Journal,2000,110(460):91-159.

[19] Olson Dennis, Mossman Charles. Neural network forecasts of Canadian stock returns using accounting rations. International Journal of Forecasting,2003,19(3):453-465P.

[20] Quah Tong-Seng,Srinivasan Bobby. Improving returns on stock investment through neural network selection.Expert Systems with Applications,1999,17(4):295-301P.

[21] Motiwalla Luvai, Wahab Mahmoud.Predictable variation and profitable trading of US equities:A trading simulation using neural networks.Computers and Operations Research,2000,27(11-12):1111-1129P.

[22] Yu Shang-Wu.Forecating and Arbitrage of the Nikkei stock Index Futures:An Application of Backpropagation Networks.Asia-Pacific Financial Markets,1999,6(4):341-354P.

[23] Malliaris,M.E. Modeling the behabior of the S&P 500 index: a neural network approach. Artificial Intelligence for Applications, Proceedings of the Tench Conference on 1994.San Antonia, TX USA.1994:86-90P.

[24] Yiwen Yang,Guizhong Liu.Multicariate time series prediction based on neural networks applied to stock market . Systems,Man,and Cybernetics,2001 IEEE International Conference ,Tucson,AZ USA ,2001:2680-2685p.

[25] Roman,J. Jameel,A. Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns.System Sciences, 1996.Proceeedings of the Twenty-Ninth Hawaii International Conference,Wailea,HI,USA,1996:454-460P.

[26] 武振、郑丕谔.基于遗传神经网络的股价波动预测[J].天津大学学报，2004,6（4）：307-310.

[27] 王建军、牛东晓、李莉.基于相似度与神经网络的协同短期符合预测模型[J].华东电力，2009（10）：254-256.

[28] 江弋、林永鹏.RBF 神经网络在股价预测中的应用[J].心智与计算，2007,4（1）：413-419.

[29] 蔡红,陈荣耀.基于 PCA-BP 神经网络的股票价格预测研究[J].计算机仿真,2011,38(3):365~368.

[30] 尹璐．基于 GA-BP 神经网络的股票预测理论及应用[D].北京:华北电力大学,2010.

[31] L. Takeuchi and Y.-Y. A. Lee, “Applying deep learning to enhance momentum trading strategies in stocks,” 2013.

[32] Avraam Tsantekidis, Nikolaos Passalis，”Forecasting Stock Prices from the Limit Order Book using Convolutional Neural Networks”,2017 IEEE 19th Conference on Business Informatics:7-12P.

[33] M. Ugur Gudelek，S. Arda Boluk，”A Deep Learning based Stock Trading Model with 2-D CNN Trend Detection”2017 IEEE.

[34] Feng Li.The Information Contet of Forward-Looking Statements in Corporate Filings-A Naive Bayesian Machine Learning Approach[J].Journal of Accounting Research,2010,48(5):1049-1102.

[36] Geca T,Zahavi J.Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news[J].Decision Support Systems,2014,57(3):212-223.

[36] Fortuny E J D,Smedt T D,Martens D ,et al.Ecaluating and understanding text-based stock price prediction models[J].Information Processing & Management,2014,50(2):426-441.

[37] Skuza M,Romanowski A.Sentiment Analysis of Twitter data within big data distributed environment for stock prediction[C].Computer Scicence and Information Systems.IEEE,2015:1349-1354.

[38] X.Ding,Y.Zhang,T.Liu and J.Duan,“Deep learning for event-driven stock prediction.” in IJCAI, 2015, pp. 2327–2333.

[39] M.C. Chan, C.-C. Wong, and C.-C. Lam, “Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization,” in Computing in Economics and Finance, vol. 61, 2000.

[40] 张世军.基于网络舆情的 SVM 股票价格预测研究[D].南京信息工程大学，2014.

[41] Xiao Ding,Yue Zhang.Deep Learning for Event-Driven Stock Prediction.Harbin Institute of Technology,2015.

[42] 朱梦珺，蒋红迅，许伟.基于金融微博情感与传播效果的股票价格预测简[J].山东大学学报（理学版），2016,51（11):13-25.

[43] 鲁训法, 黎建强. 中国股市指数与投资者情绪指数的相互关系[J]. 系统工程理论与实践, 2012, 32(3):621-629.

[44] 赵茉莉. 网络爬虫系统的研究与实现[D]. 电子科技大学, 2013.

[45] 涂小琴. 基于 Python 爬虫的电影评论情感倾向性分析[J]. 现代计算机, 2017(35):52-55.

[46] 熊畅. 基于 Python 爬虫技术的网页数据抓取与分析研究[J]. 数字技术与应用, 2017(9):35-36.

[47] 周中华, 张惠然, 谢江. 基于 Python 的新浪微博数据爬虫[J]. 计算机应用, 2014, 34(11):3131-3134.

[48] 夏火松, 李保国. 基于 Python 的动态网页评价爬虫算法[J]. 软件工程, 2016, 19(2):43-46.
[49] 刘开瑛. 自然语言处理[M]. 科学出版社, 1991.
[50] Daniel Jurafsky, James H. Martin. 自然语言处理综论[M]. 电子工业出版社, 2005.
[51] 姚天顺. 自然语言理解:一种让机器懂得人类语言的研究-第2版[M]. 清华大学出版社, 2002.
[52] 郑捷.NLP 汉语自然语言处理原理与实践[M].北京：电子工业出版社，2017.
[53]Elshourbagy M, Hemayed E, Fayek M. Enhanced bag of words using multilevel k -means for human activity recognition[J]. Egyptian Informatics Journal, 2016, 17(2):227-237.
[54]Martins C A, Monard M, Matsubara E T.Reducing the dimensionality of bagofwords text representation used by learning algorithms[J]. Actapress Com, 2003.
[55] 周练. Word2vec 的工作原理及应用探究[J]. 图书情报导刊, 2015(2):145-1
[56] 路永和, 李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作, 2013, 57(3):90-95.
[57] 张瑾. 基于改进TF-IDF算法的情报关键词提取方法[J]. 情报杂志, 2014(4):153-155.
[58] 纪思捷, 胡豪杰. 基于机器学习算法的大数据处理[J]. 电子技术与软件工程, 2015(23):202-202.
[59] 褚昆. 基于互信息的统计语言模型数据平滑算法[D]. 哈尔滨工程大学, 2009.
[60] 刘同明. 数据融合技术及其应用[M].国防工业出版社, 1998.
[61] G.Batres-Estrada,“Deep learning for multivariate financial time series,” ser. Technical Report, Stockholm, May 2015.
[62] H.White,Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns, ser. Discussion paper- Department of Economics University of California San Diego. Department of Economics,University of California, 1988.
[63] B.G.Malkiel,“Efficient market hypothesis,”The New Palgrave: Finance. Norton, New York, pp. 127–134,1989
[64] Sreelekshmy Selvin;Stock price prediction using LSTM, RNN and CNN-sliding window model.International Conference on Advances in Computing, Communications and Informatics (ICACCI),2017:1643-1647.
[65] W. Xu,S. Ji, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. Pattern Analysis and Machine Intelligence,IEEE Transactions on, 35(1):221– 231,2013.2,5.

[66] LeCun Y, Bottou L,Bengio Y,et al.Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[67] 伊恩.古德菲洛（Lan Goodfellow）.深度学习[M].北京：人民邮电出版社，2017.

[68] Chen X, Wang C,Xiao B,et al.Learning Convolutional Domain-Robust Representations for Cross-View Face Recognition[J]. Ieice Transactions on Information & Systems, 2014, E97D(12):3239-3243.

[69] Jeff .Donahua,Lisa.Hendricks.Long-term Recurrent Convolutional Networks for visual Recognitions and Description.2014.IEEE

[70] Gu J, Wang Z, Kuen J. Recent Advances in Convolutional Neural Networks[Online]. arXiv:1512.07108, 2017.

[71] Andrearczyk V, Whelan P F. Using filter banks in Convolutional Neural Networks for texture classification[J]. Pattern Recognition Letters, 2016, 84:63-69.

[72] David M. Q. Nelson;Stock market's price movement prediction with LSTM neural networks International Joint Conference on Neural Networks,(IJCNN)2017 :1419-1426.

[73] 孙瑞奇. 基于 LSTM 神经网络的美股股指价格趋势预测模型的研究[D]. 首都经济贸易大学, 2015.

[74] Graves A, Jaitly N, Mohamed A R. Hybrid speech recognition with Deep Bidirectional LSTM[C]// Automatic Speech Recognition and Understanding. IEEE, 2014:273-278.

[75] Gers F A, Schmidhuber E. LSTM recurrent networks learn simple context-free and context-sensitive languages[J]. IEEE Trans Neural Netw, 2001, 12(6):1333-1340.

[76] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10):2451-2471.

[77] Shi X, Chen Z, Wang H, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. 2015, 9199:802-810.

# 攻读硕士学位期间发表的论文和取得的科研成果

[1] 王科俊，李洪强. 基于新闻的股票价格波动的预测.哈尔滨工程大学自动化学院 2017 年学术年会

# 致　　谢

时间如白驹过隙，转眼间研究生生活已经接近尾声，在这两年时光里，我有幸能够在自己感兴趣的人工智能领域探索和研究，研究期间遇到很多困难与不解，在老师、亲人、同学的帮助下，一步一步不断提高自己学习知识、独立思考的能力，慢慢懂得了滴水穿石，勤能补拙的道理。

在这两年的研究生生活里，首先要感谢的就是王科俊老师。王老师在我整个研究生生涯中对我起到了很大的指导作用。想当初，懵懵懂懂的考研结束之后，进入到实验室，最早接触到王老师，我对王老师的印象就是一个生活中很随和的人。但当真正的接触的课题开始进行研究之后，王老师对于研究又是非常的严格。每次开会的时候，都会针对我们正在做的工作提出看法和存在的问题所在，不断引导着我们完成整个课题。同时，王老师深厚的理论知识以及认真负责的态度也深深的影响着我，不断的督促着我向前迈进。在完成这个论文期间，王老师针对我的论文又不断的指导着我，启发着我，这才使得我能够顺利的完成论文。感恩王老师！

同时还要感谢408所的所有老师，郝学森博士，孙庆刚博士还有我同届的同学们，每次遇到一些问题的时候，我们总是能够一起商量来解决问题，共同面对问题。感谢徐丹丹、郭俊垚、陈静、王浩霖、曹逸、高瑞岐，我们在这两年的时间里一起进步，一起成长，收获颇多。

最后感谢我的家人，无论何时都给予我最大的支持和鼓励。