涉密论文 □　公开论文 □

浙江大学

本 科 生 毕 业 论 文

题目　基于深度增强学习的量化投资研究

姓名与学号　马啸远

指导教师　郑小林

合作导师

年级与专业　计算机科学与技术 2014 级

所在学院　计算机科学与技术学院

提交日期

浙江大学本科生毕业论文承诺书

1.本人郑重地承诺所呈交的毕业论文，是在指导教师的指导下严格按照学校和学院有关规定完成的。

2.本人在毕业论文中除了文中特别加以标注和致谢的地方外，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 浙江大学 或其他教育机构的学位或证书而使用过的材料。

3.与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文选题和研究内容过程中没有伪造相关数据等行为。

5. 在毕业论文中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6.本人完全了解 浙江大学 有权保留并向有关部门或机构送交本论文的复印件和磁盘，允许本论文被查阅和借阅。本人授权 浙江大学 可以将本论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

致 谢

时光荏苒，白驹过隙，大学四年已经临近尾声。经过几个月时间的磨砺，毕业论文终于完稿，回首这么长时间来收集整理文献资料，学习相关技术，并最终顺利完成了项目和论文，我得到了非常多的帮助和关心，在这里我要向他们表达最诚挚的谢意。

首先感谢我的导师，计算机科学与技术学院的郑小林副教授。郑老师为人谦和，平易近人。在选题、搜集资料和写作阶段，都给予了我莫大的关怀和鼓励以及悉心的指导。每当我碰到问题，老师总会耐心地给我解答，提出中肯的意见。他严谨求是的治学之风和对事业的孜孜追求将会一直影响并激励我。借此机会，我谨向郑小林老师致以深深的谢意。

其次，我要感谢我在实习阶段公司中的同事们。在我毕业论文的写作过程中，向他们提过许多专业相关问题，他们都会一一进行解答，将自己的经验传授给我。尤其在最近项目上线阶段，有同事因为我在写毕业论文承担了一部分我的工作。在这里，我谨向我的同事们致以我最真诚的感谢，谢谢你们的帮助，谢谢你们的包容。

另外，我还要感谢我的父母。焉得艾草，言树之心。你们的养育之恩我无以为报，只能不停努力，提升自己，才对得起你们的栽培。你们健康快乐是我的最大心愿！

最后，感谢我的母校浙江大学，临近离校之际，心中有千言万语无法表达！只愿自己走上社会之后不辜负您的教育之恩，今天我以母校为荣，希望明日母校能以我为荣！

摘 要

量化资产配置属于量化投资领域的一个重要方向，旨在通过管理一个资产组合，并合理配置其中的资产比重，达到分散风险，提高投资效率的目的。如今，在大数据时代背景下，机器学习技术发展迅速，尤其是近几年增强学习算法的进步，被越来越多得应用在实际场景下。因此，尝试使用机器学习技术来解决量化投资问题是很有研究意义的。在本文中，笔者会尝试使用DDPG深度增强学习算法来构建一个量化投资模型。

本文从数据收集处理、具体算法的实现和网络结构等方面来对本研究进行深入探讨。最后，使用多个测试基准来对模型进行评估。另外，为了更好地对算法进行比对研究，笔者还会使用另一种算法——模仿学习，同样来构建一个模型，以此作为对比。在最后的测试中，深度增强学习算法训练的模型表现出了出色的效果并优于其他模型，最优秀的模型在不到2年内可以达到650%以上的收益率，这表明利用深度增强学习来训练量化投资模型是合理并且有价值的。

关键字：量化投资、投资组合管理、深度学习，增强学习，模仿学习

Abstract

Quantitative asset allocation is one of the important directions in the field of quantitative investment. It is mainly a process of managing one portfolio and allocate the proportion of assets quantitatively, to spread risk and increase investment efficiency. Nowadays, with the boost of machine learning, new ideas have been come up with to solve portfolio problems, especially with the improvement of reinforcement learning algorithms. In this work, I'm introducing a quantitative investment model with DDPG deep reinforcement learning algorithm.

In this paper, I will detail the data processing, algorithm and the framework of the networks. Finally, the model will be tested based on multiple benchmarks. In addition, in order to prove this model has a great performance, I will also build a model with imitation learning algorithm. In the final test, the DRL model performs excellently and out distances the other models. The best model even reaches a rate of return over 650%. Our work proved that it is reasonable and significant to train quantitative investment models using DRL algorithm.

Key words: Quantitative Investment、Portfolio、Deep Learning、Reinforcement Learning、Imitation Learning

目　　次

第一部分　毕业论文

第一部分

毕业论文

第一章 绪论

1.1 背景介绍

1.1.1量化投资

在金融业如此繁荣的今天，“投资”（Investment）的概念早已经深入人心，大到银行公司，小到个人散户，或多或少都会接触到股票、证券、期货等金融产品的投资交易。而随着信息化时代的来临，一门新的理论“量化投资”（Quantitative Investment）被提出并发展壮大。其旨在通过程序化和数量化的手段，由计算机自动发出买卖指令，以获取稳定收益。到目前为止，量化投资已经发展近 30 年，由于业绩稳定，市场份额一直在不断扩大，也越来越得到投资人的认可。

量化投资不同于传统定性投资观念的一个重要特点就是“概率取胜”，即，一方面定量投资会不断从历史数据中挖掘有用信息并加以利用，另一方面是依靠一组资产组合取胜，而不是单个或几个。因此，资产组合在量化投资中显得尤为重要。“投资组合管理”（Portfolio）即是指投资人按照合理的资产选择理论对资产进行多元化管理，以实现分散风险、提高投资效率的目的。投资组合主要是通过选择各种各样的证券和其他资产组成资产组合，并管理这些投资组合，达到最大化回报，最小化风险的目的。投资组合的管理主要由两类活动构成：资产配置；调整主要资产间的权重。这是一个综合性的问题，需要对证券的微观统计、宏观政策、舆论影响等多个方面进行分析，尽量挖掘出更多的有效信息，并做出多元化的投资方案，组建相对合理的组合。

本文所研究的重点就是将传统投资组合思想结合和计算机量化分析技术相结合的一种投资策略，量化资产配置。量化资产配置属于量化投资的一个重要方向，通过对各类证券公开数据进行统计分析，进而确定投资组合的配置比例。在当今大数据背景下，金融行业的数据量是巨大的，包括各类金融产品的价格、交易情况，加上公司的基本面情况等等，都为量化投资分析提供了有效的数据支持。

1.1.2深度增强学习

随着大数据时代的来临，机器学习技术得到了学术界和工业界的广泛关注和重视——机器学习技术。机器学习是一门利用概率论、统计学、线性代数等多领域解决问题的学科，其主要研究对象是人工智能，旨在通过经验自动改进计算机算法近年来。其中，深度学习和增强学习在近几年尤为热门，尤其是在 2016 年 Google AlphaGo 一战成名之后。AlphaGo 就是使用了深度增强学习和蒙特卡洛树搜索算法结合来构建了一个围棋 AI[17]。

深度学习是机器学习中一种基于对数据特征进行学习的方法，主要灵感来源于人工神经网络的研究，希望通过模仿人脑的机制来解释数据。深度学习的理论证明指出，具有一定深度的神经网络通过适当训练可以模拟出任意的线性或非线性函数[7]，这个特性是深度学习的核心。其后，卷积神经网络（Convolutional Neural Network, CNN）、递归神经网络（Recurrent Neural Network, RNN）、长短期记忆模型（Long Short Term Memory, LSTM）等等新的网络结构陆续被提出，可以被应用在各自适用的领域中。

增强学习可以被描述为一个“马尔科夫决策过程”（Markov Decision Process, MDP），即，智能体与环境不断地进行交互并使环境发生变化，学习一个从环境到动作的映射。。从早期的蒙特卡洛增强学习算法，TD（Time Differential）算法，Q-Learning 算法，到近几年的 DQN（Deep Q-Learning Network）、策略梯度（Deterministic Policy Gradient, DPG）、actor-critic 等算法的提出和实践，增强学习方法得到了越来越多的发展和应用。

深度增强学习（DRL）是将这两者结合的一种方法，利用深度神经网络挖掘环境（state）的特征，让增强学习能力大增。

由于机器学习的强大能力，将其应用在量化投资领域将一直都是一个长久的研究课题。

1.2 本文研究目标和内容

近几年 DeepMind 提出了很多新的深度增强学习算法，但是目前应用在量化投资上的研究还不算多，这让笔者想到可以利用这些算法来进行量化投资算法的研究。2016 年 DeepMind 在其发表的论文中提出了一种基于策略的 off-policy 增强学习算法 DDPG（Deep Deterministic Policy Gradient），并证明该算法具有更好的收敛性和稳定性。在本文中，笔者将介绍如何用 DDPG 算法来训练一个量化投资模型。模型基于以下两个目标：

1）自动化管理投资组合

模型会管理某个股票组合（在本实验中将从美国 S&P 500 指数中选取 16 支典型股票）的比重，并在每个交易时段的末尾根据历史数据配置新的比重。

2）收益最大化

模型需尽可能地使一个投资周期内的回报收益最大化。

研究主要针对以下几点展开：

1）研究对比现有模型

对目前的深度增强学习算法及在量化投资上的应用研究进行详细研究和对比，可以有助于我了解如何将算法与投资应用相结合，并且借鉴各种算法的优点，认识缺点并尝试改进。

2）证券数据的收集和处理

目前网络上有大量证券数据的接口可供使用。在本文中，证券数据的来源是 kaggle，这是一个全球性的机器学习竞赛和代码托管平台。

3）实现基于 DDPG 的量化投资模型

DRL 深度增强学习是一种典型的无监督学习方法。在 Deep Mind 的论文中，详细介绍了 DDPG 算法的原理[24]。在此论文基础上，我使用 TensorFlow 深度学习框架具体实现了该算法，并且应用在量化资产配置问题上。我选择了一个由 16 支股票构成的投资组合，并将投资回报单步化。

4）实现基于模仿学习的量化投资模型

模仿学习也是近些年新提出的一种为了解决增强学习难以收敛问题的算法。不同于增强学习算法，它是有监督的。用模仿学习算法针对同样的投资组合训练一个模型，和 DDPG 训练的模型综合比较。模仿学习同样也是解决智能体学习问题的一种重要算法，因此对它们进行比较是有意义的。

5）分析评估模型性能

选择合适的测试基准，来评估模型的能力。本研究中主要通过三个指标来评估模型的性能，收益率、夏普比率——一个常用的基金绩效评价标准，以及最大回撤。另外为了对比分析，设置 6 组不同的超参数，单独训练 6 个模型，并且对它们的性能进行对比评估找出最好的一组参数

1.3 本文结构安排

本文正文将按照以下结构进行安排：

第一章绪论，将对本文内容进行总述。

第二章相关工作，将介绍在该问题上的其他相关研究，并分析各自优缺点。

第三章研究方案，将分为 5 小节对研究的具体算法和方案进行介绍。

第四章实验结果与分析，展示了本研究在测试中的实验结果，以图和表的形式给出。并结合理论分析，得出结论。

第五章结论，将对本文进行总结，并提出未来的工作方向。

第二章 相关工作

2.1 工业界现状

目前在工业界，做量化投资比较成熟的做法依然是利用统计学进行套利。大多数机构进行投资主要是运用一些传统理论，如 CAPM（Capital Asset Pricing Model，资产定价模型）、Ross 资产定价定理、Fama,French 三因子模型等，对资产进行定价分析，再结合配对交易以及一些对冲交易模型从中套利，本质上是利用了价格与价值的差价。计算机的应用主要体现在进行统计学分析，得到一些数据特征帮助人做决策。另外，基本面分析在投资策略中占很大一部分，技术分析更多是作为辅助手段。不过在美国华尔街等金融业比较发达的地方，已经越来越多地用计算机替代交易员，不依靠人的任何策略。

目前，深度学习在量化投资上的应用依然处在研究中，实用比较少。因此，利用深度学习和增强学习进行量化投资的研究是有必要和有意义的。

2.2 学术界相关研究

2.2.1 通过预测价格

最直接的方法就是通过证券市场的历史价格数据，预测其未来数据[6]。早期有大量的研究希望通过深度神经网络直接对证券的价格进行涨跌预测，进而做出反应。可以通过建立简单的线性神经网络（多层感知器）或稍微复杂的 CNN，输入某支股票的历史数据（5 天或 10 天或更多），来预测下一天的价格[13]，这是一个典型的回归问题。这种方法简单直接，并且很容易实现，只要数据量足够大即可，但是，从测试结果来看，预测的效果却并不理想。后来，LSTM 的应用，又给了价格预测新的思路。这种长短期记忆模型，可以有选择地记忆或者遗忘以往的数据特征，非常适合证券价格这种具有很强时序性的数据。但是利用 RNN 或者 LSTM 来预测价格的实验效果依然有限，这很可能是因为容易对训练数据过拟合，在测试时的准确率下降非常严重。而且，股票市场过于复杂，影响因素众多，直接预测价格本身就是非常难的。最重要的是，这没有体现出量化投资的多元化投资思想。因此基于预测价格的思路目前是不太可行的。

2.2.2 计算固定投资组合

投资组合可以有效分散风险，提高投资效率，因此有很多研究希望能够计算出一个固定的投资组合以此打败市场。

马科维茨在其论文 *Portfolio Selection* 中提出了著名的均值-方差模型马科维茨均值-方差模型[1]，其大致思想可以概括为，统计出各支股票的期望收益率和方差，画出它们的投资机会集并找到一条有效边界。但是在证券数量多的情况下计算量巨大，马科维茨模型不能直接应用。有研究者希望利用神经网络来估算出这条有效边界曲线。在 Alberto et al.的研究中，作者将马科维茨模型视为一个“组合优化”（Combinatorial Optimisation）问题，并尝试用 Hopfield 网络来求解[19]。虽然该方法效果较为理想，但是马科维茨模型本身是存在缺陷的[4]。

AE（Auto Encoding, 自编码模型）是深度学习的另外一个重要内容，主要思想是希望对数据降维，找到其“隐变量”，由一个编码器（Encoder）和一个解码器（Decoder）组成。其中编码器负责找到数据的“隐变量”——代表数据的底层特征，即减少多余特征维度后的数据，而解码器则负责将“隐变量”还原成原数据。在 Hu et al.[22]的研究中，利用 AE 尝试训练了一个对股票 K 线图编解码的模型。可以认为编码代表了某支股票的底层特征。然后，他们对所有股票的“隐变量”进行聚类，将整个指数中的股票分成了多类。在构成投资组合时，挑选每个类中“最好”的那一支股票，这个“最好”用夏普比率（Sharpe Ratio）来衡量。该方法取得了不错的效果。在 Heaton et al.[10]的研究中同样借鉴了 AE 模型，试图将某个指数的各股权重输入并输出一个新的权重，企图“打败指数”。

这类通过计算固定的投资组合的方法不具有实用性，因为无法根据市场变化即时进行调整，缺乏灵活性。

2.2.3 增强学习的应用

随着增强学习的发展，开始有研究者将其应用在量化投资领域。

在增强学习发展早期，Q-Learning 算法非常流行。Dempster et al.[15]和 Moody et al.[14]在其研究中都是使用了朴素的 Q-Learning 算法，使用 epsilon-greedy（spsilon 贪心策略），企图迭代多次来找到一个动态平衡。这些研究都只用了一支股票，没有实际意义。2013 年，Deep Mind 团队的一篇论文提出了 DQN 算法，将 Q-Learning 和深度学习结合，并在电脑

游戏中得到了异常好的效果（Atari[16], AlphaGo[17]）。很快，又有研究者希望将 DQN 算法引入量化投资研究。Deng et al.[20]和 Jin et al.[12]都在其工作中尝试了 DQN 算法，并且扩展了证券范围，使用了两支股票，其中一支具有较高波动性，另一支则较低，以此来模拟市场真实情况。上文提到，Q-Learning 以及一些早期的增强学习算法，都要求动作空间必须是离散的，这完全不符合市场投资的特点。而且，DQN 算法是一种基于值函数估计（Value-Based）的算法，收敛非常慢，而且无法学习到随机的策略。

DPG（Deterministic Policy Gradient，确定性策略梯度）算法直接对策略空间建模，也就是说，输入一个状态，可以直接输出一个动作，而不是从离散的动作空间中选择。Jiang et al.[9]在其研究中，分别使用 CNN、RNN、LSTM 三种神经网络结合 DPG，构造了一个自动交易系统。DPG 虽然避免了基于值函数估计难以收敛的问题，但它无法单步更新参数，只能回合更新。训练效率非常低。另外，基于策略估计的算法很容易陷入局部最优解。

以上介绍了一些相关的工作，都有其各自的优缺点。最近几年，Deep Mind 团队发表了一些新的深度增强学习算法，如 actor-critic，A3C、DDPG，目前笔者还没有找到研究者应用在量化投资上面。DDPG（Deep Deterministic Policy Gradient， 深度确定性策略梯度）是在 actor-critic 算法的基础上改进的，除了包含 2 个主要部分，actor——用于在每个决策步骤制定对应的策略，以及 critic——用于估计状态的价值外，同时，DDPG 还结合了 DQN 的思想，引入了 target network，并且模仿 DQN 算法，记录一个 Replay Buffer，用于参数延迟更新。因此，DDPG 算法综合了多种优势，兼有基于价值和基于策略的思想，同时引入 DQN 的思想，既提升了训练效率，也让算法更容易收敛。本文的研究课题就是利用 DDPG 算法来构建量化投资模型。

第三章 研究方案

在本章中，我会详细介绍针对该研究我做了哪些工作，具体的研究方案，以及算法实现的细节。该节将分为 5 个小节，第 1 节将对研究进行概述并对问题进行数学形式上的描述，第 2 节将介绍数据的收集和预处理，第 3 节将对研究中如何进行环境的抽象进行介绍，第 4 节和第 5 节分别介绍 DDPG 算法和模仿学习算法，以及如何用算法构建量化投资模型。

3.1 概述

量化资产配置具体来讲，就是在一个投资周期中，连续地将总资产重新分配给多个金融产品，并在每个交易时段不停重复这个过程。在本文的研究中，我会将问题做一定程度的简化：

第一，所有研究只针对股票这一种金融产品，并且所有股票均来自美国 S&P500（Standard & Poor's Index，美国标准普尔指数）。

第二，将交易周期设置为固定值——一天。投资是时间驱动性非常强的行为，因此需要比较频繁地对市场进行观察并作出反应。遵循基本的交易逻辑，再加上获取到的数据大多是以天为基本周期的，因此在本文中就将一天作为一个交易周期。而这一天最重要的数据就是一支股票的开盘价、收盘价、最高价、最低价这四个特征。有且只有这四个特征将会成为资产配置的依据。

第三，所有的研究都是基于以下两个基本假设：

1) 所有交易行为均能完成且迅速完成。在实际的市场中，并非所有交易请求都会成功，即使成功也很难没有延迟地完成。在本研究中，将假设这些问题都不存在。
2) 所有交易对市场本身都没有影响。在实际的投资活动中，尤其是资金数量较大的情况下，所有交易行为都会对市场造成影响。而在本研究中，将假设这些影响太小因此忽略不计。

第四，实际上所有股市的交易都不是无成本的，这涉及到印花税、交易佣金等，因此，每一次交易都会让总资产按照一定比率缩减。在本研究中，将用μ_t来表示第t个交易时段产生的交易成本比率。简便起见，我将把缩减比率设为固定值，$\mu_t \equiv 0.005$。

在研究中，为了方便定义和推导，必须对问题进行数学形式上的描述。在形式定义上，我参考了这篇论文[9]。

定义N为我们要投资的股票的总数量。不失一般性，在下面的所有数学符号中，下标带有t的都是有时序性的属性，表示在t个交易时段末尾的值，而T表示整个投资周期的最后一个时段。另外，所有的向量的长度都是$\mathrm{N}+1$，这是因为添加了现金。

$\boldsymbol{v_t}$表示所有股票的价格，这是一个向量。定义$\boldsymbol{y_t}$为 收盘价 / 开盘价 的比值，即：

$$\boldsymbol{y_t} = \boldsymbol{v_t} \oslash \boldsymbol{v_{t-1}} \tag{3-1}$$

其中$\oslash$表示逐位相除。注意到$\boldsymbol{y_t}[0] \equiv 1$，因为现金价值不变。

定义p_0为初始总资产，p_t表示第t交易时期末期的资产总值。

定义$\boldsymbol{\omega_t}$表示 t时刻各股价值的权重，$\omega_{t,i}$就代表持有第 i支股票的价值占总资产的比重。实际上，在本研究中，我们的目标就是要用历史交易数据来得到$\boldsymbol{\omega_t}$作为投资建议。

现在，可以推导得到这个公式：

$$p_t = (1 - \mu_t) p_{t-1} \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} \tag{3-2}$$

其中·表示向量点积。另外，$p_0 = 1$。

因此可以推导得出这个公式：

$$\frac{p_T}{p_0} = \prod_{t=1}^{T} (1 - \mu_t) \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} \tag{3-3}$$

定义收益率为ρ_t。则可以推导得出下面三个公式：

$$\rho_t = \frac{p_t}{p_{t-1}} - 1 = (1 - \mu_t) \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} - 1 \tag{3-4}$$

$$log\rho_t = log(1 - \mu_t) \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} - 1 \tag{3-5}$$

$$\log \frac{p_T}{p_0} = \log \prod_{t=1}^{T} (1 - \mu_t) \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} = \sum_{t=1}^{T} \log (1 - \mu_t) \boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} \tag{3-6}$$

之所以要定义对数收益率，这是因为在下面的算法实现中需要使用到，下面会具体介绍。

在t时段开始时，权重向量为ω_{t-1}，在末尾时，由于这一天中价格的变化，权重向量将变为下面这个值：

$$\boldsymbol{\omega_t}' = \frac{\boldsymbol{y_t} \odot \boldsymbol{\omega_{t-1}}}{\boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}}} \tag{3-7}$$

其中$\odot$表示逐位相乘。

这个时候，量化投资模型会直接给出一个新的ω_t，然后对股票进行买和卖的操作，重新配置资产。这个过程用下图可以更清晰地展示：

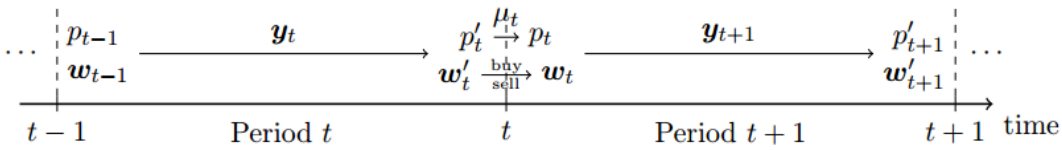

图 3.1 资产总值和权重向量的变化示意图

3.2 数据收集和处理

本研究中用到的所有股票数据集，均来自 kaggle。kaggle 是一个世界著名的机器学习竞赛和代码托管平台，上面有大量的数据集，并且可以保证真实性、有效性。

date	open	high	low	close	volume	Name
2013/2/8	15.07	15.12	14.63	14.75	8407500	AAL
2013/2/11	14.89	15.01	14.26	14.46	8882000	AAL
2013/2/12	14.45	14.51	14.1	14.27	8126000	AAL
2013/2/13	14.3	14.94	14.25	14.66	10259500	AAL
2013/2/14	14.94	14.96	13.16	13.99	31879900	AAL

图 3.2 kaggle 股票数据集展示

数据集本身是 CSV 格式的，可以方便地读取到内存，用 Python 的 numpy 库整理成 ndarray 的形式，利用 h5py 库保存到硬盘。最后处理完成的数据是这样的：

```
CMCSA
[[ 34.91    35.2     34.73    34.83  ]
 [ 35.01    35.56    34.81    35.4   ]
 [ 35.37    35.43    34.68    34.71  ]
 ...,
 [ 93.48    95.06    90.3172  90.72  ]
```

图 3.3 处理后的股票数据示意图

可以看到每支股票的数据是以多个四元组组成的。四元组中的数字就表示 open，high，low，close。

股票历史数据是模型进行资产配置的唯一依据。每次决策时，模型会读取一个 window_length 长度的向量，即当天的前 window_length 天的数据将会作为输入。需要注意

的是，并不是将 4 个数据全部输入，需要对数据进行处理。为了消除数据纲量，另外由于每支股票的价格量级是不同的，我们更关心的是价格的变化关系，输入价格本身是没有意义的，因此只需要保留 close / open 的比值。这个比值是一个范围为$(0, \infty)$的数，在数据量非常大的情况下可以认为其数学期望值为 1，因此这里我们将对这个比值进行标准化（Normalization）处理，让处理后的数据均值变为 0。最终作为网络输入的数据由如下公式计算得到：

$$d = \left(\frac{close}{open} - 1\right) \times 100 \qquad (3-8)$$

3.3 Portfolio Env

3.3.1 马尔科夫决策过程

在增强学习的技术领域中，有一个最基本的概念是马尔科夫决策过程（Markov Decision Process，MDP）。这是一个概率论和统计学范畴中的数学模型，用来描述在部分随机、部分可控的状态下，如何进行决策的过程。简单来说，在生活中的大多数决策过程都可以被视为一个马尔科夫决策过程。在 MDP 中，我们可以用四个基本要素来对决策进行描述：智能体（agent）；环境（state / env）；动作（action）；奖励（reward），如图 3.4 所示。一个增强学习算法其实就是一个智能体与环境不断交互的过程，通过不断得到环境的反馈——奖励，来调整自己的决策策略，最终能够学习到一个环境到动作的映射，或者说，决策。可以看到，任何学习的过程其实都是类似的。

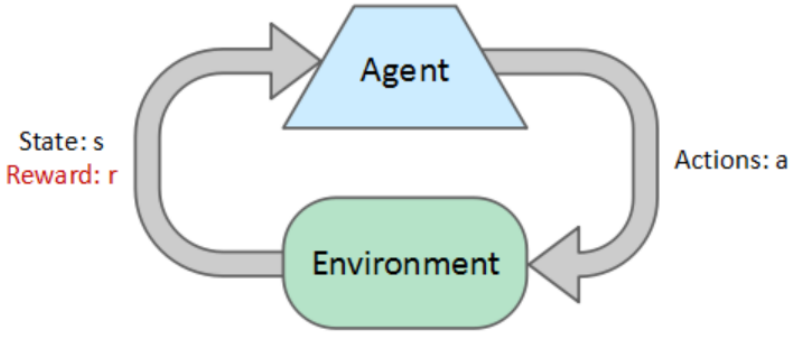

图 3.4 马尔科夫决策过程

MDP 的关键问题在于寻找到一个策略，可以在状态s下选择依据策略π选择动作a，并且目标是可以让累积收益R(T)最大化：

$$\pi = argmax\{\mathrm{R(T)}\} = argmax\{\sum_{t=0}^{T}\gamma^t R_{a_t}(s_t, s_{t+1})\}\, s.t.\ a_t = \pi(s_t), 0 \le \gamma < 1 \quad (3-9)$$

显然，投资也可以被视为一个典型的马尔科夫决策过程，并且各个部分的含义如下：

Action：动作，即投资组合中各股的权重，也即$\boldsymbol{\omega_t}$。

Agent：代表一个量化投资模型，可以自动给出动作。如本研究中实现的 DDPG 模型和模仿学习模型。

State：agent 在每一步决策是观察到的市场情况。具体来讲，每一步决策时，都是依据一个 window_length 长的窗口中的股票历史数据，所以形状为$(N+1, windw_length, 1)$（N 加 1 是因为加上现金）。

Reward：代表每一步交易带来的单步收益回报。在投资活动中，投资者一般更多关注的是一个周期中的总收益率，也就是$\frac{p_T}{p_0} - 1$，没有针对单个交易行为的收益的表述。实际上，最大化$\frac{p_T}{p_0}$也就等价于最大化$\log\frac{p_T}{p_0}$。从公式$(3-6)$可以看到，$\log\frac{p_T}{p_0}$被写成了累加和的形式，形式上与公式$(3-9)$给出的累积收益R(T)相符，这里可以认为$\gamma = 1$。这也算是对问题的一个假设，我们假定整个投资周期中的所有交易决策对于当下都是等同重要的，不存在随时间推移而降低的因素。因此，每一步交易的单步 reward 可以被定义为：

$$r_t = \log(1 - \mu_t)\boldsymbol{y_t} \cdot \boldsymbol{\omega_{t-1}} \quad (3-10)$$

3.3.2 Environment

在本研究中，环境代表股票市场。为了方便地进行数据读取交互和调试，需要对环境进行抽象。这里我参考了 openAI 的 gym 环境库。gym 是一个用于开发和比较增强学习算法的工具包。在 gym 中，所有环境都有一套统一的 api，包含三个基本函数，reset，step，render。我借鉴了这个设计思路，设计了 Portfolio Env，作为本研究中所有算法的环境。该环境为交易、收益率计算、绘图等提供了统一的解决方案。

```
class PortfolioEnv(gym.Env):
    metadata = {'render.modes': ['human', 'ansi']}
    def __init__(self,…

    def step(self, action):…
    def reset(self):…
    def render(self, mode='human', close=False):…
```

图 3.5 Portfolio Env 代码示意

reset：重置环境，将初始时间随机地初始化。

step：执行一步动作，并得到回报。即进行一次交易，根据公式(3 − 10)计算得到单步回报。

render：封装绘图函数，将所有量化投资模型的结果以曲线形式画在一张图上。

3.4 DDPG 投资模型

3.4.1 DDPG 算法

在增强学习理论中，目前有两种不同的算法方向。第一种是基于价值估计的算法，旨在希望对某一个状态采取的动作进行价值的评估，决策时只需选择对应价值最大的动作。另一种是基于策略估计的算法，旨在直接给出基于某个状态下的动作空间的概率分布，决策时从此分布选取概率密度最大的动作或是随机选取一个动作。

作为基于值函数估计的算法代表，DQN（Deep Q-Learning Network，深度 Q-learning 神经网络）在 2013 年被 DeepMind 提出，这是首次将深度学习和传统增强学习相结合。DQN

后来被成功应用了在 Atari 游戏上，取得超越人类的优异表现[16]。但是，DQN 算法本身存在几个严重缺陷。首先，DQN 只能应用在离散动作空间上，缺乏灵活性；第二，基于值函数估计的算法被证明比较难以收敛，容易陷入局部最优解，而且这种方法只能学习到固定的策略，所以任何时候只要面对同一个环境状态都将输出确定的策略，在“状态重名”发生时这将是灾难性的；第三，有时候定义一个状态的价值是困难、模棱两可的[23]，比如一个球在空中下落需要我们接住时，左右移动的价值并不具备太大的意义，而直接基于策略的方法则可以直接地给出具体动作。

在确定性策略梯度算法（DPG）的基础上[23]，DeepMind 于 2016 年提出了一种基于策略、off-policy 的 actor-critic 算法 DDPG（Deep Deterministic Policy Gradient，深度确定性策略梯度算法），并且说明了该算法具有更好的收敛性[24]。相较于传统的 actor-critic、DPG 算法，DDPG 主要解决了四个问题：

1) DDPG 算法实现参数单步更新，提高了效率。传统的 actor-critic 算法要求每一次参数更新都需要一个完整的回合然后得到最终的收益，比如游戏需要完成完整的一局，投资需要完成一个投资周期，大大降低训练效率，这是完全无法接受的。DDPG 像 DQN 一样，可以单步更新。
2) DDPG 算法借鉴了 DQN 参数延迟更新的思想。参数延迟更新的做法是维护一个 Replay Buffer，每一次做出决策后，将(s_t, a_t, s_{t+1}, r_t)这样一个元组记录下来，其中s_t表示当前状态，a_t表示采取的动作，r_t表示获得的单步收益。而在参数更新时从中随机抽取出一个 batch。由于所有记录都是随机并且没有时序相关性的，因此可以在一定程度上将估值的方差相互抵消。
3) DDPG 算法引入 target network，并一脉相承地继承了 off-policy 的特性。传统 actor-critic 算法每次都是依据策略梯度，直接对策略网络和价值网络进行参数更新，这样做会让网络变得不稳定[24]，难以收敛。而类似 DQN 等算法的做法是引入一个 target network，即目标网络，策略网络和价值网络每次参数更新都有一个明确的目标，虽然这个目标可能暂时是不完善、还没有被训练好的。但是在长期的训练之后，两者可以同时趋于收敛并稳定。引入 target network 将训练过程分为了两个部分，相当于将不稳定因素减半分给双方，因此一定程度上提高了算法稳定性，当然代价是训练时间会变长。target network 是典型的 off-policy 的方法，我们还可以从此类算法得到一个额外的好处，即可以将已有的 policy 作为 target network 的初始化，可以用人类决策有监督训练的，也可以是其他模型训练好的。2016 年大放

异彩的 AlphaGo 就是先用人类棋谱训练了一个 policy，作为增强学习策略网络训练的初始化[17]。

4) DDPG 算法加入了随机策略。连续动作空间的问题有一个很大的挑战是在开发（exploit）和探索（explore）之间难以找到平衡。DDPG 采取了一个折中的方式，将探索过程与网络学习本身分离。在生成动作时，会给动作添加一个正态分布的随机噪声。这有助于提高网络的稳定性和容错性，并且可以一定程度上避免陷入局部最优解。

可以看到，DDPG 算法包含了 4 个网络，1 个策略网络$\mu(\mathrm{s}|\theta^{\mu})$，1 个目标策略网络$\mu'(\mathrm{s}|\theta^{\mu'})$，1 个价值网络$\mathrm{Q}(\mathrm{s},\mathrm{a}|\theta^{\mathrm{Q}})$，以及 1 个目标价值网络$\mathrm{Q}'(\mathrm{s},\mathrm{a}|\theta^{\mathrm{Q}'})$。在训练过程中，每一次做决策都依据如下公式：

$$a_t = \mu(s_t|\theta^{\mu}) + N_t \tag{3-11}$$

其中N_t是加入的噪声。

在参数更新时，随机从 Replay Buffer 中选取 N 个记录(s_i, a_i, s_{i+1}, r_i)。令：

$$y_i = r_i + \gamma \mathrm{Q}'\left(s_{i+1}, \mu'\left(s_{i+1}\middle|\theta^{\mu'}\right)\middle|\theta^{\mathrm{Q}'}\right) \tag{3-12}$$

其中γ为衰减率。

然后通过最小化下面这个 TD error（Time Differential Error）对价值网络进行参数更新：

$$L = \frac{1}{N}\sum_i \left(y_i - \mathrm{Q}'\left(s_i, a_i\middle|\theta^{\mathrm{Q}'}\right)\right)^2 \tag{3-13}$$

依据策略梯度定理，可以计算得到下面的策略梯度：

$$\nabla_{\theta^{\mu}} J = \frac{1}{N}\sum_i \nabla_{\theta^{\mu}} \mu(\mathrm{s}|\theta^{\mu})|_{s_i} \cdot \nabla_a \mathrm{Q}(\mathrm{s},\mathrm{a}\left|\theta^{\mathrm{Q}}\right)|_{s=s_i, a=\mu(s_i)} \tag{3-14}$$

最后，更新 2 个目标网络：

$$\theta^{\mathrm{Q}'} = \tau\theta^{\mathrm{Q}} + (1-\tau)\theta^{\mathrm{Q}'} \tag{3-15}$$

$$\theta^{\mu'} = \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{3-16}$$

其中τ是学习率，$\tau \ll 1$。

3.4.2 网络结构

在这一节中，我会详细介绍 DDPG 量化投资模型的具体网络结构。网络分为 2 个大部分，策略网络和价值网络。

3.4.2.1 策略网络

策略网络$\mu(\mathrm{s}|\theta^{\mu})$，在 DDPG 算法中又被称为 actor，主要负责得到一个 state 即市场状态，输出应当采取的 action，对应于具体的投资组合权重。

由于证券市场具有很高的复杂性，信息量巨大，所以很难对环境进行完全建模。但是，“有效市场”假说指出，如果市场是严格有效的，那么价格会随着新信息的出现迅速变动找到新的平衡。也就是说，所有股市的公开信息，都可以从价格中反映出来[2, 18]。最近有研究将股市分为三个级别，弱势、半强势、强势，越强势则表明市场越有效，价格更能充分反映信息，而美国股市属于半强势市场，因此价格可以很大程度上反映市场的状态。所以，在这里我们可以直接将股票的历史价格作为 state 输入。网络看到的历史价格的天数由超参数$windw_length$决定，所以策略网络的输入将是一个形状为$(N+1, windw_length, 1)$的 tensor（N 为投资组合的大小）。在本研究中，$windw_length$将作为一个超参数，我将设置三组不同的值 3、7、14，来对比网络看到不同的历史天数得到的不同的表现能力。

整个策略网络的架构如图 3.6 所示：

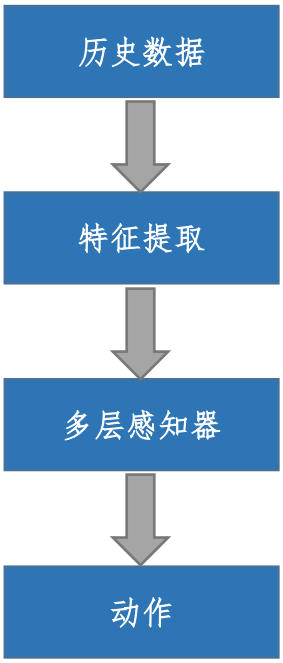

图 3.6 策略网络架构示意图

网络得到一个输入后，通过特征提取部分得到历史数据的特征，这个特征的维度一定比原始输入要高很多，我们认为这个特征可以充分表达市场的状态信息。特征被输入一个多层感知器（Multi-Layer Perceptron, MLP）。多层感知器是一种简单的神经网络模型，能够拟合复杂的非线性函数，我们利用它做的事情是拟合一个动作映射函数，计算得到投资组合的权重向量。为了保证权重向量的元素和为1，需要在最后一层的输出后添加一个softmax作为激活函数，不过这里我们的目的不是为了分类，而是得到一个归一化的向量。例如投资组合有4支股票，那么最后的action就有可能为$[0, 0.5, 0.2, 0.3, 0]$这样一个向量（第一个位置代表现金），它表示需要将50%的资金配置在股票1号，20%的资金配置在股票2号，剩下的全部配置在股票3号，4号股票将不被持有。

我们的重点在于提取股票的特征。目前在深度学习中，有2种重要的网络模型可以用来提取数据特征，CNN（Convolutional Neural Network， 卷积神经网络）和LSTM（Long Short Term Memory， 长短期记忆模型）。CNN被广泛应用于计算机视觉领域，因为它擅长提取图像的边缘特征进而组合成为上层的复杂特征，非常符合人类对图像的观察理解过程。而LSTM被广泛使用在时序性数据中，比如文本、股票、天气情况等等，也就是说，LSTM擅长从一个序列中挖掘出数据的前后相关性和规律。LSTM 是 RNN（Recurrent Neural Network， 递归神经网络）的改进版，它被赋予了能够选择性地记忆和遗忘以往数据的能力，因此能力更强大。在本研究中，我把特征提取的网络选取设为一个超参数，分别使用CNN和LSTM作为特征提取网络，对比得到的不同的表现。结果将第四章中给出。

如图3.7给出的是CNN特征提取的网络示意图。输入是一个$(N+1, windw_length, 1)$的tensor，输出将是一个$(N+1, 64)$的tensor。

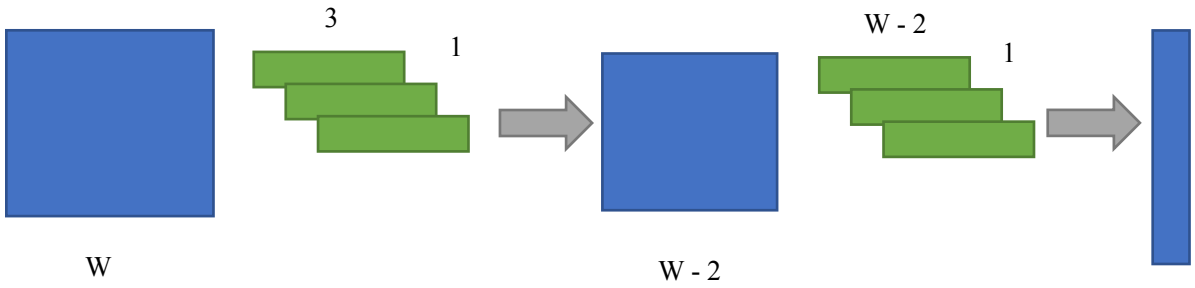

图 3.7 CNN特征提取

如图3.8给出的是LSTM特征提取的网络示意图。图示中，每一个圆圈表示一个cell，或单元，是LSTM的基本单位。历史窗口的数据将依次被输入，最终输出一个$(N+1, 64)$

的 tensor。

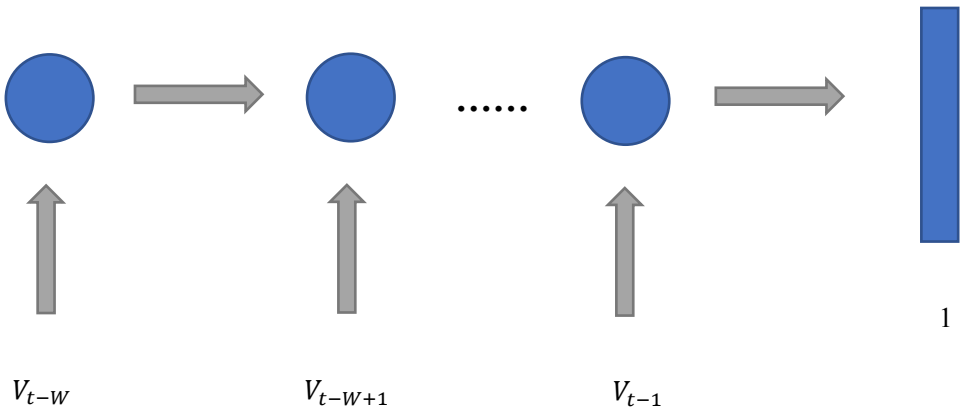

图 3.8 LSTM 特征提取

3.4.2.2 价值网络

价值网络$Q(s,a|\theta^Q)$，在 DDPG 算法中又被称为 critic，是用来评判采取某个动作的好坏的。输入当前 state 和 actor 生成的 action，输出一个值作为给这个动作的评分。

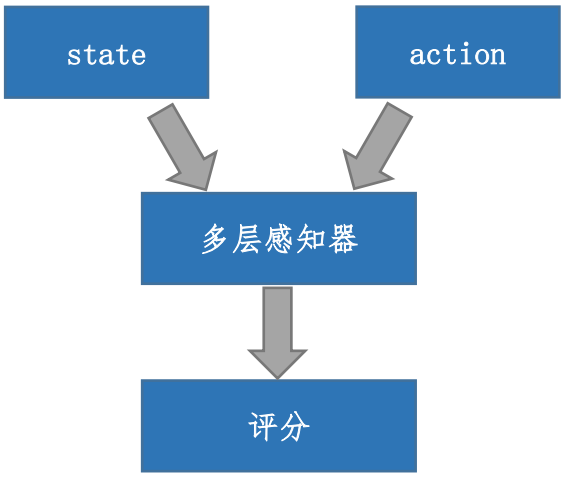

图 3.9 价值网络架构示意图

由于价值网络需要 2 个输入，所以需要一种方式将 2 个 tensor 连接在一起。这里我直接将 2 个张量相加得到一个新的张量，然后输入一个多层感知器，经过几层全连接层，得到一个标量。

3.5 模仿学习投资模型

不同于增强学习，模仿学习是一种监督式学习算法，同样也被用来解决智能体的学习问题。用模仿学习算法同样训练一个投资模型，是为了对比增强学习和模仿学习在量化投资问题上的表现有什么不同。在第四章中将会展示实验结果并进行分析讨论。

3.5.1 模仿学习算法

在增强学习任务中，正如公式$(3-8)$所示，通常是通过最大化累积折扣回报在学习最优策略，这种方式简单直接，而且在训练数据足够充分并认为一定程度上能够覆盖状态空间时，确实有较好的表现，但是在多步决策中，学习器有时不能频繁地得到奖励，这一点被称为“奖励稀疏性”。比如，一盘象棋或者围棋，可能要下非常久，才能得到一个关于胜负的反馈，而这时可能已经走了几百步棋，在这整个过程中，都不能获得奖励。虽然像很多游戏或者本文的研究内容——量化投资，我们可以通过合理的手段得到单步回报，但是让算法从 0 开始训练依然是非常困难的。实际上，大多数的学习过程都不是从 0 开始而是在一定的基础上的。比如让一个人去学习下围棋，在此之前他可能已经学会下五子棋或其他棋类，已经具有一定的策略性，或者他已经看过一些围棋比赛的录像，对于围棋已经有了初始概念。传统的增强学习算法训练过程类似于让一个刚刚出生的小孩去学习开车，不可避免地会出现一些难以训练的问题。另外，基于累积回报最大化的学习方式存在巨大的探索空间，如果不能在 exploit（开发）和 explore（探索）间找到合理的平衡，也很难得到良好的结果。

模仿学习（Imitation Learning）是一种监督式的学习算法，是希望从已有的经验中学习决策方法。类似于回归算法，模仿学习算法会用大量的 state 作为输入，action 作为标签来进行训练，这里认为所有训练集中的 action 都是最合理或是非常合理的。AlphaGo[17]在最一开始，就是用大量人类棋谱训练了初始的策略网络。如果说增强学习是让一个孩子自己从头开始慢慢摸索在吃亏中成长，模仿学习更像是一个母亲用自己的经验手把手教会孩子如何解决问题，因此也被称为“行为克隆”。

但是，模仿学习也有很严重的问题，因为存在“复合误差”。这类似于蝴蝶效应，一开

始非常微小的误差将导致之后巨大的误差。如图 3.10 所示，黑线表示训练集的期望策略，而红线是学到的期望策略，虽然误差很小，但是累积到最后将是不可想象的。本质上讲，出现这个问题的原因是训练集中的状态和动作空间依然有限，网络在学习过程中只能看到一个部分，在实践中则需要面临所有的情况。

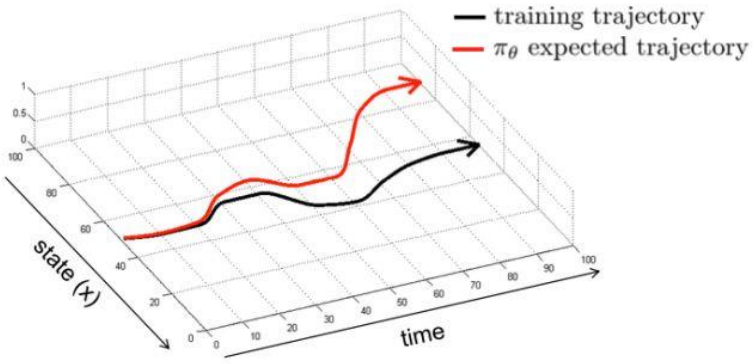

图 3.10 模仿学习复合误差示意图

模仿学习的算法对比增强学习要简单很多，可以被视为一个回归问题。只需要构建大量$(state, action)$数据对作为训练集。由于找不到人类交易员的决策数据，所以在本文的量化投资问题中，最优的策略可以贪心地去选择收盘价与开盘价比值最高的那支股票，这样“人为”地构造出一个决策集。比如在某个时刻，$\boldsymbol{y_t} = [1, 1.5, 0.8, 2.3, 0.5]^{\boldsymbol{T}}$，那么在进行资产配置时，将所有资金都买入 3 号股票即可，这样可以在下一轮获得最大的收益。

3.5.2 网络结构

不同于在上一节介绍的策略网络，模仿学习的网络将只使用 LSTM 进行训练。

整个网络的架构如图 3.11 所示：

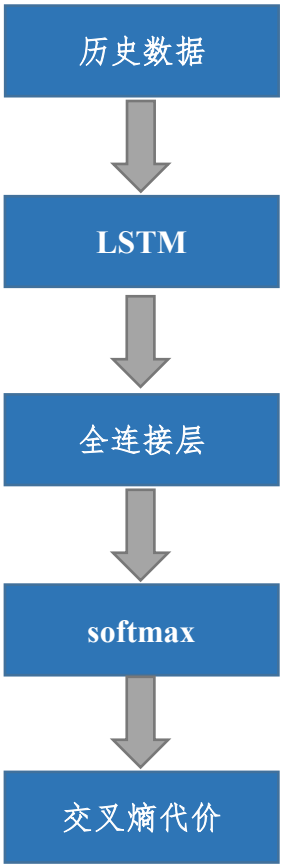

图 3.11 模仿学习网络架构示意图

LSTM 网络部分类似于图 3.8 的结构，但是这里隐藏层的大小被设为 20，而不是 64。经过 LSTM 之后，会经过 3 个全连接层，大小分别是 64、64、(N+1)。值得注意的是，这里我在每一个全连接层之后都加了 30%的随机失活（Dropout），这是为了避免过拟合。最后，通过 softmax 激活函数归一化后，得到了预测的 action，之后通过交叉熵代价函数计算损失。在机器学习中，交叉熵代价函数被广泛应用于衡量 2 个概率分布之间的差异，另外在经过 softmax 层之后，多分类的交叉熵代价等价于对数似然概率最大化。在训练时，用 Adam 作为参数更新优化器，它会计算针对交叉熵代价的梯度然后通过反向传播更新网络参数。

第四章 实验结果与分析

4.1 实验环境

本文的实验基于以下环境完成：

表 4.1 实验环境

环境	参数（版本）
操作系统	Win7 专业版 64 位（DirectX 11）
CPU	Intel® Core™ i5-4590 @ 3.30GHz 四核
GPU	Nvidia GeForce GTX 750Ti(2GB)
内存	24GB（DDR3 1600MHz）
Python	3.6.4
TensorFlow	1.2.1
tflearn	0.3.2
Keras	2.1.6
gym	0.10.5

4.2 数据集和测试基准

上文已经介绍了，本研究使用 kaggle 的股票数据集，该数据集涵盖美国 S&P 500 指数中的所有 500 家上市公司的股票，时间跨度从 2013-02-08 到 2018-02-07 整整 5 年时间。

在投资模拟中，不可能选取所有 500 支股票，这是不现实的。只需要从其中选择一个固定的股票组合，达到分散风险的目的即可。为了覆盖多数的行业领域，尽量提高多样性，对冲风险，我选择了 16 支有代表性的股票：

“AAPL”（苹果），“ATVI”（动视暴雪），“CMCSA”（康卡斯特），“COST”（好市多），“CSX”（CSX 运输），“DISH”（DISH 网络科技），“MSFT”（微软），“EBAY”（易贝），“FB”（Facebook），“GOOGL”（谷歌），“HAS”（孩之宝），“ILMN”（Illumina 生物科技），“INTC”（英特尔），“MAR”（万豪国际酒店），“REGN”（Regeneron 生物制药），“SBUX”（星巴克）。

以上股票组合涉及网络、生物、服务、餐饮、工业、电信等多个行业领域，覆盖面广，因此是作为投资组合的良好选择。

另外有 3 个基准将作为与深度增强学习训练的模型的对比：S&P 500、随机选择模型、平均选择模型，如表 4.2。

表 4.2 本实验选取的四个测试基准

基准	描述
S&P 500	S&P 500 是美国标准普尔指数，与常用的道琼斯工业指数相比，它涵盖的公司更多，因此风险更为分散，也更能够反映市场的变化。这里选用 S&P 500 作为测试基准反映市场的整体变化
随机选择模型	每次做决策时将资金随机分配
平均选择模型	每次做决策时将资金平均分配到各支股票上

在时间跨度选取上，前三年将被用来进行网络的训练，后两年将被用来进行测试，如表 4.3。

表 4.3 训练和测试的时间跨度选取

	时间跨度
训练集	2013-02-08 ~ 2016-02-07
测试集	2016-02-08 ~ 2018-02-07

4.3 评估指标

为了衡量不同的模型之间的表现差异，需要合适的评价指标，有许多现成的标准可以供我们选择。在本实验中，将选取 4 个指标进行评判：股票总价值、收益率、夏普比率、最大回撤率。如表 4.4。

表 4.4 本实验选取的四个测试指标

指标	描述
Portfolio value	投资组合的总价值，只计算比例关系，不关心具体数值
Return	最终收益率
Sharpe	夏普比率
MDD	最大回撤率

Portfolio value 即总价值p_t，其定义在公式$(3-2)$给出，衡量了整个投资周期中我们的总资本的变化，实验中将会画成折线图来表现其变化情况。

Return 是投资周期的最终收益率，即$\frac{p_t}{p_0}-1$，是最基本的基金衡量指标。

Sharpe ratio，夏普比率，是一个非常常用的基金绩效评价指标，其计算公式为：

$$\text{Sharpe} = \frac{E(R_p) - R_f}{\sigma_p} \qquad (4-1)$$

其中R_p是投资组合的预期收益率，R_f是无风险利率，一般用国债利率表示，σ_p是投资

组合的收益标准差。因此，夏普比率计算的就是投资组合每承受一单位的风险，能获得多少的超额收益，显然夏普比率越大越好。

MDD（Max Drawdown），最大回撤率。其计算公式为：

$$\text{MDD} = \text{Min}\ \frac{P_y - P_x}{P_x}, y > x \qquad (4-2)$$

最大回撤计算的是在一个投资周期内，所能产生的最大的亏损率，显然最大回撤率越接近 0 越好。

4.4 测试结果

4.4.1 DDPG 模型

DDPG 模型在训练上一共有 2 组超参数作为对比：window_length（[3, 7, 14]）、特征提取网络选取（[cnn, lstm]），因此总共会训练产生 6 个不同的模型。模型的命名格式为：DDPG_window_[3, 7, 14]_predictor_[cnn, lstm]，再加上 3 个测试基准，总共 9 条曲线。

所有 DDPG 投资模型在测试集上的表现在图 4.1 中展示：

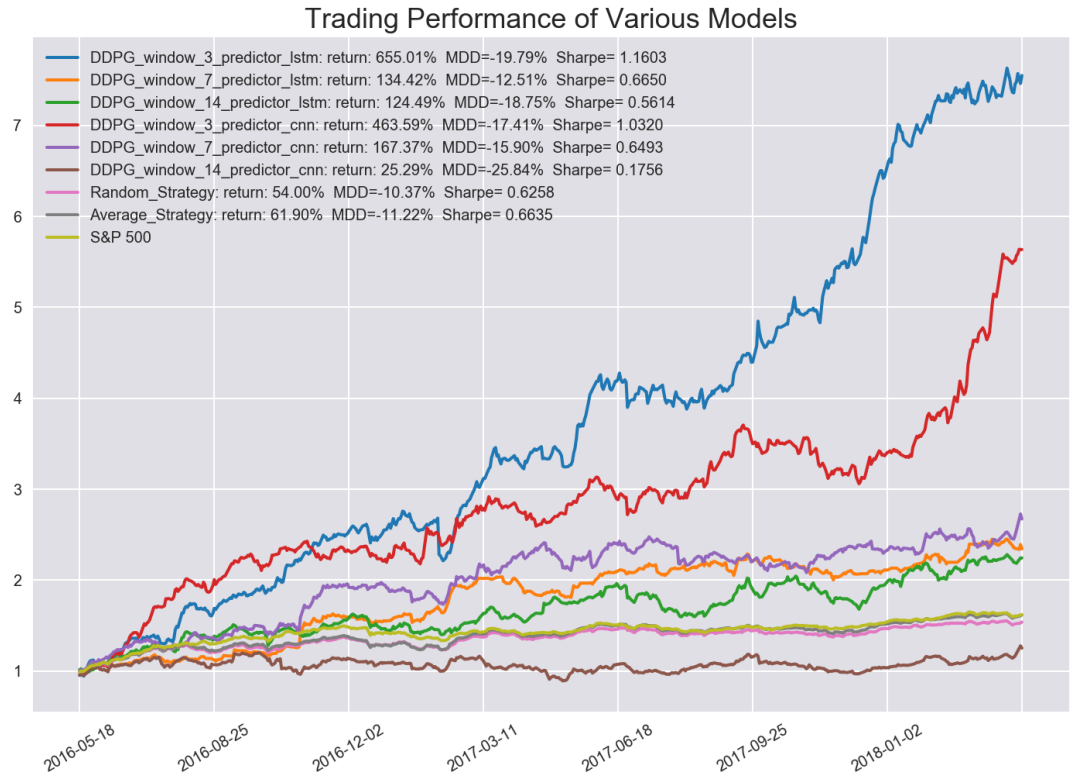

图 4.1 DDPG 模型在测试集上的表现结果

所有 DDPG 投资模型在测试基准上的表现在表 4.5 中展示：

表 4.5 DDPG 模型在测试基准上的表现

模型	训练时间（min）	收益率（%）	夏普比率	最大回撤（%）
Window_3_lstm	260	655.01	1.1603	-17.41
Window_7_lstm	500	134.42	0.6650	-15.90
Window_14_lstm	900	124.49	0.5614	-25.84
Window_3_cnn	150	463.59	1.0320	-19.79
Window_7_cnn	240	167.37	0.6493	-12.51
Window_14_cnn	460	25.29	0.1756	-18.75
Random	/	54.00	0.6258	-10.37
Average	/	61.90	0.6635	-11.22

4.4.2 模仿学习

模仿学习模型在训练上只有一组超参数作为对比：window_length（[3, 7, 14]），总共会训练产生 3 组模型。在命名格式上为：imitation_window_[3, 7, 14]_predictor_lstm。

所有模仿学习投资模型在测试集上的结果在图 4.2 中展示：

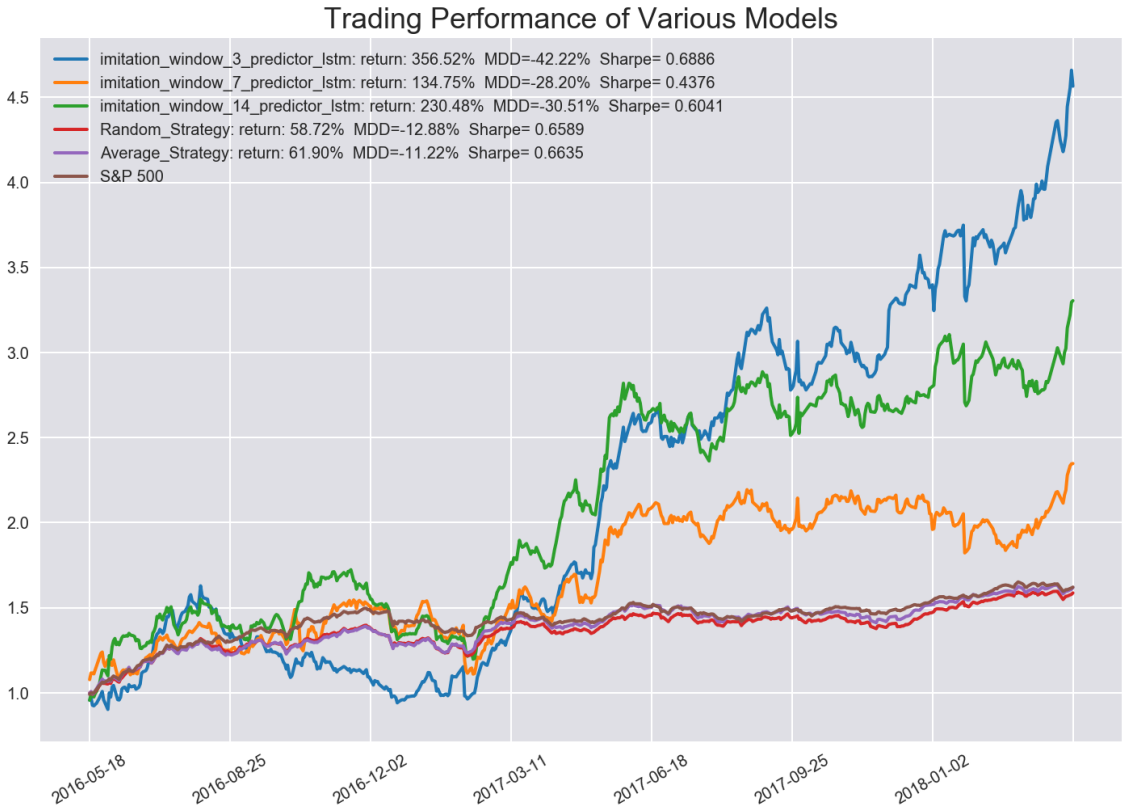

图 4.2 模仿学习模型在测试集上的表现结果

所有模仿学习投资模型在测试基准上的表现在表 4.6 中展示：

表 4.6 模仿学习模型在测试基准上的表现

模型	训练时间（min）	收益率（%）	夏普比率	最大回撤（%）
Window_3_lstm	265	356.52	0.6886	-42.22
Window_7_lstm	270	134.75	0.4376	-28.20
Window_14_lstm	280	230.48	0.6041	-30.51
Random	/	58.72	0.6589	-12.88
Average	/	61.90	0.6635	-11.22

4.4.3 DDPG 和模仿学习对比

为了直观体现 2 种不同的算法在量化投资上的能力差异，选取 DDPG 模型中 LSTM 作为特征提取的 3 个模型，和模仿学习的 3 个模型绘制在一张图上作为对比，如图 4.3。

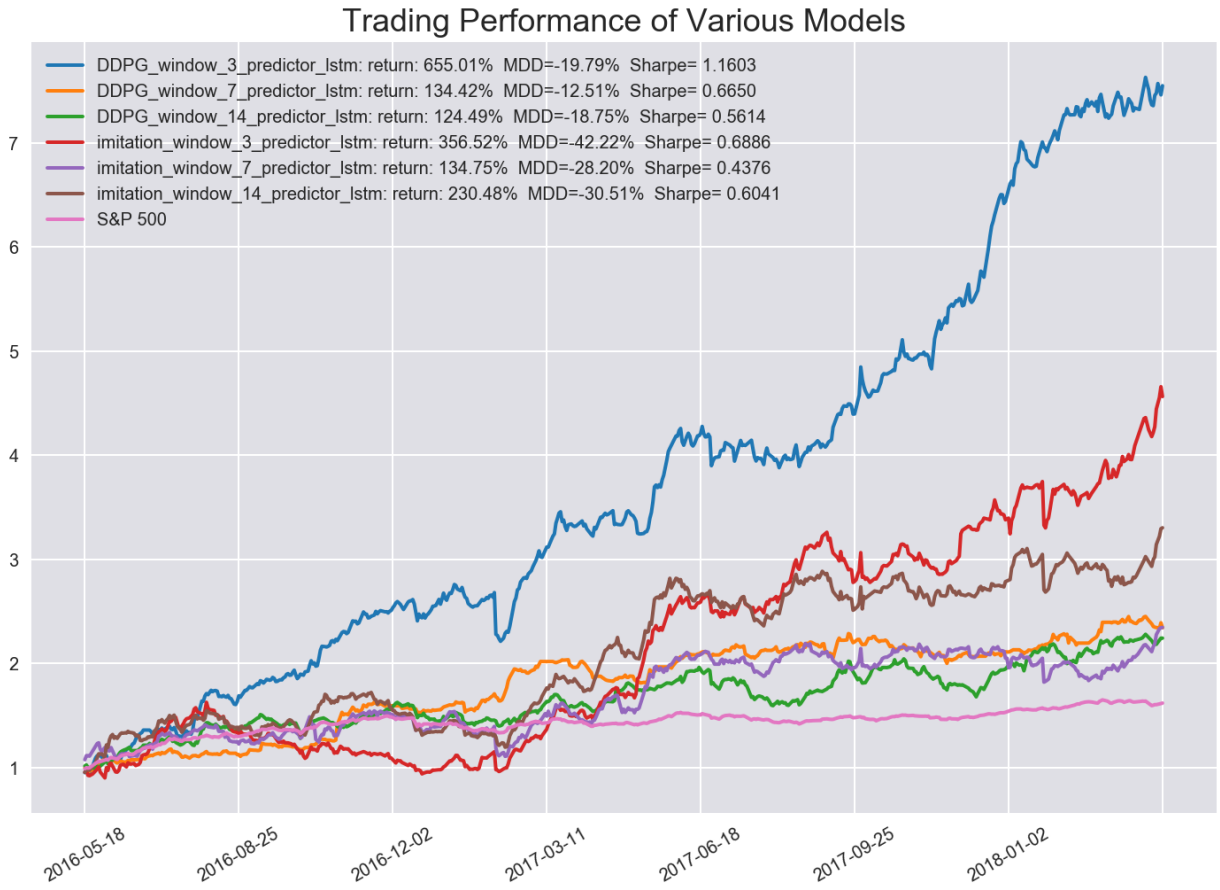

图 4.3 DDPG 和模仿学习模型对比

4.5 分析与讨论

在上一节，展示了本研究的所有实验结果，在本节将进行详细的分析和讨论。

可以看到，作为测试基准的随机策略和平均策略，走势基本和市场是完全一致的，在2年时间内也是可以达到 50%以上的收益率的，这说明两点：第一，整个市场的价值随着时间的发展一定是在上涨的，虽然很缓慢，因为从大趋势上来看，宏观经济一定是在稳定增长的；第二，即使是没有任何投资技巧的随机和平均化投资策略，依然能够基本跟随市场甚至有时候可以稍微战胜市场，说明多元化的投资一定程度上确实可以对冲风险，即使我们这里只选择了 16 支股票，但是已经基本足够概括 500 支股票，这证明了投资组合的价值。

从图 4.1 和 4.2 可以看到，在整个投资周期内，所有的模型均呈现出稳定的增长趋势，并且大多数都远远超过市场，说明算法是有效的，神经网络能够学习到股票的特征。

4.5.1 LSTM 和 CNN

从图 4.1 和表 4.5 中可以发现，在 window_length 相同的情况下，使用 LSTM 训练的模型要比 CNN 的强大很多，尤其是在 window_length 为 3 时，用 LSTM 训练的模型 2 年收益率可以达到 650%以上，折合成年化收益率大约在 270%以上，这是非常优秀的表现，目前我在网上没有找到任何公开的基金能够达到这个收益率。

在最大回撤表现上，两者并没有太大区别。而夏普比率方面则明显可以看到 LSTM 模型要比 CNN 模型好很多，也就是说，每多承担一单位的风险就能获得更多的收益。总体来讲，LSTM 模型优于 CNN 模型。

这一点并不出乎意料，因为上文提到过，LSTM 擅长处理时序性数据，而股票价格是时序性很强的，我们人类在进行分析时也会去找到价格的前后关系，分析走势。而 CNN 擅长使用在图像领域，提取图片边缘特征。因而在这里 LSTM 有更好的表现。不过，可以看到的是，LSTM 网络的训练时间普遍要比 CNN 长一些，但是依然是可以接受的。

4.5.2 历史窗口大小的影响

从图 4.1 和表 4.5 可以明显发现，越小的 window_length 使得模型有更优的表现，毫无例外。在本实验的参数设置中，3 天的历史窗口大小是最优的，而 14 天的窗口大小是最差的，甚至 window_length 为 14 的 CNN 网络模型都没有战胜市场，如图 4.1。

随着窗口的变大，收益率和夏普比率都呈现出明显的递减趋势。这说明，有时候看到越多的信息并不是好事。离当前越近的几天的信息越有效，对当前的影响也越大。虽然直觉上告诉我们，看到越久的价格，越能够综合分析得到更准确的判断，但是实际的表现却恰恰相反，可能是因为越少的信息对判断的干扰越小。实际上，在股票技术分析中，3 天也是一个非常常用的回看天数。

4.5.3 DDPG 和模仿学习

在图 4.2 和表 4.6 中，展示了模仿学习训练的 3 个模型的表现结果。模型也表现出了不错的性能，最好的可以达到 2 年 350%的收益率。但是模仿学习模型的最大回撤率都非常大，最严重的是-42.22%，这意味着在整个投资周期中有可能亏损达到 40%以上。而且 3 个模型的夏普比率都非常低，甚至比随机策略和平均策略要低。说明模仿学习训练的模型的风险是非常大的，从这个角度讲，其实模型的表现是比较差的，并没有达到通过投资组合分散风险的目的。

在图 4.3 中，我展示了同样用 LSTM 作为网络结构的 DDPG 和模仿学习训练的模型表现的对比。除了表现最好的增强学习模型外，其他模型的差异其实并不大，甚至模仿学习训练的模型的收益率还要高一些，但是如上所说，后者的风险远远大于前者。

因此，可以得出结论 DDPG 模型在量化投资问题上的表现要优于模仿学习的表现。我认为这主要是因为以下几点：

1) 增强学习可以比模仿学习看到更广的状态空间。上文提到，模仿学习是一种典型的“行为克隆”，希望通过给予网络人类经验来训练一个模仿人类决策的模型。但是一方面因为训练集不够大，无法覆盖状态空间和动作空间，另一方面存在“累积误差”，会让决策偏移越来越大。而增强学习则不同，模型的训练完全基于最大化累积收益，因而其目标更加明确，并且由于合适的 exploit 和 explore 之间的权衡，可以看到更广的状态空间，通俗来讲，增强学习训练的模型经验会更丰富，能

够看到更多信息。这就是为什么虽然有些 DDPG 模型的收益率可能并不如模仿学习，但是却更稳定，夏普比率更高，更适合投资。

2) 模仿学习训练的过拟合程度较大。由于只有 3 年数据，因此一定程度上给模仿学习训练的样本量确实比较小。这样就容易造成过拟合，在测试集上表现比较差。而 DDPG 模型主要有 2 点可以有效减少过拟合：第一，如公式(3 – 11)所示，每一步动作都会加入一个随机噪声，不是固定的，这样就增强了模型的鲁棒性；第二，上文提到，Portfolio Env 在初始化时，会随机选择一天，而不是固定的某一天，所以每一次训练中，算法都能看到基本不同的情况。

第五章 结论

5.1 本文总结

本文从机器学习在量化投资上的应用出发，在概述相关背景之后，通过大量的调研，对现阶段国内外相关研究进行了介绍，分析了当前存在的问题和未来的发展方向。并提出本文的研究课题，基于 DDPG 深度增强学习的量化投资模型研究。

利用 kaggle 提供的股票数据，用 DDPG 算法成功训练了一个自动交易的量化投资模型，并且有很好的测试表现。为了找到合适的网络结构和历史窗口大小，分别使用 CNN 和 LSTM，以及 3、7、14 的窗口大小训练多个模型进行对比，最终确定 LSTM 和窗口为 3 的最佳参数，该模型表现出超过 650%的 2 年收益率。另外，作为比对研究，用模仿学习算法同样训练了一个模型，并对比了两种模型的表现差异，从算法层面对差异进行了分析。

最后，得出结论，DDPG 深度增强学习算法应用在量化投资问题上是成功的，并且其表现要优于模仿学习。

5.2 未来工作

本研究是成功的，但是依然还有很大的提升空间。目前根据笔者的大量调研以及根据笔者自身的理解，目前可以想到如下几个未来工作的方向：

1) 改进网络结构本身。在深度学习中，网络本身的结构对于其最终表现效果影响是非常大的，有研究表明，越深的网络结构通常意味着更好的表现，当然同样也有可能导致梯度弥散等问题[25]。通过改变策略网络和价值网络的结构，以及配置不同的超参数，是改进现有网络的重要方向。
2) 将模仿学习网络作为目标网络。在上文提到，off-policy 的增强学习算法有一个很大的优势，可以将已有的 policy 作为 target network 的初始化。在本文的训练中，target network 也是被完全随机初始化的，但是类似于 AlphaGo[17]，完全可以尝试将已经训练好的一个策略作为其初始化，这里可以使用模仿学习学到的策略。当然，这要求两个网络具有相同的结构。这相当于先让网络学到一些人类的经验，然后再以累积收益最大化为目标进行自我探索提升，理论上讲，这样可以比完全随机初始化有更好的效果。
3) 引入新闻等舆论影响。众所周知，股市并不是完全受公司财政情况支配的，很大

程度上要受到舆论政策的影响，尤其是经济相关的。自然语言的处理可以用 LSTM 模型，当然真正要实用可能情况会很复杂。在本研究中，环境是单一的，只有股票的历史价格。未来的工作可以考虑将经济类的新闻标题进行分析，并作为环境的一部分输入策略网络。

参考文献

[1] Markowitz, Harry. Portfolio Selection[J]. Journal of Finance. 1952, 7(1):77~91.

[2] Eugene fama. Efficient market hypothesis: A review of theory and empirical work[J]. The journal of finance. 1970, 5(2):383~417.

[3] 韩正宇. 现代投资组合理论评述[J]. 经济研究参考. 2013 (60):53~61.

[4] 龙先文, 邓纯阳. 对马科维茨投资组合理论的反思[J]. 特区经济. 2005(11).

[5] Tom Mitchell. Machine Learning[M]. 曾华军译. 北京：机械工业出版社，2008

[6] Ferson W. Changes in expected security returns, risk, and the level of interest rates[J]. Journal of Finance 1989, 5(44):1191~1217.

[7] Geoffrey E.Hinton, Simon Osindero. A Fast Learning Algorithm for Deep Belief Nets Neural computation. Neural Computation. 2006, 18(7):1527~1554

[8] C. J. Watkins and P. Dayan. Technical note: Q-Learning[J]. Machine Learning, 1992 (8):279~292.

[9] Zhengyao Jiang, Dixing Xu. A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem[J]. Papers, 2017, arXiv:1706.10059.

[10] J. B. Heaton et al. Deep learning for finance: deep portfolios[J]. Applied Stochastic Models in Business and Industry. 2017, 33(1):13~15.

[11] Xin Du, JinJian Zhai, Koupin Lv. Algorithm Trading using Q-Learning and Recurrent Reinforcement Learning[D]. Stanford University, 2017.

[12] Olivier Jin, Hamza El-Saawy. Portfolio Management using Reinforcement Learning[D]. Stanford University, 2017.

[13] Seyed Taghi, Akhavan Niaki, Saeid Hoseinzade. Forecasting S&P 500 index using artificial neural networks and design of experiments[J]. Journal of Industrial Engineering International, 2013, 9(1):1.

[14] J. Moody, M. Saffell. Learning to trade via direct reinforcement[J]. IEEE Transactions on Neural Networks, 2001, 12(4):875~889.

[15] M.A.H. Dempster, V. Leemans. An automated trading system using adaptive reinforcement learning[J]. Expert Systems with Applications, 2006, 30(3):543~552.

[16] David Silver et al. Human-level control through deep reinforcement learning[J]. Nature, 2015.2, 518 (7540):529~533, ISSN 0028-0836.

[17] David Silver et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587):484~489.

[18] D Charles, II Kirkpatrick, Julie R Dahlquist. Technical analysis: The complete resource for financial market technician[M]. 2006, ISBN:13.

[19] A. Fernndez, S. Gmez. Portfolio selection using neural networks[J]. Computers and Operations Research, 2007, 34(4):1177~1191.

[20] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading[J]. IEEE transactions on neural networks and learning systems, 2017, 28(3):653~664.

[21] 李文鹏, 高宇菲, 钱佳佳, 陈曦. 深度学习在量化投资中的应用[J].统计与管理, 2017(8):104-106.

[22] Guosheng Hu et al. Deep Stock Representation Learning: From Candlestick Charts to Investment Decisions[J]. Papers, 2017, arXiv:1709.03803.

[23] David Silver et al. Deterministic Policy Gradient Algorithms[J]. International Conference on Machine Learning. 2014:387-395.

[24] David Silver et al. Continuous Control with Deep Reinforcement Learning[J]. Computer Science, 2015, 8(6):A187.

[25] Kaiming He, Xiangyu Zhang et al. Deep Residual Learning for Image Recognition[J]. Computer Vision and Pattern Recognition. 2016, 5(3):770~778 arXiv:1512.03385.

作者简历

姓名：马啸远　性别：男　民族：汉　出生年月：1996-07-28　籍贯：山西省怀仁县

2011.09-2014.07　太原市第十八中学

2014.09-2018.07　浙江大学攻读学士学位

获奖情况：无

参加项目：无

发表的学术论文：无

本科生毕业论文任务书

一、题目：基于深度增强学习的量化投资研究

二、指导教师对毕业论文的进度安排及任务要求：

该毕业论文要求对量化投资以及增强学习相关国内外研究现状与相关的平台进行深入分析的基础上，实现一个基于增强学习的量化投资模型。并通过对比分析，运用合适的工具和可行的技术路线进行实现。在此基础上，要实现开题报告中提出的目标和任务，并按照开题报告中制定的实施计划，按时完成论文工作。毕业论文要求做到结构清晰，语句通顺。

进度安排如下：

2017.12.1~2018.1.1：确定课题与相关文献整理

2018.1.2~2018.3.1：确定研究的技术方案和关键思路

2018.3.2~2018.4.5：完成开题报告和文献综述

2018.4.6~2018.5.10：完成模型实现与测试

2018.5.11~2018.5.30：撰写毕业论文

起讫日期 20　年　月　日 至 20　年　月　日

指导教师（签名）＿＿＿＿职称＿＿＿＿

三、系或研究所审核意见:

负责人（签名）＿＿＿＿＿＿

年　月　日

毕 业 论 文 考 核

一、 指导教师对毕业论文的评语：

指导教师(签名）

年 月 日

二、答辩小组对毕业论文的答辩评语及总评成绩：

成绩比例	文献综述占（10%）	开题报告占（15%）	外文翻译占（5%）	毕业论文质量及答辩占（70%）	总评成绩
分值					

答辩小组负责人（签名）

年 月 日