北京大学

本科生毕业论文

题目：基于双升游戏的多智能体环境与人工智能算法

A Multi-Agent Environment and AI Algorithm for Tractor

姓　　名：王星博

学　　号：2000017794

院　　系：元培学院

专　　业：人工智能

导师姓名：李文新

二〇二四年五月

北京大学本科毕业论文导师评阅表

<table>
<tr><td>学生姓名</td><td>王星博</td><td>本科院系</td><td>元培学院</td><td rowspan="2">论 文 成 绩
（等级制）</td><td rowspan="2"></td></tr>
<tr><td>学生学号</td><td>2000017794</td><td>本科专业</td><td>人工智能</td></tr>
<tr><td>导师姓名</td><td>李文新</td><td>导师单位/
所在学院</td><td>计算机学院</td><td>导师职称</td><td>教授</td></tr>
<tr><td rowspan="2">论文题目</td><td>中文</td><td colspan="4">基于双升游戏的多智能体环境与人工智能算法</td></tr>
<tr><td>英文</td><td colspan="4">A Multi-Agent Environment and AI Algorithm for Tractor</td></tr>
<tr><td colspan="6">导师评语
王星博同学的论文研究的是如何构筑一个多智能体博弈平台，并支持在平台上进行各种智能体的研发。论文选择“双升”游戏作为博弈环境，这是一个 2v2 的具有随机手牌和隐藏信息的复杂环境，为人工智能的前沿方向。论文成功完成了环境构建，并在其上开发了几个有效的智能体。该环境成功应用到“多智能体系统”课程中作为学生作业。论文工作达到本科论文水平，是一篇优秀的本科论文。

导师签名：
年　　月　　日</td></tr>
</table>

版权声明

摘　要

游戏作为人工智能领域的研究对象和平台历史已久。从传统的棋牌类游戏到现当代基于各种媒介的电子游戏，不同的游戏以其不同的特点而具有了不同的研究价值。双升游戏作为一种中国流行棋牌游戏，具备多智能体多轮混合动机博弈的特征，且策略上具有较高的复杂度，蕴含着新的研究问题，对人工智能研究具有独特价值。

将双升游戏作为人工智能研究对象的基础，是一个可以高效运行的双升游戏虚拟平台或环境。通过一个开放式的在线人机对战平台，可以筛选高水平玩家和算法并收集行为数据，满足了监督学习对大量专家数据的需求；依托一个支持高效并行的游戏环境，则可以实现大规模对局采样，为强化学习提供条件。

游戏人工智能算法，是游戏人工智能研究的重要课题，也是游戏人工智能其他方向研究的前提条件。游戏人工智能算法涉及监督学习、强化学习、模仿学习等多个方面的算法和理论，旨在利用人工智能技术针对特定的游戏模型提出决策水平接近人或超越人的解决方案。

本文从博弈模型和复杂度的视角分析了双升游戏的特征和研究价值，介绍了基于 Botzone 开放人机对战平台开发的 Botzone 双升在线游戏和双升强化学习环境，并在此基础上提出了一种综合了监督学习和强化学习等方法的双升游戏人工智能算法。

关键词：多智能体环境，在线游戏平台，强化学习，双升

ABSTRACT

Game has a long history as a research object and platform in the field of artificial intelligence. From traditional chess and card games to modern electronic games based on various media, different games show different research values with their different characteristics. As a popular card game in China, Tractor game is a multi-agent multi-round mixed motive game, which has high complexity in strategy. Tractor game has unique value for research in artificial intelligence and provides new research problems.

An efficient virtual platform or game environment is the basis for conducting AI researches on Tractor game. An open online platform supporting human-machine matches can be used to collect expert data required by supervised learning from high-level players and algorithms; by leveraging a game environment that supports efficient parallel processing, large-scale sampling can be achieved, facilitating reinforcement learning.

Game AI algorithms are an important research topic in game AI, and are also a prerequisite for other fields of game AI research. Game AI algorithms involve various algorithms and theories in supervised learning, reinforcement learning, imitation learning and so on, aiming to propose AI solutions that approach or surpass human intelligence for specific game models.

This paper analyzes the research potential of Tractor game from the perspective of game model and complexity, introduces Botzone-Tractor, an online game developed on Botzone and RLTractor, a Tractor reinforcement learning environment. The paper also proposes an AI algorithm for Tractor games that integrates supervised learning and reinforcement learning methods.

KEY WORDS: [Multi-Agent Environment, Online Game Platform, Reinforcement Learning, Tractor Game]

目 录

1. 引言

游戏在人类社会中已有相当长的发展史。从传统棋牌游戏到电子游戏，以及未来可以展望的虚拟现实与增强现实游戏，游戏既是人类智慧和智能技术的产物，也是其发展进步的推动者。

在人工智能的视域下，游戏对监督学习和强化学习的发展发挥了尤为重要的作用。大量的游戏（包括传统游戏与电子游戏）为强化学习环境的构建提供了原型和启发。这些环境或利用本身蕴含的多种博弈模型和独具特色的场景为学习算法提出新的研究问题，或依靠游戏规则和决策的复杂度对算法性能提出挑战，或依托其对真实场景的高还原度提供真实世界的模拟平台。游戏之所以能够成为强化学习环境的灵感源泉，一是因为游戏的规则基于人为设计，具有相对简明的逻辑规律，因而相比于真实世界，游戏作为构建强化学习环境的原型或基础时，在建模上更为简易快捷；二则是因为从游戏性的角度，游戏的规则一般具有一定复杂度，能够检验人类智能和机器智能的水平。

1.1 游戏强化学习环境及特征

在现今的强化学习基准环境中，有相当一部分环境是基于游戏构建的。其中代表性的环境包括基于星际争霸(StarCraft)的 SMAC[1]，基于我的世界(Minecraft)的 Minedojo[2]，搭载了德州扑克(Texas Hold'em)的棋牌游戏环境 RLCard[3]，包含 59 款雅达利 2600 家用机游戏(Atari 2600)的 ALE(Arcade Learning Environment)[4]，以及以传统桌游《外交风云(Diplomacy)》为蓝本设计的 Diplomacy 环境[5]。除此之外，围棋、国际象棋、麻将等传统棋牌游戏也广受关注，并被用于构建不同的强化学习环境。然而，包括斗地主、双升、掼蛋在内的一类中国流行棋牌游戏却没有得到足够的重视。实际上，这类棋牌游戏在推理和决策上均具有与现行的一些强化学习环境相当的复杂度，且这类游戏独特的规则玩法提供了与其他游戏不同的博弈抽象，也使得对应的强化学习环境具有了与现有基线环境不同的特征（组合），从而能够一定程度填补现有强化学习研究的空缺。

我们对游戏强化学习环境作划分和评估时参考的主要特征包括：

(一) 随机性：环境是否包含随机性（随机初始条件、（马尔可夫前提下）状态转移方程是否具有随机性等）。随机性会提升环境的复杂度和多样性。

(二) 单人/多人：同时参与环境交互的智能体数量。更多的玩家（智能体）会提升环境的决策复杂度，同时多智能体环境相比单智能体环境能够为更多领域（例如智能体间博弈、智能体通讯、社交关系和社会结构等[6]）的研究提供平台。

(三) 单轮/多轮：游戏过程是否包含多个相同/相似的轮次。每个轮次单独结算，每轮结果对后续轮次或最终结果具有影响。多轮次的游戏为智能体进行对手建模、在线

学习和动态调整策略创造了条件。

(四) 博弈动机：对于多智能体环境，一般可从中抽象出智能体间的博弈关系（博弈动机）。博弈动机包括合作、竞争以及二者兼有的混合动机。在合作博弈中，智能体可以通过合作达到比不合作情况下更高的总收益；在竞争博弈中，智能体则必须互相竞争来获得更高的收益。混合动机博弈中则同时具有合作和竞争的元素。复杂的博弈型对人工智能算法的认知、推理和决策能力是一种挑战，为人工智能算法在人机协作中的应用奠定了基础，也对社交关系、社会结构等问题的建模和研究提供了平台。

(五) 完美信息/非完美信息：即游戏过程中智能体是否能够获得全部的局面信息。非完美信息问题一般要求智能体具有更强的记忆或推理能力。

(六) 是否允许语言交流：在多智能体游戏环境中，智能体之间存在信息的交流。这些交流可以通过智能体的行为和观察完成，也可能直接通过自然语言完成（如果环境允许）。对于人类玩家而言，是否允许语言交流，对游戏的玩法可能产生本质的改变；而对人工智能而言，自然语言交流往往对算法和模型提出新的挑战。

强化学习环境的分类依据当然不止于以上列出的六种，但是此处我们仅对列出的特征作讨论，且如表 1.1 所示，这六项指标已经可以对现有的基线强化学习环境作较好的划分。

表 1.1 游戏强化学习环境与特征

游戏	环境	随机性	单人/多人	单轮/多轮	混合动机	非完美信息	自然语言交流
星际争霸	SMAC	√	多人	单轮	√	√	×
我的世界	MineDojo	√	单人	-	-	√	×
德州扑克	RLCard	√	多人	单轮	×	√	×
围棋	-	×	多人	单轮	×	×	×
麻将	-	√	多人	多轮	×	√	×
雅达利 2600	ALE	√	单人	单轮	-	×	×
外交风云	Diplomacy	×	多人	单轮	√	√	√
双升	-	√	多人	多轮	√	√	×

表 1.1 列出了当下被用作强化学习环境的几种主要游戏、对应的（使用最多的）基线环境名称及具有的特征。接下来我们将围绕表 1.1 详述双升游戏作为强化学习环境的优势。

1.2 双升游戏与双升强化学习环境

双升是起源于中国的一种扑克游戏的变体，由四名玩家两两组队进行游戏。双升具有特殊的“升级”规则，每轮游戏后，根据两队的得分情况，进行由 2，3，4，……直到 A 的

升级。扑克游戏通用的基础玩法赋予了双升随机性和非完美信息的特征，双升特有的规则变式又使得这个游戏成为一个多轮的混合动机多人博弈环境。这些特征组合起来，使得双升相比于已有的基线强化学习环境具有了问题建模上的独特性，并拥有潜在的强化学习研究价值。

1.2.1 双升的规则

在全国范围内，双升游戏的规则大同小异，在不同地区有不同的细化规则，现行较为统一并由部分官方赛事采用的游戏规则可以简要描述为：

（一）游戏使用 2 副牌（108 张），四名玩家两两组队。起牌时四个玩家逆时针摸牌，每人 25 张，保留 8 张底牌，庄家在发牌结束后获得 8 张底牌并盖覆 8 张牌作为新底牌。

（二）报主与反主：发牌过程中，任意拥有级牌的玩家可以通过报主和反主来改变本局的主花色。报主和反主均只能进行一次。第一局的报主和反主同时具有确定庄家的功能。

（三）牌序与得分：每轮出牌四名玩家必须按顺序打出等量的牌，牌序大小与本局主花色和级牌有关。四名玩家本轮打出的所有牌中，5、10、K 称为“分”（具有一定分值：5 为 5 分，10、K 各 10 分）。根据牌序大小，获胜的玩家（所在队伍）获得所有“分”的分值，并成为下一轮首先出牌的一方。

（四）结算与升级：所有玩家将手牌出完的最后一回合，获胜方可以获得底牌中的所有“分”的分值，若闲家本局得总分超过 80 则获胜并“上台”，成为下一局的庄家，反之则由庄家获胜并继续坐庄。同时按照得分，获胜方进行 2，3，4，……直到 A 的升级，每局的级牌由庄家的等级决定，等级率先达到 A 并再胜一局的一方获得最终胜利。

1.2.2 双升作为强化学习环境的优势

目前尚未有依托双升游戏开发的强化学习环境，但是双升因其独特的规则玩法，在强化学习领域或有潜在的研究价值。

（一） 双升游戏的博弈特征

双升游戏是多轮的混合动机多人博弈。其多轮特征体现在升级规则上，完整游戏流程至少包含 5 局，至多可以包含 27 局甚至更多。各局本质上相似，但每一局的结果会改变下一局的初始状态，并一定程度上影响智能体的决策风格。多轮性为游戏带来了足够的变化，也给智能体足够的交互数量来进行对手建模和决策调整。

混合动机的特点则是基于双升的玩家分组，四名玩家两两组队对抗，既存在组内合作，也存在组间竞争。人类玩家的经验策略表明，两名玩家可以通过互相配合来争取牌权和得分；两队玩家之间也可以通过打出“拖拉机”等套牌有策略地打乱对方的牌组，以获得更高的分数。混合动机特征使得智能体的博弈场景更加多样，为智能体间合作与竞争的研究提供了空间，也为学习算法提出了新的挑战。

（二） 双升游戏的复杂度

双升游戏具有足够大的行动空间。实际上，在不考虑甩牌的前提下，每轮先手行动的可选动作数大致为 16441 种（如表 1.2），

表 1.2 双升动作空间

牌型	动作数量
单牌	54
对子	54
拖拉机	16333
总计	16441

考虑到甩牌规则（特定情形下可以打出多种牌型的复合），双升的实际动作空间不低于表中统计的数值，可以接近斗地主的动作空间大小[7]（27472），并且远大于麻将的动作空间大小（10^2）。这为双升游戏的复杂度提供了保证。

双升作为非完美信息游戏，其对局复杂度可通过信息集来分析。对于围棋类的完美信息游戏，其对局复杂度一般通过博弈树复杂度来分析[8]，完美信息游戏的博弈树复杂度计算建立在智能体能够完全观测到所有局面信息的假设之上。而非完美信息游戏中，智能体一般无法观测到全部局面信息，因而每个时刻在某个智能体的视角下，可能存在多个不同的游戏状态是难以区分的，可以认为这些状态包含在同一个信息集中，信息集构成了智能体基于自身观测对状态空间作出的划分。在此基础上，对非完美信息游戏而言，智能体的策略应当建立在信息集上而非游戏状态上才更为合理，信息集的概念也为我们分析非完美信息游戏复杂度提供了途径。

用信息集衡量对局复杂度[9]涉及到两个参数：信息集数量和信息集大小。信息集数量，即整个游戏状态空间可以被划分为多少个集合；信息集大小，即信息集内包含多少种不同的游戏状态。

信息集数量：仅考虑出牌阶段，双升游戏中每名玩家发给 25 张牌，并有 8 张底牌。考虑庄家已知 8 张底牌的情况，则第一轮信息集数量为C_{100}^{25}。为方便估算假设每一轮各家均出单牌，则第二轮信息集数量为$C_{100}^{25}C_{25}^{1}C_{75}^{1}C_{74}^{1}C_{73}^{1} = C_{100}^{25}A_{25}^{1}A_{75}^{3}$。类推可知第三轮信息集数量为$C_{100}^{25}C_{25}^{1}C_{24}^{1}C_{75}^{1}C_{74}^{1}C_{73}^{1}C_{72}^{1}C_{71}^{1}C_{70}^{1} = C_{100}^{25}A_{25}^{2}A_{75}^{6}$。按此计算方式最后一轮信息集数量为$C_{100}^{25}A_{25}^{24}A_{75}^{72}$。为简化计算我们没有考虑两副牌的重复性。则可以计算出双升游戏的平均信息集数量为$\frac{1}{25}C_{100}^{25}(1 + A_{25}^{1}A_{75}^{3} + \cdots + A_{25}^{24}A_{75}^{72}) = O(10^{133})$，大于麻将的信息集数量（$10^{121}$）和桥牌的信息集数量（$10^{67}$）[7]。

信息集大小：保持上述假设，第一轮信息集大小为$C_{75}^{25}C_{50}^{25}C_{25}^{25}$，仍假设所有玩家只出单

牌，第二轮信息集大小为$C_{72}^{24}C_{48}^{24}C_{24}^{24}$。以此类推，可计算双升游戏的平均信息集大小为$\frac{1}{25}(C_{75}^{25}C_{50}^{25}C_{25}^{25}+C_{72}^{24}C_{48}^{24}C_{24}^{24}+\cdots+C_3^1C_2^1C_1^1)=O(10^{33})$。作为参考，桥牌的平均信息集大小为$10^{15}$，麻将的平均信息集大小为$10^{48}$。

综上所述，双升游戏具有较大的动作空间、信息集数量和信息集大小，三者综合，使得双升游戏具有较高的游戏复杂度，能够检验算法的性能，并对学习算法提出了一定的挑战。从游戏复杂度视角看，双升游戏是具备作强化学习环境的必要条件的。

1.3 Botzone：开放式在线 AI 对抗平台

解决一个多智能体问题或在一个较为复杂的多智能体环境上实现水平较高的智能，一般可采用的方法并不局限于强化学习。基于规则和搜索的算法、监督学习、包含人机交互的模仿学习等方法都具备解决多智能体问题的潜力。而上述各种方法的综合应用则对多智能体系统提出了更高的需求：监督学习要求相当规模的数据集作为基础，则系统本身至少需要具备批量运行的能力，同时要求系统能够评测数据本身或数据来源（人类/产出数据的算法）是否可靠且具备一定程度的智能；人机交互则需要系统提供清晰便捷的人机接口（包括图形界面等）。

Botzone(www.botzone.org.cn)是北京大学人工智能实验室开发的开放式在线 AI 对抗平台[10]。该平台目前包含多达 45 种公开游戏，允许用户自由创建对局。该平台的主要功能特色包括：

(一) 支持人机交互。用户既能以人类玩家的身份通过便捷的交互接口参与对局，也能上传 AI 进行对局。

(二) 算法评测。Botzone 上每个游戏都有 AI 排名功能。用户可以提交 AI 参与天梯排名，由系统定期安排天梯赛，所有参与天梯的 AI 通过对局累积分数和进行排名，对于一个在天梯榜上较长时间的 AI，可以通过其天梯分数和排名表征其算法的智能水平。

(三) 算法比赛。Botzone 用户可以加入和创建小组。以小组为基础可以举办各种游戏算法比赛，Botzone 承办了 IJCAI 2020，2022，2023 和 2024 的麻将人工智能比赛，同时也承办了国内多所高校多门课程历年的竞赛性和考核性的算法比赛。目前 Botzone 网站上已有超过 100 个公开的小组。

(四) 游戏开发。Botzone 允许用户自行创建游戏并提供简单的接口。

(五) 回合制接口。Botzone 游戏对局的交互逻辑遵循回合制，在游戏进程中，由一个中心处理程序整合对局信息并依次与每个玩家程序（人类）和裁判程序交互，因此适用于传统玩法的棋牌类游戏的开发。

Botzone 平台的功能特点，为双升游戏在线对局系统的快速开发奠定了基础。也使得基于双升游戏的人机交互、数据采集和算法评测成为可能。接下来我们将从 botzone 双升

的开发入手，介绍基于双升游戏的在线对局和评测系统、强化学习环境和强化学习算法。本课题的贡献包括：

1. 基于 Botzone 对战平台开发了双升游戏的在线对局、评测系统，为人机交互、数据采集和算法评测创造了条件；
2. 设计和开发了双升游戏强化学习环境；
3. 综合特征工程、监督学习和强化学习设计了双升游戏人工智能算法。

1.4 小结

本节介绍了游戏对人工智能领域研究的重要意义、现有的基于游戏的强化学习基线环境及其特点，并在此基础上引入双升游戏作为一个新的多智能体环境原型，介绍了双升游戏的规则玩法，并从博弈模型和复杂度两个方面评估了基于双升游戏搭建强化学习环境的创新性和可行性。最后介绍了在线 AI 对抗平台 Botzone 作为双升游戏在线评测环境的开发基础。

2. Botzone 双升：双升游戏的在线对局与评测系统

2.1 Botzone 游戏框架

这一部分将介绍 Botzone 平台的对局运行机制和游戏代码构成。Botzone 游戏的运行环境由中央控制器、裁判程序、Bot 程序/人类接口、播放器组成。

2.1.1 Botzone 对局运行机制

Botzone 对所有游戏设定一个统一的中央控制器，并将裁判程序、Bot 程序/人类接口（统称为“玩家程序”）视为同质的单元。对局开始时，中央控制器接收对局初始化信息，并唤醒裁判程序，根据裁判程序的输出，控制器开始唤醒下一个玩家的 Bot 程序或等待人类接口的输出。此后，控制器依次先接收玩家程序输出，再调用裁判程序，并根据裁判程序输出唤醒对应的玩家程序，直到裁判程序判定对局结束为止。对局过程中，裁判程序和玩家程序均被唤醒多次，每个程序被唤醒的生命周期均为一次输入和一次输出。

2.1.2 Botzone 交互方式

Botzone 各程序间通过 JSON 与控制器进行交互。控制器发给裁判程序的信息包括初始化信息、历史交互及上一回合玩家的动作；裁判程序需要返回控制器的信息包含对局状态（进行中/结束）、下一个行动的玩家 id、发给玩家的信息及发给播放器的信息。控制器发给玩家的信息包含控制器和玩家交互的历史、裁判程序本轮发给玩家的信息；玩家需要返回本回合的动作（Bot 程序生成或由人类接口包装）。

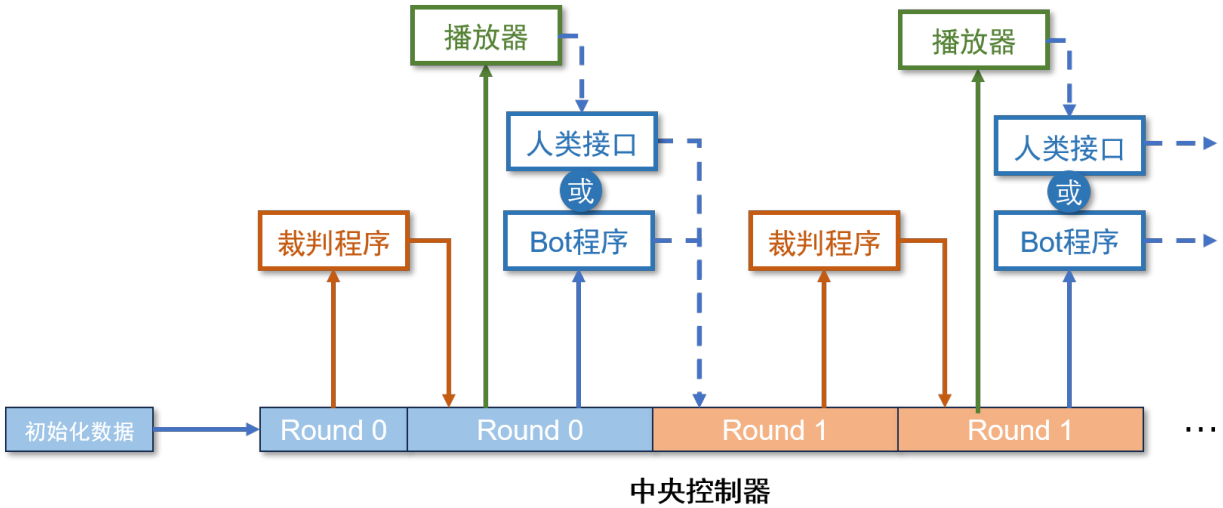

图 2.1 Botzone 架构与对局运行机制

2.2 Botzone 双升

基于 Botzone 游戏架构和对局机制，这一部分探讨以 Botzone 为平台开发双升在线人

机对战游戏。基于 Botzone 的双升在线人机对战游戏（下称“Botzone 双升”）包含控制器（由 Botzone 平台统一提供）、裁判程序及播放器（包含人类接口）。

2.2.1 Botzone 双升裁判程序

Botzone 双升裁判程序的主要功能包括接收对局初始化数据并开始对局、处理玩家动作、判断对局状态及结束对局并计算分数。每一轮裁判程序接收的 JSON 输入包含了全局初始化参数及历史交互（裁判程序的历史输出和所有玩家程序的历史输出）。 裁判程序的运行流程可以用伪代码表示为：

Algorithm 1 Judge for Tractor Game

```
1.  Input: Initialization data I, History H, Last response R from player P
2.  G ← Init(I) ◁ Initializing gamestate G.
3.  for log in H, do: ◁ Recovering gamestate
4.      G ← Update(G, log)
5.  if not isLegal(R): ◁ Check if the response is legal
6.      Punish(P)
7.      Terminate(G)
8.  G ← Update(G, R)
9.  if isEnd(G):
10.     Terminate(G)
11. else:
12.     Message ← MakeMessage(G) ◁ requests to the next player and info for video player
```

Update 模块读取对局日志并更新游戏状态，过程主要包含对局阶段的判定（双升游戏可分为发牌（报主）、盖底牌、出牌三个阶段）、局面更新、分数结算三个部分。局面更新函数被调用时，将先基于局面信息判定游戏处于哪一阶段，在不同的阶段下调用不同的函数处理本回合玩家动作并更新局面，如当前回合为一个出牌轮次（四人各出牌一次）或一局的末尾，裁判程序将额外进行分数结算。每次裁判程序的更新流程可以表示为如下的流程图：

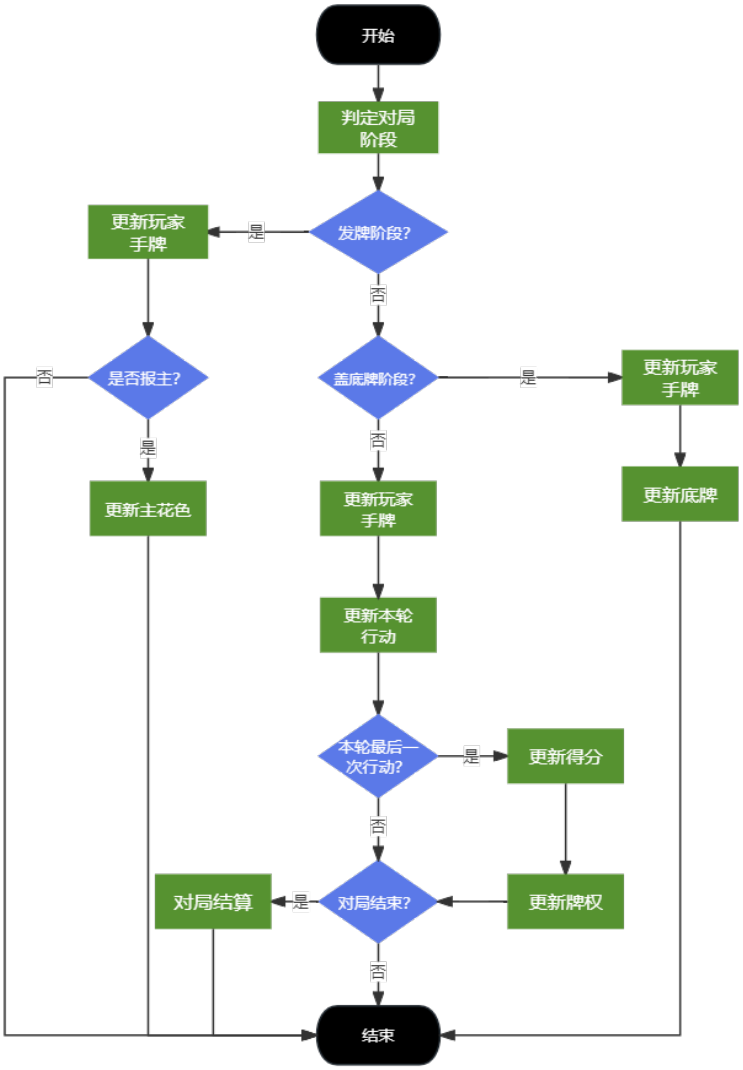

图 2.2 双升裁判程序局面更新模块运行流程

2.2.2 Botzone 双升播放器

Botzone 双升播放器程序接收中央控制器传递的信号，并将局面信息展示在图形界面上，同时提供人类操作的接口，接收人类动作并回传给控制器。播放器程序由一个 html 文件（图形界面框架）、一个 JavaScript 文件（网页事件逻辑）和一个 css 文件（图形界面效果）构成。Botzone 平台为所有游戏播放器提供了统一的 player_api.js 接口，以传递对局信息和回传玩家行动数据。本节重点介绍播放器网页事件的处理流程。

Botzone 双升播放器中处理网页事件的程序由 JavaScript 语言编写。和裁判程序一样，这个程序需要接收中央控制器发送的局面信息，处理并向人类玩家发送请求，接收人类玩家的动作并回传给中央控制器。与裁判程序不同的是，播放器程序提供了一系列类及函数供 Botzone 接口 player_api.js 实例化并调用，实现了对局过程中的长时运行。而 Botzone 双升裁判程序每次运行时仅处理一步的对局信息，因此每次处理当前步动作前需要根据中央

控制器提供的交互历史从头恢复局面。

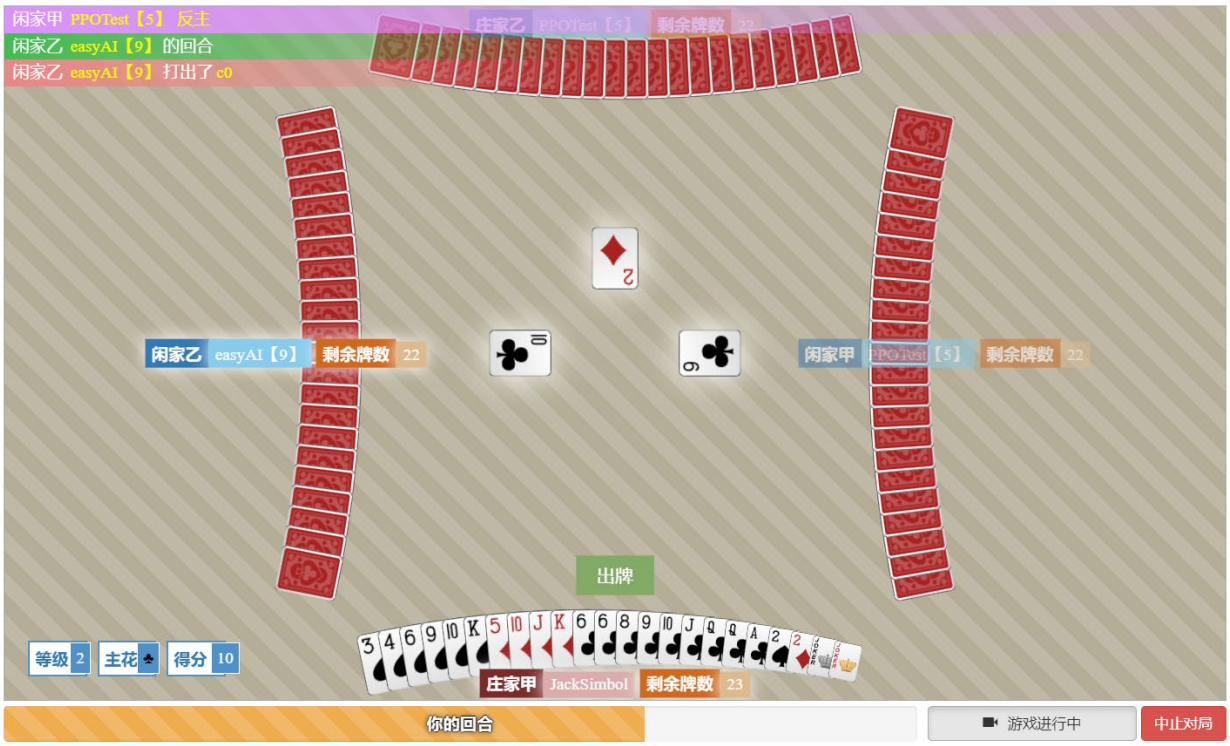

图 2.3 Botzone 双升播放器交互示例

长时运行加快了播放器程序响应的速度，使得播放器能快速流畅地通过图形界面与人类玩家进行交互，优化了用户体验。Botzone 双升播放器的响应流程涉及与中央控制器和人类玩家的双向交互。具体流程为：播放器启动并初始化对局和图形界面，初始化完成后播放器将向中央控制器发送信号通知初始化已完成，并等待中央控制器的指令。播放器收到中央控制器的信号后，根据信号更新对局状态和图形界面。若本轮需要人类玩家行动，播放器会展示人类接口并等待人类玩家操作。人类玩家提供操作信息后，播放器会进行简单的判定（人类动作是否合法），若人类动作不合法，播放器会要求人类玩家重新操作。人类玩家提供合法操作后，播放器会将人类动作包装回传至中央控制器，并等待下一个指令，如此循环直到中央控制器发送终止信号为止。Botzone 双升播放器的交互流程可以描述为如图 2.4 的流程图。

2.3 难点分析与解决

基于双升游戏的规则特点，相比于其他牌类游戏，Botzone 双升的开发过程具有独特的难点。本节将介绍 Botzone 双升裁判程序和播放器开发过程的主要难点和解决方案。

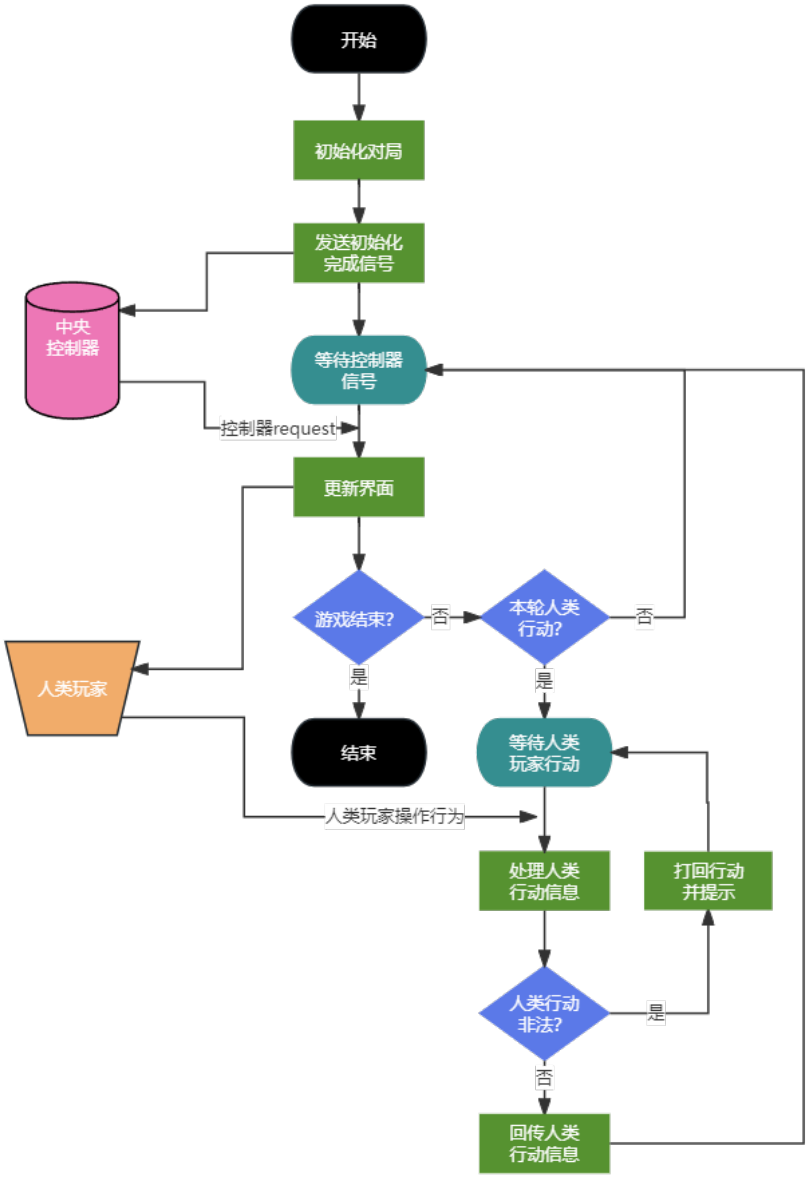

图 2.4 Botzone 播放器交互流程

2.3.1 多轮游戏流程的实现

与斗地主等单轮牌类游戏不同，双升游戏具有独特的升级玩法。四名玩家之间可以连续进行多场对局，每场对局结束后，根据两方得分情况进行“升级”，“升级”会改变下一场对局的初始状态（庄家、等级等）。多轮玩法使得 Botzone 双升相比其他 Botzone 平台游戏在开发过程中需要考虑到更多问题（包括对局间信息的传递、对局初始化方式等）。

Botzone 平台在创建游戏桌时允许房主定义一部分的初始化数据。这些数据会在对局开始前传递至中央控制器。利用这部分初始化数据，可以实现跨对局的信息传输。在多轮比赛模式下，每场对局结束时，比赛调度程序将从对局信息中读取本局庄家、等级和胜负信息，并输出下一轮对局的初始化信息，自动创建下一轮对局的游戏桌。如此即实现了双

升游戏的多轮玩法。

2.3.2 报主/反主规则的实现

牌类游戏的行动方法一般为回合制。而双升游戏中，报主/反主规则具有即时动作性，玩家只要手牌中有符合要求的牌，即可在发牌阶段的任一时刻进行报主或反主。报主/反主规则使得双升游戏兼具了即时动作与回合制的特点。如 1.3 节所言，Botzone 平台的对局交互逻辑为回合制，中央控制器在轮到玩家行动才会唤醒对应的程序或向播放器传递信息，因此实时监听玩家的行动是不可能实现的。因此，Botzone 双升的开发需要在平台的回合制机制和双升游戏的即时动作之间找到折中点，以依托现有平台最大程度还原游戏本身的规则。

报主/反主过程的即时动作，本质上是允许玩家在发牌过程中的任意阶段根据已经获得的手牌决定是否需要报主。因此，如果玩家在一定时间段中手牌的信息没有变化，那么可以认为这段时间内任何一个时刻做出的决定实际是等价的。在发牌阶段，每个玩家的手牌信息只有被发给牌时才会发生变化，因此在给同一个玩家的两次发牌之间，让该玩家进行一次决策即可。按照 Botzone 的交互逻辑，可以按如下方式组织发牌阶段的交互：从庄家开始，以顺时针顺序依次给四名玩家发牌，每次发一张。玩家收到牌后，若手牌满足报主/反主的条件，则可以立即进行报主/反主。玩家选择不报/报主/反主后，再给下一名玩家发牌。控制器发牌给玩家（request）和玩家选择是否报/反主（response）构成一回合的交互。

这种交互形式带来的效果是，若控制玩家的为 bot 程序，那么 bot 程序将在玩家被发给牌时唤醒，如玩家的手牌满足报主/反主条件，那么 bot 程序需要输出动作选择是否报/反主（否则默认不报）；若控制玩家的是人类接口，那么人类玩家在手牌满足报主/反主条件的每一回合都需要选择是否报/反主。这种处理方式客观上增加了人类玩家的操作数量，但是基于 Botzone 平台的交互机制最大程度地保留了双升游戏本身的玩法。

2.4 小结

在这一部分我们介绍了基于 Botzone 平台开发的双升在线人机对战游戏：Botzone 双升。我们简要介绍了 Botzone 平台的对局运行机制和交互方式，并详细阐述了 Botzone 双升的两大核心部分：裁判程序和播放器的功能和开发过程。最后，对开发过程中的几个难点进行了分析并提出了经过验证有效的解决方案。

3. 双升游戏强化学习环境

由于 Botzone 双升支持双升游戏的人机对局，可以利用其进行算法智能水平的评测和人类玩家行为数据的收集。但是，由于 Botzone 双升依托 Botzone 平台进行开发，且包含了控制器、播放器等程序，体量较大，难以胜任双升游戏强化学习需要的快速大规模采样。因此，需要开发双升游戏的强化学习环境。下面将介绍基于双升游戏的强化学习环境 RLTractor。

3.1 双升游戏环境建模

双升强化学习环境的开发以双升游戏的多智能体环境建模为基础。每一轮双升对局（不考虑升级）可以建模为一个马尔可夫博弈（Markov Game）过程。一般而言，一个马尔可夫博弈过程可描述为一个五元组$< N, S, A, T, R >$，N代表玩家的集合$N = \{1,2, \dots, n\}$，S代表所有可能的状态集合$S = \{s_1, s_2, s_3, \dots\}$，$A$代表可选的动作空间$A = \{a_1, a_2, a_3, \dots\}$，$T$为状态之间的转移函数$T: S \times A \to S$，$R$为奖励函数$R: S \to \mathbb{R}$。下面即围绕这五个部分对双升游戏进行建模。

(一) 玩家集合：一局双升游戏涉及到四名玩家，这四名玩家对家为一队两两组队，形成 2v2 的合作对抗关系。因此有$N_{Tr} = \{1, 2, 3, 4\}$。

(二) 状态空间与动作空间：表示双升游戏的状态和动作空间，是以牌的基本表示为基础的。双升游戏使用两副共计 108 张牌。对于一副牌而言，其中的每一张牌都可以映射为一个 4*14 的 0-1 矩阵上（下称“牌矩阵”），该矩阵的 1 元素所在行表示花色，所在列表示点数（第一行和第二行的最后一位分别表示小王和大王）。由于双升游戏使用两副牌，我们可以使用一个 2*4*14 的张量表示任意一张牌（如图 3.1）。如此，我们可以进一步将若干张牌的组合表示为这些牌对应的牌矩阵的和。由于双升游戏中可能出现的每一张牌在牌矩阵上都有唯一对应的 1 的位置，因此对于双升游戏中任意可能的牌组合（玩家手牌、一次行动中打出的牌、底牌、牌堆中的牌等），其对应牌矩阵的和也是一个 0-1 矩阵。我们会在 3.4 节中更详细地阐述这种设计的相关细节和目的。

双升作为一个局部观测的游戏，在游戏过程中每个玩家仅能观测到一部分的游戏状态。不失一般性，双升游戏过程中，任意局面下从任意玩家的角度观测到的游戏状态包含以下信息：

1. 本局等级。在对局开始时确定。
2. 主牌花色。在发牌阶段结束时确定。发牌阶段主花色会随着报/反主的进行而改变。
3. 当前手牌。玩家视角观测到的自己的手牌

4. 本局内玩家的行动历史。出牌阶段中，为了简化局面信息，仅保留最近一个轮次（四名玩家各行动一次）的行动历史，其余历史信息整合为本局中每一名玩家已出过的牌的集合。

利用上述的四部分内容，可以描述双升游戏中任意玩家的局面信息。

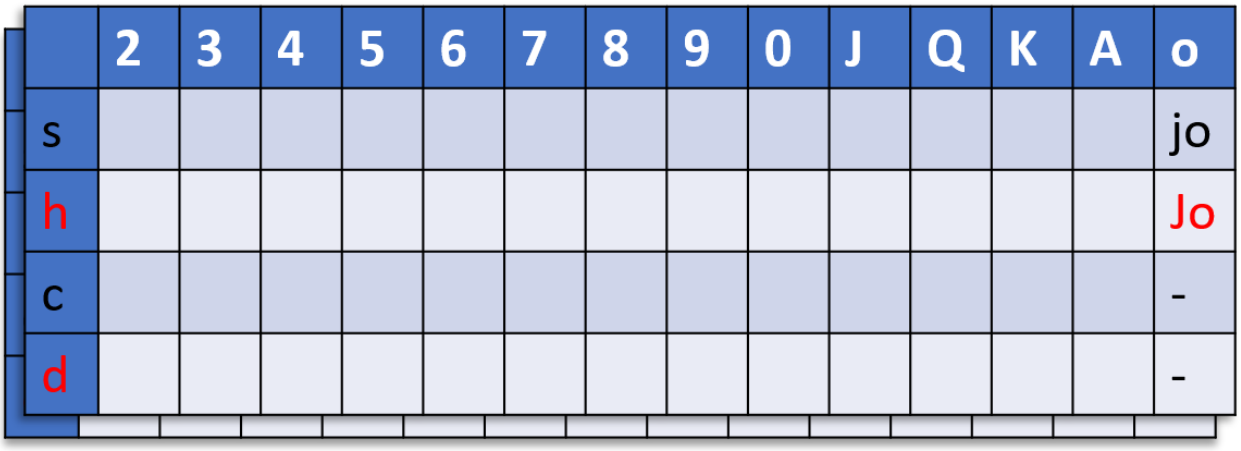

图 3.1 RLTractor 的牌矩阵

(三) 动作空间：双升游戏根据阶段不同，可行的动作包含了报主（亮出一张等级牌）、反主（亮出两张同花色级牌）、盖底牌（庄家盖覆八张手牌）、出牌。每次出牌根据是否为本轮次第一次行动以及上家出牌，可以打出不同的牌型组合（如 1.2 分析）。

(四) 状态转移：双升游戏的状态转移为确定性函数，通过双升游戏规则来完成。

(五) 奖励函数：双升游戏对局中带有积分机制（见 1.2.1），基于此可实现较为稠密的奖励设计。每一轮次行动后，都可以按照本轮获得的分数对玩家进行奖励：若本轮自家得分，则获得分数等量的正向奖励；若本轮对手得分，则获得分数等量的负向奖励。

3.2 算法框架

RLTractor 提供一个环境运行和算法训练的通用性框架。框架由并行的多个进程（多个执行器和一个学习器）、样本池和模型池组成。各部分间的通信关系如图 3.2 所示。

(一) 执行器：执行器负责运行大量对局并采样数据。执行器在每个对局开始前，从模型池中加载最新的模型，作为玩家添加到环境中。在游戏过程中，执行器将每个玩家的每个状态输入到网络中，获得每个玩家一整局的状态-动作-奖励组，将采样到的数据保存到样本池中。为提高采样效率，一般使用多个执行器并行采样。

(二) 样本池。样本池以队列形式储存执行器采集到的样本，并供学习器抽取样本更新模型参数。通过样本池，实现了执行器和学习器之间的跨进程训练数据传输。

(三) 学习器。学习器从模型池中读取当前最新的模型，并基于这个模型开始训练（如模型池为空，学习器会初始化一个新的模型）。当样本池中训练数据达到一定数量的时候，

学习器才会启动，并从样本池中乱序抽取训练数据更新模型参数。每隔一定时间，学习器会将当前的模型参数保存到模型池中。

(四) 模型池。模型池储存学习器初始化和训练得到的模型参数，并供执行器调用。通过模型器，实现了学习器和执行器之间的跨进程模型数据传输。

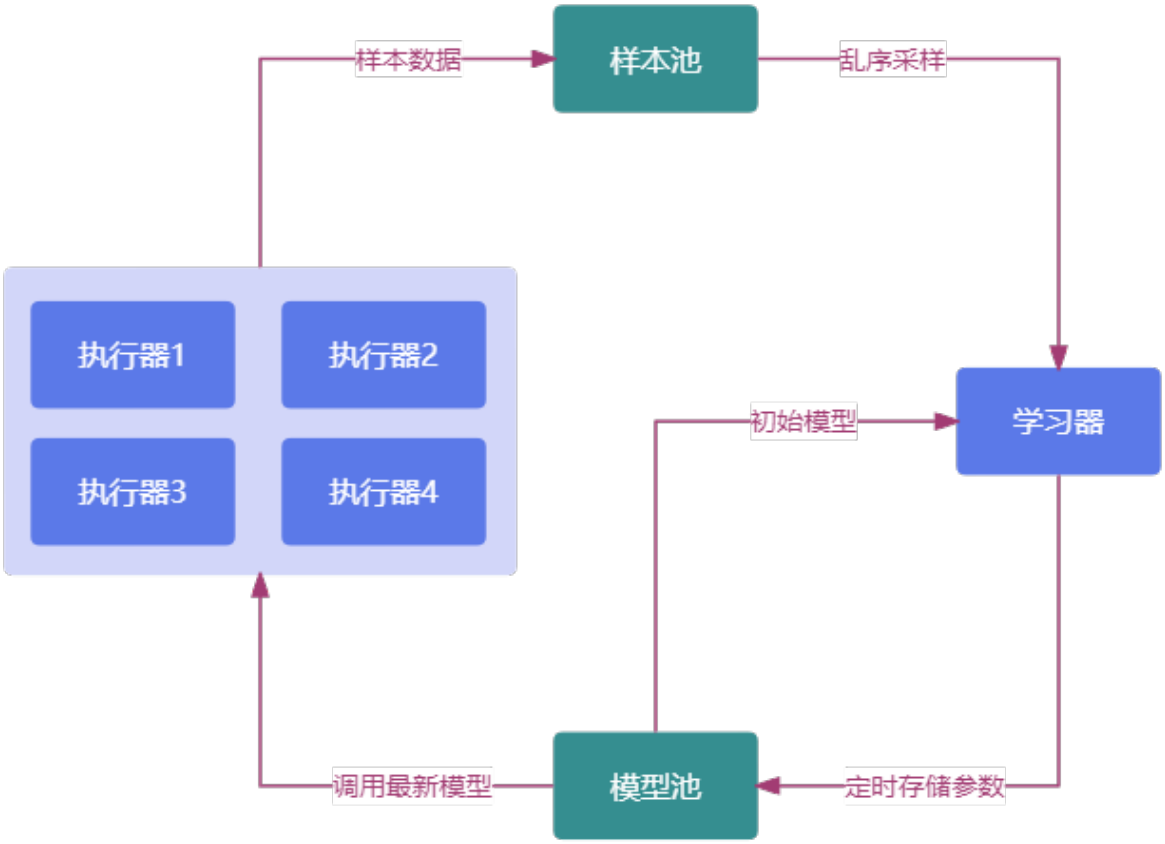

图 3.2 RLTractor 的算法框架

3.3 代码结构

双升强化学习环境代码由 python 语言编写，主要包含如下几部分：

1. model_pool.py 实现模型池功能
2. replay_buffer.py 实现样本池功能
3. actor.py 执行器组件，供调用和创建执行器进程
4. learner.py 学习器组件，供调用和创建学习器进程
5. env.py 双升环境代码
6. wrapper.py 特征处理（将环境输出的游戏状态处理成张量供模型输入）
7. model.py 模型结构
8. train.py 训练入口，控制训练流程，创建和调度执行器、学习器进程

为了便于评测模型和算法表现，双升强化学习环境中还包括了和 Botzone 双升进行交

互的__main__.py。该程序调用由用户上传至 Botzone 的模型文件，读取 Botzone 双升局面信息，通过模型做出决策并将决策翻译为符合 Botzone 双升规则的玩家行动。

3.4 难点分析与解决

传统棋牌游戏的特性和双升本身的玩法对双升强化学习环境的开发提出了一系列挑战。本节将介绍双升强化学习环境开发过程中的难点和解决方案。

3.4.1 牌的基本表示

双升游戏使用两副牌（每副 54 张）。Botzone 双升将这些牌按照点数和花色顺序编号为 0~107，并用编号表示这些牌。裁判程序、播放器和 Bot 程序在进行局面还原、判定和决策时需要通过编号还原牌面的具体信息。这种表示的问题在于，其一，并不能完全地反映牌面本身带有的花色和点数信息。对于在线人机对战程序而言，无论是裁判程序、播放器还是玩家程序都可以通过人为编写的代码来解译这些编号并获得牌面本身的信息，但是对于神经网络而言，单纯以编号表示牌面意味着模型需要额外学习编号和牌信息的映射关系。其二，由于双升游戏每回合出牌的数量可以不同，用编号表示牌意味着模型需要处理序列化的不定长信息。鉴于神经网络更加擅长处理矩阵运算，将牌面信息表示为矩阵或为更加有效的方式。

基于扑克牌的特点，可以用一个 4*14 的 0-1 矩阵刻画一副牌内的任何一张牌。该矩阵只有一个位置的元素为 1，其余位置均为 0。1 所在的行代表花色，列（1-13）代表点数，第十四列的第一、二行为小王和大王。由于双升使用两副一样的扑克牌，因此一局双升游戏中任何一张牌都可以用 2*4*14 的张量表示。这种表示方法的优势有二：首先，将牌面本身的信息储存在了矩阵结构中，方便神经网络训练；其次，可以用一个 2*4*14 的张量，表示双升游戏涉及到的任意若干张牌的组合，避免了序列化信息带来的影响。由于任何一个这样的牌组合都是两副扑克牌的子集，因而可以将这些牌对应的牌矩阵加和，得到表示牌组合的 0-1 张量。

3.5 小结

双升强化学习环境是开发双升游戏人工智能算法的基础。这一部分我们介绍了双升强化学习环境的建模，并按照训练需求设计了环境的算法框架、代码结构。最后，我们分析了双升强化学习环境开发过程中的难点并提出了解决方案。

4. EasyTractor：双升游戏人工智能算法

基于 Botzone 双升在线对战游戏和双升强化学习环境，我们得以进行双升游戏对局数据的采集、算法能力的评测和监督学习及强化学习训练，为双升游戏人工智能算法的开发奠定了基础。本部分将介绍面向双升游戏的人工智能算法：EasyTractor。该算法结合了特征工程、监督学习和强化学习等多种方法，实现了双升游戏上较高的决策水平。

4.1 特征工程

和斗地主类似，双升游戏的特点是，根据游戏阶段和局面的不同，每次行动出牌的数量可以不同。如何让模型打出符合规则的牌组合，是双升人工智能算法开发面临的问题之一。强化学习领域处理这一问题的通用方法是引入可行动作空间来限定和引导模型输出，即人为限定当前状态下可行的动作集合，要求模型在集合内进行选择。该方法应用于双升强化学习环境中，可以解决序列化输出的问题；对应地，需要对当前局面进行额外地处理以获得可行动作信息。由于双升游戏的出牌规则相对复杂，局面对行动空间造成的约束较多，人为的特征提取可以规避模型对复杂规则的学习，使得模型能够

EasyTractor 提供 move_generator 类来进行动作空间的生成。根据当前局面（综合全局信息、历史行动和玩家当前手牌），move_generator 按照特定的生成当前可以打出的所有牌组合。这些牌组合与 3.1 提到的其他局面信息一起作为状态(state)输入给模型。训练过程中，模型实际接收的输入信息为：

- major_mat: (2*4*14) 本局主牌信息
- deck_mat: (2*4*14) 玩家手牌
- hist_mat: (8*4*14) 本轮历史出牌信息（依次为 4 名玩家）
- played_mat: (8*4*14) 全局已出牌信息（4 名玩家）
- option_mat: (360*4*14) 本回合可选牌组信息（180 个可选动作）

如 move_generator 生成的可行动作数量不足 180 个，余下的通道将以 0 矩阵填充，同时会向模型提供一个 180 维的 action_mask，表示当前实际有多少可行动作。以上各项信息对应的张量在 0 维上堆叠，得到 380*4*14 的状态张量作为模型输入。

4.2 算法架构

EasyTractor 的训练过程结合了监督学习和强化学习方法。首先使用专家数据对决策网络进行监督学习预训练，再将预训练网络添加到双升强化学习环境中进行自对弈强化学习。下面将介绍监督学习和强化学习过程的模型和算法细节。

4.2.1 网络模型结构

EasyTractor 采用改良后的 ResNet18[11]作为模型结构，以适应矩阵化的牌面信息。针

对 RLTractor 环境提供的状态信息的形状和结构，对网络模型进行了如下的改动：

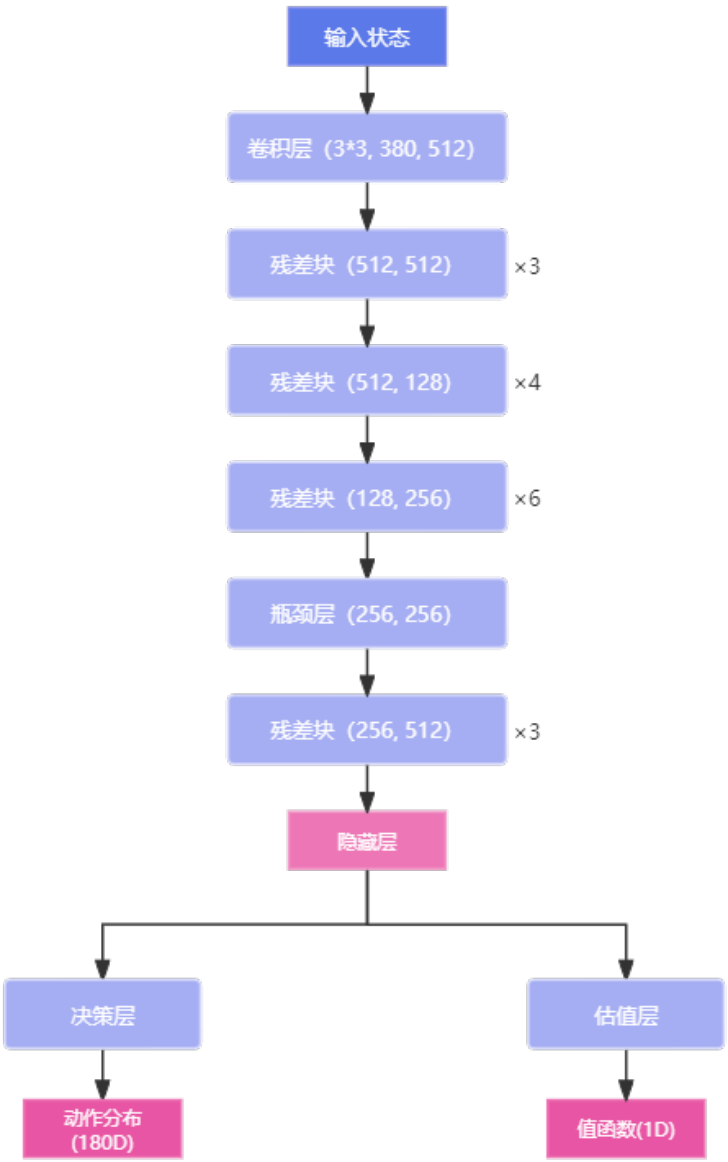

图 4.1 EasyTractor 决策网络结构

(一) 输入输出通道数：ResNet18 通常用于处理图片信息（类型多为 3 通道 RGB 矩阵等），并通过第一层卷积将 3 通道扩增至 64 通道。而 RLTractor 将游戏状态建模为 380 通道的张量，通道数远高于 64。因此，在 EasyTractor 决策网络的第一层卷积处将输出通道数增加到 512 通道，避免损失信息，并在后续残差块中再将输出通道数降低。

(二) 降采样规模：基于图片的分辨率特征，图片信息的矩阵尺寸一般较大（例如每一通道为 64*64 矩阵）。ResNet18 在第一层卷积使用步长为 2 的 7*7 卷积核进行快速降采样，这对于每通道尺寸仅为 4*14 的双升游戏状态似乎是不太现实的。因此，EasyTractor 决策网络将第一层卷积的卷积核大小设为 3*3，步长设为 2 以适应游戏状态的尺寸。

(三) 隐藏层、决策层和估值层：为适应基于优势理论[12]的强化学习算法的潜在需求，网络模型在输出决策（动作分布）的同时还需要提供一个对于当前状态的估值。针对这一需要对模型作出的改动为：移除 ResNet 末尾原有的全连接层，模型接收游戏状态输入后输出一个隐藏层。这个隐藏层将分别输入到决策层和估值层中，得到决策分布和状态估值，如图 4.1 所示。

4.2.2 监督学习预训练

监督学习预训练采用的数据来源于 Botzone 双升游戏中实力较强的 Bot。该 Bot 采用基于规则和搜索的策略，决策上已经达到一般玩家的水平。通过 Botzone 平台，获得了该 Bot 程序 4096 局自对弈的对局日志，经过数据处理获得 242202 个状态-动作对。利用这些数据，按 4：1 划分训练集和测试集，选取批量大小为 256， 学习率 3e-4，对 EasyTractor 的网络模型进行预训练。模型在测试集上达到 54%准确率。由于缺乏对应的监督数据，在监督学习中对于网络模型中的估值层不做训练，仅训练隐藏层前的部分和决策层。

4.2.3 强化学习

基于双升强化学习环境，可以对预训练模型进行强化学习训练。强化学习使用 PPO 算法[13]，采用自对弈的训练模式。PPO 算法的损失函数包含三个部分：策略损失函数、值损失函数和交叉熵损失函数。策略损失函数和值损失函数的计算均涉及当前估值层的输出，交叉熵损失函数则涉及到决策层输出的动作分布。

训练过程中，网络模型初始化为监督学习得到的预训练模型，五个并行的执行器进程调用初始化模型并开始进入对局采样，采样得到的数据存入样本池中，样本池中样本数量达到一定数值后，学习器将启动，从样本池中随机采样并更新模型参数。学习器进程和执行器进程并行运行，每隔一定时间，学习器会将模型参数保存至模型池中，供执行器在运行新的对局时调用。每个执行器的采样对局数设置为 2000，学习率设置为 3e-5。

4.3 性能评测

以下将通过简单的实验验证该算法架构的可行性并评测模型的决策水平。维持模型结构不变，我们以纯监督学习和纯强化学习方法作为基线算法。为比较 EasyTractor 训练的模型和这两种基线算法所得模型的智能水平，我们让 EasyTractor 模型与两种基线模型分别进行多轮对局。评测比赛的组织方式为：由两个模型控制四名玩家进行游戏，每个模型分别控制两名对家（决策时，两个对家虽然由同一模型控制，但独立地进行决策，且玩家间没有通讯）。每个对局结束，按照升级数给予分数（例如：闲家升两级，则控制闲家的模型获得两分）。纯监督学习方法采用和 EasyTractor 预训练过程相同的超参数和数据集，纯强化学习方法采用 PPO 算法进行自对弈训练，超参数与 EasyTractor 强化学习过程一致。将两种基线算法得到的模型与 EasyTractor 分别进行 1024 局对局，评测结果如表 4.1 所示。

由结果可以看出，EasyTractor 在 1024 局的表现强于纯监督学习和纯强化学习算法，证

明 EasyTractor 架构相比基线算法能够提升模型性能。

表 4.1 EasyTractor 评测结果

基线算法	EasyTractor 得分	基线模型得分
监督学习	840	696
强化学习	738	624

4.4 算法表现和局限性分析

EasyTractor 提出了一种针对双升游戏的结合监督学习和强化学习的训练架构，并使模型达到了接近人类玩家的智能水平，但是其算法尚存在一定的局限性，性能仍然有提升空间。

(一) 监督数据的质量和多样性。监督学习预训练的目的是为强化学习提供一个良好的初始模型以指导其采样，避免自对弈强化学习漫长且盲目的探索和采样。目前用于监督学习预训练的专家数据全部来源于同一个基于规则和搜索的 Bot 程序的自对弈对局，在质量和多样性方面都有所欠缺。首先，Bot 程序本身的决策水平仅能够达到初级玩家的水平，其产出的数据也只能对模型提供较为基础的指导。其次，由于 Bot 程序是基于规则和搜索的，因而在决策上表现出明显的风格特点，只使用单一的 Bot 程序进行自对弈，产生的数据也服从一个高度特异化的分布。综合来看，由于监督数据本身的质量和多样性偏低，模型很难通过单纯的监督学习获得预期的知识并达到预期的智能水平。

(二) 监督学习预训练的准确度。由于预训练模型仅为强化学习提供了探索起点，因此对监督学习预训练准确度的要求并不苛刻。作为一个预训练模型，我们同样不希望该模型在决策上与基于规则的算法过度相似，而更希望其保留一部分参数优化和随机探索的空间。目前监督学习预训练模型在测试集上准确率为 54%，接近同样采用监督学习预训练的 AlphaGo[14]在预训练测试集上的最优准确率 57%，但是仍然有可能通过训练参数和模型结构的调整来提升其性能。

(三) 多轮游戏特征的利用。双升游戏具有多轮的特点。在一整局比赛中，玩家之间会进行多个相似轮次的游戏，EasyTractor 处理多轮对局的方案是将每一轮对局分立地处理。即前几局的信息并不会影响模型本局的决策，模型的决策风格也不因整体游戏阶段（例如：当前双方等级、庄家信息）的改变而改变。实际上，多轮游戏在延长了游戏进程的同时，为玩家动态调整决策和进行对手建模创造了条件。为了充分利用多轮游戏的特征，可以考虑让模型预测对手的决策偏好并动态调整自身的策略[15]。

(四) 多智能体间的信息交流。作为一个多人游戏，双升游戏虽然没有为玩家提供自然语言交流的渠道（一部分官方比赛在规则中明确禁止玩家间的语言交流），但是玩家仍然可

以通过出牌等游戏行为来进行信息交流。一些特殊花色和点数的牌被玩家公认或事先约定为“信息牌”，玩家在游戏中通过打出这些牌来向队友传递自己的手牌信息或行动意图（例如某种花色的牌已经出完，或希望打出某些花色的牌）。目前 EasyTractor 并未考虑玩家间信息交流，但是可以推断，利用这种潜在的交流将会对模型的推理能力提出更大的挑战，当然也会提高模型的决策表现。

4.5 小结

这一部分我们介绍了双升游戏的人工智能算法：EasyTractor。该算法整合了特征工程、监督学习和强化学习等多种方法，依托 Botzone 双升在线平台和双升强化学习环境完成了数据采集、处理、训练以及性能评测。我们详细介绍了该算法的模型架构、训练过程，并分析了其性能表现和局限性。

5. 总结与展望

长期以来，游戏一直是机器学习和强化学习重要的研究对象和灵感来源。双升游戏作为中国流行棋牌游戏的一种，以其独特的多轮多人混合博弈模式、非完美信息的特征和相当高的复杂度成为潜在的强化学习环境开发原型，并为游戏人工智能提出新的研究问题和挑战。本文基于双升游戏，依托 Botzone 在线人机对战平台开发了 Botzone 双升在线对战游戏，并搭建了双升强化学习环境。在以上工作的基础上，本文进一步提出了一种面向双升游戏，整合了监督学习和强化学习等方法的人工智能算法，为双升人工智能问题提出了初步的解决方案，并分析了算法的局限性和提升方式。

长期以来，一部分人工智能领域的研究者选择将游戏作为其研究的对象、平台，或者从游戏中发现研究问题和获取灵感，证明了游戏具有特殊的研究价值。从游戏本身的角度而言，游戏的研究价值可以概括为二：策略的高度复杂性和对现实世界的模拟性。被用作强化学习环境或环境开发原型的游戏也大抵可以被划分到这两个视域。从这个角度看，传统的棋类和牌类游戏的研究价值很明显在于前者。

棋类游戏和牌类游戏，在一段时间内曾广受深度学习和强化学习研究领域的关注。被广泛应用于强化学习环境的游戏包括了国际象棋、围棋、跳棋、德州扑克、21 点等。而今天，连 AlphaGo 在围棋领域碾压人类玩家的功绩也已经成为历史，在足够庞大的模型和算力面前，棋牌类游戏高度抽象的规则和策略复杂度已经不再构成新的挑战，对于大多数游戏，当模型和算法的性能超越人类之后，其决策水平的进一步提升也没有太大的价值和必要。在游戏领域，我们已然可以宣布自己制造出了很多强大的智能体；但从其他领域的视角，我们一方面希望游戏能够作为现实世界的投射，让智能体通过虚拟世界的训练获取知识和智能以完成现实世界的任务；另一方面也希望在游戏中过关斩将的强大智能体除了站在人类玩家的对立面外，还能够成为教练、解说、陪练，甚至队友，和人类紧密配合、相互理解。《我的世界》《星际争霸》《文明》等一系列游戏以其对现实世界不同程度的模拟和更多的智能体数量成为了新一轮游戏人工智能研究的核心对象。

然而，上述研究问题并非《文明》等当代电子游戏独有的，很多研究者低估了双升、斗地主、麻将等一系列传统棋牌游戏在这类问题上的研究价值。这些棋牌游戏不仅保留了多智能体博弈的特征并提供良好的博弈抽象，而且通过人为定制的规则抽象简化了状态和动作空间，相比《我的世界》等控制上较为复杂的游戏，实际上节约了智能体学习控制的算力开销，使得算法能够专注于与其他智能体的交互、对齐和配合，为人机协同方面的研究奠定基础。也正因如此，我们期待以本工作为起点，能够基于 Botzone 双升和双升强化学习环境开展更多涉及到多智能体竞争与合作、多智能体信息交流等领域的研究，并对麻将、双升、掼蛋等中国传统棋牌游戏在人工智能领域的研究潜力作更加充分的论证，做出更有价值的成果。

参考文献

[1] M. Kim *et al.*, “The StarCraft Multi-Agent Exploration Challenges: Learning Multi-Stage Tasks and Environmental Factors Without Precise Reward Functions,” *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3266652.

[2] L. Fan *et al.*, “MINEDOJO: Building Open-Ended Embodied Agents with Internet-Scale Knowledge,” in *Advances in Neural Information Processing Systems*, 2022.

[3] D. Zha *et al.*, “RLCard: A platform for reinforcement learning in card games,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2020. doi: 10.24963/ijcai.2020/764.

[4] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The Arcade Learning Environment: An evaluation platform for general agents,” *Journal of Artificial Intelligence Research*, vol. 47, 2013, doi: 10.1613/jair.3912.

[5] A. Bakhtin *et al.*, “Human-level play in the game of Diplomacy by combining language models with strategic reasoning,” *Science*, vol. 378, no. 6624, 2022, doi: 10.1126/science.ade9097.

[6] E. A. Duéñez-Guzmán, S. Sadedin, J. X. Wang, K. R. McKee, and J. Z. Leibo, “A social path to human-like artificial intelligence,” *Nature Machine Intelligence*, vol. 5, no. 11, 2023, doi: 10.1038/s42256-023-00754-x.

[7] D. Zha *et al.*, “DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning,” in *Proceedings of Machine Learning Research*, 2021.

[8] Wikipedia, “Game Complexity.” [Online]. Available: en.wikipedia.org/wiki/Game_complexity

[9] M. Johanson, “Measuring the size of large no-limit poker games,” *Technical Report, Department of Computing Science, University of Alberta*, vol. TR13-01, 2013.

[10] H. Zhou, H. Zhang, Y. Zhou, X. Wang, and W. Li, “Botzone: An online multi-agent competitive platform for AI education,” in *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 2018. doi: 10.1145/3197091.3197099.

[11] J. W. Yun, “Resnet - Deep Residual Learning for Image Recognition (cvpr16 best),” *Cvpr*, vol. 19, no. 2, 2016.

[12] V. Mnih *et al.*, “Asynchronous methods for deep reinforcement learning,” in *33rd International Conference on Machine Learning, ICML 2016*, 2016.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” arxiv.org/abs/1707.06347

[14] D. Silver *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, 2016, doi: 10.1038/nature16961.

[15] J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, “Learning with opponent-learning awareness,” in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2018.

致　谢

从很久以前，就在盘算着致谢的事情，应当写些什么，应当感谢什么人，该如何起笔，又要以哪一首诗或是格言收尾。想了很多个版本，终于真的拿起笔来，却又不知道该说什么。我们中的很多人，也许久已疏离于修辞，也许久已不曾把思想和情感直白地付诸文字，以至于面对一张白纸，写出的任何一个字都仿佛是矫情刻意、笨拙生硬，或者为人不齿。或许另有一些人，也许担心遗忘，因而十分隆重地不敢轻慢了这一篇小文，寄予额外的期望，希望以此致谢为自己四年的生活作传。但是说到底，致谢就是致谢，就好比运动会前，无用的开场词当然是越短越好。

感谢我的导师李文新老师，不仅给本工作大量的指点和支持，而且无时无刻不以她的言谈和行动引导着我。令我叹服的不止是导师对科研的深刻理解，自加入课题组的那一天起，老师积极自信的心态、率性自然的气质，以及对学生理解、宽容、负责的态度，始终影响着我，使我如沐春风。

感谢我的父母，拊我蓄我，长我育我。这么多年，爸妈始终是我最忠实的倾听者，也是我最坚强的后盾。感谢他们的支持，让我一路走到今天。我们并不是完美无缺的家庭，但这么多年相互扶持，也走过了不少风雨，感谢他们教会我隐忍与自立，也感谢他们教会我宽和与从容。

感谢恩师们。感谢孔雨晴老师和邓小铁老师，保留了我对理论研究的兴趣，让我见识到另一个领域研究的视界。感谢赵国栋老师、梅申友老师、毕明辉老师、尤泓斐老师等，让我领略到北大课程的丰富多彩，发现更多的可能。感谢胡俊老师，始终支持和关心江淮的同学们，在我迷茫时为我指明方向。

感谢鲁云龙、李昂、郑启帆、汪永毅、李凌峰、齐藤、王楚才、陈泊舟、刘涵宇、周亦楷，到实验室一年有余，时间不长，却有许多珍贵的奇遇，更有全新的收获。有你们在，AI lab 总是充满活力。感谢范金楷，参与了 Botzone 双升的早期测试和 Bot 程序开发工作，为本工作的推进做出了重要的贡献。

感谢通院多智能体组的朋友们，感谢封雪老师、綦思源老师、黄奕喆、孔繁奇、刘好、唐敏、毕铭杰、秦傲洋，你们的学术理解和科研经验都使我收益颇丰，在和你们的合作中，我开始触摸到科研的形状。

感谢元培学院的朋友们。感谢我的舍友孙皓然、刘宇川，感谢你们在我每个敲键盘的深夜无言的体谅，感谢我们一起度过的许多个日夜；感谢孙英博、唐明川、赵浩男、曹吕林、张一帜、罗元辰、陈彦伯、冯奥博、豆佳豪、肖霄、黄玉龙腾、朱逸轩、吕天、翟与宁、蒋双羽、龚毅、安昱达、叶子琦、张珏、高骞、廖振宇、王君宇，还有其他“三五四”的伙伴们，认识你们之前，我不曾想过自己会拥有这么多朋友，我又是何等幸运地生活在这样一个融洽的团体之中，与你们一同分享大学生活。感谢熊辰浩、陶凝骁、王瑞诚、罗逸龙等同学，为我提供了许多学业和科研上的帮助，让我的生活有了更多色彩。感谢林

昊苇、尹禹同、叶航、姜广源、古云天、杜逸恒、陈骏清、杨昊翔、李珩立、陈旭峥还有许许多多的通班同学们，很荣幸能和你们一起探索出一条新路。

感谢韩晗、陈本、缪辰、张矣可、Vicky、李佶丰、叶湾韵、闫天行、罗泽权、万智鹏，还有编辑部的其他朋友们，你们是欢乐的源泉。在一个如此神奇的地方，遇到了这么多神奇的人，说了许多神奇的话，写出了许多神奇的故事，对我而言真是如同美梦一般。

感谢江淮的朋友们。感谢王子澳，在我尚未踏入校门时为我画出了关于北大和未来的第一幅画，感谢韦楚韵、刘鑫，在我最无助的时候给了我支持和力量。感谢马亦然、沙熠、许新哲、黄佳慧、王康安、杨乐山、谢子俊、王杨依、张天行还有许许多多江淮的朋友们，让北大江淮像家一样吵闹而温馨。感谢江剑飞，陪我冶游放歌，陪我走过冻风和积雪的山道，陪我一点点精进球技，陪我在赶不动文章的深夜有一搭没一搭地闲侃。

我想要感谢很多很多人，我结识的、陪伴我的可爱的人们，我恨不得把所有人的名字写进这篇致谢，虽然已经列举很多。回头看去，我像个自说自话的人，絮叨着一个又一个名字，絮叨着我们共同的故事，但我又是如何充满自豪和欣喜地写下这一个又一个名字，是你们成就了我，也是你们定义了我，因为有你们，我度过了独一无二的大学生活。很快，我也将带着你们给我的陪伴与帮助、力量与鼓舞开始新的旅途。

这篇文章完成的前一天，燕园有很美的晚霞。许多人驻足昂首，许多人涌向未名湖畔，或是爬上教学楼的楼顶，只是为了一睹这可遇不可求的美景。夕阳西下，风起云散，有的人并肩前行，有的人分道扬镳，但是我们终将找到自己的风景，正如我们曾相聚在这片薰衣草般的紫霞之下。到那时，愿我们都能想起那片晚霞，愿我们都还保有一颗雀跃的心。

My heart leaps up when I behold
A Rainbow in the sky:
So was it when my life began;
So is it now I am a man;
So be it when I shall grow old,
Or let me die!

—— ***The Rainbow***, *William Wordsworth*

王星博

2024 年 5 月 15 日于燕园

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。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。本声明的法律结果由本人承担。

论文作者签名：

日期：　　年　　月　　日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：　　导师签名：

日期：　　年　　月　　日