中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于深度强化学习的多无人机电子对抗决策算法研究

作者姓名：高　远

指导教师：郭立红　研究员　　孙守红　研究员

中国科学院长春光学精密机械与物理研究所

学位类别：工学博士

学科专业：机械电子工程

培养单位：中国科学院长春光学精密机械与物理研究所

2024 年 6 月

中国知网　https://www.cnki.net

Research on multi-UAV electronic countermeasure decision algorithm based on deep reinforcement learning

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

In Mechatronic Engineering

By

Yuan Gao

Supervisor: Professor Lihong Guo

Professor Shouhong Sun

Changchun Institute of Optics, Fine Mechanics and Physics,

Chinese Academy of Science

June 2024

中国知网 https://www.cnki.net

摘 要

在现代战争中，除了陆、海、空、天等四个传统作战域以外，电磁频谱也是一个重要的作战域。电子战是剥夺敌方使用电磁频谱，同时保护己方使用电磁频谱的科学与艺术。从电子战最初诞生一直到 21 世纪，电子战一直呈指数增长，且随着无人化作战逐渐成为电子战的重要作战样式，其在现代战争中所扮演的角色越来越重要。军事智能化是人工智能的重要应用方向，而决策博弈是智能化战争对抗的核心和中枢，借由电子对抗仿真技术，采用深度学习与强化学习相结合的方法研究电子对抗作战中多执行单元间的完全合作任务，是解决装备协同对抗问题的有效途径。

当前，电子对抗决策研究仍存在一些问题，首先，作战是复杂的多维度任务，而目前大部分电子对抗决策研究都聚焦于智能感知、干扰资源分配、认知干扰决策以及抗干扰决策等子领域，更关注于局部技术，缺乏完整电子战的决策研究；其次，目前集群智能中的个体决策内容较为简单，缺乏个体面对复杂任务复杂决策内容时的多智能体决策研究，并且当个体决策内容复杂化之后多智能体之间如何进行高效协作发挥更高的作战效能也是电子对抗策略研究的重要内容之一。

本文以典型电子对抗任务场景——多无人机协作护航任务为例进行策略研究。在多无人机协作护航任务中，多无人机需要在蓝方的电子侦察覆盖以及火力打击覆盖下，通过电子对抗的手段，削弱蓝方侦察能力，避免被蓝方火力打击，并为后续轰炸动作做好铺垫。针对任务中无人机需要同时决策飞行、侦察、干扰等多种类动作以及多智能体之间协作性不强的问题，本文构建了多无人机协作护航任务仿真系统模型，提出了基于任务分解部分集成的多智能体强化学习决策方法（Task Decomposition and Partial Assembly Reinforcement Learning Algorithm，TDPA）、基于动作依赖的近端策略优化算法（Action Dependence of Proximal Policy Optimization，AD-PPO）、带基线的多智能体多种类动作改进 Transformer 序列决策算法（Multi Agent Multi Action DeLighT Baseline Transformer，MA2DBT）智能决策算法。本文的研究内容如下：

（1）针对典型电子对抗场景中的多无人机协作护航任务，分析其任务流程以及任务细节，以红蓝对抗为背景，设计了多无人机协作护航任务仿真系统模型。将多无人机协作护航任务建模为部分可观测马尔可夫决策过程，以红方无人机作为智能体，设计了智能体的观测空间、动作空间，并且根据任务目标设计了任务环境的奖励函数。此外，对多无人机协作护航任务中实体的功能模型进行了构建，包括红方无人机侦察模型、蓝方雷达探测模型、蓝方雷达受干扰

中国知网 https://www.cnki.net

模型等。

（2）多无人机协作护航任务中，由于智能体无人机需要在多种类动作组成的高维动作空间中进行决策，而任务包含侦察、干扰等多个任务目标，不同任务目标对于多种类动作的选择有不同需求，容易产生互相干扰的情况，进而影响智能体的整体决策。基于分层决策的思想，提出了基于子任务分解的多智能体强化学习决策算法，将多无人机协作护航任务分解为侦察和干扰两个子任务，分别进行策略学习，并根据任务的进展情况综合两个策略，得到完整的智能体决策结果。本文为两个子任务分别设计了不同的奖励函数，解决了策略学习中的任务冲突问题，加强其对子任务的针对性学习，并通过仿真实验验证算法的有效性。

（3）在前述的基于子任务分解的多智能体强化学习决策算法中，两个子策略对飞行动作进行了重复决策，存在决策冗余，而且综合策略结果中的飞行动作对某一任务目标存在倾向性，容易造成针对另一任务目标的动作孤立甚至失效。针对该问题，提出了一种基于动作依赖的多智能体强化学习决策算法，对于决策模型的神经网络结构进行了特定性的优化，将前置动作的决策结果与观测向量一起作为后置动作的决策输入，建立动作与动作之间的依赖关系，不孤立任何一类动作，同时降低了智能体决策的难度。最后仿真实验结果表明，该算法有效降低了动作失效率，具有更好的收敛特性和任务评估结果。

（4）在前述的基于动作依赖的多智能体强化学习决策算法中，虽然加强了动作与动作之间的协作性，并通过决策模型网络结构的优化降低了单个决策模型决策的动作维度，但面对不同的任务场景时，需要对任务实际进行深入分析并根据实际情况重新设计网络结构，不具备通用性，且智能体之间的协作性没有受到重点关注。针对上述问题，本文提出了一种面向复合动作空间的多智能体强化学习序列决策算法，将序列模型引入智能决策，每一步的决策过程拆分为一个决策序列，每次决策只选择一个智能体的一种动作，且决策的输入包括当前观测信息以及已决策出的动作信息，加强多智能体之间动作之间协作性的同时降低单次决策的维度，并且考虑到每个时间步决策序列的延长，通过优化网络模型、优化价值估计模型的方法提升决策模型的性能。最后，通过仿真实验验证了 MA2DBT 算法在复合动作空间场景下的有效性，并通过消融实验探索算法各部分对整体的贡献。

关键词：电子对抗仿真，智能决策，多智能体强化学习，分层决策，动作依赖，序列模型

中国知网 https://www.cnki.net

Abstract

In modern warfare, in addition to the four traditional combat domains of land, sea, air and space, the electromagnetic spectrum is also an important combat domain. Electronic warfare is the science and art of depriving the enemy of the electromagnetic spectrum while protecting its own use. From the birth of electronic warfare until the 21st century, electronic warfare has been growing exponentially, and as unmanned warfare gradually becomes an important operational style of electronic warfare, it plays an increasingly important role in modern warfare. Military intelligence is an important application direction of artificial intelligence, and decision game is the core and center of intelligent war confrontation. By using electronic countermeasure simulation technology, the method of combining deep learning and reinforcement learning to study the complete cooperation task between multiple execution units in electronic countermeasure operations is an effective way to solve the problem of equipment cooperative confrontation.

At present, there are still some problems in the research of electronic warfare decision-making. First, combat is a complex multi-dimensional task, and most of the current research on electronic warfare decision-making focuses on sub-fields such as intelligent perception, interference resource allocation, cognitive interference decision-making and anti-interference decision-making, and pays more attention to local technologies, lacking complete decision-making research on electronic warfare. Secondly, individual decision-making content in current cluster intelligence is relatively simple, and there is a lack of multi-agent decision-making research when individuals face complex tasks and complex decision-making content. Moreover, when individual decision-making content is complicated, how to cooperate efficiently among multiple agents to achieve higher combat effectiveness is also one of the important contents of electronic countermeasures strategy research.

In this paper, a typical electronic countermeasures task scenario-multi-UAV cooperative escort mission is taken as an example to study the strategy. In the multi-UAV cooperative escort mission, the multi-UAV needs to weaken the reconnaissance capability of the blue side by means of electronic countermeasures under the coverage of the blue side's electronic reconnaissance and fire attack, avoid being hit by the blue side's fire, and pave the way for subsequent bombing operations. In order to solve the problem that UAVs need to make decisions on multiple kinds of actions such as flight, reconnaissance and interference at the same time, and the cooperation between multiple agents is not strong, this paper constructs a simulation system model of multi-UAVs cooperative escort mission. A multi-agent Reinforcement Learning Algorithm based on Task Decomposition and Partial Assembly (TDPA) is proposed, Action Dependence of Proximal Policy Optimization (AD-PPO) and Multi Agent Multi Action DeLighT

中国知网 https://www.cnki.net

Baseline Transformer intelligent decision algorithm (MA2DBT). The research content of this paper is as follows:

(1) Aiming at the multi-UAV cooperative escort task in typical electronic countermeasures scenarios, the task process and task details are analyzed, and the multi-UAV cooperative escort task simulation system model is designed against the background of red-blue confrontation. The multi-UAV cooperative escort task is modeled as a partially observable Markov decision process, taking the red UAV as the agent, the observation space and action space of the agent are designed, and the reward function of the task environment is designed according to the mission objective. In addition, the functional models of the entities in the multi-UAV cooperative escort mission are constructed, including the red UAV reconnaissance model, the blue radar detection model, and the blue radar interference model.

(2) In the multi-UAV cooperative escort task, because the UAV needs to make decisions in the high-dimensional action space composed of multiple types of actions, and the task contains multiple task objectives such as reconnaissance and interference, different task objectives have different requirements for the selection of multiple types of actions, and it is easy to interfere with each other, thus affecting the overall decision of the agent. Based on hierarchical decision making, a multi-agent reinforcement learning decision algorithm based on sub-task decomposition is proposed. The multi-UAV cooperative escort task is divided into two sub-tasks: reconnaissance and interference. The strategy learning is carried out respectively, and the two strategies are integrated according to the progress of the task to obtain the complete agent decision result. This paper designs different reward functions for two subtasks respectively, solves the task conflict problem in strategy learning, strengthens the targeted learning of subtasks, and verifies the effectiveness of the algorithm through simulation experiments.

(3) In the aforementioned multi-agent reinforcement learning decision algorithm based on subtask decomposition, the two sub-strategies make repeated decisions on flight actions, resulting in decision redundancy. Moreover, the flight actions in the integrated strategy result are biased towards one task objective, which is easy to cause the actions against the other task objective to be isolated or even ineffective. A multi-agent reinforcement learning decision algorithm based on action dependence is proposed, which specifically optimizes the neural network structure of the decision model, takes the decision result of the pre-action and the observation vector together as the decision input of the post-action, establishes the dependency relationship between the action and the action, does not isolate any kind of action, and reduces the difficulty of the agent decision. Finally, the simulation results show that the proposed algorithm can effectively reduce the action failure rate, and has better convergence characteristics and task evaluation results.

(4) In the aforementioned action-dependent multi-agent reinforcement learning

中国知网 https://www.cnki.net

decision algorithm, although the cooperation between actions is strengthened and the action dimension of decision-making in a single decision model is reduced by optimizing the network structure of the decision model, it is necessary to conduct in-depth analysis of the actual task and redesign the network structure according to the actual situation when facing different task scenarios. It is not universal, and the cooperation between agents has not received much attention. In order to solve the above problems, this paper proposes a multi-agent reinforcement learning sequential decision-making algorithm for complex action space, introduces the sequence model into intelligent decision-making, divides the decision-making process of each step into a decision sequence, and selects only one action of one agent for each decision. The input of the decision includes the current observation information and the decided action information, which strengthens the cooperation between actions while reducing the dimension of a single decision, and takes into account the extension of each time step decision series, the performance of the decision model is improved by optimizing the network model and optimizing the value estimation model. Finally, the effectiveness of the MA2DBT algorithm in the composite action space scene is verified by simulation experiments, and the contribution of each part of the algorithm to the whole is explored by ablation experiments.

Key Words: Electronic Warfare simulation, Intelligent Decision making, multi-agent reinforcement learning, hierarchical decision making, action dependence, sequential model

中国知网 https://www.cnki.net

目　录

第 1 章 绪论.....1
1.1 研究背景及意义.....1
1.2 电子对抗仿真概述.....3
1.2.1 电子对抗仿真.....3
1.2.2 电子对抗仿真技术的组成.....5
1.2.3 电子对抗仿真技术的应用.....6
1.3 电子对抗仿真中智能决策的研究现状.....8
1.3.1 智能决策系统概述.....8
1.3.2 电子对抗仿真中智能决策技术.....9
1.3.3 电子对抗智能决策问题研究现状.....10
1.4 论文主要工作及章节安排.....12
1.4.1 主要研究内容.....12
1.4.2 研究路线与章节安排.....14
第 2 章 深度强化学习相关理论基础.....16
2.1 引言.....16
2.2 深度强化学习理论.....16
2.2.1 概述.....16
2.2.2 基于值函数的方法.....19
2.2.3 策略梯度方法.....23
2.2.4 演员-评论家方法.....25
2.2.5 分层强化学习.....27
2.3 多智能体强化学习理论.....29
2.3.1 概述.....29
2.3.2 多智能体近端策略优化算法.....29
2.3.3 多智能体确定性深度策略梯度算法.....31
2.4 强化学习与序列模型.....32
2.4.1 序列模型.....32
2.4.2 多智能体 Transformer.....34
2.5 本章小结.....35
第 3 章 典型电子对抗任务系统模型设计.....37
3.1 引言.....37
3.2 典型电子对抗任务系统模型设计.....37
3.2.1 侦察任务.....38
3.2.2 干扰任务.....39

中国知网 https://www.cnki.net

3.2.3 动作、观测及奖励.....40
3.3 任务流程介绍.....43
3.4 本章小节.....44
第 4 章 基于子任务分解的多智能体强化学习决策方法.....45
4.1 引言.....45
4.2 子任务分解及综合决策.....45
4.2.1 子任务与动作的对应关系.....45
4.2.2 侦察子任务.....46
4.2.3 干扰子任务.....48
4.2.4 综合决策.....50
4.2.5 算法流程.....51
4.3 实验结果与分析.....54
4.3.1 侦察任务比较实验.....54
4.3.2 干扰任务比较实验.....56
4.3.3 完整电子对抗任务比较实验.....57
4.4 本章小结.....59
第 5 章 基于动作依赖的多智能体强化学习决策算法.....61
5.1 引言.....61
5.2 基于动作依赖的多智能体强化学习决策算法.....61
5.2.1 多种类动作间的决策依赖性分析.....61
5.2.2 双向动作依赖.....62
5.2.3 网络结构.....63
5.2.4 算法流程.....65
5.3 实验结果与分析.....67
5.3.1 算法比较实验.....67
5.3.2 动作决策顺序实验.....69
5.4 本章小结.....70
第 6 章 面向复合离散动作空间的多智能体序列决策算法.....71
6.1 引言.....71
6.2 面向复合离散动作空间的强化学习序列决策算法.....71
6.2.1 分解高维动作空间.....71
6.2.2 序列模型.....72
6.2.3 动作价值函数估计优化.....75
6.2.4 算法流程.....77
6.3 实验结果与分析.....79
6.3.1 单智能体任务决策比较实验.....79
6.3.2 多智能体任务决策比较实验.....81
6.3.3 决策顺序实验.....82

中国知网 https://www.cnki.net

6.3.4 算法消融实验.....83
6.4 本章小结.....85
第 7 章 总结与展望.....87
7.1 总结.....87
7.2 创新性工作.....88
7.3 工作展望.....89
参考文献.....90

中国知网 https://www.cnki.net

图目录

图 1-1 典型电子对抗场景示意图.....2
图 1-2 研究路线图.....14
图 2-1 强化学习流程示意图.....16
图 2-2 游戏场景示意图.....18
图 2-3 价值函数的回溯图.....21
图 2-4 A3C 异步训练框架图.....26
图 2-5 H-DQN 智能体分层架构.....28
图 2-6 MADDPG 架构图.....32
图 2-7 RNN 结构示意图.....33
图 2-8 LSTM 结构示意图.....33
图 2-9 MAT 网络架构.....35
图 3-1 交叉定位示意图.....39
图 3-2 多无人机护航任务示意图.....44
图 4-1 侦察距离示意图.....48
图 4-2 红方可干扰范围示意图.....50
图 4-3 基于子任务分解的多智能体强化学习决策流程.....51
图 4-4 TDPA 算法的基础算法框架.....52
图 4-5 TDPA 算法中侦察子策略的收敛特性.....55
图 4-6 TDPA 算法中干扰子策略的收敛特性.....56
图 4-7 TDPA 算法综合策略的收敛特性.....58
图 4-8 TDPA 算法与现有算法的累计奖励曲线比较.....58
图 5-1 无人机决策场景示例.....62
图 5-2 原始 MMDP 和转换后 SE-MDP 之间的比较.....63
图 5-3 决策模型观测-动作信息表征示意图.....64
图 5-4 基于动作依赖的多智能体强化学习算法网络结构.....65
图 5-5 基于动作依赖的多智能体强化学习算法收敛性比较.....68

中国知网 https://www.cnki.net

图 5-6 动作决策顺序实验结果.....69
图 6-1 MA2DBT 算法的网络结构.....74
图 6-2 序列决策模型状态、动作嵌入网络示意图.....75
图 6-3 不同算法在 academy_empty_goal_close 场景中的收敛特性...79
图 6-4 MA2DBT 算法在同一场景下不同决策复杂度下的收敛特性 80
图 6-5 MAT 和 MAPPO 算法在同一场景下不同决策复杂度下的收敛特性.....80
图 6-6 不同复杂度任务下的算法收敛特性.....81
图 6-7 不同序列决策顺序的收敛特性.....83
图 6-8 不同基线收敛特性的比较.....84

表目录

表 2-1 DQN 算法的伪代码.....23
表 2-2 REINFORCE 算法伪代码.....24
表 2-3 A3C 算法伪代码.....26
表 4-1 TDPA 算法伪代码（子策略）.....53
表 4-2 TDPA 算法（综合策略）.....53
表 4-3 侦察子任务下不同策略性能比较.....55
表 4-4 干扰子任务下不同策略性能比较.....57
表 4-5 完整任务下不同策略性能比较.....59
表 5-1 基于动作依赖的多智能体强化学习算法伪代码.....66
表 5-2 AD-PPO 算法与不同算法策略性能的比较.....68
表 6-1 MA2DBT 算法的伪代码.....78
表 6-2 决策模型任务执行情况.....84

中国知网 https://www.cnki.net

第1章 绪论

1.1 研究背景及意义

未来战争将是在陆、海、空、天、网络和电磁空间开展的全域作战。电磁空间是所有作战实体都要共享的物理空间，电磁空间的对抗将成为战争的焦点、决定战争的走向。电子对抗（又称电子战）是剥夺敌方使用电磁频谱，同时保护己方使用电磁频谱。从电子战最初诞生一直到 21 世纪，电子战一直呈指数增长，且其在现代战争中所扮演的角色越来越重要[1]。根据俄罗斯官方发布的《军事百科辞典》的定义，“电子战指的是利用电子手段来攻击敌其他资产以改变作战环境状态的过程。电子战的目标是降低敌方部队的作战效能（包括指控能力及武器系统应用能力）。电子战的一般流程为，首先对战场进行智能感知，然后根据感知结果确定需要干扰的目标，接着对现有的干扰资源进行任务分配，并进行干扰决策压制敌方，同样的受干扰的一方也可能执行抗干扰策略，然后继续进行感知，形成完整的“感知环境(observe)－适应环境(orient)－做出决策(decide)－采取行动(act)”(OODA)环路。传统电子战系统主要基于人工经验知识，缺乏足够的自学习能力，随着战场电磁环境的日趋复杂和对抗目标抗干扰能力的日渐增强，传统电子战对新型电磁目标、未知目标难以准确、快速识别，干扰措施难以快速生成的问题制约着电子战能力的提升。随着信息技术、人工智能技术的快速发展，未来战场上包括雷达和通信等军事信息系统逐渐呈现一体化、数字化、网络化、智能化的特点，无人化作战逐渐成为电子战的重要作战样式。未来军事对抗具有环境高复杂、信息不完整、博弈对抗强、响应高实时、自主无人化的突出特征，军事作战任务规划、无人系统集群自主对抗和作战仿真推演等领域都迫切需要人工智能技术。

近年来，以深度学习和强化学习为代表的人工智能技术取得了巨大突破，军事智能化成为了人工智能的重要应用方向。多智能体强化学习是当前机器学习研究领域的前沿技术，采用深度学习与强化学习相结合的方法研究多智能体间的完全合作任务，为解决装备协同对抗问题提供了有效途径。多智能体系统是由多个智能体组成的系统，每个智能体都具有自主决策能力。在多智能体系统中，智能体之间可以相互交互、通信和协作，以实现共同的目标。多智能体系统广泛应用于无人机协同控制、自动驾驶、机器人协作等领域。基于深度强化学习的多智能体决策算法旨在解决多智能体系统中的协同决策问题。该算法通过训练智能体的策略网络，使得智能体能够根据环境状态选择最优的行动。在训练过程中，智能体通过与环境的交互来收集数据，并使用深度神经网络进行策略的更新和优化。

中国知网 https://www.cnki.net

随着人工智能向战争的全要素渗透、与作战的全过程融合，现代战争形态正在悄然发生改变，智能化战争初现端倪，情报侦察、作战规划、指挥控制等作战过程呈现崭新的趋势。战斗机器人、自主飞行器和无人潜艇等智能化装备将日益成为实战的重要组成部分。这些装备拥有自主任务执行能力，减少了对人力资源的依赖。自主飞行器可以提供空中巡视和作战支援，而自主潜艇则能够进行水下侦察和攻击。这些人工智能装备的到来将大大提高作战效率，也使得军事行动更加隐蔽灵活。在人工智能赋能下，电子战系统需要能够自适应地感知战场的瞬息万变，并能够在极短的时间内自适应生成对抗策略，并自主评估对抗效果，进一步根据评估结果调整下一步对抗策略。智能决策模型可以为作战决策提供高效、精准的支持，为指挥员提供科学、有效、及时的决策依据。除了在战略层面上的决策能力，人工智能还在战术执行中发挥重要作用。在现代战争中，战场情况瞬息万变，需要快速做出反应和决策。人工智能可以通过感知技术和自主学习，自动化地识别和分析战场情况，并提供军事指挥官所需的多种决策选择，有效提高了军事行动的效率，降低了安全风险。

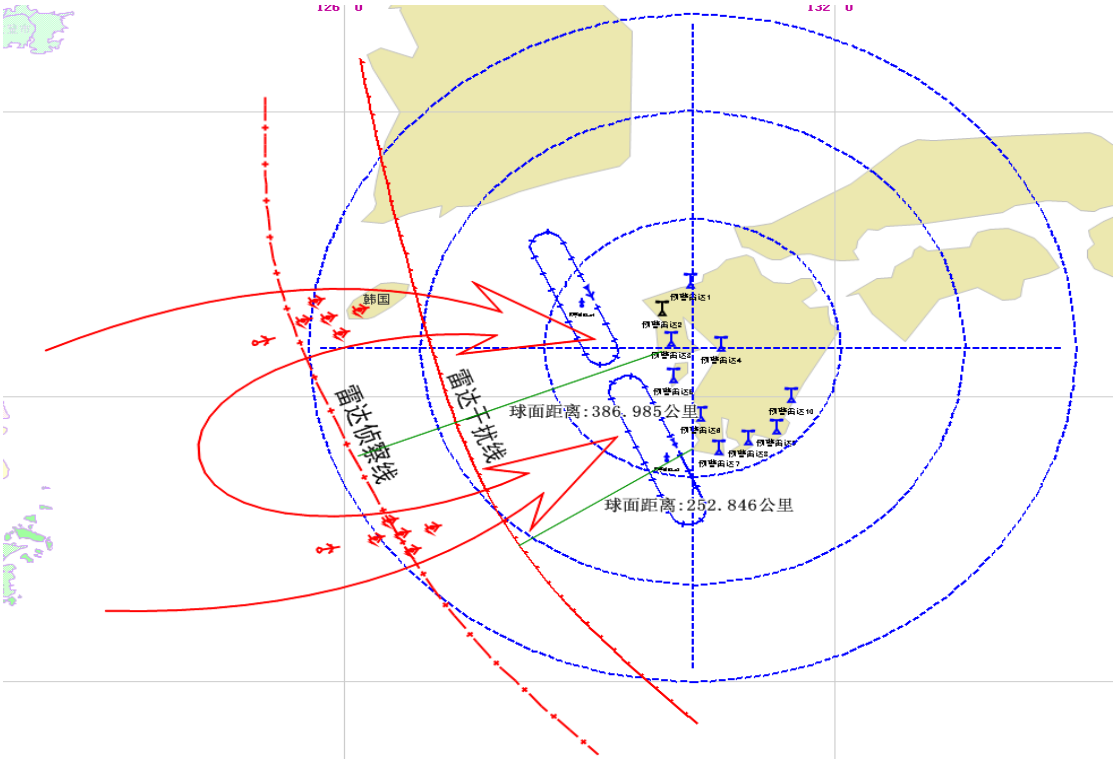


图 1-1 典型电子对抗场景示意图

Figure 1-1 Schematic diagram of typical electronic countermeasures scenario

多智能体强化学习技术在无人系统集群自主对抗、军事作战任务规划和仿真推演等领域有广阔的应用前景。在无人机群自主对抗领域，采用多智能体强化学习智能决策技术可以为无人系统集群中的每个智能体即时生成对抗行动，协调指挥智能体集群的攻击、防御等作战行为，逐步实现自主作战。在军事作

中国知网 https://www.cnki.net

战任务规划领域，针对任务规划面临的信息不完全、随机性大、策略复杂等挑战，多智能体强化学习智能决策技术可以探索强对抗环境下，基于不完全信息任务规划问题的有效求解方法，为作战任务规划提供支撑，为指挥员作战方案的制定提供辅助。在作战仿真推演系统、兵棋推演系统等仿真推演领域，采用多智能体强化学习智能决策技术，可以作为敌方与我方进行模拟对抗，避免了由我方指挥员担任敌方指挥员所带来的思维定式、与敌方指挥员风格不符等问题。通过多智能体决策模型控制红方，与环境进行交互，学习红方策略，支撑红方指挥决策系统的构建。

现阶段，电子战策略研究主要集中在智能感知、干扰资源分配、认知干扰决策以及抗干扰决策等子领域，但针对于完整电子战 OODA 环路的复杂任务决策研究较少。

本文研究的电子对抗多智能体强化学习算法，以典型电子对抗场景为基础，构建包含协同控制、协同侦察、协同干扰的完整 OODA 复杂任务决策模型，研究基于子任务分解、基于动作依赖、面向复合离散动作空间的多智能体强化学习决策算法，并进行基于仿真场景的算法验证。

1.2 电子对抗仿真概述

1.2.1 电子对抗仿真

电子对抗（Electronic Warfare, EW）是现代军事冲突中的关键领域，它涉及利用电磁频谱进行攻防行动，以获取战场优势。在现代战争中，电子系统如雷达、通信、导航和武器控制系统扮演着至关重要的角色。这些系统的有效运作对于情报收集、指挥控制、目标定位和精确打击至关重要。电子对抗的目的是通过干扰、欺骗、破坏或压制敌方的电子系统，同时保护自己的系统免受敌方的类似攻击，从而在信息战中占据上风[2]。

电子对抗的重要性不仅体现在军事领域，它还对国家安全、经济稳定和公共安全有着深远的影响。例如，民用通信系统、电力网和交通控制系统等关键基础设施的电子系统，如果遭受电子攻击，可能会导致严重的社会经济后果。电子对抗技术的发展和应用，对于维护国家安全和公共秩序具有不可替代的作用。

随着电子技术的不断进步，电子对抗的手段和策略也在不断演变。传统的电子对抗方法，如电子干扰和反辐射导弹，虽然仍然有效，但现代战场的复杂性和动态性要求更为先进和灵活的电子对抗手段。这就催生了电子对抗仿真技术的发展。

电子对抗仿真技术提供了一个虚拟的实验平台，允许军事和研究人员在不实际部署武器系统的情况下，测试和评估电子对抗策略和设备的有效性[3]。通

中国知网 https://www.cnki.net

过仿真，可以模拟各种复杂的电磁环境和对抗场景，从而预测和评估不同电子对抗行动的可能结果。这种技术的应用，不仅可以提高电子对抗训练的质量和效率，还可以在新系统开发和战术研究中发挥关键作用。

除此以外，电子对抗仿真技术同时节约资源与降低风险。实际的电子对抗行动可能涉及昂贵的设备和资源，且存在一定的风险。仿真技术使得这些活动可以在计算机上进行，大大降低了成本和风险。仿真技术还允许进行更多的实验和测试，因为可以重复模拟相同的场景，或者探索不同的策略组合，以找到最有效的电子对抗方法。

电子对抗仿真技术是现代军事战术训练和装备研发的重要工具，它通过模拟电子对抗环境中的各种电磁行为，为军事指挥官和工程师提供了一个无风险的实验平台。这种技术的发展经历了从简单的模拟到高度复杂的数字仿真的演变，反映了电子对抗领域对精确模拟和分析需求的增长。

电子对抗仿真技术的早期发展可以追溯到 20 世纪中叶，当时的主要目标是模拟雷达系统和通信干扰。最初的仿真系统基于简单的数学模型和模拟电路，这些系统能够提供基本的电磁环境模拟和对抗效果评估。在这一时期，仿真技术主要用于军事训练和系统性能评估，帮助军事人员理解电子对抗的基本原理和操作。随着计算机技术的进步，仿真模型开始变得更加复杂，能够模拟更多的电子对抗设备和战术行为[4]。

现代电子对抗仿真技术已经发展成为一个高度复杂和多学科交叉的领域。当前的仿真系统不仅能够模拟各种电子对抗设备，如雷达、干扰器、电子支援措施（ESM）和反辐射武器（ARM），还能够模拟复杂的电磁环境，包括地形、大气条件、电磁干扰和敌方电子对抗行动。这些系统通常基于高性能计算机和先进的软件平台，能够提供实时或近实时的仿真能力。此外，现代仿真技术还集成了人工智能和机器学习算法，以提高仿真的自适应性和决策支持能力。在军事训练、系统测试、战术开发和科学研究中，电子对抗仿真技术都发挥着不可或缺的作用。

电子对抗仿真技术的未来发展趋势预示着更高的仿真精度、更强的交互性和更广泛的应用范围。随着计算能力的进一步提升，仿真系统将能够处理更加复杂的场景和更大规模的仿真任务。云计算和大数据分析的应用将使得仿真数据的处理和分析更加高效，同时也能够支持更大规模的分布式仿真环境。人工智能和机器学习技术的进一步集成将使仿真系统能够自动适应新的战术和策略，提供更加精确的预测和决策支持。此外，虚拟现实（VR）和增强现实（AR）技术的发展将为电子对抗仿真带来更加沉浸式的用户体验，提高训练的现实感和效果。未来，电子对抗仿真技术将继续在军事和民用领域发挥重要作用，成为电子对抗能力发展的关键驱动力。

中国知网 https://www.cnki.net

1.2.2 电子对抗仿真技术的组成

电子对抗仿真的精确性和可靠性依赖于其核心组成部分：仿真模型、仿真环境以及仿真控制与管理。这些组成部分共同构成了一个多维度的仿真系统，能够复现真实世界的电磁战场环境，从而支持军事决策者和工程师在虚拟空间中进行深入分析和策略制定[5]。

仿真模型构成了电子对抗仿真系统的核心，它负责精确模拟电子对抗设备和电磁环境的行为。模型的复杂性跨度广泛，从基础的线性系统到高度非线性和动态的复杂系统，其准确性直接关系到仿真结果的有效性和可靠性。在电子对抗设备仿真方面，模型需详尽地模拟雷达、干扰器、通信系统等的功能和性能，包括但不限于设备的发射接收特性、信号处理算法、抗干扰能力以及系统间的交互方式。例如，雷达仿真模型需精确模拟波束形成、扫描模式、目标检测与跟踪算法；干扰器模型则需模拟其干扰信号的特性，如频率、功率、调制方式及其对特定雷达系统的干扰效果。电磁环境的模拟则涉及对电磁波在不同介质中传播特性的理解，包括大气、地形、建筑物等因素对信号的影响，以及不同电子设备在同一环境中工作时可能产生的电磁兼容性问题。

为了提升仿真模型的准确性，研究人员通常会利用实际设备的测试数据来校准模型参数，使模型更真实地反映设备的实际行为。同时，仿真模型的持续更新也是必要的，以适应新型电子对抗设备和战术的发展。

仿真环境为电子对抗仿真提供了必要的背景信息，涵盖了地形、天气、电磁干扰等多种因素。这些因素对电子对抗行动有着直接的影响，必须在仿真中得到准确的再现。地形因素，如山脉、平原、海洋、城市等，对电磁波传播的影响各不相同，例如山脉可能阻挡或反射电磁波，城市环境可能导致多径效应。天气条件，如雨、雾、云层等，也会对电磁波的传播特性产生影响。电磁干扰（EMI）是仿真环境中的另一个关键因素，仿真环境需模拟敌方电子对抗行动或友方设备无意干扰的特性及其对电子对抗设备性能的影响。为了实现这些环境因素的准确模拟，仿真系统通常需要集成地理信息系统（GIS）数据、气象数据和电磁兼容性分析工具，以创建一个真实且动态变化的电磁环境。

仿真控制与管理是电子对抗仿真系统中的关键组成部分，负责协调和监督仿真活动的进行。它确保仿真按照预定脚本进行，同时允许用户根据需要进行实时干预和调整。仿真控制包括启动仿真、监控仿真进度、调整仿真参数以及终止仿真，这些功能对于确保仿真活动的顺利进行至关重要。仿真管理则涉及仿真资源的分配、仿真数据的记录和仿真结果的分析，包括对仿真过程中产生的大量数据进行有效管理，确保数据的完整性和可用性。此外，仿真管理还涉及仿真结果的后处理，如数据分析、可视化和报告生成，这些对于从仿真活动中提取有价值的信息至关重要。现代仿真系统通常采用用户友好的界面和自动

中国知网 https://www.cnki.net

化工具来提高仿真控制与管理的效率，支持多用户协作，以便在团队环境中共享仿真资源和结果。随着技术的发展，仿真控制与管理工具将变得更加智能和自动化，以适应不断变化的仿真需求。

1.2.3 电子对抗仿真技术的应用

1.2.3.1 军事训练

电子对抗仿真技术在军事训练中的应用是其最直接和广泛的用途之一。其为军事人员提供了一个安全、可控且成本效益高的培训环境，使得他们能够在不实际部署的情况下，模拟和练习各种电子对抗技能和战术。在这种环境中，军事人员可以面对各种复杂的电磁威胁，学习如何操作电子对抗设备，执行干扰任务，以及进行有效的电子对抗与攻击。

仿真技术使得军事人员能够在不同的战场条件下，如城市环境、山区、海洋等，进行针对性的训练。这些环境的复杂性要求军事人员具备高度的适应性和决策能力。通过仿真，军事人员可以在没有实际风险的情况下，反复练习和完善这些技能。此外仿真技术还可以模拟敌方的电子对抗行动，为军事人员提供对抗真实对手的经验。在战术知识方面，仿真技术可以帮助军事人员理解电子对抗的基本原理，如信号传播、干扰原理、电子对抗设备的工作原理等。通过模拟不同的战术场景，军事人员可以学习如何在特定的电磁环境中制定和执行战术计划。这种训练不仅提高了他们的操作技能，也增强了他们的战术思维和决策能力。

军事训练中的仿真技术还涉及到团队协作和指挥控制的训练。在现代战争中，电子对抗往往需要多部门、多单位的协同作战。仿真技术可以模拟这种多单位的作战环境，训练军事人员如何在复杂的指挥体系中有效地沟通和协作。这种训练对于提高整体作战效能至关重要。随着技术的发展，虚拟现实和增强现实技术的应用为军事训练带来了新的机遇。这些技术可以提供更加沉浸式的仿真环境，使军事人员能够以更加直观和真实的方式进行训练。

电子对抗仿真技术在军事训练中的应用极大地提高了训练的效率和质量。它不仅为军事人员提供了一个无风险的训练环境，还使得他们能够在复杂的电磁环境中提高自己的技能和战术知识。随着仿真技术的不断进步，未来的军事训练将更加高效、真实和全面。

1.2.3.2 系统测试与评估

电子对抗仿真技术在系统测试与评估中扮演着至关重要的角色，它是确保新电子对抗系统达到预期性能的关键步骤。在系统部署前，通过高度逼真的仿真环境进行全面的测试和评估，可以验证系统是否能够适应多变的战场环境并满足预定的性能标准。这种仿真测试不仅涵盖了系统功能的验证，还包括了对

中国知网 https://www.cnki.net

其在多样化战场条件下的效能进行深入分析。

在功能测试方面，仿真技术能够模拟复杂的电磁环境，包括敌方的干扰信号，以检验电子对抗设备的信号处理能力、抗干扰性能、目标检测和跟踪精度等关键性能指标。例如，西安电子科技大学电子信息对抗与仿真技术教育部重点实验室的研究方向之一就是复杂环境中的异类多源数据仿真技术，这表明了仿真技术在模拟真实战场条件下的电子对抗系统性能方面的重要性。

在性能评估方面，仿真技术能够模拟不同的地形、天气和电磁环境，以测试电子对抗系统在这些条件下的适应性和稳定性。例如，山区的多径效应和城市环境中的建筑物可能会对雷达和通信系统产生影响，仿真技术可以在这些复杂环境中对系统性能进行全面测试。

可靠性和稳定性测试也是系统测试与评估的重要组成部分。通过在仿真环境中对系统进行长时间的连续运行测试，可以评估系统在持续工作状态下的表现，并模拟各种故障情况，如设备故障、软件错误等，以测试系统的容错能力和恢复机制。

此外，仿真技术在系统测试与评估过程中还能帮助发现和解决潜在的设计缺陷。通过模拟不同的作战场景，可以在系统投入使用前识别并改进可能存在的问题，从而提高系统的作战效能和降低后期维护成本。

随着仿真技术的不断进步，系统测试与评估的自动化和智能化水平也在不断提升。自动化测试工具减少了人工操作的需求，提高了测试效率；智能化评估系统能够自动分析测试数据，生成详细的性能报告。这些技术的发展使得系统测试与评估更加高效、准确，为电子对抗系统的性能保障提供了强有力的支持。

电子对抗仿真技术的应用不仅提高了测试的效率和准确性，而且在系统投入使用前，通过仿真测试可以发现和解决潜在问题，确保电子对抗系统在实际部署时能够发挥最大的效能。

1.2.3.3 战术研究与开发

电子对抗仿真技术在战术研究与开发中的应用是推动军事战术创新的重要力量。通过仿真技术，研究人员可以在虚拟环境中探索新的战术和策略，评估现有战术在新环境下的有效性，以及测试新战术对现有电子对抗系统的影响。

在战术研究方面，仿真技术提供了一个实验平台，使得研究人员可以在不受实际条件限制的情况下，自由地尝试和验证新的战术思想。这包括对电子对抗行动的策略、战术动作、作战流程等进行模拟和分析。例如，研究人员可以模拟在特定电磁环境下，不同的干扰策略对敌方雷达系统的影响，从而优化干扰行动的执行。

在战术开发方面，仿真技术可以帮助研究人员设计和测试新的电子对抗设

中国知网 https://www.cnki.net

备和系统。这包括对新设备的信号特性、干扰效果、抗干扰能力等进行评估。通过仿真，可以在设备实际制造和部署之前，对其性能进行预测和优化，从而降低研发风险和成本。

仿真技术还可以用于评估现有战术在新环境下的有效性。随着敌方电子对抗能力的不断进步，现有的战术可能需要调整以适应新的威胁。通过仿真，研究人员可以在不同的电磁环境中测试现有战术的效果，从而确定是否需要进行战术调整。

此外，仿真技术在战术研究与开发中的应用还包括对电子对抗行动的决策支持。在复杂的电子对抗环境中，指挥官需要做出快速而准确的决策。仿真技术可以模拟不同的决策路径，预测其可能的结果，从而为指挥官提供决策支持。

随着人工智能和机器学习技术的发展，仿真技术在战术研究与开发中的应用将更加智能化。这些技术可以帮助研究人员分析大量的仿真数据，发现新的战术模式和策略。同时它们还可以自动调整仿真参数，以适应不断变化的战术需求。

电子对抗仿真技术在战术研究与开发中的应用对于推动军事战术创新具有重要意义。它不仅提供了一个自由探索新战术的平台，还有助于优化现有战术，提高电子对抗行动的决策质量。

1.3 电子对抗仿真中智能决策的研究现状

1.3.1 智能决策系统概述

智能决策支持系统（IDSS）是一种集成了人工智能技术的计算机辅助工具，旨在辅助决策者在复杂和不确定的环境中做出更加合理和高效的决策。IDSS 的核心功能是模拟人类决策者的思考过程，通过分析和处理大量信息，提供决策建议。这些系统通常包含知识库、推理引擎、用户界面和数据库等关键组件，能够处理包括战略规划、资源分配、风险评估等复杂的决策问题[6]。IDSS 的设计使得决策者能够更好地理解问题，评估不同的解决方案，并选择最佳行动方案。

IDSS 的概念最早在上个世纪 70 年代提出，随着计算机技术和人工智能的快速发展，IDSS 经历了几个阶段的发展。在早期阶段，IDSS 主要基于规则的专家系统，依赖于预设的规则和知识库进行决策[7]。这些早期系统虽然在特定领域取得了成功，但它们的应用范围受限于知识库的规模和复杂性。进入发展阶段，随着机器学习和数据挖掘技术的进步，IDSS 开始整合这些技术，提高了系统的自适应性和学习能力[8]。这一时期的 IDSS 能够处理更复杂的数据，并在一定程度上模拟人类的学习过程。到了成熟阶段，深度学习、大数据分析和云计算等技术的发展，使得 IDSS 能够处理更大规模的数据，提供更精准的决策支持[9]。

中国知网 https://www.cnki.net

这些技术的进步极大地扩展了 IDSS 的应用范围和能力。

现代 IDSS 采用的技术包括机器学习、数据挖掘、自然语言处理（NLP）、深度学习和云计算等。机器学习技术，尤其是监督学习和无监督学习，使得 IDSS 能够从数据中学习和预测，如支持向量机（SVM）和随机森林等算法在分类和回归问题中的应用[10]。数据挖掘技术用于发现数据中的模式和关联，如聚类分析和关联规则挖掘。NLP 技术使系统能够理解和生成自然语言，提高用户交互的自然性，例如 BERT 和 GPT 等预训练语言模型在自然语言理解方面的应用[11]。深度学习技术通过神经网络模拟人脑处理信息的方式，处理复杂的数据模式，云计算技术提供强大的计算资源和数据存储能力，支持大规模的 IDSS。

IDSS 可以根据其功能和应用领域分为不同的类别。规则型 IDSS 依赖于预设规则进行决策，如早期的 DENDRAL 系统。模型驱动型 IDSS 使用数学模型和算法来模拟和预测决策结果，如经济模型在金融决策中的应用。知识驱动型 IDSS 结合专家知识和数据驱动的方法，提供决策支持，如医疗诊断系统。混合型 IDSS 结合以上多种技术，以适应更复杂的决策环境，如军事指挥控制系统。

尽管 IDSS 具有许多优点，如提高决策质量、效率、可解释性和适应性，但它们也存在一些缺点。决策的准确性高度依赖于输入数据的质量和完整性，系统复杂性随着功能的增加而增加，过度依赖 IDSS 可能导致决策者忽视直觉和经验。

1.3.2 电子对抗仿真中智能决策技术

在电子对抗仿真领域，智能决策技术的发展正推动着战术决策的革新。这些技术模仿人类决策过程，使系统能够在复杂和动态的电子战环境中迅速做出有效反应。李理等人利用深度强化学习方法多智能体深度确定性策略梯度（MADDPG）算法，构建了一个多智能体决策模型[12]。该模型在国产硬件和操作系统环境下进行仿真，有效地支持了地空协同作战决策，展示了深度强化学习在多智能体协作问题上的应用潜力。章胜等人的研究进一步将深度强化学习应用于双机近距空战机动智能决策模型，并通过软硬件实现与人机对抗飞行试验[13]。这一成果不仅验证了智能决策模型的有效性，也证明了深度神经网络在空战决策中的有效性。然而智能决策技术在电子对抗仿真中的应用也面临着挑战。首先，强化学习算法通常需要大量的训练数据，这在实际应用中可能难以获得。周攀等人的研究中提到，尽管在无人机近距空战格斗自主决策模型中取得了进展，但数据稀缺问题仍然是一个难题[14]。其次深度学习模型往往需要大量的计算资源，可能限制了其在资源受限环境下的应用。此外智能决策模型在特定环境下表现良好，但在新环境中的泛化能力有待提高，Chebotar 等指出强化学习在许多场合下并不能解决实际问题[15]。

尽管存在这些挑战，强化学习在电子对抗仿真中的优势仍然明显。它能够

中国知网 https://www.cnki.net

在没有明确模型的情况下学习最优策略，适用于动态和不确定的环境，能够处理复杂的决策问题。深度强化学习被用于无人机集群追击任务[16]，以及基于启发式强化学习的空战机动智能决策[17]。这些研究表明，强化学习在电子对抗仿真中的应用前景广阔，尤其是在提高决策效率和适应复杂环境方面。

未来的研究需要在数据依赖、计算资源和泛化能力等方面取得突破，以实现智能决策技术在电子对抗领域的广泛应用。同时多智能体协作策略的设计、信用分配问题以及实时性要求等也是亟待解决的关键问题。随着算法的不断进步，强化学习作为一种在智能决策中展现出巨大潜力的技术，面临着多方面的挑战。

首先数据依赖性问题使得在安全关键领域如军事和航空航天中难以获得足够的交互数据。为了解决这一问题，研究者们正在探索无监督或半监督学习方法，并开发模拟环境来生成训练数据。其次深度强化学习模型对计算资源的需求很高，这限制了它们在资源受限环境中的应用。为此研究者们正在开发更高效的算法和模型压缩技术，并利用云计算和边缘计算资源。

泛化能力是另一个关键挑战，因为智能体在特定环境中训练得到的策略可能无法适应新环境。研究跨任务学习和元学习方法，以及提高智能体的泛化能力，是解决这一问题的方向。探索与利用的平衡也是实际应用中的一个难题，需要设计更智能的探索策略。在多智能体环境中，如何设计有效的协作策略和处理竞争关系也是一个复杂问题，需要研究多智能体系统的理论基础和开发分布式强化学习算法。实时性和延迟奖励问题在需要快速响应的应用中尤为重要，优化算法和研究基于预测的决策方法是解决这一问题的途径。

最后，可解释性和透明度对于提高用户对智能决策的信任至关重要。开发可解释的人工智能技术和提高模型透明度是当前的研究重点。解决这些挑战将有助于强化学习在智能决策领域的进一步发展，并在更广泛的应用场景中发挥其潜力。

1.3.3 电子对抗智能决策问题研究现状

电子战中对抗双方博弈斗争是一个相互识别、相互躲避的动态过程，电子战装备只有具备“边对抗边学习”的能力，通过对手反馈状态的辨识及时调整己方的应对策略，才能掌握未来电子战中的主动权。为此，国内外提出了具有学习能力的认知电子战系统的概念[18]。

认知电子战的定义是，以具备认知能力的电子战装备为基础，注重运用自主交互式电磁环境学习能力与动态智能化对抗能力的电子战作战行动。主要包含认知侦察、认知干扰以及认知防御三个基本要素[19]。

目前，研究人员的电子战策略研究重点分为智能电磁感知任务、干扰资源分配任务、认知干扰决策任务以及认知抗干扰决策任务。智能感知任务的研究

中国知网 https://www.cnki.net

关注于感知系统的设计[20,21]、针对弱信号的侦察[22]以及多侦察单位协同搜索感知[23–25]等；后面的三者则更偏向于决策任务，电子战中的干扰资源分配任务是如何在有限的干扰资源总量条件下获得整体上的最佳电子攻击效果，是典型的组合优化问题[26–32]；认知干扰决策是电子干扰的智能化发展，以保证己方单位免受敌方多功能雷达的探测和攻击[33–40]，同样也存在多个平台之间的协同干扰决策[41–43]；而认知抗干扰则是相对应的智能化防御策略，意图通过自主决策的方法使得雷达具备自适应的抗干扰能力[44–46]，也有无线通信系统上的抗干扰[47–53]，同时也有关于雷达抗干扰性能评估模型的研究[54–56]。

电子战的一般流程为，首先对战场进行智能感知，然后根据感知结果确定需要干扰的目标，接着对现有的干扰资源进行任务分配，并进行干扰决策压制敌方，同样的受干扰的一方也可能执行抗干扰策略，然后继续进行感知，形成完整的 OODA 环路。现阶段的研究集中在上述的几个子领域中，而缺乏对把各个部分整合起来的完整电子战过程进行探索研究。

本文所关注的多无人机电子对抗场景模型，包含多无人机协同控制任务、协同侦察任务、协同干扰任务，是一个完整的电子战任务场景。多无人机作战是未来战争的重要形式[57]，控制策略的优劣直接影响着无人机集群完成任务的安全、稳定等效能[58]。自然界中的群体智能是算法研究的启发之一，Duan 等研究狼群智能行为机理，应用于无人机集群协同决策问题[59]；Shen 等受鸟类行为启发设计了分层的集群控制框架[60]；Gao 等提出了一种基于态势感知共识的无人机群分布式协作方法及其信息处理机制，加强对复杂环境的适应性[61]。面对协同搜索问题，Zhang 等采用给改进的粒子群算法分配无人机侦察区域，以最大化无人机集群效用[62]；Yue 等提出了一种安全转移软 AC 算法，对最大化收益进行了安全约束[63]；Baek 等采用分布式算法解决无人机集群跟踪地面目标的任务，提出了最小化目标定位不确定性的最优传感器管理技术和以及共识决策算法[64]。Luiz 等利用行为树的方法实现了忠诚僚机的分散式决策[65]，多无人机协同任务需要面对复杂不确定战场环境，Jun 等提出了一种基于模糊认知图的多无人机协同攻击决策建模方法[66]，而 Chen 等针对不同机动能力无人机群体间的攻防对抗问题进行了研究[67]；Rahmes 等利用集群和协作波束成形技术，更有效地支持复杂的雷达干扰和欺骗任务[68]；Ma 等研究了超视距下的多无人机协同占用问题[69]；Xu 等提出了基于云信任模型的多无人机协同决策中心动态选择方法[70]；多无人机的协同作战[71,72]也是研究热点之一。

目前，多智能体强化学习在交通管控[73]、多环化学过程控制[74]、致病基因预测[75]、雷达波形设计[76]、军事作战[77]等复杂问题中进行了应用研究，是研究复杂任务决策的有力工具。Ibrahim H. Ahmed 等对智能体之间的有效交互系统进行了研究[78]；现实世界中获取的信息不一定是准确的，针对此问题，Chen 等提出了一种多智能体容错强化学习[79]；Riley 等提出了一种安全的多智能体强化学

中国知网 https://www.cnki.net

习方法，以便应用于安全关键或任务敏感场景[80]；Malysheva 等提出了 MAGNet 方法，在强化学习中引入了环境关联图的概念[81]；Liu 等提出用于多智能体强化学习的封建潜在空间探索，通过学习潜在结构指导多个智能体协调探索[82]；Kim 等针对懒惰智能体所带来的协作悖论问题进行了分析[83]；Kuba 等针对策略梯度方法的估计高方差问题进行了严谨的数学分析，推导出了实现最小方差的最优基线[84]；Chen 等引入了分层注意力图的概念，增强了智能体模型的可扩展性[85]；Jiang 等提出了一种去中心化探索与选择性记忆策略梯度算法，将记忆信息纳入了决策考虑[86]。

针对多智能体任务中智能体的高维动作空间导致策略收敛困难的问题，最直接的方法就是降低单次决策的维度，Alonso 等提出了一种骨架级的多智能体仿真控制方法来应对[87]；Selvakumar 等通过离散抽象的方式来降低动作空间维度[88]；Wang 等采用分布式决策的方法来降低联合决策空间[89]；此外，采用确定性策略进行学习也可以缓解策略难以收敛的问题[90–93]。

针对多智能体任务中的智能体协作问题，Shi 等提出了多智能体强化学习的知识重用方法，共享经验策略，避免知识的浪费[94]；Ma 等聚焦于顶层分配策略的学习，准确刻画作战单元之间的协同演化内因，有效地实现了大规模协同目标分配方案的动态生成[95]；Fukumoto 等通过将自私经验替换为合作经验，加强智能体之间协作的学习[96]；Ikeda 等针对通信延迟状况下的多智能体协作系统进行研究[97]；wang 等引入虚拟学习目标直观地进行协作的学习[98]；Sheikh 等提出了一种平衡个人奖励和团队奖励的多智能体协作框架[99]；Wang 等提出了一种基于长期行为分析的分层协同任务分配框架，旨在解决多智能体系统中完全协作任务的动态任务分配问题[100]。

Transformer 在自然语言处理、计算机视觉、音频处理等许多人工智能领域都取得了巨大的成功，是当前应用最广泛最流行的序列模型网络架构，同样也引起了强化学习领域学者的关注。Chen 等将抽象的强化学习描述为序列建模问题，将 Transformer 框架与离线强化学习相结合[101]；Zheng 等提出了基于序列建模的在线决策 Transformer 算法，将离线预训练和在线微调两种范式混合到统一框架中[102]，同样也适用于多智能体[103]；Wen 等基于 Transformer 框架提出了一种针对协作多智能体的训练框架[104]。

1.4 论文主要工作及章节安排

1.4.1 主要研究内容

本文系统性地针对典型电子对抗任务场景的智能决策算法开展研究，以多

中国知网 https://www.cnki.net

无人机协作护航任务为研究对象，主要进行了多无人机协作护航任务系统模型的设计、基于子任务分解的多智能体强化学习决策算法、基于动作依赖的多智能体强化学习决策算法、面向复合动作空间的多智能体强化学习序列决策算法这 4 个方面的研究工作，具体研究内容如下：

（1）多无人机协作护航任务的系统模型设计

针对典型电子对抗场景中的多无人机协作护航任务模型，分析多无人机协作护航任务的任务流程以及任务细节，以红蓝对抗为背景，设计多无人机协作护航任务仿真系统模型，将多无人机协作护航任务建模为部分可观测马尔可夫决策过程，以红方无人机作为智能体，设计了智能体的观测空间、动作空间，并且根据任务目标设计了任务环境的奖励函数。此外，对多无人机协作护航任务中实体的功能模型进行了构建，包括红方无人机侦察模型、蓝方雷达探测模型、蓝方雷达受干扰模型等。

（2）基于子任务分解的多智能体强化学习决策算法

针对多无人机协作护航任务中由于智能体无人机需要在多种类动作组成的高维动作空间中进行决策，而任务包含侦察、干扰等多个任务目标，不同任务目标对于多种类动作的选择有不同需求，容易产生互相干扰的情况，进而影响智能体的整体决策的问题，基于分层学习的方法，提出了基于子任务分解的多智能体强化学习决策算法。将任务分解为侦察和干扰两个子任务，分别进行策略学习，并根据任务的进展综合两个策略，实现完整的任务决策，解决了策略学习中的任务冲突问题，加强其对子任务的针对性学习。最后，设计了多组仿真实验验证算法的有效性。

（3）基于动作依赖的多智能体强化学习决策算法

在前述的基于子任务分解的多智能体强化学习决策算法中，两个子策略对飞行动作进行了重复决策，存在决策冗余，而且综合策略结果中的飞行动作对某一任务目标存在倾向性，容易造成针对另一任务目标的动作孤立甚至失效。针对该问题，提出了一种基于动作依赖的多智能体强化学习决策算法。对决策模型的神经网络结构进行了特定性的优化，将前置动作的决策结果与观测向量一起作为后置动作的决策输入，建立动作与动作之间的依赖关系，不孤立任何一类动作，同时降低智能体决策的难度。最后，通过仿真实验验证算法的适用性和合理性，并与多组算法进行比较。

（4）面向复合动作空间的多智能体强化学习序列决策算法

在前述的基于动作依赖的多智能体强化学习决策算法中，虽然加强了动作与动作之间的协作性，并通过决策模型网络结构的优化降低了决策的动作维度，但面对不同的任务场景时需要对任务实际进行深入分析并根据实际情况重新设计网络结构，不具备通用性，且没有重点关注智能体之间的协作性。针对上述问题，提出了一种面向复合动作空间的多智能体强化学习序列决策算法，将序

中国知网 https://www.cnki.net

列模型引入智能决策，每一步的决策过程拆分为一个决策序列，每次决策只选择一个智能体的一种动作，增强智能体之间、动作之间协作性的同时降低单次决策的维度，并通过优化网络模型、优化价值估计模型的方法提升决策模型的性能。最后，进行一系列比较实验验证了 MA2DBT 算法在复合动作空间场景下的有效性，并针对多无人机协作护航任务，进行消融比较实验，探索算法各部分对整体的贡献。

1.4.2 研究路线与章节安排

图 1-2 给出了本文的研究路线。

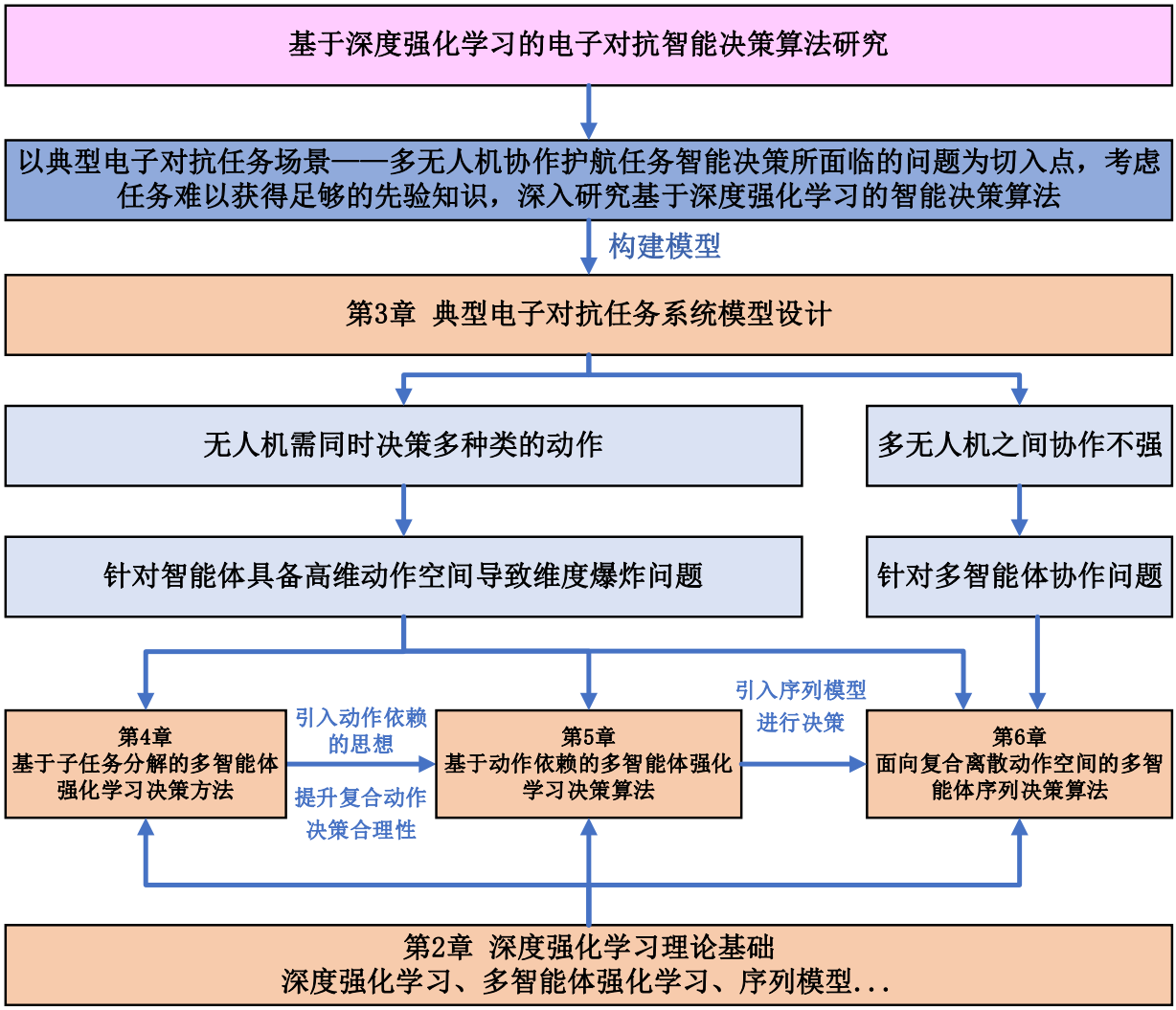


图 1-2 研究路线图

Figure 1-2 Research roadmap

根据上一节的研究内容，本文共分为七章，论文组织结构与框架如下：

第一章为绪论。阐述了本文的选题背景及意义，从电子对抗作战仿真、智能决策在电子对抗场景中的应用以及遇到的高维动作空间、多智能体协作等方面分析了国内外研究现状，总结当前研究存在的不足和发展趋势。最后介绍本

中国知网 https://www.cnki.net

文的主要研究内容及章节安排。

第二章为深度强化学习相关理论基础。描述了论文所涉及的深度强化学习、多智能体强化学习、以及强化学习与序列模型等相关基础知识。

第三章为典型电子对抗任务系统模型设计。以多无人机协作护航任务为例，分析其任务流程以及任务细节，以红蓝对抗为背景，将多无人机协作护航任务建模为部分可观测马尔可夫决策过程，以红方无人机作为智能体，设计多无人机协作护航任务仿真系统模型。

第四章为基于子任务分解的多智能体强化学习决策算法。分析了多无人机协作护航任务的任务特性，提出了基于子任务分解的多智能体强化学习决策算法，根据任务实际进行任务分解，子任务独立学习并综合决策，并设计仿真实验验证算法的有效性。

第五章为基于动作依赖的多智能体强化学习决策算法。分析了子任务分解综合决策的局限性和多种动作之间协作性不强的问题，提出了基于动作依赖的多智能体强化学习决策算法，通过单个智能体的多种动作先后决策的方式，将前置动作加入后置动作的决策输入中，形成动作之间的协作依赖。最后进行多组仿真实验验证该算法。

第六章为面向复合动作空间的多智能体强化学习序列决策算法。分析了依据任务实际进行决策网络优化的方式对于多种电子对抗场景适用性不强以及没有重点关注智能体之间协作的问题，提出面向复合动作空间的多智能体强化学习序列决策算法，将序列模型引入多智能体多维动作决策，并进行决策网络结构优化以及价值评估网络优化，最后以开源环境进行多组仿真比较实验，以多无人机协作护航任务为例进行消融实验。

第七章为全文总结与工作展望。

中国知网 https://www.cnki.net

第2章深度强化学习相关理论基础

2.1 引言

本文针对电子对抗仿真决策中遇到的高维动作空间以及多智能体协作的问题展开了研究，在问题的建模和求解过程中涉及了强化学习等相关数学基础知识。为了给后续章节的研究工作提供理论依据，本章对相关理论进行了简要介绍。首先概述了深度强化学习的相关理论，包括基于值函数的方法、策略梯度的方法、演员-评论家架构以及针对复杂任务的分层强化学习的介绍和相关典型算法。然后阐释了多智能体强化学习系统和两种经典的多智能体强化学习算法。最后阐述了序列模型和序列模型与强化学习结合的相关知识。

2.2 深度强化学习理论

2.2.1 概述

强化学习（reinforcement learning, RL）讨论的问题就是智能体如何在复杂、不确定的环境中最大化它能获得的奖励。强化学习是机器学习的一个分支，研究如何通过智能体与环境的交互，使其能够在动态环境中学习最优行为策略。强化学习的目标是通过试错和反馈来使智能体逐渐改善其行为，以获得最大化的累积奖励。在强化学习中，智能体通过观察环境的状态选择动作，并与环境进行交互。环境根据智能体的行动给予反馈，即奖励或惩罚，以指导智能体的学习。智能体的目标是通过尝试不同的行动策略，并根据奖励信号来调整自己的策略，以最大化其所获得的长期累积奖励。强化学习的核心思想是基于试错学习的智能体行为，与传统的监督学习和无监督学习有所不同。强化学习注重在与环境的交互中通过奖励和惩罚来引导智能体的学习，具有较强的探索性和适应性。

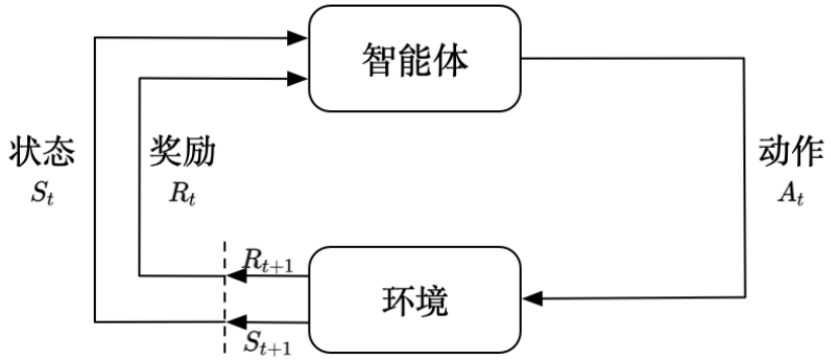


图 2-1 强化学习流程示意图

Figure 2-1 Schematic diagram of reinforcement learning processes

中国知网 https://www.cnki.net

强化学习的数学基础和建模工具是马尔可夫决策过程（Markov decision process, MDP）。MDP 是序列决策（sequential decision）的经典形式化表达，用于在系统状态具有马尔可夫性质的环境中模拟智能体可实现的随机性策略与回报，其动作不仅影响即时收益，还影响后续的情景以及未来的收益。当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么此随机过程即具有马尔可夫性质。

一个 MDP 通常由状态空间、动作空间、状态转移函数、奖励函数等组成。在每个时刻，环境有一个状态（state），可以理解为对当前时刻环境的概括。在超级玛丽的例子中，可以把屏幕当前的画面（或者最近几帧画面）看作状态。玩家只需要知道当前画面（或者最近几帧画面）就能够做出正确的决策，决定下一步是让超级玛丽向左、向右、或是向上。因此，状态是做决策的依据。再举一个例子，在中国象棋、五子棋游戏中，棋盘上所有棋子的位置就是状态，当前格局就足以供玩家做决策。假设不是从头开始一局游戏，而是接手别人的残局，只需要仔细观察棋盘上的格局，就能够做出决策。知道这局游戏的历史记录（即每一步是怎么走的），并不会提供额外的信息。相反的，在星际争霸、红色警戒、英雄联盟等游戏中，玩家屏幕上最近的 100 帧画面并不是状态，因为这些画面不是对当前环境完整的概括。在地图上某个看不见的角落里可能正在发生些事件，这些事件足以改变游戏的结局。一个玩家屏幕上的画面只是对环境的部分观测（partial observation）。

状态空间（state space）是指所有可能存在状态的集合，记作 S，状态空间可以是离散的，也可以是连续的。状态空间可以是有限集合，也可以是无限可数集合。在超级玛丽、星际争霸、无人驾驶这些例子中，状态空间是无限集合，存在无穷多种可能的状态。围棋、五子棋、中国象棋这些游戏中，状态空间是离散有限集合，可以枚举出所有可能存在的状态（也就是棋盘上的格局）。

动作（action）是智能体基于当前的状态所做出的决策。在超级玛丽的例子中，假设玛丽奥只能向左走、向右走、向上跳。那么动作就是左、右、上三者中的一种。在围棋游戏中，棋盘上有 361 个位置，于是有 361 种动作，第 i 种动作是指把棋子放到第 i 个位置上。动作的选取可以是确定性的，也可以是随机的。确定性是确定性地选取一个动作，而随机是指以一定概率选取一个动作。

动作空间（action space）是指所有可能动作的集合，记作 A。在超级玛丽的例子中，动作空间是 $A=\{\text{左},\text{右},\text{上}\}$。在围棋例子中，动作空间是 $A=\{1,2,3,\ldots,361\}$。动作空间可以是离散集合或连续集合，可以是有限集合或无限集合。

奖励（reward），也叫收益信号，是指在智能体执行一个动作之后，环境反

中国知网 https://www.cnki.net

馈给智能体的一个数值，收益信号定义了强化学习问题中的目标。智能体的唯一目标是最大化长期总收益。在生物系统中，收益与痛苦或愉悦的体验类似，这些信号定义了智能体所面对的问题即时且典型的特征。因此，收益信号是改变策略的主要基础。如果策略选择的动作导致了低收益，那么可能会改变策略，从而在未来的这种情况下选择一些其他的动作。一般来说，收益信号可能是环境状态和在此基础上所采取动作的随机函数。奖励往往由设计者来定义，奖励定义得好坏非常影响强化学习的结果。比如可以定义玛丽奥吃到一个金币，获得奖励+1；如果玛丽奥通过一局关卡，奖励是+1000；如果玛丽奥碰到敌人，游戏结束，奖励是−1000；如果这一步什么都没发生，奖励就是 0。如何定义奖励函数是一个见仁见智的问题。一般来讲应该把打赢游戏的奖励定义得大一些，这样才能鼓励玛丽奥通过关卡，而不是一味地收集金币。

策略（policy）定义了学习智能体在特定时间的行为方式。简单来说，策略是根据观测到的状态，如何做出决策，即如何从动作空间中选取一个动作，也就是环境整体到动作的映射。它对应心理学中的“刺激-反应”规则和关联关系。在某些情况下，策略可能是一个简单的函数或查询表，而在另一些情况下，它可能涉及大量的计算，例如搜索过程。策略本身是可以决定行为的，因此策略是强化学习智能体的核心。一般来说，策略可能是以环境所在状态为自变量，智能体所采取的动作为因变量的随机函数。举一个现实的例子，假设玩家在玩超级玛丽游戏，当前屏幕上的画面是图2-2时，做出何种决策是最优的？玩家有很大概率会决定向上跳，这样可以避开敌人，还能吃到金币。向上跳这个动作就是玩家大脑中的策略做出的决策。

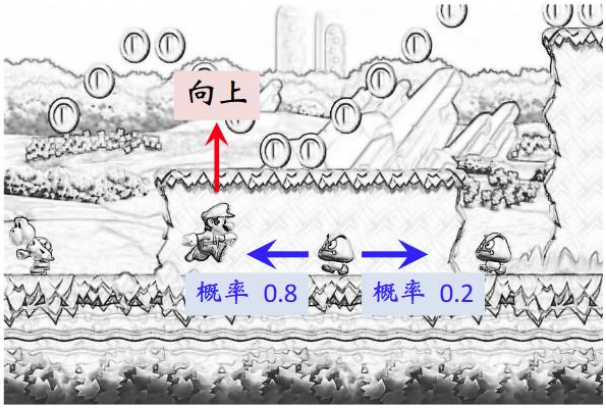


图 2-2 游戏场景示意图

Figure 2-2 Game scene diagram

价值函数也是强化学习系统的核心要素之一。收益信号表明了在短时间内什么是好的，而价值函数则表示了从长远看什么是好的。简单来讲，一个状态的价值是一个智能体从这个状态开始，对将来累计的总收益的期望。尽管收益决定了环境状态直接、即时、内在的吸引力，但价值表示了接下来所有可能状

中国知网 https://www.cnki.net

态的长期期望。例如，某状态的即时收益可能很低，但它仍可能具有较高的价值，因为其之后的状态中可能会出现高收益的状态，反之亦然。从生物学的角度解释，收益就像即时的愉悦（高奖励）和痛苦（低奖励），而价值则是当前环境与特定状态下，对未来情绪有多愉悦或多不愉悦的更具有远见的评价。

从某种意义上来讲，奖励更加重要，而作为奖励期望的价值属次等重要，没有奖励就没有价值，而评估价值的唯一目的也是获取更多的收益。然而，在制定和评估策略时，决策模型的核心是价值。强化学习中动作的选择都是基于对价值的判断做出的，其最终目标是寻求能带来最高价值而不是最高即时奖励的动作，因为这些动作从长远的角度来看会带来最大的累计收益。确定价值要比确定奖励更困难，因为奖励基本上是由环境交互后直接给出的，但价值必须综合评估，并根据智能体在整个任务执行过程中观察到的即时奖励序列重新估计。价值评估方法才是强化学习算法中的重要组成部分。

最开始，强化学习对于价值函数的表征采用建立表格的方式，但当面对更接近现实任务更复杂的强化学习时，其具有组合性的、巨大的状态空间所面临的问题不仅仅是大型表格所需要的内存，还在于精确地填充它们所需的时间和数据。在许多任务中，遇到的很多状态都是样本经验以外的。为了在这些状态下做出合理的决策，就必须从以往经历的与当前状态在某种程度上相似的状态中去归纳，也就是泛化能力。从样本中进行泛化的方法已经得到了广泛的研究，将深度神经网络强大的表征学习能力与强化学习的有效决策策略相结合，旨在解决复杂环境下序列决策问题的深度强化学习应运而生。

强化学习分为基于模型的强化学习（Model-Based Reinforcement Learning, MBRL）和无模型的强化学习（Model-Free Reinforcement Learning, MFRL）。在基于模型的方法中，首先智能体尝试学习环境的模型，即了解状态转移概率和奖励函数，然后利用模型进行策略优化或规划，如动态规划方法。无模型的方法中智能体不直接学习环境模型，而是通过与环境直接交互生成样本，基于样本来学习最优策略，它又分为基于值函数的方法（如 Q-learning、SARSA、DQN）、基于策略梯度的方法（如 REINFORCE）以及结合两者的 Actor-Critic 方法（如 AC、A2C、A3C[105]、DDPG[106]、PPO[107]、TD3[108]、SAC[109]等）。本文重点关注无模型的强化学习方法。

2.2.2 基于值函数的方法

决策模型的核心是价值，如何估计价值进行强化学习决策的核心任务。

基于值函数方法（Value-based method）的核心思想是通过估计一个或一组值函数来指导智能体的学习过程。值函数描述了在给定状态下采取某种策略所能得到的未来累积奖励期望。

未来累计奖励期望，也称之为回报，记作G_t，在最简单的情况下，

中国知网 https://www.cnki.net

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T \tag{2-1}$$

其中，T 是最终时刻。

式(2-1)在有“最终时刻”的应用中是有意义的。但在许多现实情况下，智能体与环境的交互不一定存在“最终时刻”而是一直持续或者持续很久不断发生的事件，那这式子就存在未来累计期望趋于无穷的问题。针对这一问题，引入了折扣的概念，即智能体尝试选择动作，使得它在未来收到的经过折扣系数加权后的奖励总和是最大化的，

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2-2}$$

其中，γ 是折扣系数（也叫折扣率），$0 \le \gamma \le 1$。

折扣率决定了未来收益的现值：未来时刻 k 的奖励值只有它当前值的 γ^{k-1} 倍。如果 $\gamma < 1$，那么只要收益序列 $\{R_k\}$ 有界，式(2-2)中累计期望总和就是有限的。如果 $\gamma = 0$，那么智能体是“目光短浅的”，其只关心最大化即时奖励，但一般来说，最大化即时奖励会减少未来的收益，以至于实际上的累计奖励是减少的。而随着 γ 越接近 1，意味着折扣回报将更多的考虑未来的收益，也就是说，这是一个有远见的价值评估。

临近时刻的回报可以通过递归的方式相互联系起来，这对于强化学习的理论和算法是至关重要的，

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{2-3}$$

在强化学习中，最常见的值函数是状态价值函数 $V(s)$ 和动作价值函数 $Q(s,a)$。状态价值 $V(s)$ 表示智能体从状态 s 开始，遵循当前策略，直到结束所能获得的累计奖励期望。

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] \tag{2-4}$$

类似地，动作价值函数 $Q(s,a)$ 表示智能体在状态 s 下采取动作 a 后，后续遵循当前策略所能得到的未来累积奖励期望。

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right] \tag{2-5}$$

智能体期望未来能得到的收益取决于智能体所选择的动作，因此价值函数是与特定的行为方式相关的，称之为策略，记作 π。基于值函数的方法通过比较价值函数来定义策略，一般有贪心策略以及 ϵ-贪心策略。贪心策略是指每次

中国知网 https://www.cnki.net

都选取价值最高的动作，但这种方法容易陷入局部最优，因为其对策略空间的探索是不足的。而ϵ-贪心策略是在贪心策略的基础上，加入了一个系数ϵ，表示动作选择的概率，即决策动作时有ϵ的概率随机选取动作，$1-\epsilon$的概率执行贪心策略，这种方法加强了对于策略空间的探索。

在强化学习和动态规划中，价值函数的一个基本特性是满足递归关系，对于任何策略π和任何状态s，s的价值于其可能的后续状态的价值满足

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a)\left[r + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right] \\ &= \sum_a \pi(a \mid s) \sum_{s', r} p(s, r \mid s, a)[r + \gamma v_\pi(s')] \end{aligned} \tag{2-6}$$

其中，式(2-6)即为v_π的贝尔曼方程，表示状态价值和后续价值之间的关系。以图 2-3 所示的回溯图来看，从一个状态向后观察未来的所有可能，根据策略π决策出动作a给到环境执行，环境根据其自身的特性函数p，反馈出即时奖励r和下一时刻状态s'。贝尔曼方程就是当前状态下根据其出现概率对未来所有可能进行加权平均，这也就说明了当前时刻的价值等于后续价值与即时奖励之和的期望。

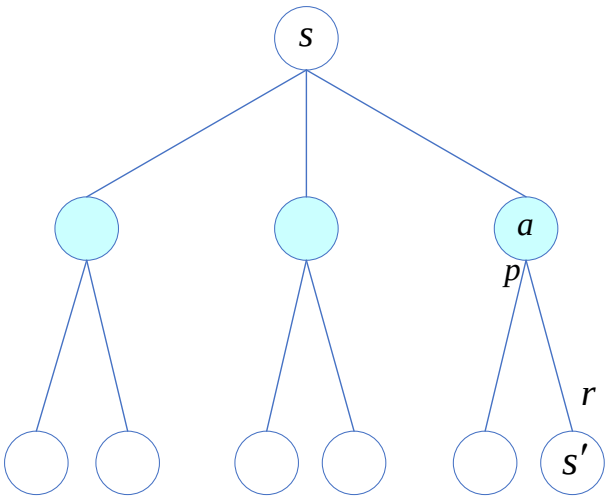


图 2-3 价值函数的回溯图

Figure 2-3 Backup diagram for value function

强化学习的一般步骤为策略评估、策略迭代，而基于值函数的方法是根据值函数的结果进行动作决策的，所以这类算法的重点在于如何更好地进行对于价值函数的评估和迭代。主要的求解方法有三种，分别是动态规划（Dynamic Programming, DP）、蒙特卡洛方法（Monte-Carlo, MC）和时序差分（Temporal-Difference, TD）方法。动态规划方法具有严格清晰的数学基础且已经被深入研

中国知网 https://www.cnki.net

究，但它需要完整、精确的环境模型，属于基于模型的方法。蒙特卡洛方法和时序差分方法不需要环境模型，是无模型的方法。

蒙特卡洛算法本质上是一种基于大量随机实验的数据驱动型解决方案，它利用随机抽样手段来应对那些理论上具有确定解但在实际运算上复杂度较高的问题。在强化学习领域中，智能体通过不断与环境互动积累起一系列包含状态、行为及对应奖励的数据记录，这些连续的互动过程被称作“经验轨迹”。该方法特别关注完整且有时间限制的经验轨迹片段。借助蒙特卡洛法，智能体通过反复采样并记录下不同策略 π 下的多条经验轨迹，进而对各个状态下平均收益进行估算。这意味着，在每一条经验轨迹完成后，系统能够统计出从当前状态开始直到某个未来时间步长内所有累计奖励的平均值，从而作为对未来长期累积奖励期望值的近似估计。简而言之，只有在完整地经历了一个经验轨迹后，才能对该轨迹对应的累积奖励做出准确的统计计算。

蒙特卡洛方法中价值函数的更新过程为

$$V(s_t) \leftarrow V(s_t) + \alpha[G_t - V(s_t)] \tag{2-7}$$

其中，α 是学习率，G_t 是 t 时刻后的累计奖励，是价值的无偏估计。

时序差分方法同样是基于反复采样得到的经验样本来进行价值函数更新的，但其结合动态规划中的自举思想，无需等待完整的经验轨迹，而是每一时间步基于已得到的其他状态的估计值来更新当前状态的价值函数。

时序差分方法的价值函数更新过程为

$$V(s_t) \leftarrow V(s_t) + \alpha[R_t + \gamma V(S_{t+1}) - V(S_t)] \tag{2-8}$$

其中，R_t 是当前 t 时刻的即时奖励，γ 是折扣率。

Q-Learning 是一种典型的基于值函数的强化学习算法，对于状态-动作对的价值函数进行估计，采用贪心策略进行动作决策，其更新过程为

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma max Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{2-9}$$

而面对复杂任务中存在的大规模高维状态空间下的值函数近似问题时，DQN（Deep Q-Network）算法成功地利用深度学习改进 Q-Learning 的算法，是深度强化学习的起点。DQN 算法中采用 Q 网络来表示动作价值函数，然后通过迭代更新的方式进行函数逼近。神经网络的迭代更新目标是最小化损失函数，并且需要大量的样本，因此算法在与环境交互的过程中，智能体持续将采集的经验样本存入回放经验池（replay buffer），每次更新时从经验池中随机取出部分样本，然后利用小批量样本随机梯度下降法（stochastic gradient descent, SGD）训练网络。在经验池中随机抽取样本可以打乱经验之间的相关性，使得神经网络更新更有效率。此外，DQN 的 Fixed Q-targets 也是一种打乱相关性的机理，DQN 中使用到两个结构相同但参数不同的神经网络，预测 $Q_{估计}$ 的神经网络具备最新的参数，而预测 $Q_{现实}$ 的神经网络，也称为 Target 网络，使用的参数则是很

中国知网 https://www.cnki.net

久以前的。

DQN 中的损失函数为预估值与真实值的均方差，

$$\begin{aligned} L(\theta) &= \mathbb{E}_{\pi}\left[\left(Q_{\text{现实}} - Q_{\text{估计}}\right)^2\right] \\ &= \mathbb{E}_{\pi}\left[\left(r + \gamma Q_{\text{target}}(s', a') - Q(s, a)\right)^2\right] \end{aligned} \tag{2-10}$$

DQN 算法的伪代码如表 2-1 所示。

表 2-1 DQN 算法的伪代码

Table 2-1 Pseudocode of DQN

算法：DQN 算法
1:初始化回放经验池 $\mathcal{D}$；
2:初始化 Q 网络的参数 θ 和 target Q 网络的参数 θ'；
3:循环 $\text{episode} = 1 \to M$ ：
4: 循环 $t = 1 \to T$ ：
5: 根据概率 ϵ 选择一个随机动作 a_t 或者选择 $a_t = argmaxQ(\phi(s_t), a; \theta)$；
6: 执行动作 a_t 并获得即时奖励 r_t 和下一时刻状态 s_{t+1}；
7: 保存经验样本 (s_t, a_t, r_t, s_{t+1}) 到经验池 $\mathcal{D}$；
8: 从经验池中随机抽取批量样本；
9: 根据公式(2-10)计算梯度并更新 Q 网络的参数 θ；
10: 每隔 C 步更新 Target Q 网络的参数 $\theta' = \theta$；

2.2.3 策略梯度方法

基于值函数的方法都是先学习动作价值函数，然后根据估计的动作价值函数选择动作，如果没有价值函数的估计，策略也就不存在了。另一种强化学习的方向就是直接学习参数化的策略，动作的选择也就不依赖于价值函数，这就是策略梯度方法。

强化学习的目标是找到一个可以让智能体(agent)获取最大收益的最优策略。而策略梯度算法着眼点在于直接对策略本身进行建模和优化。策略通常用一个带参数(通常用 θ 表示)的函数来建模，而后可以使用各种各样的算法来对参数进行优化，以达到收益（目标函数）最大化的目标。在策略梯度方法中，策略可以用任意的方式参数化，只要 $\pi(a \mid s, \theta)$ 对参数可导。

策略梯度算法的一个优势是可以以任意的概率来选择动作，在有重要函数近似的问题中，最好的近似策略可能是一个随机策略，越接近现实任务的应用

中国知网 https://www.cnki.net

场景越是如此。例如，在非完全信息的纸牌游戏中最优的策略一般是以特定的概率选择两种不同的玩法，比如德州扑克中的虚张声势。基于价值函数的方法没有一种自然的途径来求解随机最优策略，但策略梯度方法可以。

另外，策略参数化形式的选择有时是在基于强化学习的系统中引入专家先验知识的一个好方法。

策略参数化相比于ϵ-贪心策略除了实践上的优势外，还有重要的理论优势。ϵ-贪心策略中只要估计的动作价值函数变化导致了最大动作价值函数对应动作发生了变化，则选择某个动作的概率就会突然发生很大变化，即使估计的动作价值函数没有那么大变化，而对于策略参数化的方法，选择动作的概率作为被优化参数的函数会平滑地变化。很大程度上，与基于动作价值函数的方法相比，基于策略梯度的方法有更强的收敛保证。

策略梯度方法基于某种性能度量$J(\theta)$的梯度，是标量$J(\theta)$对策略参数的梯度。根据策略梯度定理，有

$$\nabla J(\theta) \propto \sum_{s} \mu(s) \sum_{a} q_{\pi}(s,a) \nabla \pi(a \mid s,\theta) \tag{2-11}$$

其中，梯度是参数向量θ每个元素的骊到组成的列向量，π表示参数θ对于s的策略。

REINFORCE（蒙特卡洛策略梯度）算法是一种经典的策略梯度算法。算法依赖于蒙特卡洛方法生成样本来估计期望收益，以更新策略的参数。由于使用蒙特卡洛方法得到的梯度在期望意义上与实际的梯度是相等的，因此REINFORCE可以正常工作，有

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \mathbb{E}_{\pi}[Q^{\pi}(s,a) \nabla_{\theta} \ln \pi_{\theta}(a \mid s)] \\ &= \mathbb{E}_{\pi}[G_t \nabla_{\theta} \ln \pi_{\theta}(A_t \mid S_t)] \end{aligned} \tag{2-12}$$

REINFORCE 算法通过实际采样得到的样本路径（sample trajectories）中来估计策略梯度，并使用进行参数更新。该方法依赖于完整的样本路径，是典型的蒙特卡洛方法。REINFORCE 的伪代码如表 2-2 所示。

表 2-2 REINFORCE 算法伪代码

Table 2-2 Pseudocode of REINFORCE

算法：REINFORCE	
1:	输入：一个可导的参数化策略$\pi(a \mid s,\theta)$；
2:	算法参数：步长$\alpha > 0$；
3:	初始化策略参数$\theta \in \mathbb{R}^{d'}$；
4:	无限循环（对于每一幕）；
5:	根据$\pi(\bullet \mid \bullet,\theta)$，生成一幕序列$S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$；

中国知网 https://www.cnki.net

6: 对于幕的每一步循环，$t=0,1,\ldots,T-1$：

7: $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$；

8: $\theta \leftarrow \theta + \alpha \gamma^t G \nabla ln\pi(A_t \mid S_t, \theta)$；

2.2.4 演员-评论家方法

演员-评论家（Actor-Critic）算法，也叫行动器-评判器，结合了基于值函数的方法以及策略梯度方法，旨在综合两种方法的优点。AC 算法共包含两种网络，其中 Actor 是一个策略网络，它负责学习在给定状态下应该采取什么样的动作；Critic 是一个值函数网络，它负责评估 Actor 选择的动作是否好，以及对动作价值进行更新。在 Actor-critic 算法中，Actor 和 Critic 是分开训练的，它们相互影响，通过反馈来不断提升整体性能。Actor-critic 算法的优点包括较低的方差、快速的收敛速度以及对连续动作空间的支持。

A2C（Advantage Actor-Critic）算法是 Actor-critic 算法的一种变体，它进一步改进了原始的 Actor-Critic 算法。在 A2C 算法中，主要有两个关键的改进：首先 A2C 算法引入了优势函数（Advantage Function），用于评估 Critic 的输出与 Actor 输出动作的优势差异。这有助于更准确地指导 Actor 的学习，使其朝着更优的方向更新策略。其次 A2C 算法通过同步更新 Actor 和 Critic 网络来减少训练的方差，同时提高训练的效率，这样可以更快地收敛到较好的策略。A2C 算法在稳定性和效率上比传统的 Actor-Critic 算法有所提高，在很多强化学习任务中取得了较好的表现。它被广泛应用于解决连续动作空间的强化学习问题，比如机器人控制、游戏玩法等领域。通过引入优势函数和同步更新，A2C 算法能够更准确、更快速地学习到最优策略。

A3C（Asynchronous Advantage Actor-Critic）算法重点关注并行训练。在 A3C 中，由 Critics 学习价值函数，多个 Actor 进行并行训练并且不断和全局参数进行同步。A3C 是专门针对并行训练所设计的，如图 2-4 所示，A3C 可以使用多个智能体(agent)来进行并行训练。

中国知网 https://www.cnki.net

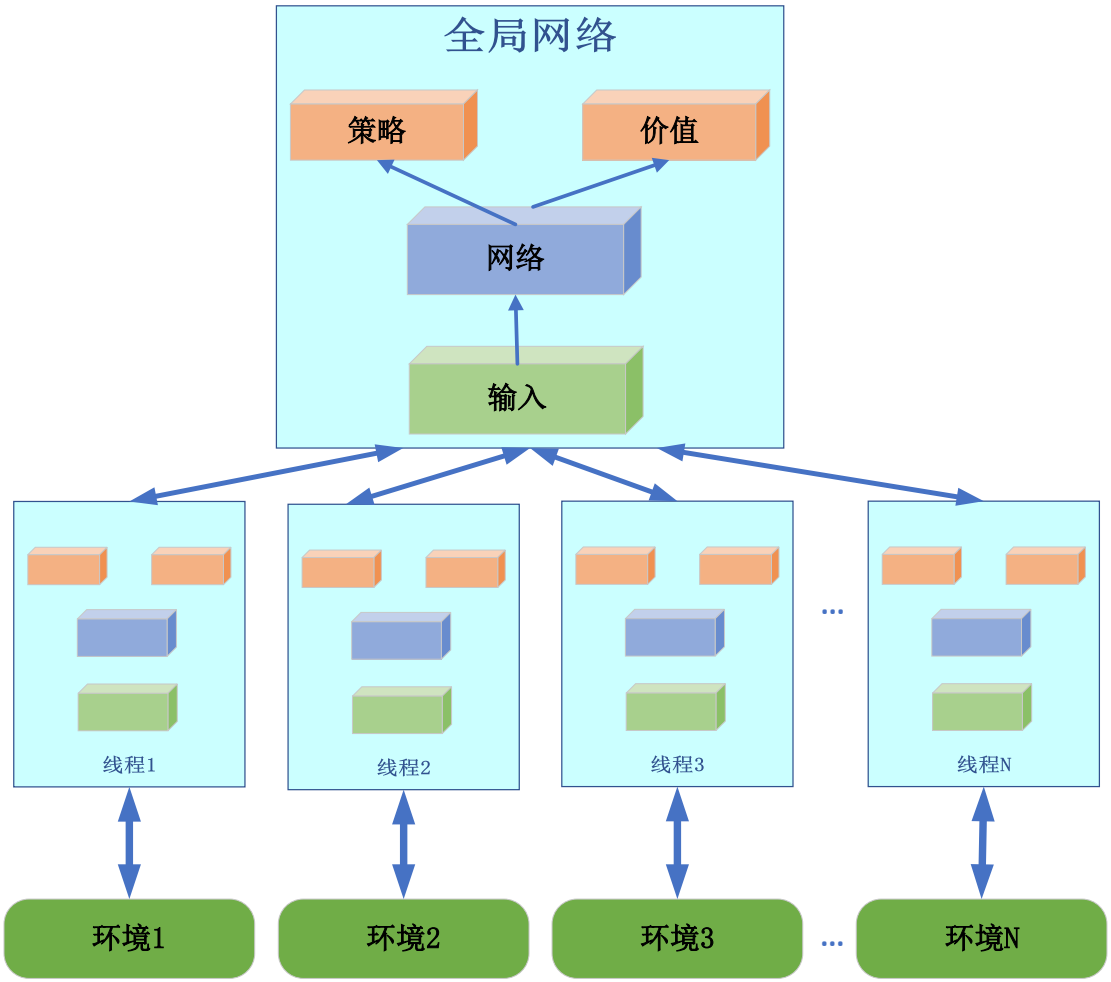


图 2-4 A3C 异步训练框架图

Figure 2-4 Asynchronous training frame diagram of A3C

A3C 算法的伪代码如表 2-3 所示

表 2-3 A3C 算法伪代码

Table 2-3 Pseudocode of A3C

算法：A3C
1: 设定全局参数θ,ω，以及特定线程的参数θ',ω'；
2: 初始化时间步数$t=1$；
3: 循环 当$T<T_{max}$：
4: 重置梯度值：$d\theta=0,d\omega=0$；
5: 线程相关参数使用全局参数同步：$\theta'=\theta,\omega'=\omega$；
6: 令$t_{start}=t$，并且对初始状态进行采样s_t；
7: 循环 当s_t不是终止状态且$t-t_{start}\le t_{max}$：
8: 选择动作$A_t\sim\pi_{\theta'}(A_t\mid S_t)$，执行$A_t$得到即时奖励$R_t$和下一时刻状态$s_{t+1}$；

中国知网 https://www.cnki.net

9:	更新 $t=t+1, T=T+1$；
10:	计算最后状态的收益估计，若 s_t 为终止状态，则 $R=0$，否则
11:	$R=V_{\omega'}(s)$；
12:	循环 $i=t-1,\ldots,t_{start}$：
13:	$R \leftarrow \gamma R + R_i$；
14:	计算 θ' 累计梯度：$d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi_{\theta'}(a_i \mid s_i)(R - V_{\omega'}(s_i))$；
15:	计算 ω' 累计梯度：$d\omega \leftarrow d\omega + 2(R - V_{\omega'}(s_i))\nabla_{\omega'}(R - V_{\omega'}(s_i))$；
16:	使用累计梯度 $d\theta, d\omega$ 更新参数 θ, ω；

此外，后续学者对于深度强化学习算法的研究基本都基于 Actor-Critic 的架构来进行，包括 DPG、DDPG、PPO、SAC、TD3 等等。

2.2.5 分层强化学习

分层强化学习（Hierarchical Reinforcement Learning）是一种旨在通过层次化的结构来解决复杂任务的强化学习方法。在传统的强化学习中，智能体需要直接学习从状态到动作的映射关系，然而对于复杂任务，直接建模这种映射可能会面临高维度空间、稀疏奖励等问题。因此，分层强化学习提出将任务分解成多个层次，每个层次负责解决一部分子任务，从而简化问题的复杂度。

当任务环境存在稀疏奖励（sparse reward）问题时，智能体可能长期都无法获得具有正奖励的样本，给值函数和策略的学习带来了困难。而通过分层把策略分为不同层级的子策略，每个子策略在学习的过程中会得到来自上一层级传递来的奖励，这样可以大大提升样本的利用效率。

目前的分层强化学习主要可以分为两大类，第一类是基于选项（option）的，第二类是基于目标（goal）的。option 表示的是一种具有时序抽象的策略，可以说是抽象出来的上层策略，这是策略层面上的定义；而 goal 则是目标层面上的定义，即智能体需要达到什么目标，每一层的不同目标同样对应着一个不同的子策略。比如去一个地方需要搭乘地铁、公交车、共享单车，这里地铁、公交车、共享单车都属于 option，而搭乘三种交通工具抵达的目的地则是 goal，可以看出两者是紧密联系的。总体来说，几乎所有的分层强化学习都是上层控制器在较长的时间跨度里选择 option/goal，而下层控制器在较短的时间跨度里根据 option/goal 选择动作 action。分层之所以能够提升样本效率，是因为上层控制器给下层控制器提供 goal/option 的同时还会根据下层控制器的策略好坏反馈一个对应的内在奖励（intrinsic reward），这就保证了即便在外部奖励为 0 的情况下，下层控制器依然能够获得奖励，从而一定程度上缓解了奖励稀疏的问题。

H-DQN[110]算法是通过任务（goal）来分层的，属于基于子任务（subgoal）

中国知网 https://www.cnki.net

的学习。算法用 option 的时间扩展来定义每个目标 g 的策略 π_g，在学习如何选择最优 goal 序列的同时，智能体还将学习这些 goal 所对应的 option 策略。为了学习到每个 π_g，智能体有一个 critic，它能够提供内在奖励，根据这些奖励智能体能够实现目标。图 2-5 是 H-DQN 算法的智能体架构。

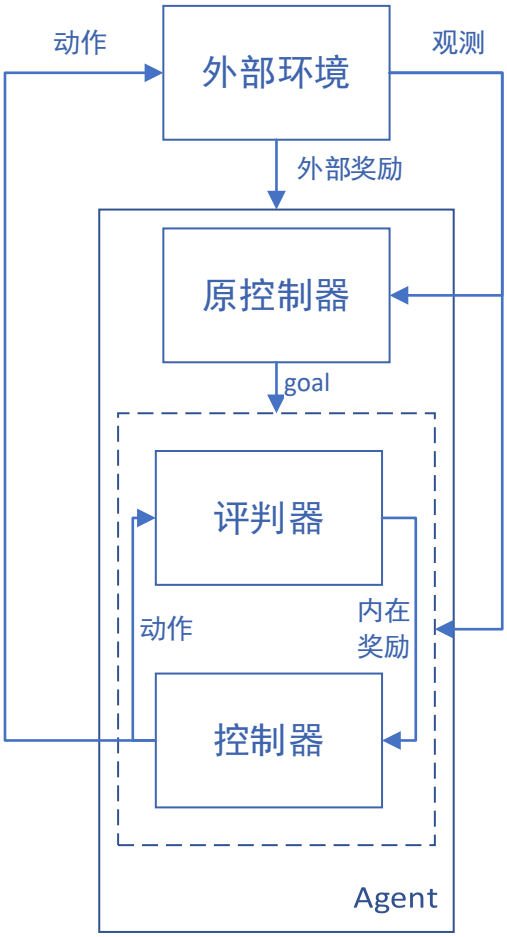


图 2-5 H-DQN 智能体分层架构

Figure 2-5 Agent layered architecture of H-DQN

内在 critic 负责评估目标是否达到并传递给控制器一个合适的奖励。控制器的目标函数是最大化内在奖励所对应的收益：

$$R_t(g)=\sum_{t'=t}^{\infty}\gamma^{t'-t}r_{t'}(g) \tag{2-13}$$

类似的，meta-controller 的目标是最大化外在奖励（即环境奖励）所对应的收益：

$$F_t=\sum_{t'=t}^{\infty}\gamma^{t'-t}f_{t'} \tag{2-14}$$

其中 f_t 表示从环境接收到的外部奖励信号。

中国知网 https://www.cnki.net

2.3 多智能体强化学习理论

2.3.1 概述

多智能体系统（multi-agent system, MAS）中包含 m 个智能体，智能体共享环境，智能体之间会互相影响。在多智能体系统中，各个智能体之间通常存在信息交换、协作、竞争等关系，它们共同协同工作以完成特定的任务或解决问题。多智能体系统的设计旨在实现系统整体性能的最大化，并利用分布式智能的优势来应对复杂、多样化的环境。

多智能体强化学习（multi-agent reinforcement learning, MARL）是指让多个智能体处于相同的环境中，每个智能体独立与环境交互，利用环境反馈的奖励改进自己的策略，以获得更高的回报（即累计奖励）。在多智能体中，一个智能体的策略不能简单依赖自身的观测、动作，还需要考虑其他智能体的观测、动作。因此，多智能体强化学习比单智能体强化学习更困难。多智能体强化学习通常可以应用于以下许多领域和问题，比如在博弈论中，多智能体强化学习可以用于研究博弈论中的均衡解，帮助理解多个智能体在竞争或合作情境下的最优行为策略。在多智能体协作中，多智能体强化学习可以应用于实现多个智能体之间的协作，例如在多机器人系统、多智能体交互系统等领域。在资源分配与调度中，在资源有限且竞争激烈的环境下，多智能体强化学习可以用于实现有效的资源分配和调度，以最大化整体效益。

MDP 的多智能体版本通常被称为马尔科夫博弈（Markov games）。假设一共有 N 个智能体，并且有一个状态集合 S 。每个智能体 $1,2,\ldots,N$ 拥有各自行为的集合 $\mathcal{A}_1,\ldots,\mathcal{A}_N$ ，以及各自的观测状态集合 $\mathcal{O}_1,\ldots,\mathcal{O}_N$ 。定义状态转移函数包含所有智能体的状态、行为以及观测值空间：$\tau:\mathcal{S}\times\mathcal{A}_1\times\ldots\times\mathcal{A}_N\to\mathcal{S}$ 。对于每一个智能体，它的策略仅会作用于它自己的观测空间上。如果策略是随机的，将其表示为：

$$\pi_{\theta_i}:\mathcal{O}_i\times\mathcal{A}_i\mapsto[0,1] \tag{2-15}$$

如果策略是确定性的，将其表示为：

$$\mu_{\theta_i}:\mathcal{O}_i\mapsto\mathcal{A}_i \tag{2-16}$$

2.3.2 多智能体近端策略优化算法

近端策略优化（Proximal Policy Optimization, PPO）算法是一种用于强化学习的优化算法，旨在提高策略学习的稳定性和效率。PPO 算法基于策略梯度方法，通过对策略的更新进行限制，以确保每次更新不会偏离太远，从而保持训练的稳定性。PPO 算法在 2017 年由 OpenAI 提出，并在许多领域取得了良好的效果。

PPO 算法的核心思想是通过最大化经验轨迹的加权优势估计来更新策略，

中国知网 https://www.cnki.net

同时引入一个近端约束，限制每次更新的幅度，以避免更新过大导致训练不稳定。具体来说，PPO 算法同时优化两个策略：一个是当前的策略（新策略），一个是之前的策略（旧策略），通过比较两个策略的比率来确定更新方向，并根据近端约束来调整更新幅度。

PPO 算法相对于传统的策略梯度方法，如 TRPO（Trust Region Policy Optimization），在实现上更简单，并且更容易扩展到大规模的问题。它在训练过程中的稳定性和高效性使其成为当前强化学习领域的研究热点之一，被广泛用于各种类型的强化学习任务，包括机器人控制、游戏玩法等领域。

但是，在多智能体环境下，PPO 算法在多智能体系统中的样本效率明显低于非策略方法。而多智能体近端策略优化算法（MAPPO）仅使用最小的超参数调优，没有在算法原理或架构的修改，使得 PPO 算法在多智能体系统中获得惊人的性能[111]。

DEC-POMDP 问题可以定义为：$<S,A,O,R,P,n,\gamma>$，S 为状态空间，A 为每个智能体共享的动作空间，$o_i=O(s;i)$ 为每个智能体 i 的局部观测；$P(s'|s,A)$ 是在给定的联合动作 $A=(a_1,a_2,a_3,\cdots,a_n)$ 条件下的状态由 s 转移 s' 的状态转移概率，其中 n 为智能体的个数；$R(s,A)$ 为共享的奖励，γ 是折扣因子，a_i 是由依据局部观测 o_i 和策略 $\pi_\theta(a_i|o_i)$ 决策的智能体动作。利用某一时刻 t 的联合动作更新折扣累积奖励，即联合动作，联合状态，共享奖励。

通过学习一个策略 π_θ 和一个值函数 $V_\varphi(s)$，实现了多智能体环境下 PPO 的结构，$V_\varphi(s)$ 利用智能体中除了自身局部观测外的其他全局信息作为输入，减少单个智能体训练的误差，评价智能体动作选择的优劣。当只关注合作型多智能体场景时，此奖励为共享奖励。

根据策略梯度方法，策略依据下列公式进行多次采样更新：

$$\nabla\bar{R}_\theta=E_{\tau\sim p_\theta(\tau)}[R(\tau)\nabla\log p_\theta(\tau)] \tag{2-17}$$

每次更新参数 θ 后，这些数据样本就不再是新策略的样本，无法去更新新的策略，这样样本的使用效率就会很低。

由此引出了重要性采样（Importance Sampling）：

$$E_{x\sim p}[f(x)]=E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \tag{2-18}$$

也就是说，x 服从 p 分布，那么 $f(x)$ 的期望等于 $f(x)p(x)/q(x)$ 的期望，其中，$p(x)/q(x)$ 称为重要性权重。

重要性采样忽略策略 π，用策略 μ 采集的样本求出策略 π 下的期望，但前提条件是两个分布不能相差太多，即重要性权重不能太大，否则估算出来的方差会很大。方差大意味着估计值的不稳定和不准确，可能会导致误差太大或者

中国知网 https://www.cnki.net

收敛很慢。为了减小方差，我们需要选择一个与目标分布相似的采样分布，或者使用一些改进方法。

2.3.3 多智能体确定性深度策略梯度算法

确定性深度策略梯度算法（Deep Deterministic Policy Gradient, DDPG）是无模型的、off-policy 的 actor-critic 算法，它将 DPG 算法和 DQN 算法结合了起来[112]。DQN 通过经验回放以及目标网络冻结来稳定 Q 函数的学习过程。原始的 DQN 工作在离散空间中，DDPG 结合了 actor-critic 框架，将 DQN 扩展到连续空间上，并且用其学习一个确定性的策略。

为了更好的进行探索，算法通过给确定性策略 $\mu_\theta(s)$ 添加噪声 N 来得到一个新的探索策略 μ'：

$$\mu'(s) = \mu_\theta(s) + \mathcal{N} \tag{2-19}$$

此外，DDPG 对于 actor 和 critic 参数的更新都是软更新。软更新不会直接使用新策略的参数，而是将新旧策略进行加权求和，也就是在参数更新时，引入了一个参数 $\tau \ll 1$，并使用该参数进行更新操作：

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \tag{2-20}$$

使用此方法，可以限制目标网络的更新速度。这与 DQN 中的设计不一样，在 DQN 中目标网络会在一定的时间段内被冻结。

多智能体确定性深度策略梯度算法（Multi-Agent DDPG, MADDPG）扩展了 DDPG 算法，以使其可以工作在一个新的环境中。在此环境中，多个智能体需要合作来完成某项任务，但是这些智能体仅能观测到自身的本地信息。在单个智能体的观察中，它看到的环境是非平稳的，这种非平稳性是由各个智能体策略不断更新而这些更新又不是全局可知所造成的。MADDPG 是一个 actor-critic 模型，它被特别设计用于处理这种可以与多个智能体进行交互的、策略不断变化的环境。

MADDPG 中的 critic 是一个中心化的 critic，中心化指的是 critic 可以使用所有智能体的相关数据（比如：执行的动作、观测到的状态等）。每个智能体会学习一个自己的行为价值函数：

$$Q_i^{\vec{\mu}}(\vec{o}, a_1, \ldots, a_N) \tag{2-21}$$

可以看到这里的 Q 的输入使用到了所有其它执行体相关的数据，其中 $a_1, \ldots, a_N$ 是 N 个智能体所执行的动作。每个智能体的 $Q_i^{\vec{\mu}}$ 函数都是独立进行训练学习的，它们奖励的结构可能会出现任意的结构，包括它们可能会去竞争奖励。

每个智能体 i 也拥有一个 actor，用于进行策略探索以及策略参数 θ_i 的更新。算法的架构图如图 2-6。

中国知网 https://www.cnki.net

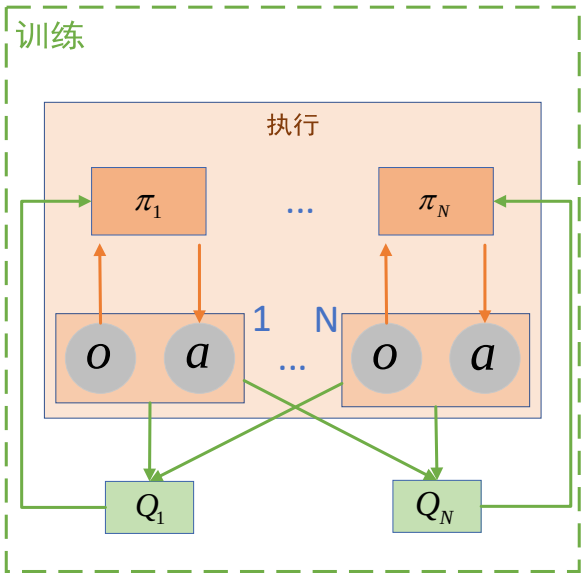


图 2-6 MADDPG 架构图

Figure 2-6 Architecture diagram of MADDPG

此外，为了缓解不同智能体之间的竞争或者合作带来的高方差，MADDPG 提出了引入一个新方法——策略集成 (policy ensembles)，即每个智能体训练 K 个策略，然后每一轮随机选择一个策略公开，智能体使用 K 个策略集成的结果进行策略更新。

2.4 强化学习与序列模型

在强化学习问题中，越接近真实的决策问题越有可能需要对于时序进行建模。以 dota2 游戏为例，在整个游戏过程中，智能体需要涉及资源收集、科技发展这种需要长期记忆的事件，还有战场侦察、技能衔接等需要短期记忆的事件，以及部队阵型、兵种配合等需要多智能体进行协作的事件。智能体想要更好的在这些任务中进行决策，就需要对时序有更好的认知，也就是做一个“有记忆”的智能体。而这恰好是基于循环神经网络的序列模型所擅长的领域。强化学习与序列模型的结合是迈向真实决策问题的关键一步。

2.4.1 序列模型

RNN 是一种特殊的神经网络结构，它是根据“人的认知是基于过往的经验和记忆”这一观点提出的。它与 DNN、CNN 不同的是：它不仅考虑前一时刻的输入，而且赋予了网络对前面内容的一种“记忆”功能。RNN 之所以称为循环神经网络，是因为一个序列当前的输出与前面的输出是有关的。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层节点之间不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。其网络结构如下，

中国知网 https://www.cnki.net

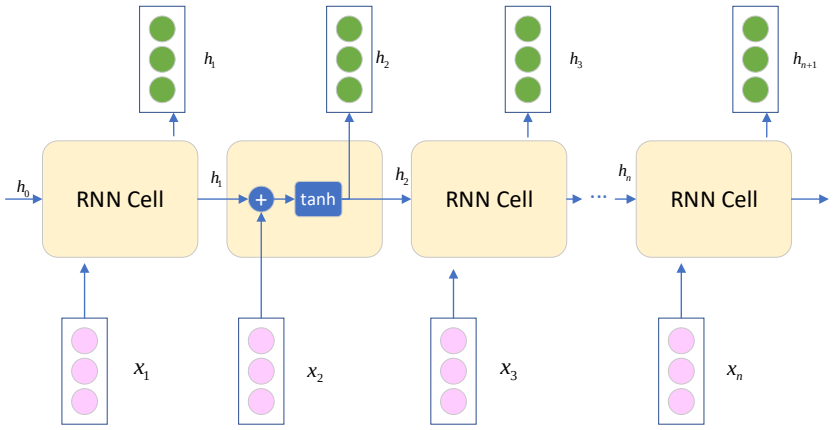


图 2-7 RNN 结构示意图

Figure 2-7 Structure diagram of RNN

LSTM（Long Short-Term Memory）是一种长短期记忆网络，是一种特殊的 RNN（循环神经网络）。与传统的 RNN 相比，LSTM 更加适用于处理和预测时间序列中间隔较长的重要事件。

传统的 RNN 结构可以看作是多个重复的神经元构成的“回路”，每个神经元都接受输入信息并产生输出，然后将输出再次作为下一个神经元的输入，依次传递下去。这种结构能够在序列数据上学习短时依赖关系，但是由于梯度消失和梯度爆炸问题，RNN 在处理长序列时难以达到很好的性能。

而 LSTM 通过引入记忆细胞、输入门、输出门和遗忘门的概念，能够有效地解决长序列问题。记忆细胞负责保存重要信息，输入门决定要不要将当前输入信息写入记忆细胞，遗忘门决定要不要遗忘记忆细胞中的信息，输出门决定要不要将记忆细胞的信息作为当前的输出。这些门的控制能够有效地捕捉序列中重要的长时间依赖性，并且能够解决梯度问题。其结构示意图如下，

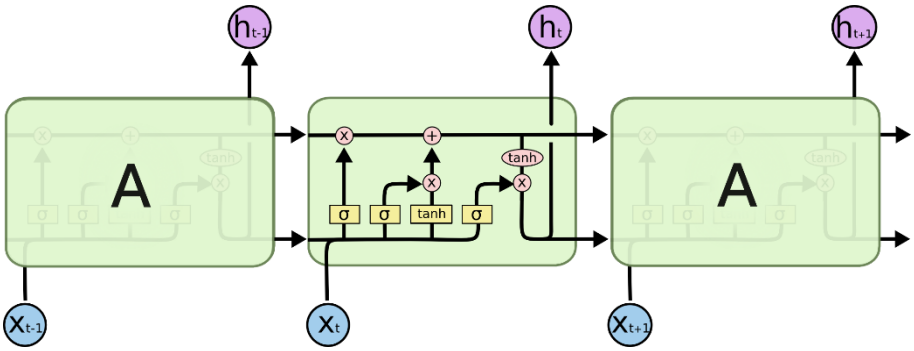


图 2-8 LSTM 结构示意图

Figure 2-8 Structure diagram of LSTM

Transformer 模型采用的是编码器-解码器架构，其编码器和解码器不是传统的循环神经网络（RNN）结构，取而代之的是编码器栈（encoder stack）和解

中国知网 https://www.cnki.net

码器栈（decoder stack）。编码器栈和解码器栈中分别为连续 N 个具有相同结构的编码器和解码器。每一个编码器中包含一个自注意力模块和一个前馈神经网络，每个子网络都具有残差连接，并且在每个残差合成后都对结果进行归一化操作。每一个解码器中除了包含与解码器类似的自注意力机制和全连接前馈神经网络外，还在两个子网络之间添加了一个注意力模块，同样三者均具有残差连接，且残差合成后进行归一化操作。残差连接可以防止梯度消失和网络深度提高带来的网络退化问题。

Transformer 基于注意力机制建立数据之间的关系，值越高则两者相关性越高。Transformer 并没采用 RNN 结构，因此可以进行并行计算，这大大提高了其运行速度。

2.4.2 多智能体 Transformer

在多智能体强化学习中，存在许多算法来解决协作问题。例如，A2C 和 MAPPO 在所有智能体之间分配相同的参数，然后通过信任区域方法进行训练。PR2 和 GR2 方法在集中式训练分布式执行框架下进行递归推理。然而，它们的局限性在于智能体的策略不知道开发合作的目的，并且仍然依赖于精心制定的最大化目标。理想情况下，智能体团队应该通过设计意识到其培训的联合性，从而遵循一种整体有效的范例，一个理想的解决方案。

多智能体强化学习是具有挑战性的，因为其不仅需要识别每个智能体的策略改进方向，而且为了全局还需要将智能体的更新策略联合起来。集中式训练分布式执行的模式在一定程度上使得这种困难有所缓解，它允许智能体在训练阶段能访问全局的信息与对手的动作，这种方法使得单智能体的算法能成功扩展到多智能体。然而这些拓展的方法不能涵盖多智能体交互的全部复杂性；事实上，在最简单的合作任务中，它们中的一些被证明是失败的。

为了解决这个问题，研究人员提出了多智能体优势分解定理，该定理捕捉了不同的智能体对回报的贡献，并通过顺序决策过程方案提供了合作产生背后的直觉。基于此，提出了一种名为 Multi-Agent Transformer（MAT）的新架构，它有效地将协作式多智能体强化学习应用到序列模型问题中。算法在多智能体强化学习和序列模型中搭建了一个桥梁，使得现代序列的建模能力在多智能体强化学习领域得以展现。MAT 架构利用多智能体优势分解定理将联合策略搜索问题转换为顺序决策过程，这使得多智能体问题的时间复杂度仅为线性，最重要的是，使 MAT 具有单调的性能改进保证。与 Decision Transformer 等现有技术只适合预先收集的离线数据不同，MAT 是通过基于策略的方式从环境中进行在线试验和错误训练的。

为了建立多智能体强化学习和序列模型之间的联系，优势分解定理提供了一个从序列模型角度理解多智能体强化学习问题的新角度。如果每个智能体以

中国知网 https://www.cnki.net

任意决策顺序知道其前面的行为，则智能体的局部优势 $A_{\pi}^{i_j}(\boldsymbol{o},\boldsymbol{a}^{i_{1:m-1}},a^{i_m})$ 的总和恰好等于联合优势 $A_{\pi}^{i_{1:n}}(\boldsymbol{o},\boldsymbol{a}^{i_{1:n}})$。这种跨智能体的有序决策设置简化了其联合策略的更新，其中最大化每个智能体自身的局部优势等同于最大化联合优势。因此，在策略更新过程中，智能体不再需要担心其他智能体的干扰；局部优势函数已经捕获了智能体之间的关系。优势分解定理揭示的这一特性并启发研究人员为多智能体强化学习问题提出了一种多智能体顺序决策范式，以任意决策顺序分配智能体（每次迭代一个排列）；每个智能体都可以访问其前辈的行为，然后在此基础上做出最优决策。这个顺序范例利用序列模型，例如 Transformer，来显式地捕获优势分解定理中描述的智能体之间的顺序关系。

对于多智能体使用这个 Transformer 架构的想法来源于这样一个事实：智能体的观察序列的输入之间的映射 $(o^{i_1},o^{i_2},...,o^{i_n})$ 和智能体动作序列的输出 $(a^{i_1},a^{i_2},...,a^{i_m})$ 是类似于机器翻译的序列建模任务。正如优势分解定理所描述的那样，行动目标 a^{i_m} 依赖于所有先前智能体的决策。

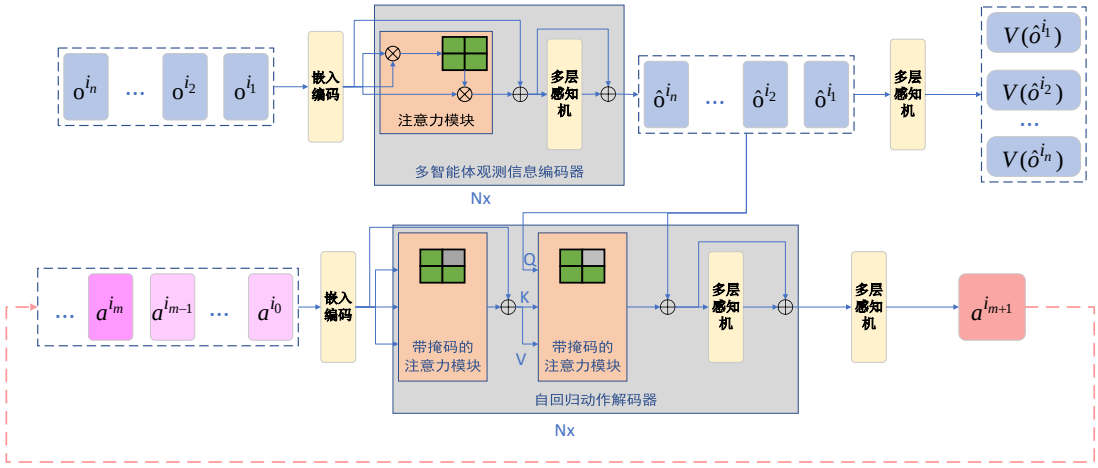


图 2-9 MAT 网络架构

Figure 2-9 Network architecture of MAT

如图 2-9 所示，MAT 网络框架由一个编码器（Encoder）和一个解码器（Decoder）组成，编码器学习联合观测的表示以及其状态价值函数，解码器以自回归的方式为每个单独的智能体输出动作。通过将智能体团队视为一个序列，Transformer 架构能够对具有可变数量和类型的智能体团队进行建模，并且智能体可以有目的地去进行协作。

2.5 本章小结

本章对后续章节涉及的深度强化学习理论、多智能体强化学习理论以及序列模型与强化学习的结合进行了阐述，为后续研究内容奠定了理论基础。关于

中国知网 https://www.cnki.net

深度强化学习理论，我们首先从基本的马尔可夫决策过程出发，研究了强化学习理论中基于值函数的方法、策略梯度方法、以及集两者所长的演员评论家方法，进一步研究了针对于复杂问题的分层强化学习方法的相关内容，并阐述了其中经典算法的原理。其次，研究了扩展到多智能体领域的强化学习，并详细阐述了两种经典的多智能体强化学习算法。最后，针对强化学习向更真实任务决策前进的重要伙伴序列模型，研究了经典的序列神经网络模型以及一种结合Transformer进行协作决策的智能体建模新范式。

中国知网 https://www.cnki.net

第3章 典型电子对抗任务系统模型设计

3.1 引言

本文针对典型电子对抗仿真决策中遇到的高维动作空间以及多智能体协作的问题展开了研究，为给后续章节的研究工作提供仿真实验所需的仿真环境系统模型，本章以多无人机协作护航任务为例，研究了典型电子对抗任务的系统模型设计。本章针对多无人机协作护航任务，对于其侦察任务模型、干扰任务模型以及作为马尔可夫决策过程的动作空间、观测空间以及奖励函数进行分析设计，最后阐述理想的任务执行流程。

3.2 典型电子对抗任务系统模型设计

现阶段，电子战策略研究主要集中在智能感知、干扰资源分配、认知干扰决策以及抗干扰决策等子领域，但针对于完整电子战 OODA 环路的复杂任务决策研究较少。无人机作战是未来战争的重要形式，无论是无人机电子对抗，还是多无人机协同控制，控制策略的优劣直接影响着多无人机完成任务的安全、稳定等效能。利用搭载在无人机上的统一小型化侦察和干扰模块，在敌方区域内进行无线电压制以生成可控区域是电子战发展的重点研究对象。因此，本文以多无人机协作护航任务这一典型电子对抗场景为例，进行环境建模，为后续采用多智能体强化学习算法进行策略探索研究做铺垫。

在多无人机协作护航任务场景中，多无人机需要在蓝方的电子侦察覆盖以及火力打击覆盖下，通过电子对抗的手段，削弱蓝方侦察能力，避免被蓝方火力打击，从而护送轰炸机抵近到可攻击距离，完成轰炸任务。

基于此，本文针对多无人机协作护航任务进行任务场景模型设计。环境包含红蓝两方，其中红色方由六架侦干一体无人机组成，该无人机同时配备侦察载荷以及干扰载荷，任务要求其在一定的时间和能量条件下，发现并干扰蓝方雷达，压制其侦察能力，为轰炸机安全抵近至攻击位置创造条件（轰炸机不作为智能体）；蓝方包含重要目标实体、雷达实体以及火力实体，七个雷达实体将对附近海域进行周扫侦察警戒，警戒范围将完全覆盖红方针对重要目标的可攻击位置，当发现敌方时，将引导临近的火力实体对敌进行火力打击。本文以无人机为智能体，该无人机同时配备有侦察和干扰载荷，即智能体同时具备认知侦察以及认知干扰能力，可以完成完整的电子战 OODA 环路，且多个无人机之间需要通过协作完成任务，对多智能体之间的协作性提出了挑战。下面将对多无人机协作护航任务场景模型进行进一步阐述说明。

中国知网 https://www.cnki.net

3.2.1 侦察任务

首先，红方无人机需要侦察获得蓝方雷达的位置、频段等信息，无人机采用定向侦察模式，该模式侦察距离远但范围小。当任意无人机在其侦察方向发现雷达时，整个无人机小组将得到一条侦察结果，视为对该雷达信息的一次捕获，雷达信息被捕获三次及三次以上即被视为红方完成对该雷达频段的确定。满足一定角度条件的两条侦察结果之间可进行交叉定位，从而精准定位雷达位置，完成红方对雷达信息的侦察任务。

在无人机不断侦察的过程中，系统模型对于已定位位置信息并截获频段信息的雷达进行面向红方的威胁度排序。多无人机协作护航任务的目的是开辟安全可攻击位置为红方对蓝方重要目标的打击作铺垫，因此蓝方雷达的威胁度排序依照的准则是雷达对于可攻击位置的覆盖程度，考虑的参数包括雷达探测范围对于可攻击位置区域的覆盖完整度以及可攻击位置区域与雷达的距离。雷达对可攻击位置区域的覆盖完整度越大，则该雷达威胁度越大；雷达距离可攻击位置区域越近，则压制其探测能力至不覆盖可攻击位置的难度越大，威胁度越大。

红方侦干一体无人机所搭载的侦察载荷采用无源侦察的方式，本身不发射电磁波，利用接收目标上电子信息设备的电磁辐射信号，对目标进行探测定位。其侦察距离满足

$$R_r = \left[\frac{P_t G_t G_r \lambda^2}{(4\pi)^3 P_{rmin}} \right]^{\frac{1}{2}} \tag{3-1}$$

其中，P_t 表示雷达发射功率，G_t 表示雷达天线增益，G_r 表示机载侦察载荷接收机天线增益，λ 表示雷达发射波长，P_{rmin} 表示机载侦察接收机灵敏度。

侦察概率满足下列公式：

设第 i 次观察发现目标的概率为 $q_i (i = 1, 2, \cdots, n)$，令随机变量 X 表示首次发现目标的观察次数（序号）。显然，在前 $n-1$ 次观察中未发现而恰在第 n 次观察发现目标的概率为

$$P(X = n) = q_n \prod_{i=1}^{n-1} (1 - q_i) \tag{3-2}$$

发现目标观察次数的期望值为

$$E(X) = \sum_{n=1}^{\infty} nP(X = n) \tag{3-3}$$

其方差为

$$\sigma_X^2 = \mathrm{E}\{[X - E(X)]^2\} = E(X^2) - E(X)^2 = \sum_{n=1}^{\infty} n^2 P(X = n) - E(X)^2 \tag{3-4}$$

中国知网 https://www.cnki.net

如果 q_i 可看为常数 $(q_i = q, i = 1,2,\cdots,n)$，即在搜索期间条件没有较大变化的情况下，则有

$$P(X = n) = q(1-q)^{n-1} \tag{3-5}$$

$$E(X) = \sum_{n=1}^{\infty} nq(1-q)^{n-1} = \frac{1}{q} \tag{3-6}$$

$$\sigma_X = \sqrt{\sum_{n=1}^{\infty} n^2 q(1-q)^{n-1} - \frac{1}{q^2}} = \frac{\sqrt{1-q}}{q} \tag{3-7}$$

即随机变量 X 服从几何分布。

为了进一步确定雷达的位置，每两条侦察信息之间进行交叉定位。侦察信息中包括无人机在某一位置向某一方向进行定向侦察获得雷达信号，针对同一雷达的两条满足一定角度条件的侦察信息即可实现该雷达的交叉定位，示意图如下所示：

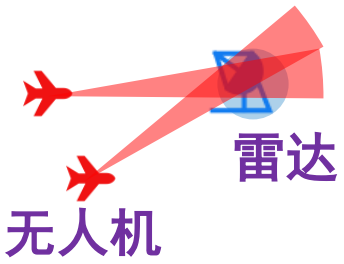


图 3-1 交叉定位示意图

Figure 3-1 Cross location diagram

3.2.2 干扰任务

其次，在多无人机协作护航任务中，多无人机需要以边侦察边干扰的方式来实现最终目标——开辟出供轰炸机投弹的安全可攻击位置，因此，无人机在得到一定侦察结果时，即可同步开始执行干扰任务。

电子干扰是使敌方电子设备和系统丧失或降低效能所采取的电波扰乱措施，是电子对抗的组成部分。其目的是削弱或破坏敌方使用各种电子设备和系统战场侦察、作战指挥、通信联络和兵器控制与制导的能力，为隐蔽己方企图和提高己方飞机、舰艇的生存能力创造有利条件。

在本文所设计的系统模型中，无人机所配备的电子干扰载荷主要影响蓝方雷达对于红方的探测距离和探测概率。

雷达对无人机的探测距离满足：

中国知网 https://www.cnki.net

$$R_{\max}=\left[\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 P_r}\right]^{\frac{1}{4}} \tag{3-8}$$

其中，σ 表示无人机在探测方向的截面积。

雷达对无人机的探测概率依据下列公式计算：

$$P_{d_i}=\int_0^{\infty} e^{-t}\left\{1-\varphi\left[\frac{Y_0-n_0(1+S_{n_i}t)}{\sqrt{n_0(1+2S_{n_i}t)}}\right]\right\}dt \tag{3-9}$$

在有干扰条件下，雷达发现概率 P_{d_i} 主要取决于信号能量和干扰能量之比的大小其中，函数 $\varphi(x)$ 的形式为

$$\varphi(\mathrm{x})=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mathrm{x}} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{dt} \tag{3-10}$$

其中，S_{n_i} 为第 i 次与目标接触时单个脉冲的信干比，计算式为

$$S_{n_i}=\frac{P_t G_t^2 \Delta f_j \sigma}{4\pi P_j G_j \gamma_j G_i(\theta)\Delta f_r L}\cdot\frac{R_j^2}{R_i^4} \tag{3-11}$$

其中，$G_i(\theta)$ 为雷达天线在干扰机方向上的增益，有

$$\mathrm{G_i}(\theta)=\begin{cases} G_t\left(0\le|\theta|\le\dfrac{\theta_{0.5}}{2}\right) \\ k\left(\dfrac{\theta_{0.5}}{\theta}\right)^2 G_t\left(\dfrac{\theta_{0.5}}{2}<|\theta|\le 90^\circ\right) \\ k\left(\dfrac{\theta_{0.5}}{90}\right)^2 G_t(90^\circ<|\theta|\le 180^\circ) \end{cases} \tag{3-12}$$

k 为常数。

3.2.3 动作、观测及奖励

本文所设计的多无人机协作护航任务模型中，将整个任务过程建模为分布式部分可观测马尔可夫决策过程（Decentralized Partially Observable Markov Decision Process, Dec-POMDP），马尔可夫决策过程的重要三元素即为智能体的观测空间、动作空间以及奖励函数。

首先是动作空间，为了减小策略模型学习的难度，本文的任务模型中采用离散动作空间，即智能体只需在有限个离散动作中进行选择做出决策。任务模型将侦干一体无人机视为一个智能体，多无人机即组成多智能体系统，多智能体需要在各自的动作空间中做出决策，从而组合成完整的联合动作。在多无人机协作护航任务中，智能体需要对无人机的飞行方向、飞行速度、定向侦察旋转方向、干扰目标、干扰强度以及干扰频段等做出决策。其中，

中国知网 https://www.cnki.net

1) 飞行方向离散为向东、向西、向南、向北、悬停（盘旋），共五个动作；
2) 飞行速度分为低速、中速以及高速，共三个动作；
3) 定向侦察旋转方向离散为左转、右转和不变，每次旋转的角度固定，即定向侦察的角度范围；
4) 干扰目标为七个选择动作（蓝方共七个雷达），分别对应蓝方威胁度排序中的排位，从第一位到第七位，共七个动作；
5) 干扰强度分为零度、轻度、中度、重度，其中零度表示不进行干扰，共四个动作；
6) 干扰频段将蓝方雷达的所有可能频段分为高、中、低三个频段，共三个动作。

在多无人机协作护航任务中，智能体在每个时间步需要同时对上述六种动作做出决策，经过排列组合，离散动作空间的维度达到了 3780，过于庞大的动作空间对于策略模型的训练是不利的。针对该问题，本文对动作空间进行精简。

可以发现，当干扰强度为零度时，即不进行干扰，那么选择干扰目标的动作就失去了意义。此外，在侦察任务中无人机通过对雷达信号的截获已经确定其工作频段，因此将干扰频段作为智能体决策的动作是非必要的。基于上述思考，对原有的动作空间进行了合理的合并、删减，将干扰频段的选择动作以规则的形式给出，将干扰强度为零度的动作进行合并，得出：

1) 飞行方向离散为向东、向西、向南、向北、悬停（盘旋），共五个动作；
2) 飞行速度分为低速、中速以及高速，共三个动作；
3) 定向侦察旋转方向离散为左转、右转和不变，每次旋转的角度固定，即定向侦察的角度范围，共三个动作；
4) 干扰动作分为零强度干扰，即不进行干扰动作以及第一到第七威胁度排序的干扰目标和弱度、中度、强度的干扰强度动作之间排列组合，共二十二个动作。

最终将离散动作空间维度精简至 990。

其次是观测空间，智能体决策模型是根据环境给出的观测向量为依据做出决策的，所以观测空间的合理性也是提升模型决策水平的关键因素。在马尔可夫决策过程中，智能体的观测向量需要体现出自身的变化以及环境的变化，使得智能体更好的认知，进而更准确的做出决策。本文所设计的观测向量分为两个部分：

其一，是智能体无人机自身的信息，包括自身的位置、当前定向侦察的角度；

其二，是蓝方相对于该无人机的信息，包括无人机距雷达的距离、无人机相对于雷达的角度以及雷达的受干扰情况，受侦察任务的影响，仅包含位于威胁度排序中的雷达；

中国知网 https://www.cnki.net

其中，为了使无人机具备一定的环境适应性，适应蓝方重点目标的不同位置，本文选择用相对坐标的形式，坐标原点为向重点目标直线飞行过程中第一次侦察到雷达信息的无人机位置。观测向量中的距离信息均采用欧氏距离，角度信息均采用以正北方向为 0 度，顺时针旋转的绝对角度。雷达的受干扰情况以雷达探测范围的被影响程度来表示。

最后是奖励函数，在强化学习中智能体策略学习以最大化累计奖励值为目标，如何合理设置奖励函数是智能体策略学习的重点关注问题。在本文所设计的多无人机协同护航任务模型中，采用团队奖励的形式，多智能体的奖励主要分为三个部分，接近奖励、侦察奖励以及干扰奖励。

1) 智能体需要接近到侦察距离才能侦察到蓝方雷达，蓝方雷达在干扰范围内才能被无人机干扰，因此需要鼓励无人机接近蓝方雷达，依据无人机接近的程度给予接近奖励，所有无人机的接近奖励相加即为团队接近奖励；

$$R_{\text{approach}}=\begin{cases}0.02x, & d>d_i\\ \max(0.01x,0), & d_i>d>0.6d_i\\ \max(x/0.6d_i,0)\times 0.5, & d<0.6d_i\end{cases} \tag{3-13}$$

其中，x 表示无人机当前时刻相比上一时刻行进的距离（若后退则为负），d 表示无人机与蓝方雷达的距离，d_i 表示无人机的最大干扰距离。

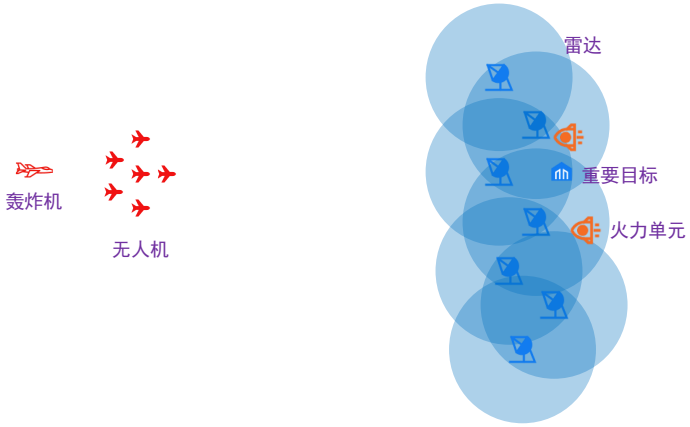
2) 多智能体相互协作下每捕获、定位到一个雷达，即可获取该雷达的位置、频段信息，可获得团队侦察奖励，每个雷达+1 分；

3) 智能体对雷达实施干扰，影响雷达的侦察能力，根据被影响的程度以及安全投弹位置的出现时间计算得出干扰奖励，每个时间步雷达探测距离与其最大探测距离相比减小的比例即为干扰奖励，当开辟出安全投弹位置并保持时，则额外按保持时间给予奖励；

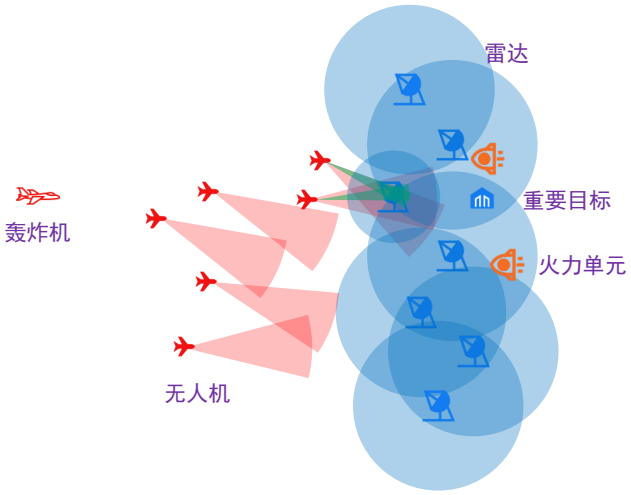
4) 无人机应在保全自身的情况下完成任务，避免无谓的牺牲，因此若红方无人机被蓝方雷达发现并击毁，则多无人机团队获得罚分，即负奖励。

中国知网 https://www.cnki.net

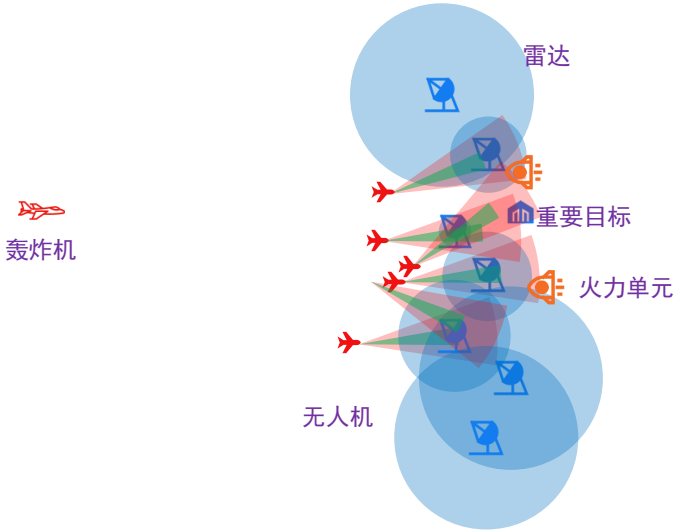
3.3 任务流程介绍



(a) 任务流程 1：无人机起飞并抵近蓝方雷达

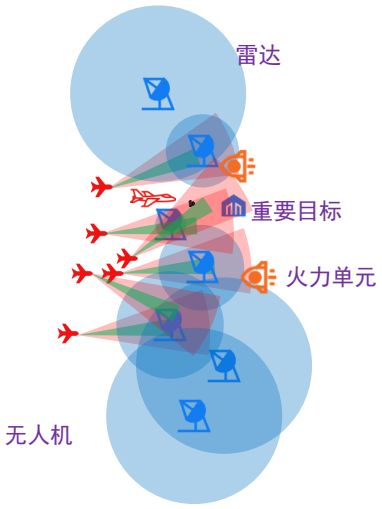


(b) 任务流程 2：无人机执行侦察、干扰任务



(c)任务流程 3：无人机开辟安全攻击位置

中国知网 https://www.cnki.net



(d) 轰炸机（NPC）进行打击

图 3-2 多无人机护航任务示意图

Figure 3-2 The ideal electronic warfare process of multi-UAVs escort mission

首先，对场景进行初始化，包括红方的六架无人机（智能体）以及一架轰炸机（NPC），无人机通过电子对抗手段开辟出安全区域帮助轰炸机完成打击任务，而蓝方则包括七个雷达以及两个火力单元守护在重要目标（target）附近，如图 3-2(a)。然后无人机对雷达进行侦察，并对侦察到的影响轰炸机打击目标的雷达进行干扰，如图 3-2(b)。当无人机侦察到所有侦察范围覆盖攻击点位的雷达时，无人机干扰有威胁的雷达抑制其探测，为轰炸机开辟出安全通路，如图 3-2(c)。最后由轰炸机抵达安全的攻击位置完成轰炸（这一步属于既定策略，以规则的形式实现），如图 3-2(d)，至此完成多无人机协作护航任务。在整个任务中，无人机也将面对被雷达侦察并打击的危险。

3.4 本章小节

本章对于后续研究内容仿真实验所需要的典型电子对抗场景——多无人机协作护航任务系统模型进行了设计，详细阐述了多无人机协作护航任务中的侦察任务模型、干扰任务模型以及整个任务过程作为马尔可夫决策过程的动作空间设计、观测空间设计、奖励函数的设计等。最后，描述了整个任务的理想流程。本章所描述的场景设计属于较为宏观的设计，没有涉及具体的控制参数，对比真实场景仍存在差距，未来可采用更接近真实的模型进一步优化任务场景，论文关注的重点是保证环节的完整性，为后续研究工作的开展打下了基础。

中国知网 https://www.cnki.net

第4章 基于子任务分解的多智能体强化学习决策方法

4.1 引言

本文所设计的多无人机协作护航任务具备由多种类动作排列组合形成的智能体的高维动作空间且包含侦察、干扰多个任务目标，属于复杂决策任务。首先，当任务中的智能体需要在高维动作空间中进行决策时，高维的动作空间容易形成维度灾难，不利于决策模型的优化，极大地增加了强化学习决策模型学习的难度。其次，当智能体需要同时决策多个目标的问题时，这些目标相互之间对于动作的需求可能存在冲突。如何平衡不同目标间的权重、优先级和折衷是一个挑战，特别是在多无人机协作护航任务这种动态环境中，目标的重要性可能会随时间变化。多无人机协作护航任务中的侦察结果虽然是干扰任务的基础，但也是影响侦察和干扰两个任务目标之间权重平衡的重要因素，在状态空间中很可能会存在某一状态下倾向于侦察目标而不适合干扰目标的情况，存在潜在的冲突风险。

为了使得智能体学习到更优的策略，在进行智能体策略推演学习之前，通过对本文所设计的多无人机协作护航任务进行分析，受分层强化学习中子任务分解以及强化学习值分解方法的启发，将完整的多无人机协作护航任务分解为两个平行的子任务，即侦察子任务和干扰子任务，将潜在的冲突风险规避。针对两个子任务分别训练不同的决策智能体，再根据任务的执行情况综合决策结果，从而避免了不同任务目标之间可能存在的冲突。通过层级化的方式，智能体能够更高效地探索环境，并利用抽象层次上的信息减少搜索空间，降低了智能体学习策略的难度，从而更快地找到最优解决方案。

4.2 子任务分解及综合决策

4.2.1 子任务与动作的对应关系

在多无人机协作护航任务系统模型设计中，共存在三种奖励值来源，分别为接近奖励、侦察奖励和干扰奖励，这也是在完整任务中，期望智能体达成的三个目标——接近蓝方、侦察蓝方以及干扰蓝方的最直接体现。

以分层强化学习的视角来看，当一个智能体处于一个包含多个特殊房间和通道的迷宫中寻找出口时，可以将整个寻找出口的任务分解为在每个房间的寻找出口（可能是迷宫的出口，也可能是通往另一个房间的门）的子任务，每个子任务学习不同的子策略。当智能体处于某一房间中时，由高层次策略决策在该房间中智能体执行哪一个子策略，低层次策略则调用相应的子任务策略模型

中国知网 https://www.cnki.net

进行决策。通过子任务独立学习子策略的方式，减小了子策略的探索空间，智能体在子任务中能够更高效地探索环境，从而更快地找到最优解决方案。智能体完成一个接一个的子任务，进而最终完成寻找出口的任务。基于分层强化学习的思想，理应将多无人机协作护航任务分解为飞行子任务、侦察子任务以及干扰子任务，通过在子任务中独立学习子策略，然后由高层次策略根据当前状态决策执行某一子策略的方式来完成任务。

但是，经过对多无人机协作护航任务的分析发现，在侦察任务侦察结果的判定中，需要依赖于无人机的位置、无人机侦察载荷的侦察方向以及蓝方雷达的位置来进行判定，同样在后续进一步精确定位的交叉定位中，也依赖于两条侦察信息中的无人机位置以及无人机在该位置的侦察方向。因此，无人机的侦察任务是无法脱离无人机的位置而独立决策的，也就是说，无人机的侦察任务中必须包含无人机的飞行动作以控制无人机到达更有利于执行侦察任务的位置，两者具备较强的相关性。

在干扰任务干扰结果的计算中，需要根据无人机与干扰目标之间的距离、无人机所采用的干扰强度等进行计算。其中，作为干扰目标的雷达位置固定，所以影响无人机与干扰目标之间距离的就是无人机自身的位置，无人机距离干扰目标越近则干扰效果越强，因此，无人机的干扰任务也是无法脱离无人机的位置而独立决策。也就是说，无人机的干扰任务中必须包含无人机的飞行动作以控制无人机到达更有利于执行干扰任务的位置，两者具备较强的相关性。

由上述分析可知，根据智能体无人机的三大类动作（飞行动作、侦察动作以及干扰动作）划分，分为飞行任务、侦察任务、干扰任务三个子任务是不合理的，严重割裂了动作之间的相关性，且不符合任务要求的每个时间步同时执行三种动作的条件。因此，在保证动作之间的相关性的同时，将完整的多无人机协作护航任务分解为侦察子任务和干扰子任务，并且分别训练得到相应的子策略，任务执行过程中同时应用两个子任务的子策略保证每个时间步执行完整的动作，对于重合的决策内容则通过高层次的策略模型进行整合，从而实现完整的多无人机协作护航任务执行策略。

4.2.2 侦察子任务

侦察子任务的目标是无人机通过自主的飞行和定向侦察方向的调整，直接或间接地与其他智能体协作侦察，实现对蓝方雷达的侦察定位以及捕获。作为多无人机协作护航任务的子任务，需要根据实际的决策需求从完整任务中拆解设计出侦察子任务作为马尔可夫决策过程的三要素——观测/状态空间、动作空间以及奖励函数。

中国知网 https://www.cnki.net

4.2.2.1 观测/状态空间

侦察子任务是一个环境探索任务，对于环境的情况是未知的，因此，更加关注于自身的观测信息，配合相应的奖励函数反馈，从而进行策略的学习。在侦察子任务的任务过程中，每个时间步智能体无人机需要关注其自身的位置以及当前的侦察载荷所定向侦察的方向，并依此做出决策。其中，无人机自身的位置为相对坐标位置，坐标原点为解算出的多无人机向重点目标直线飞行过程中第一次侦察到雷达信息的位置，侦察方向采用以正北方向为 0 度，顺时针旋转的绝对角度。

4.2.2.2 动作空间

考虑到侦察任务中侦察动作与飞行动作之间较强的关联性，采用飞行动作加侦察动作的组合作为侦察子任务的动作空间，即

1) 飞行方向动作为向东、向西、向南、向北、悬停（盘旋），共五个动作；
2) 飞行速度分为低速、中速以及高速，共三个动作；
3) 定向侦察旋转方向动作为左转、右转和不变，每次旋转的角度固定，即定向侦察的角度范围，共三个动作；

将上述动作进行排列组合，形成侦察子任务的离散动作空间，动作空间维度为 45。

4.2.2.3 奖励函数

从侦察子任务的动作空间分析，通过飞行动作可控制无人机的位置获取接近奖励，通过侦察动作可控制无人机侦察载荷的定向侦察方向获取侦察信息，进而多无人机团队协作完成定位和捕获获取侦察奖励。因此，侦察子任务的奖励函数由接近奖励和侦察奖励两部分组成。

1) 为了更好地引导无人机接近到侦察距离，根据蓝方雷达的位置及无人机侦察载荷的侦察距离，绘制出无人机可获得侦察信息的区域，如图 4-1 所示（图中的无人机侦察距离与雷达探测距离比例不准确，仅为示意清晰），其中蓝色区域为雷达探测区域，绿色区域表示无人机可获得侦察信息的区域，红色的弧线则被视为无人机为完成侦察任务必须抵近到的距离线，以 d_s 表示。

根据下列公式计算接近奖励：

$$R_{approach}=\begin{cases}0.02x, & d>d_s\\ \max(0.01x,0), & d_s>d>0.8d_s\\ 0, & d<0.8d_s\end{cases} \tag{4-1}$$

其中，x 表示无人机当前时刻相比上一时刻行进的距离（若后退则为负），d 表示无人机与蓝方雷达的距离，在抵近到最大侦察距离时即可获取侦察信息，

中国知网 https://www.cnki.net

可以适当前进一些以保证侦察效果，但不鼓励过度接近雷达，过度接近雷达反而限制了无人机的侦察范围。

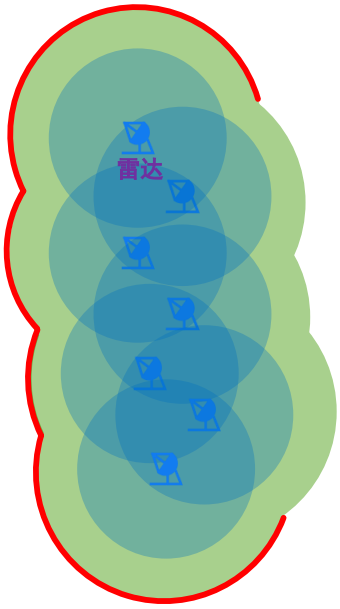


图 4-1 侦察距离示意图

Figure 4-1 Reconnaissance range diagram

2) 多无人机团队协作下每捕获、定位到一个雷达，即可获取团队侦察奖励，每个雷达+1 分；

4.2.3 干扰子任务

干扰子任务的目的是无人机通过自主的飞行和干扰目标、干扰强度的调整，与团队中其他智能体协同干扰，实现对蓝方雷达探测能力的压制以开辟出针对蓝方重要目标的安全可攻击位置。作为多无人机协作护航任务的子任务，需要根据实际的决策需求从完整任务中拆解设计出干扰子任务作为马尔可夫决策过程的三要素——观测/状态空间、动作空间以及奖励函数。

4.2.3.1 观测/状态空间

干扰子任务是一个已知环境下的实时决策任务，智能体对于环境的情况是已知的，来源于侦察的结果。因此，在关注自身信息的同时，还需要关注环境的变化情况，特别是蓝方雷达探测能力的情况，配合相应的奖励函数反馈，进行干扰策略的学习。在干扰子任务的任务过程中，每个时间步智能体无人机根据其自身的位置、当前环境状态中与蓝方雷达的距离、角度以及蓝方雷达探测能力被压制程度做出决策。其中，无人机自身的位置为相对坐标位置，坐标原点为解算出的多无人机向重点目标直线飞行过程中第一次侦察到雷达信息的位置，与蓝方雷达的距离采用欧式距离，角度采用以正北方向为 0 度，顺时针旋

中国知网 https://www.cnki.net

转的绝对角度，探测能力被压制的程度采用当前探测距离与最大探测距离的差与最大探测距离的比值。

4.2.3.2 动作空间

考虑到干扰任务中干扰动作与飞行动作之间较强的关联性，采用飞行动作加干扰动作的组合作为干扰子任务的动作空间：

1) 飞行方向动作为向东、向西、向南、向北、悬停（盘旋），共五个动作；
2) 飞行速度分为低速、中速以及高速，共三个动作；
3) 干扰动作分为零强度干扰，即不进行干扰动作以及第一到第七威胁度排序的干扰目标和弱度、中度、强度的干扰强度动作之间排列组合，共 22 个动作。

将上述动作进行排列组合，形成干扰子任务的离散动作空间，动作空间维度为 330。

4.2.3.3 奖励函数

从干扰子任务的动作空间分析，通过飞行动作可控制无人机的位置获取接近奖励，通过干扰动作可控制无人机的干扰目标和干扰强度，进而多无人机团队协作完成蓝方雷达探测能力的压制，开辟出针对蓝方重要目标的安全可攻击位置以获取干扰奖励。因此，干扰子任务的奖励函数由接近奖励和干扰奖励两部分组成。

1) 为了更好地引导无人机接近到干扰距离，根据蓝方雷达的位置及无人机干扰载荷的干扰距离，绘制出无人机可干扰雷达压制其探测能力的区域，如图 4-2 所示，其中橙黄色圆圈的半径为无人机干扰载荷最大作用距离，以 d_i 表示，无人机在该范围内至少可对一个雷达进行干扰。

根据下列公式计算接近奖励：

$$R_{approach}=\begin{cases}0.02x, & d>d_i\\ \max(0.01x,0), & d_s>d>0.6d_i\\ \max(x/0.6d_i,0)\times 0.5, & d<0.6d_i\end{cases} \tag{4-2}$$

其中，x 表示无人机当前时刻相比上一时刻行进的距离（若后退则为负），d 表示无人机与蓝方雷达的距离，无人机距离干扰目标越近，则干扰效果越好，但过于接近时容易被敌方发现，因此鼓励智能体有限度的靠近雷达。

中国知网 https://www.cnki.net

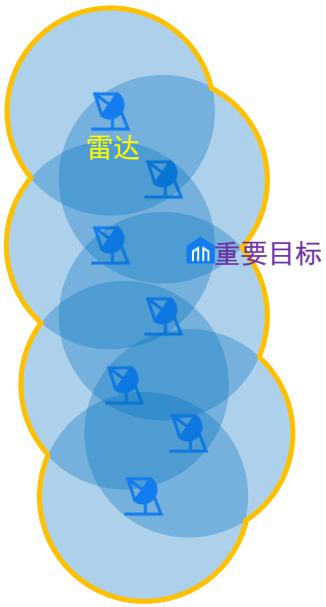


图 4-2 红方可干扰范围示意图

Figure 4-2 Interference range diagram of red

2) 智能体无人机对雷达实施干扰，影响雷达的探测能力，根据被影响的程度计算干扰奖励，每个时间步雷达探测距离与其最大探测距离相比减小的比例即为干扰奖励，当开辟出安全投弹位置并保持时，则额外按保持时间给予奖励；

4.2.4 综合决策

在完成侦察子任务和干扰子任务的子策略学习之后，需要将两个策略得出的结果综合起来，以实现完整的动作决策。由于两个子策略都对于无人机的飞行动作做出了有利于自身目标的决策，因此，需要将两者的飞行动作决策有机结合起来。本文所采用的综合依据是整体任务的进展，重点关注侦察任务的结果，即侦察到的雷达信息。多无人机协作护航任务系统模型中，蓝方雷达的数量是有限的，侦察的结果是执行干扰任务的基础，随着侦察任务的进展不断增加，干扰任务的权重也应该不断增加。

本章算法采用如下图所示的方法进行结合侦察子策略和干扰子策略的综合决策。

中国知网 https://www.cnki.net

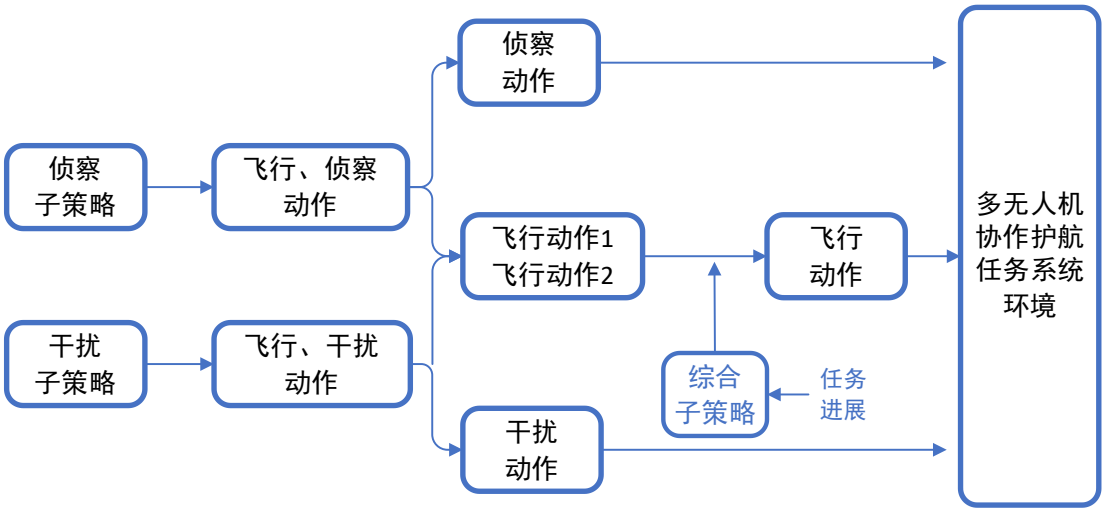


图 4-3 基于子任务分解的多智能体强化学习决策流程

Figure 4-3 Multi-agent reinforcement learning decision process based on subtask decomposition

完整策略由综合子策略、侦察子策略以及干扰子策略组成。首先调用侦察子策略和干扰子策略，分别决策出飞行与侦察的动作组合和飞行与干扰的动作组合，针对重复决策的飞行动作部分，由综合子策略根据任务进展情况决策两飞行动作的权重并综合，再加上独立的侦察动作和干扰动作，组成完整的智能体动作给到多无人机协作护航任务系统模型进行策略推演。

4.2.5 算法流程

本章算法的策略模型均采用经典的 AC（Actor-Critic）架构，如图 4-4 所示，分别为 Actor（演员）网络和 Critic（评论家）网络设计了两层隐藏层，其中第一层由 128 个神经元组成，第二层由 64 个神经元组成。Actor 网络输入当前时刻的状态/观测向量 S_t，通过一个 sigmoid 激活函数输出最终决策结果，即动作 a_t，与环境交互得到即时奖励 r_t 和下一时刻状态 S_{t+1}，以 (S_t, a_t, r_t, S_{t+1}) 作为样本存入经验池（buffer）。神经网络参数的更新采用蒙特卡洛方法，存储每个 episode 的样本到容量为一个 episode 长度的经验池中，在每个 episode 结束时取出 buffer 中的样本进行训练。由于多无人机协作护航任务是一个多智能体任务，Critic 网络以全局的状态信息作为输入，输出对全局状态的价值评估，并计算得出优势函数 $\hat{A}_t$，通过优势函数来指导 Actor 网络向最大化累计奖励的方向进行优化。掌握全局信息的 Critic 网络会将多智能体之间的协作性纳入考虑，是实现多智能体协作的关键。

中国知网 https://www.cnki.net

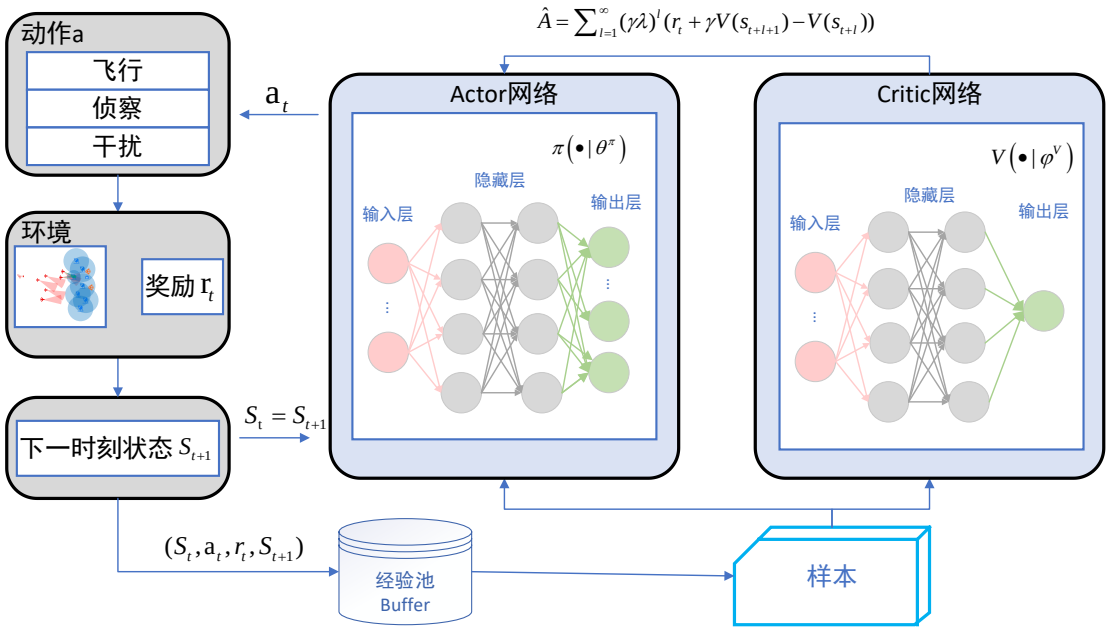


图 4-4 TDPA 算法的基础算法框架

Figure 4-4 The basic algorithm framework of TDPA

其中，Actor 网络采用重要性采样配合近端优化的方式进行策略更新，损失函数为：

$$L^{clip}(\theta)=\mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t,clip(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\right)\right] \tag{4-3}$$

$$r_t(\theta)=\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{old}}(a_t\mid s_t)} \tag{4-4}$$

Actor 网络需要输出的动作优势尽可能大，上式中优势函数 $\hat{A}_t$ 的计算依赖于 Critic 网络的评估，

$$\hat{A}_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T+1} \tag{4-5}$$

$$\delta_t=r_t+\gamma V(s_t+1)-V(s_t) \tag{4-6}$$

Critic 网络需要更准确地评估当前的状态价值，折扣回报 G_t 为无偏的状态价值，

$$G_t=r_{t+1}+\gamma r_{t+2}+\cdots+\gamma^{T-t}r_{T+1} \tag{4-7}$$

Critic 网络以样本中 $V_{真实}=G_t$ 和 $V_{估计}$ 的均方误差作为损失函数，

$$L(\varphi)=\frac{1}{N}\sum_{i=1}^{i=N}\left(G_t-v_{s_t}\right)^2 \tag{4-8}$$

虽然进行了任务分解，但仍在同一个任务环境中进行子策略的训练以保证子策略的适应性。训练侦察子策略时，侦察智能体仅决策飞行和侦察动作，干扰动作固定为不进行干扰，此时的干扰奖励为0，不会影响侦察子策略的训练。训练干扰子策略时，干扰智能体仅决策飞行和干扰动作，其侦察动作固定为不

中国知网 https://www.cnki.net

旋转，同时在环境中将侦察相关的解算包括侦察奖励的计算都屏蔽以保证不影响干扰子策略的训练。

此外，侦察结果是进行干扰任务的基础，为了使得干扰子策略对于不同的侦察结果具备适应性，在干扰子任务初始化时将按照一定的比例随机初始化当次干扰任务的侦察结果。

本章算法的伪代码如下：

表 4-1 TDPA 算法伪代码（子策略）

Table 4-1 Pseudocode of TDPA (Sub-policy)

算法：TDPA 算法（Task-decomposition partial assemble algorithm）（子策略）	
1:	初始化经验池，其容量为 episode 的长度；
2:	随机初始化 Actor 网络参数 θ^{π} 和 Critic 网络参数 φ^{V}；
3:	循环 总 episode 数：
4:	初始化状态 $\boldsymbol{s}_1$；
5:	循环 episode 长度（时间步），$t=1,2,\cdots,K$：
6:	对于智能体 $i=1,\cdots,m$ 执行：
7:	根据智能体 i 的局部观测 o_t^i 选择动作 $a_t^i=\pi_i(o_t^i\mid\theta^{\pi})$；
8:	执行联合动作 $\boldsymbol{a}_t=(a_t^1,a_t^2,\cdots,a_t^m)$ 获得奖励 r_t 和下一时间步状态 $\boldsymbol{s}_{t+1}$；
9:	将经验样本 $(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})$ 存储到经验池中；
10:	计算每一个时间步的折扣回报 G_t；
11:	从经验池 $\mathcal{D}$ 中随机采样 n 组样本经验；
12:	利用式(4-3)计算损失函数 $L(\theta)$ 并更新 Actor 网络；
13:	利用式(4-8)计算损失函数 $L(\varphi)$ 并更新 Critic 网络；

表 4-2 TDPA 算法（综合策略）

Table 4-2 Pseudocode of TDPA (Integrated Strategy)

算法：TDPA 算法（综合策略）	
1:	初始化经验池，其容量为 episode 的长度；
2:	随机初始化 Actor 网络参数 θ^{π} 和 Critic 网络参数 φ^{V}；
3:	初始化指定的侦察子策略 π_{scout} 和干扰子策略 π_{jam}；
4:	循环 总 episode 数：
5:	初始化状态 $\boldsymbol{s}_1$；
6:	循环 episode 长度（时间步），$t=1,2,\cdots,K$：

中国知网 https://www.cnki.net

7:	对于智能体 $i=1,\cdots,m$ 执行：
8:	根据智能体 i 的局部观测 o_t^i 选择综合动作 $a_t^{i-assmeble}=\pi_i(o_t^i \mid \theta^\pi)$；
9:	调用侦察子策略 π_{scout} 根据观测 o_t^i 选择侦察动作 $a_t^{i-scout}$；
10:	调用干扰子策略 π_{jam} 根据观测 o_t^i 选择干扰动作 a_t^{i-jam}；
11:	将 $a_t^{i-scout}$ 和 a_t^{i-jam} 根据 $a_t^{i-assmeble}$ 合成为完整的 a_t^i；
12:	执行完整联合动作 $\boldsymbol{a}_t=(a_t^1,a_t^2,\cdots,a_t^m)$ 获得奖励 r_t 和下一时间步状态 $\boldsymbol{s}_{t+1}$；
13:	将经验样本 $(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})$ 存储到经验池中；
14:	计算每一个时间步的折扣回报 G_t；
15:	从经验池 $\mathcal{D}$ 中随机采样 n 组样本经验；
16:	利用式(4-3)计算损失函数 $L(\theta)$ 并更新 Actor 网络；
17:	利用式(4-8)计算损失函数 $L(\varphi)$ 并更新 Critic 网络；

4.3 实验结果与分析

为了分析验证本章算法各个环节的有效性，本节进行了侦察子策略、干扰子策略以及完整策略与现有算法决策效果的比较实验。在仿真训练中，我们设置总的仿真训练步数为 5000000，episode 长度为 1000，并行线程数为 16，每次训练神经网络时，从容量为一个 episode 长度的经验池中多次随机抽取 256 组样本进行批量训练。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。

4.3.1 侦察任务比较实验

首先，图 4-5 给出了本章所提出的 TDPA 算法在侦察子任务中的收敛特性。每个 episode 结束后对神经网络进行训练并记录数据，训练时随机提取多组样本多次训练，以提升样本利用率。从图中可以观察到，运行1e6 左右时间步后收敛。

中国知网 https://www.cnki.net

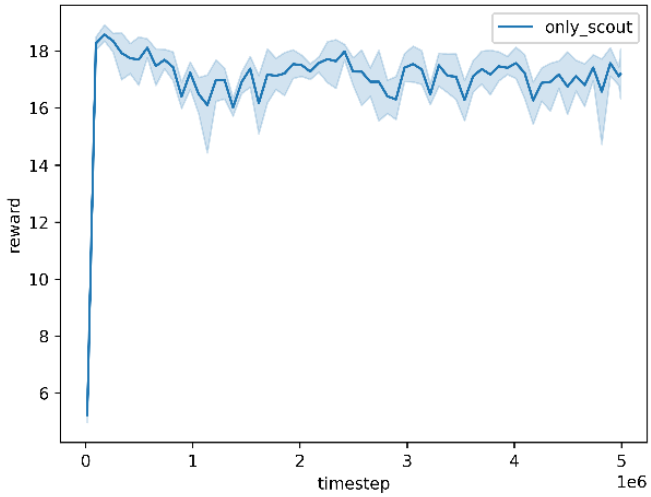


图 4-5 TDPA 算法中侦察子策略的收敛特性

Figure 4-5 Convergence characteristics of scout strategy in TDPA algorithm

表 4-3 是 MADDPG 算法在完整任务中训练所得策略与 TDPA 算法侦察子策略在侦察方面的性能对比。实验采用 MADDPG 的策略和 TDPA 算法的侦察子策略分别在环境中以随机策略的形式执行 1000 次任务，记录任务过程数据并统计分析。其中，侦察子策略执行任务时固定干扰动作为不执行干扰。

表 4-3 侦察子任务下不同策略性能比较

Table 4-3 Performance comparison of different strategies in the reconnaissance subtask

算法	侦察结果个数（侦察完整度）	平均侦察时间步数	侦察效率	侦察协作率
TDPA-scout	7（100%）	35	98.3%	29%
MADDPG	2.54（36.2%）	76	96.9%	2.1%
MAPPO	3.34(47.7%)	87	97.6%	33.2%

表 4-3 中的数据为多次多无人机协作护航任务执行的平均数据，侦察到一个雷达指多无人机通过获取的侦察信息完成了对该雷达的交叉定位以及频段捕获。其中，侦察结果个数为多无人机智能体侦察到的蓝方雷达数，侦察完整度为侦察到的雷达个数与蓝方雷达总数的比值，平均侦察时间步数为平均每侦察到一个雷达所需要的步数，侦察效率是指多无人机获取的侦察信息成功被用来交叉定位和频段捕获的比例，侦察协作率指多无人机交叉定位成功判定中的两条侦察信息分属于两个不同的智能体无人机。

由表中数据可知，专注于侦察任务的 TDPA 算法侦察子策略在各项侦察指标中相比于 MADDPG 算法均有不同程度的提升，平均侦察个数提升了 4.46，侦

中国知网 https://www.cnki.net

察完整度提升了63.8%，平均侦察时间步数提升了53.9%，侦察效率提升了1.4%，侦察协作率提升了 27%，而与 MAPPO 算法相比，平均侦察个数提升了 3.64，侦察完整度提升了 52.3%，平均侦察时间步数提升了 59.7%，侦察效率提升了0.7%，侦察协作率略低但相差不大。可以看出，通过将侦察子任务分解出来进行单独学习的方式，缓解了高维动作空间带来的维度爆炸问题，使得智能体更易学习到更好的策略。此外，专注于单一任务也使得智能体之间的协作目标更加明确，协作效率得到了保证。

4.3.2 干扰任务比较实验

其次，图 4-6 给出了每个 episode 的累计奖励值随总执行步数的变化，是本章所提出的 TDPA 算法在干扰子任务中的收敛特性。每执行 episode 长度的时间步则对神经网络进行训练并记录训练数据。

从图中可以观察到，策略明显陷入了局部最优的问题，reward 曲线在1e6时达到峰值，但之后开始下降，直到2e6左右时间步后收敛。从任务执行过程来看，智能体无人机没有保持在一个相对合理安全的位置进行干扰，而是冒着被打击的风险不断逼近雷达区域，导致无人机被击毁的概率大大提高，从而影响了干扰奖励的获取以及任务的完成。

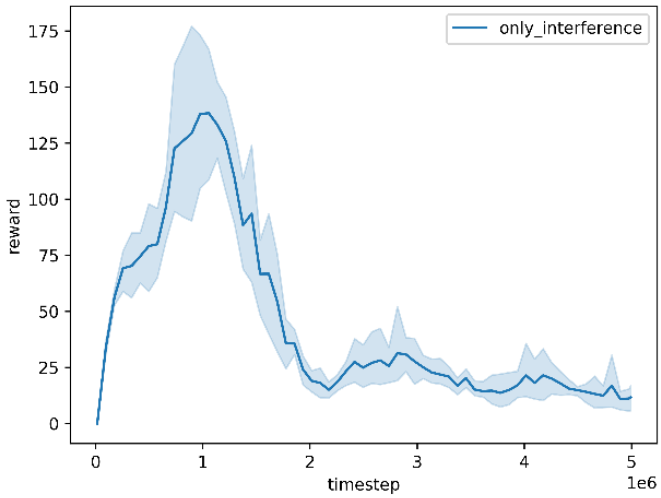


图 4-6 TDPA 算法中干扰子策略的收敛特性

Figure 4-6 Convergence characteristics of interference sub-strategies in TDPA algorithm

表 4-4 是 MADDPG 算法在完整任务中训练所得策略与 TDPA 算法干扰子策略在干扰任务上的性能对比。实验采用 MADDPG 的策略和 TDPA 算法的干扰子策略分别在纯干扰任务环境中以随机策略的形式执行 1000 次任务，记录任务过程数据并统计分析。其中，纯干扰任务不响应侦察动作，且关于侦察任务的所有判定均被屏蔽，也不会产生不为 0 的侦察奖励。

中国知网 https://www.cnki.net

表 4-4 干扰子任务下不同策略性能比较

Table 4-4 Performance comparison of different strategies in jamming subtasks

算法	平均干扰效能（最大干扰效能）	持续干扰时间	安全区域开辟时间	干扰协作率
TDPA-jamming	27.66（56.52）	72.86	1.65	8.5%
MADDPG	1.61（4.21）	11	0	0
MAPPO	21.6（48.41）	52.16	1.4	20.9%

表 4-4 中的数据为多次多无人机协作护航任务执行的平均数据，其中，干扰效能是整个任务过程中蓝方雷达探测距离受干扰缩减比例，表中统计了平均值和最大值，持续干扰时间是指多无人机对雷达进行成功干扰的时间步数，干扰的最终目的是使得红方设定的投弹位置脱离雷达的探测范围，即开辟出安全区域，安全区域开辟时间是指在多无人机的干扰压制下安全区域的出现时间，干扰协作率是指多无人机合作干扰同一雷达的时间步数占总干扰时间的比例。

由表中数据可知，MADDPG 算法对于干扰任务贡献很小，而与表现更好的 MAPPO 算法相比，专注于干扰任务的 TDPA 算法干扰子策略在干扰协作率不佳的情况下，仍在各项干扰指标中取得了不同程度的提升，平均干扰效能提升了 6.06，最大干扰效能提升了 8.11，持续干扰时间提升了 20.7，安全区域开辟时间提升了 0.25。可以看出，当智能体专注于目标更明确的子任务时更容易学习到更优的策略。

4.3.3 完整电子对抗任务比较实验

最后，图 4-7 给出了本章所提出的 TDPA 算法在完整的多无人机协作护航任务中的收敛特性。图中给出了每个 episode 的累计奖励值随总执行步数的变化。其中，TDPA 算法调用了之前单独训练的侦察子策略和干扰子策略，通过动作综合策略合成完整的多无人机动作并给到任务环境执行，并且分别调用了 1*e*6 时间步和最终 5*e*6 时间步时的干扰子策略模型进行比较。

由于有已经完成策略学习的两个子策略的加持，整体策略一开始就处于一个比较高的决策水平，并在之后的策略更新中为了平衡两个子任务整体累计奖励曲线收敛略有下降。虽然在 1*e*6 时间步时的干扰子策略的累计奖励值达到了峰值，但模型没有收敛其随机性比较大，仍旧是采用收敛后的干扰子策略模型效果要更好。

中国知网 https://www.cnki.net

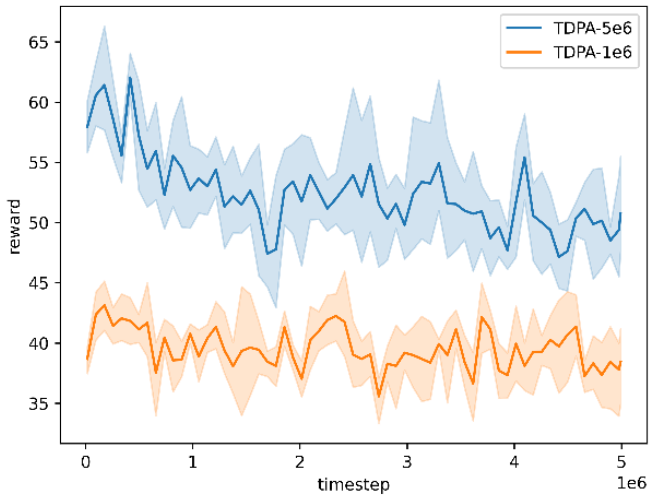


图 4-7 TDPA 算法综合策略的收敛特性

Figure 4-7 Convergence characteristics of TDPA algorithm synthesis strategy

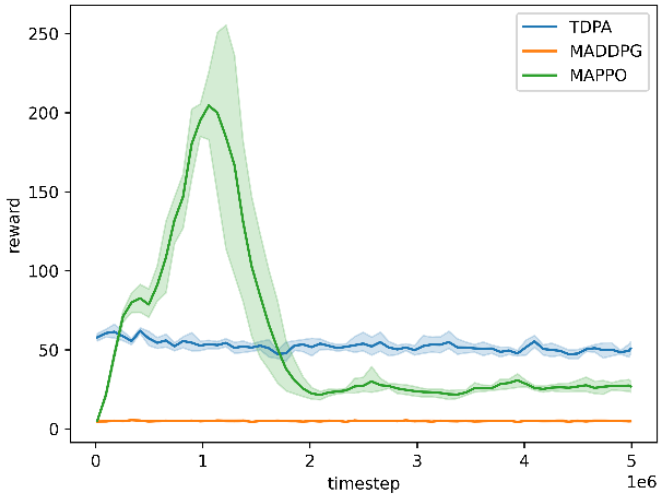


图 4-8 TDPA 算法与现有算法的累计奖励曲线比较

Figure 4-8 Comparison of cumulative reward curve between TDPA algorithm and existing algorithm

图 4-8 是本章 TDPA 算法与现有的 MAPPO、MADDPG 算法在同样的完整多无人机协作护航任务中的训练表现，每个 episode 的累计奖励值随总执行步数的变化。可以看到，分布式的MADDPG算法在多无人机协作护航任务中效果不佳，其策略探索不足，没有去进一步接近蓝方雷达，仅在初始点附近探索侦察，故整体累计奖励较低且没有较大的起伏；而 MAPPO 算法探索较为充分，不断靠近蓝方雷达区域以获取更高的接近干扰奖励，从而使得奖励曲线大幅提升，但当接近到一定程度时，智能体无人机面临被打击的风险，被打击的智能体将获得高额的负奖励，因此造成了奖励曲线的急速下降，最终在被打击的风险下

中国知网 https://www.cnki.net

收敛于局部最优；相比之下，虽然 TDPA 算法在干扰子策略的训练中也遇到了类似的问题，但在整合两个子策略后，更优的侦察策略保证了更优的累计奖励曲线收敛结果，最终收敛结果要优于 MAPPO、MADDPG 算法。

表 4-5 是 MADDPG 算法与 TDPA 算法在完整多无人机协作护航任务中学习的策略性能对比。实验采用两种算法分别在多无人机协作护航任务环境中以随机策略的形式执行 1000 次任务，记录任务过程数据并统计分析。

表 4-5 完整任务下不同策略性能比较

Table 4-5 Performance comparison of different policies in a full task

算法	侦察任务完成度	安全区域开辟时间	侦察协作率	干扰协作率
TDPA	0.78	1.2	51.5%	4.2%
MADDPG	0.36	0	2.1%	0
MAPPO	0.79	1.4	33.2%	20.9%

表 4-5 与表 4-3、表 4-4 中数据相比可以看出，TDPA 算法的完整策略相比于其专精某一任务的子策略来说，各项指标数据是下降的，这是由于子任务分解并综合决策的方法需要寻求两个子策略之间的一个平衡，平衡的代价就是互有牺牲，从降低的幅度来看，综合策略更倾向于侦察任务。尽管如此，与 MADDPG 算法相比，TDPA 智能体策略在侦察任务目标和干扰任务目标上仍有不同程度的提升。侦察任务的平均完成度提高了 0.42，干扰任务开辟安全区域的时间提高了 0.2，侦察和干扰的协作性也略有提升；与 MAPPO 算法相比，两者策略取得了相近的任务结果，但 TDPA 算法在侦察任务中的协作率要更加优秀，且在干扰协助较差的情况下取得了相近的安全区域开辟时间。可以看出，TDPA 算法通过任务分解的方法缓解了高维动作空间维度爆炸的问题，同时也缓解了复杂任务中多个任务目标多种 reward 来源造成的智能体策略学习困难的问题，有效提升了多智能体无人机对于任务的完成度。此外，相比于 MADDPG 算法分布式的智能体，TDPA 算法采用共享参数的形式，所有智能体采用同样的 Critic 网络，在同样的评估下进行学习提升了进行协作的可能性。从表中协作性数据可以看出，在专精于单一任务的子策略加持下，综合后的策略依旧对智能体之间的协作性有一定程度的保留，但完整任务中综合策略的协作性比之子任务中子策略的协作性略有下降。

4.4 本章小结

本章在前述对于多无人机协作护航任务的系统建模的基础上，对于多无人机团队进行协作电子对抗决策问题进行分析，基于分层强化学习子策略的思想，

中国知网 https://www.cnki.net

将完整的多无人机协作护航任务依据任务目标分解为侦察子任务和干扰子任务，对于动作与任务之间相关性的进行了分析，分别为子任务设计了动作空间、观测/状态空间以及奖励函数。在此基础上，提出了一种基于子任务分解的多智能体强化学习决策算法，算法分别在子任务中进行训练获取子策略，然后将子策略重复决策的部分通过综合策略进行综合，从而形成完整的对于多无人机协作护航任务的决策。仿真实验结果表明，TDPA算法有效地提升了多无人机策略对于任务目标的完成度，且共享参数的形式对于多智能体之间协作性的提升略有帮助。

中国知网 https://www.cnki.net

第5章 基于动作依赖的多智能体强化学习决策算法

5.1 引言

在多无人机协作护航任务中，前述所提出的 TDPA 算法策略通过任务分解的方法展现出了更优的任务完成效果以及更优的智能体协作性，但仍存在一定的不足。首先，该算法的子策略对于飞行动作进行了重复的决策，虽然通过综合策略将重复决策的部分进行了综合，但综合后的决策结果仍是倾向于某一子策略的，容易造成对另一动作的孤立甚至失效，进而影响整个策略的效果，容易陷入某个局部最优之中。其次，在子策略中由 Critic 网络价值评估引导下形成的智能体之间的协作性很容易被综合策略所破坏，尤其是被孤立的一方，智能体自身决策的不稳定加大了多智能体之间进行协作的难度，使得智能体之间难以生成有效的协作经验样本，对于引导多智能体向协作方向优化的 Critic 价值评估网络的学习造成了困难。

上述两个问题的核心点在于智能体自身决策的非稳态带来的影响，因此本文考虑加强多种类动作决策之间的关联性，不孤立某一类动作，在 Actor 网络进行动作决策时有意识地加入动作间协作性的考虑，稳定的自身决策可以提供更多智能体之间协作的经验样本供 Critic 网络学习，从而更好的引导 Actor 网络向多智能体协作的方向进行优化。基于此，本章提出了一种基于动作依赖的多智能体强化学习决策算法。

5.2 基于动作依赖的多智能体强化学习决策算法

在本章提出的基于动作依赖的多智能体强化学习决策算法中，受多智能体优势分解定理的启发，将多种类离散动作同时决策改为多种类的动作依顺序独立决策，后置动作的决策依赖于前置动作的决策结果，在后置动作的决策输入中加入前置动作，加强多种类动作决策的关联性，稳定自身决策策略，为智能体之间的协作奠定基础。

5.2.1 多种类动作间的决策依赖性分析

在第 4 章中进行子任务分解时，通过分析，发现无论是侦察动作还是干扰动作，都与无人机的位置有很强的相关性。换句话说，就是无论是侦察动作还是干扰动作，在进行决策时都需要考虑无人机自身的位置，而无人机执行飞行动作就会改变其位置，因此，无论是侦察动作还是干扰动作，都依赖于飞行动作的决策结果。

依托子策略进行综合决策时，无人机的决策会偏向于某一子策略，但被孤

中国知网 https://www.cnki.net

立的一方仍会执行预期的动作，但这一动作由于无人机位置的不理想已经属于无效动作，如图 5-1 左图所示，无人机选择了倾向于侦察动作的飞行动作，而该飞行动作下的新位置不适应预定的干扰动作，导致干扰动作因距离干扰目标过远而失效。相对应的，如果在进行干扰动作的决策时可以考虑飞行动作的决策结果，就可以避免动作失效的情况，如图 5-1 右图所示，无人机在决策干扰动作时考虑到了飞行动作的影响，故而选择了合理的干扰动作。

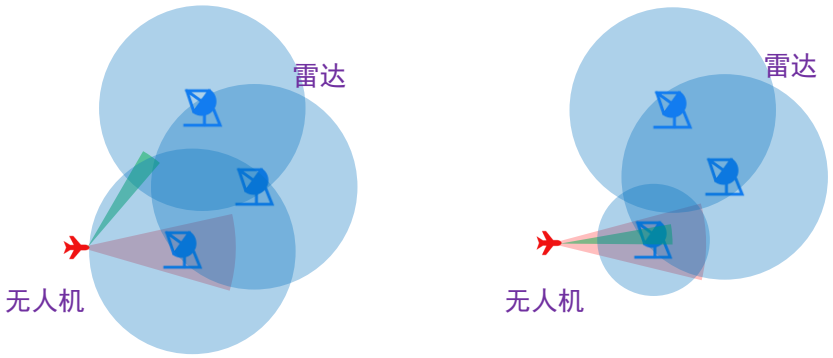


图 5-1 无人机决策场景示例

Figure 5-1 Example drone decision scenario

因此，当侦察动作和干扰动作依赖于同一飞行动作时可以保证没有被孤立的动作，可以进行更合理的决策。

5.2.2 双向动作依赖

在多智能体环境中，若全局状态为 $\boldsymbol{s}$，单个智能体 i 的局部观察为 o_i。由于智能体之间相互影响，因此智能体 i 在观察 o_i 下执行动作 a_i 后得的 r_i 与 o_i' 是由所有智能体的行为造成的，即 $r_i = R_i(\boldsymbol{s},\boldsymbol{a})$，$o_i' = T_i(\boldsymbol{s},\boldsymbol{a})$。所以对于单个智能体 i，即使在观察 o_i 下一直都执行动作 a_i，但是由于 $\boldsymbol{s}$ 未知且其他智能体的策略在不断变化，此时智能体 i 得到的 r_i 与 o_i' 可能是不同的（比如之前的 (o_i,a_i) 得到的奖励为 1，但是下一次由于队友的动作发生了变化，奖励又变成了-1）。这就是 MARL 中的不稳定性，即 reward 与 transition 存在不稳定性，破坏了强化学习算法遵循的马尔可夫假设，从而导致智能体 i 的值函数 $Q(o_i,a_i)$ 的更新十分不稳定。

双向动作依赖[113]就是来解决上述问题的，其核心思路是：不再要求多个智能体同时产生动作，而是把多智能体协作问题转化为一种稳定又高效的序列扩展马尔可夫决策过程（Sequentially Expanded Markov Decision Process, SE-MDP），建模智能体之间的双向动作依赖，抽象出最精简的协作表征，让每个智能体一个接一个地产生决策行为，最终将非平稳的多智能体决策问题转化为特殊的平稳单智能体决策问题。

中国知网 https://www.cnki.net

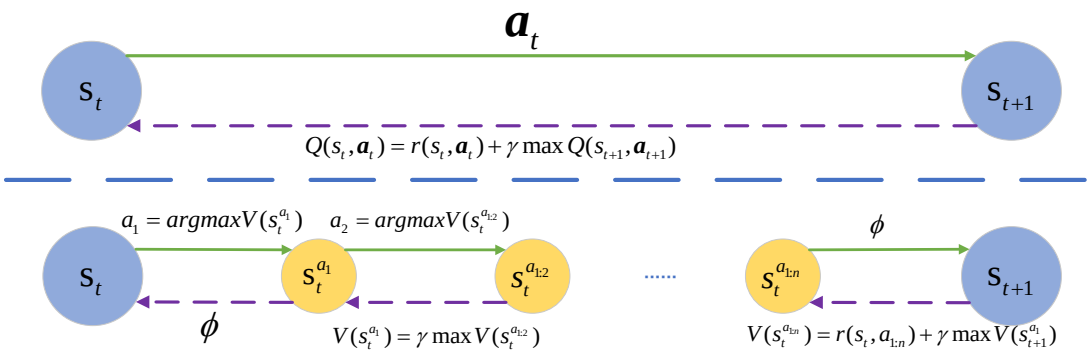


图 5-2 原始 MMDP 和转换后 SE-MDP 之间的比较

Figure 5-2 Comparison between the original MMDP and the converted SE-MDP

如图 5-2 所示，以 Q-learning 算法为例，通过引入中间状态，将原始的多智能体马尔可夫决策过程转化为一个单智能体马尔可夫决策过程。其中，中间状态 $s_t^{a_{1:i}}$ 表示在状态 s_t 下，第 1 到第 i 个智能体已经做出了动作 a_i 所到达的状态。智能体 i 接收 $s_t^{a_{1:i-1}}$ 决策出动作 a_i，下一时刻的中间状态变为 $s_t^{a_{1:i}}$，得到奖励为 0；最后一个智能体采取动作之后，中间状态变为 $s_t^{a_{1:n}}$，由一个不存在的智能体产生一个空的动作，使得状态转移到 s_{t+1}，收到环境奖励为 $r(s_t,\boldsymbol{a}_t)$。

总的来说，中间状态的引入，将原本的多智能体决策序列 $(s_t,\boldsymbol{a}_t,r(s_t,\boldsymbol{a}_t),s_{t+1})$ 扩展为单智能体的决策序列，即

$$\left(\left(s_t,a_t^1,0,s_t^{a_1}\right),\left(s_t^{a_1},a_t^2,0,s_t^{a_{1:2}}\right),\ldots,\left(s_t^{a_{1:n-1}},a_n,r(s_t,\boldsymbol{a}_t),s_{t+1}\right)\right) \tag{5-1}$$

为了解决侦察动作和干扰动作与不同飞行动作相关，而重复决策的飞行动作后综合后导致其中一方被孤立的问题，将上述双向动作依赖的思想放入本文所具体面对的任务场景中，将智能体无人机的多种类动作视为多个智能体，形成动作与动作之间的双向依赖，从而避免了某一种类动作被孤立而导致的智能体自身决策非稳态的问题。

5.2.3 网络结构

智能体的决策模型根据输入的状态向量进行决策，状态/观测如何表征对于算法的效率和性能有很大的影响，因此决策模型首先要考虑的就是 $s_t^{a_{1:n}}$ 的深度神经网络表示。

传统的马尔可夫决策过程的一个经验样本 $(s_t,\boldsymbol{a}_t,r(s_t,\boldsymbol{a}_t),s_{t+1})$ 对应序列展开后的 n 个中间样本，并且每一个中间状态样本都需要评估 $|A_i|$（中间态决策动作数）个下一时刻状态，也就是一共有 $\sum_{i=1}^{n}|A_i|$ 个状态需要评估。智能体在序列展开后如此庞大的状态空间下进行决策，对于决策模型状态的表征提出了更高

中国知网 https://www.cnki.net

的要求。

在本章算法的网络结构中，采用编码器词嵌入（Embedding）的方式对状态向量进行编码表征。

词嵌入是自然语言处理中的一种源自神经网络模型发展和分布式表示理念应用的表示技术，通过将词语映射到向量空间的方式表达词语的语义信息。词嵌入技术能够表达语义的相似性，描述词语之间的语义关联。通过计算词向量之间的夹角余弦值，可以找到语义相近的词语。此外，词嵌入向量还能通过数学关系表达词语的语义关联。

在本章算法所应对的情况中，将智能体的观测信息和所依赖的动作信息视为需要进行表示的词语，由前述对于多无人机协作护航任务系统模型的介绍可知，智能体的观测信息分为无人机自身的状态以及与蓝方的外部信息，无人机自身的状态包括无人机的位置和侦察载荷的情况，与蓝方的外部信息包括无人机当前位置与雷达的距离、角度等，两者信息内容差异较大，因此选用两个独立的编码器自身信息编码器（self encoder）和外部信息编码器（outer encoder）分别进行观测信息的表征，并且将外部信息嵌入结果（outer embedding）进行平均池化，再与自身信息嵌入编码结果（self embedding）相加，获得完整的观测信息嵌入编码结果（obs embedding），再与之前决策完成的动作进行编码得到的动作信息编码结果（action embedding）相加，得到决策模型的输入特征向量，观测-动作信息编码器结果（obs-action embedding），如图 5-3 所示。

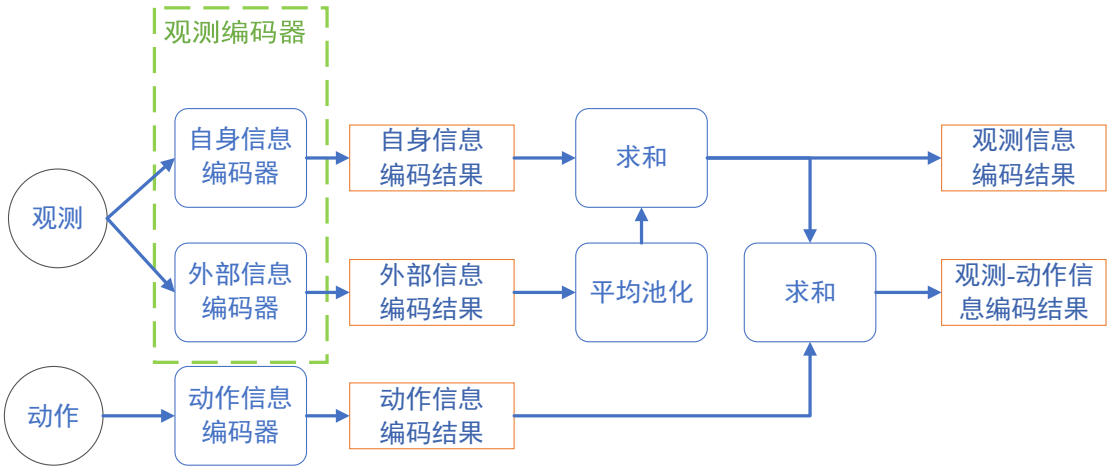


图 5-3 决策模型观测-动作信息表征示意图

Figure 5-3 Decision model observation-action information representation diagram

完成对于智能体状态的表征之后，以特征向量为输入进行决策。根据双向动作依赖的思想，视三种动作为三个智能体，每个智能体只决策一种动作，降低了决策空间的维度，将完整动作的决策过程扩展为依顺序决策。考虑到动作之间的依赖性，优先决策无人机的飞行动作，而由于无人机配备的侦察载荷和干扰载荷独立存在，且互不影响，侦察动作和干扰动作之间并没有绝对的因果

中国知网 https://www.cnki.net

关系，所以侦察和干扰动作可以进行平行的决策。此外，本章算法依旧采用 Actor-Critic 的框架，为了简化计算，Critic 网络以编码后的观测信息特征向量为输入进行状态价值的评估。具体网络架构见图 5-4。

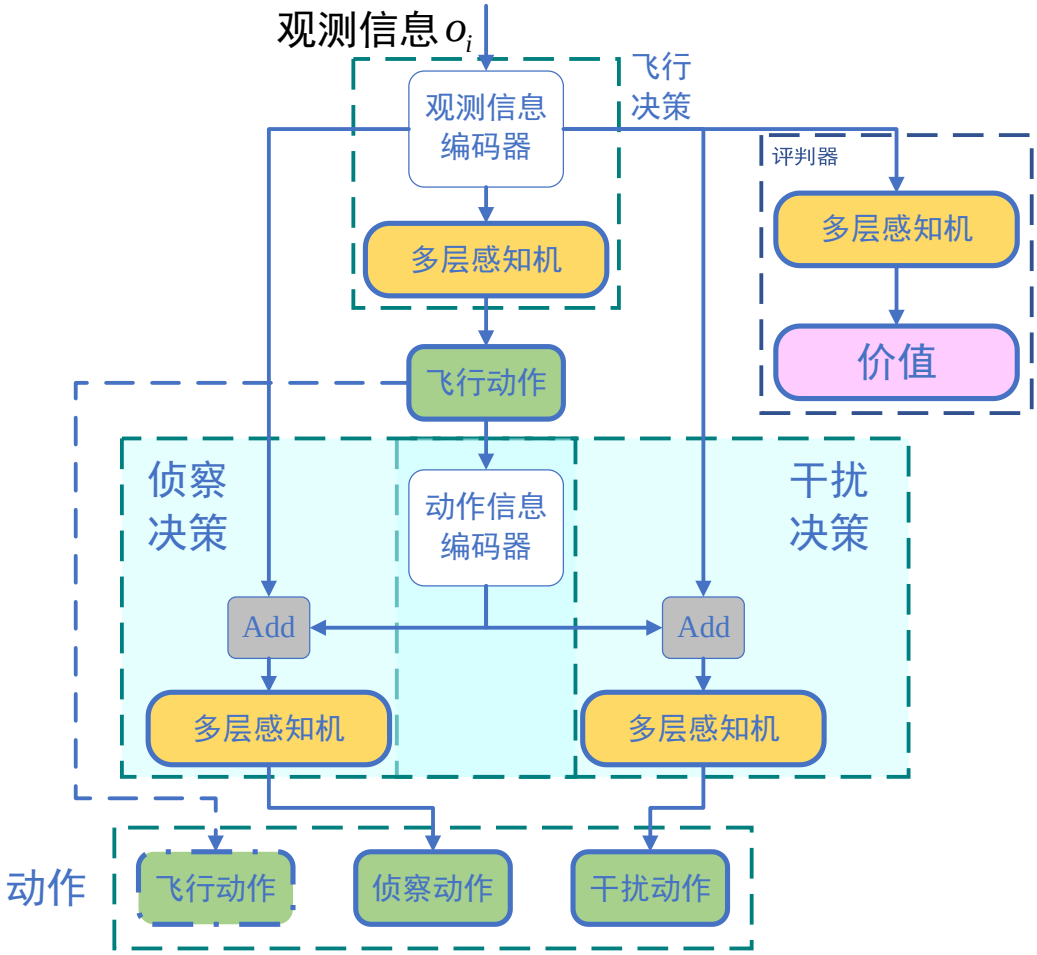


图 5-4 基于动作依赖的多智能体强化学习算法网络结构

Figure 5-4 Multi-agent reinforcement learning algorithm network architecture based on action dependence

5.2.4 算法流程

本章基于动作依赖的多智能体强化学习算法沿用 AC 架构，将 Actor 网络和 Critic 网络合并为一个多头输出的神经网络模型。决策网络模型的更新沿用近端优化的方法，其中，Actor 部分可以分为三个部分，分别进行飞行、侦察和干扰动作的决策。为了完整动作的整体性，不孤立任何一方，飞行动作的决策需要兼顾侦察和干扰的情况，因此飞行决策模型的损失函数为

$$L_{fly}(\theta) = L_{fly} + L_{scout} + L_{jamming} \tag{5-2}$$

而侦察和干扰动作之间互不影响，只需要观测信息以及飞行动作的条件下进行最有利的决策即可，因此其损失函数分别为

中国知网 https://www.cnki.net

$$L_{scout}(\theta) = L_{scout}$$
$$L_{jamming}(\theta) = L_{jamming} \tag{5-3}$$

而侦察决策模型和干扰决策模型共用的动作编码器的更新则需要兼顾两者的，故动作编码器的损失函数为

$$L_{\text{ActionEncoder}} = L_{\text{scout}} + L_{\text{jamming}} \tag{5-4}$$

为每一部分网络选取合理的损失函数可以使得网络的更新优化更为专注。上述损失函数的计算均采用，

$$L^{clip}(\theta) = \mathbb{E}_t\left[min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\right)\right]$$
$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta old}(a_t \mid s_t)} \tag{5-5}$$

优势函数 A 依据下式，采用 GAE（generalized advantage estimator）的方式进行近似计算，

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=1}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V = \sum_{l=1}^{\infty}(\gamma\lambda)^l\left(r_t + \gamma V(s_{t+l+1}) - V(s_{t+l})\right) \tag{5-6}$$

其次，Critic 的部分为了学习特征的表达，在训练阶段，使编码器近似状态价值函数，其目标是通过最小化经验 Bellman 误差，

$$L_{critic}(\phi) = R(o_t, a_t) + \gamma V_\phi(o_{t+1}) - V_\phi(o_t) \tag{5-7}$$

本章算法的伪代码如下，

表 5-1 基于动作依赖的多智能体强化学习算法伪代码

Table 5-1 Pseudo code of AD-PPO

算法：基于动作依赖的多智能体强化学习算法（AD-PPO）
1: 初始化经验池，其容量为 episode 的长度；
2: 随机初始化 Actor 网络参数 θ^π 和 Critic 网络参数 φ^V；
3: 循环 总 episode 数：
4: 初始化状态 $\boldsymbol{s}_1$；
5: 循环 episode 长度（时间步），$t = 1,2,\cdots,K$：
6: 对于智能体 $i = 1,\cdots,m$ 执行：
7: 对智能体 i 的局部观测 o_t^i 进行编码表征；
8: 飞行决策网络根据观测特征向量选择飞行动作 a_t^{fly}；
9: 对 a_t^{fly} 进行编码并与观测特征相加得到状态特征；
10: 侦察决策网络根据观测-动作特征决策侦察动作 a_t^{scout}；
11: 干扰决策网络根据观测-动作特征决策干扰动作 a_t^{jam}；
12: $a_t^i = (a_t^{fly}, a_t^{scout}, a_t^{jam})$；

中国知网 https://www.cnki.net

13:	Critic 评估状态价值$V(o_i)$并存储到经验池（buffer）中；
14:	执行联合动作$\boldsymbol{a}_t=(a_t^1,a_t^2,\cdots,a_t^m)$获得奖励$r_t$和下一时间步的状态$\boldsymbol{s}_{t+1}$；
15:	将经验样本$(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})$存储到经验池中；
16:	采用 GAE 的方法计算优势函数；
17:	从经验池$\mathcal{D}$中随机采样n组样本经验；
18:	利用式(5-5)计算损失函数$L(\theta)$并按需求更新 Actor 网络；
19:	利用式(5-7)计算损失函数$L(\varphi)$并更新 Critic 网络；

5.3 实验结果与分析

本节中，针对本章所提出的基于动作依赖的多智能体强化学习算法在多无人机协作护航任务中的有效性，进行了比较实验。在仿真训练中，设置总的仿真训练步数为 5000000，episode 长度为 1000，并行线程数为 16，从容量为一个 episode 长度的经验池中多次随机抽取 256 组样本进行批量训练，计算梯度并更新神经网络，折扣率γ设为 0.99，GAE 估计的$\lambda=0.95$。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。另外，对于多种类动作之间的决策顺序对策略结果的影响也进行了探索。

5.3.1 算法比较实验

本小节给出了本章所提出的基于动作依赖的多智能体强化学习算法（Action Dependent Proximal Policy Optimization, AD-PPO）在多无人机协作护航任务中的收敛特性，并与 MADDPG、MAPPO 等算法进行了比较。每个 episode 结束后对神经网络进行训练并记录数据，训练时随机提取多组样本多次训练，以提升样本利用率。

中国知网 https://www.cnki.net

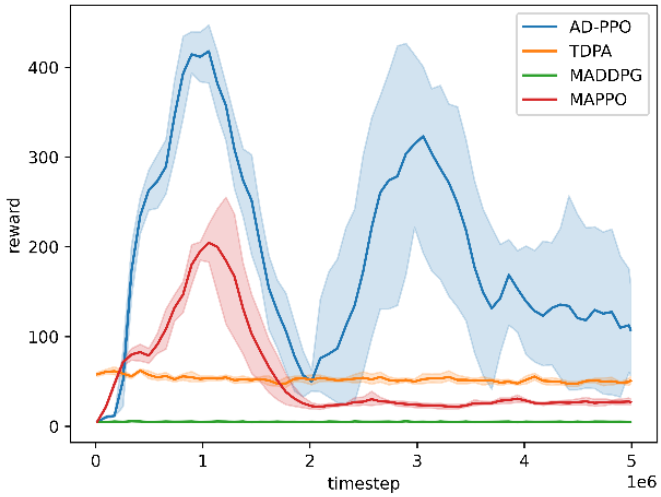


图 5-5 基于动作依赖的多智能体强化学习算法收敛性比较

Figure 5-5 Comparison of convergence of multi-agent reinforcement learning algorithms based on action dependence

图 5-5 中给出了每个 episode 的累计奖励值随总执行步数的变化。从图中可以观察到，AD-PPO 算法训练中的情况整体呈现了很大的波动性，且标准差比较大，说明算法策略的学习不是很平稳，直到 4*e*6 时间步才接近收敛，但是相比之下，确实达到了与 TDPA 算法、MAPPO 算法和 MADDPG 算法相比之下更好的效果，展现了算法对于本文中多无人机协作护航任务的有效性和适配性。。

表 5-2 AD-PPO 算法与不同算法策略性能的比较

Table 5-2 Comparison between AD-PPO algorithm and different algorithm strategies

算法	侦察任务完成度	安全区域开辟时间	侦察协作率	干扰协作率	干扰动作失效率
AD-PPO	0.97	2.1	37%	34%	23.3%
TDPA	0.78	1.2	51.5%	4.2%	38.5%
MADDPG	0.36	0	2.1%	0	24.7%
MAPPO	0.79	1.4	33.2%	20.9%	26.7%

表 5-2 是 MAPPO、MADDPG 以及前述提出的 TDPA 算法与 AD-PPO 算法在完整多无人机协作护航任务中所训练学习获得的策略的性能对比。实验采用算法分别在多无人机协作护航任务环境中以随机策略的形式执行 1000 次任务，记录任务过程数据并统计分析。

可以发现，AD-PPO 算法的在协作性方面比之 TPDA 算法略有提升，且两个任务上的协作率更加均衡，不存在大的“偏科”问题，说明自身决策的稳定性可以在一定程度上促进智能体之间的协作。并且，针对 TDPA 算法中出现的某

中国知网 https://www.cnki.net

一方动作被孤立而失效的问题，由于侦察属于探索任务，是否动作失效难以判断，因此这里统计了干扰动作的失效率，即智能体选择了作用范围以外的干扰目标的时间步在整个 episode 中的占比。数据显示，AD-PPO 算法的失效率与 TDPA 算法相比有了明显的下降，说明通过显式地建立动作之间的依赖关系，哪怕动作是分开进行决策的，动作与动作之间的联合性整体性也可以得到提升，动作与动作之间的协作意图得到了加强，这对于智能体外部与其余智能体之间的协作是有利的。

5.3.2 动作决策顺序实验

本小节对动作依赖情况下多种类动作决策顺序对策略的影响进行探索。本章 AD-PPO 算法的顺序为先决策飞行动作，然后平行决策侦察动作和干扰动作。考虑到这是针对多无人机协作护航任务的特点进行设定的，但在某些真实场景中，很可能会出现无法明确区分多种类动作之间关系的情况，在这种情况下，AD-PPO 算法是否还能够适用是本文所关注的点。

除原定的决策顺序之外，本实验还设置了飞行、侦察、干扰以及飞行、干扰、侦察两种决策顺序进行实验。在新设置的决策顺序下，算法的网络结构呈顺序结构，先根据观测特征决策第 1 个动作，然后将动作 1 编码并与观测特征相加形成状态特征 1，根据状态特征 1 决策第 2 个动作，接着将动作 2 编码与状态特征 1 叠加得到状态特征 2，再根据状态特征 2 决策第 3 个动作。

仿真实验结果如图 5-6 所示。

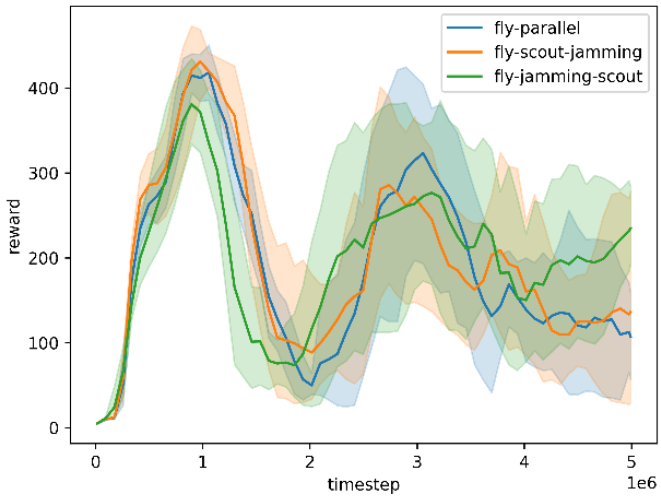


图 5-6 动作决策顺序实验结果

Figure 5-6 Experimental results of action decision order

从图 5-6 中可以发现，即使修改了多种类动作之间的决策顺序，但智能体的决策效果是相差不大的。说明在合适的状态特征下，决策模型可以从状态特征

中国知网 https://www.cnki.net

中选取对自身决策有益的信息，信息的一定程度的冗余是可以接受的。这一实验结果对于本文后续进行更通用高效的序列扩展决策模型是一个良好的开端。

5.4 本章小结

本章在前述基于子任务分解的多智能体强化学习算法的基础上，对于智能体无人机多种类动作之间协调决策的问题进行分析，借由序列扩展决策、双向动作依赖的思想，将原无人机决策多种类动作的传统马尔可夫决策过程变更为序列扩展的马尔可夫决策过程，将多种类动作同时决策变为依顺序决策，并且通过特征编码器、决策神经网络结构的设计，在后置动作的决策输入中加入前置动作的信息，形成多种类动作之间的依赖，提出了一种基于动作依赖的多智能体强化学习算法。仿真实验结果表明，AD-PPO 算法有效地降低了由于决策孤立而造成的动作失效问题，并且决策模型可以适应不同的动作决策顺序而不会对效果产生太大的影响。

中国知网 https://www.cnki.net

第6章 面向复合离散动作空间的多智能体序列决策算法

6.1 引言

多智能体协作任务中，智能体之间的协作性是决策算法达到更优的一个重要因素，而当一个智能体的动作空间是由多种类的动作计算笛卡尔积而形成的复合动作空间时，不同种类的动作之间的协作性也将是算法寻求最优策略的风向标。在多种类动作组成的复合离散动作空间的情况下，现有的强化学习决策算法如 PPO、DDPG 等，并没有考虑不同种类动作之间的协作性的重要性，故往往在巨大的动作空间中收敛效果不佳，而如果扩展到多智能体协作领域，协作性的重要性将会愈发的明显。

前述的算法通过分解任务或者分解动作空间的形式来降低单次决策的维度以缓解高维动作空间带来的维度爆炸问题，但对于多无人机协作护航任务中智能体之间的协作性还未重点关注。虽然采用共享评估网络参数的方式在一定程度上加强了智能体之间在不同任务上的协作性，但智能体 Actor 网络依赖 Critic 网络的价值评估结果的引导协作，而 Critic 网络优依赖于 Actor 网络对于决策空间的探索，Critic 网络对于智能体协作的认知是不确定的，以 Critic 网络评估价值隐性引导智能体协作的方式存在效率不高的问题。

前述提到的双向动作依赖方法可以显式地对于智能体之间的协作关系进行了建模，作为输入可以有效促进智能体之间的协作，而且智能体一个接一个地进行决策，这与自然语言处理中机器翻译的处理语句序列的形式是类似，相比于特定任务建立特定结构的神经网络模型，采用序列模型（Sequence Model，SM）无疑是更高效的方法。此外，第 5 章的动作决策顺序实验结果表明，序列扩展之后多种类动作的决策顺序并不需要严格遵守某种顺序。基于此，本章提出了一种面向复合离散动作空间决策任务的多智能体序列决策算法。

6.2 面向复合离散动作空间的强化学习序列决策算法

6.2.1 分解高维动作空间

在多智能体协作任务中，可以通过序列决策的方法使智能体之间有意识的进行一定的协作性动作选择，即根据自身观测与前一智能体的决策进行自身的决策，这种跨智能体的有序决策简化了联合策略的更新，下述定理 1 表明，最大化每个智能体自身的局部优势等同于最大化联合优势。

定理 1（多智能体优势分解定理[84]）：令 $i_{1:n}$ 表示多智能体的一个序列，对于任意的联合观测向量 $\boldsymbol{o} \in \mathcal{O}$ 和联合动作 $\boldsymbol{a} = \boldsymbol{a}^{i_{1:n}} \in \mathcal{A}$，无需进一步假设，总有

中国知网 https://www.cnki.net

下式成立：

$$A_{\pi}^{i}\left(\boldsymbol{o},\boldsymbol{a}^{i_{1:n}}\right)=\sum_{m=1}^{n}A_{\pi}^{i_m}\left(\boldsymbol{o},\boldsymbol{a}^{i_{1:m-1}},a^{i_m}\right) \tag{6-1}$$

类似的，针对多种类动作组成的复合动作空间决策任务，各种类动作之间也可以通过序列决策的方式来加强动作与动作之间的协作性。

推论 1（多种类动作优势分解）：令 i 表示任一智能体，$a_{1:h}$ 为该智能体的多动作序列。对于任一观测 o 和复合动作 $a_{1:h}\in A$，总有下列公式成立，

$$A_{\pi}^{i}(o,a_{1:h})=\sum_{c=1}^{h}A_{\pi}^{i}(o,a_{1:c-1},a_c) \tag{6-2}$$

将推论 1 进一步推广到多智能体问题，即多个智能体各自拥有多种类的动作需要选择。

推论 2（多智能体多动作优势分解）：设 $i_{1:n}$ 是多智能体的一个序列，$a_{1:h}$ 为某一智能体的多动作序列，对于任意的联合观测 $\boldsymbol{o}\in\boldsymbol{\mathcal{O}}$ 和联合复合动作 $\boldsymbol{a}=\boldsymbol{a}_{1:h}^{i_{1:n}}\in\boldsymbol{\mathcal{A}}$，则有：

$$A_{\pi}^{i}(\boldsymbol{o},\boldsymbol{a}_{1:h}^{i_{1:n}})=\sum_{m=1,c=1}^{n,h}A_{\pi}^{i_m}\left(\boldsymbol{o},\boldsymbol{a}_{1:c-1}^{i_m},a_c^{i_m}\right) \tag{6-3}$$

其中，h 为智能体数量，n 为动作种类数。

定理及推论提供了一种针对具有复合动作空间的任务决策的思路。假设，在任意状态下，智能体 i 选择了具有正优势的某种类的动作 a_c^i，$A_{\pi}^{i}\left(o,a_c^i\right)>0$，然后，想象之后所有的智能体的各部分动作都知道前面的选择 a_c^i，在这种前提下，后续的每一步选择都将趋向于选择有正优势的动作，即 $A_{\pi}^{i}\left(\boldsymbol{o},a_c^i,a_{c+1}^i\right)>0$。定理和推论保证了联合动作的正优势。另外，每一次决策只需要在该智能体的该类动作空间中进行选择，整个探索过程的复杂度是相加性的而不是相乘，选择时不需要考虑不同种类动作的联合动作，在保证协作性的同时降低了策略探索的难度。

6.2.2 序列模型

把多个具备复合动作空间的智能体决策过程分解为每次决策某一智能体的某一类动作的序列，进行序列决策的过程与自然语言处理的序列处理极其相似。例如在机器翻译领域，模型将根据待翻译句子的第一个词得出翻译结果的第一个词，然后以翻译结果的第一个词和待翻译句子的第二个词为输入，输出翻译结果的第二个词，以此类推，直到翻译完整个句子。单词的翻译结果之间是有相关性的，这一现象放到智能决策领域，可以理解为每一部分的决策是带有协作性的。

中国知网 https://www.cnki.net

以所有智能体的观测作为“待翻译句子”，所有智能体的所有种类动作作为“翻译结果”，进行序列化处理。由于任务的复杂性和动作空间的复合性，需要进一步提高序列模型的性能，通常的做法让模型扩展得更深，一般是通过增加隐藏层层数或者通过堆叠更多的 Transformer 块来实现，但这些做法会增大模型训练的难度。

这里以 DeLighT 模型为原型设计网络模型，DeLighT 模型是一种更深但较轻量级的 Transformer 结构。DeLighT 模型更高效地分配模型参数，它在每一个 Transformer 块内使用了 DExTra 变换（一种深度和轻量化的变换），并且使用了块级缩放，使得输入端的 DeLighT 块较浅较窄，而输出端的 DeLighT 块更深更宽，在加深网络模型的同时保持了较少的参数量。

编码器（Encoder）负责对输入进行编码，一个观测值序列 $(o^{i_1},\ldots,o^{i_n})$ 以任意的顺序进入，经过由 DeLighT 模块、自注意力模块以及前馈神经网络组成的编码器模块，得到编码结果，之后将传输到 Decoder 模块中的注意力模块。解码器（Decoder）负责得出动作序列，由一个初始动作（类似于文字序列的开始标识），经过 DeLighT 模块、自注意力模块、带 mask 的注意力模块以及前馈神经网络组成的解码器模块，计算得出下一顺位的动作，不断循环这一过程直至得出整个决策动作序列。其中，注意力模块的输入还包括 Encoder 的编码结果和自注意力模块的残差输入，而 mask 的作用是保证当前动作的决策不会受到未来数据的影响。

训练该网络所采用的损失函数如下：

$$L_{Actor}(\theta)=-\frac{1}{Tnd}\sum_{m=1}^{n}\sum_{c=1}^{d}\sum_{t=0}^{T-1}\min(r_t^{i_m^c}(\theta)\hat{X}_t, clip(r_t^{i_m^c}(\theta),1\pm\epsilon)\hat{X}_t) \tag{6-4}$$

$$r_t^{i_m^c}(\theta)=\frac{\pi_\theta^{i_m^c}(a_t^{i_m^c}\mid \hat{o}_t^{i_{1:n}}, a_t^{i_{1:m-1}^{1:c-1}})}{\pi_{\theta_{old}}^{i_m^c}(a_t^{i_m^c}\mid \hat{o}_t^{i_{1:n}}, a_t^{i_{1:m-1}^{1:c-1}})} \tag{6-5}$$

其中，n 为智能体的数量，d 为一个智能体的动作种类数量，T 为 episode 的长度。$\hat{X}$ 为价值优势函数，由 Critic 部分评估后计算得出。

中国知网 https://www.cnki.net

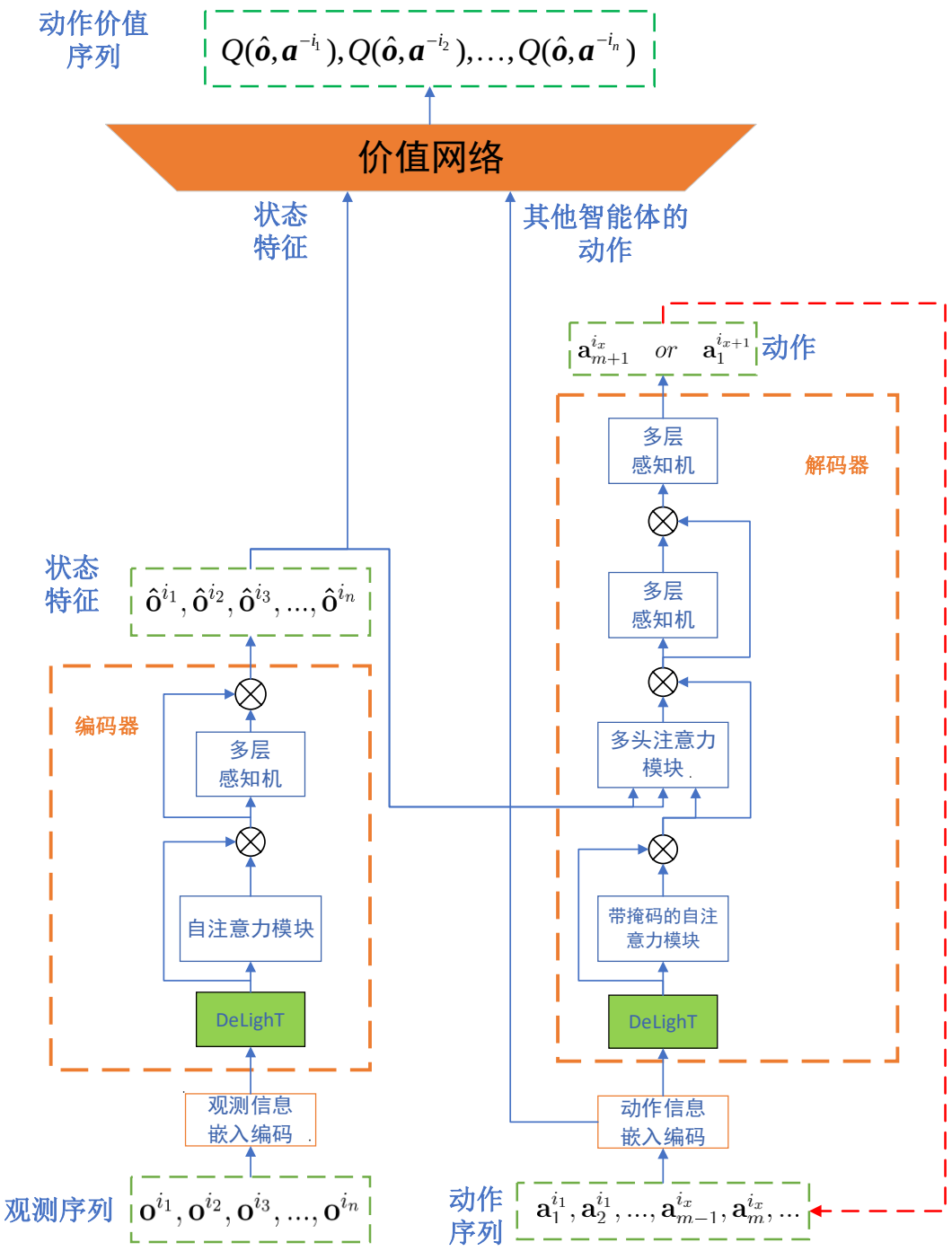


图 6-1 MA2DBT 算法的网络结构

Figure 6-1 The network architecture of MA2DBT.

将一个时间步中多智能体多种类动作的决策拆解为一个决策序列后，在降低决策复杂度的同时也由于决策战线的拉长对动作与动作之间的协调性提出了挑战。为了让决策模型决策时有据可依，因此在决策之前需要将决策可能会需要的信息进行词嵌入表征，信息的冗余是可以接受的。

中国知网 https://www.cnki.net

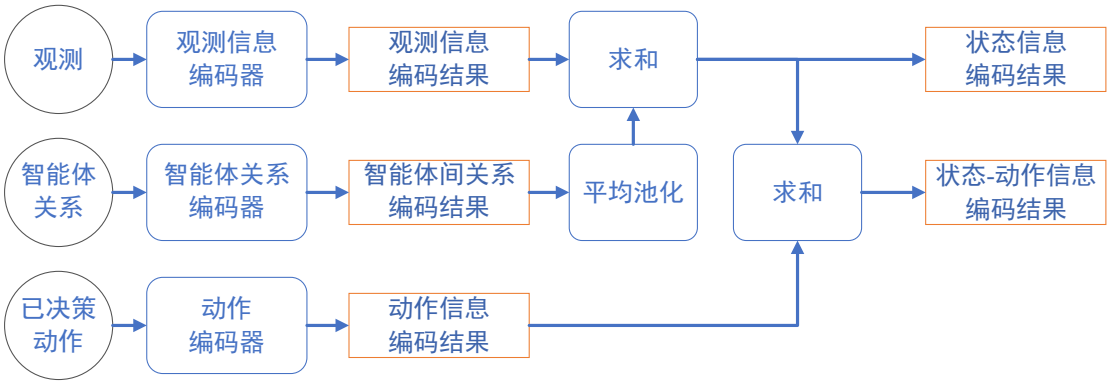


图 6-2 序列决策模型状态、动作嵌入网络示意图

Figure 6-2 Sequence decision model state and action embedded network diagram

如图 6-2 所示，决策模型的决策依据信息共分为三个部分，智能体自身的观测、智能体与其余智能体之间的关系、之前已经完成决策的动作。智能体自身的观测包括其自身的位置信息、侦察信息、干扰信息等；智能体之间的关系包括智能体之间的距离、相对方向等，这两部分组成了针对某一智能体的状态信息。将状态信息的嵌入结果作为序列决策模型中编码器的输入，在注意力机制下进一步挖掘信息的关联性。为了建立这一虚拟决策序列中的所有动作之间的关联性，需要把每次决策的动作都进行词嵌入表征，然后输入到解码器中进行下一动作的决策。对于图 6-1 中的编码模块，每个模块都简单地使用全连接层来现实。

6.2.3 动作价值函数估计优化

在多智能体协作任务中，由于只有团队奖励，所以每个智能体虽然做的动作不一，但得到的回报是一致的，这就导致有些智能体可能会“偷懒”，做出巨大贡献的智能体和没有什么贡献的智能体得到了相同的回报。这种回报的分配是很不公平的，智能体的回报分配也应该是“多劳多得”。这就是多智能体协作任务中典型的置信分配问题（credit assignment problem）。

本文所提出的算法依旧采用的是 AC 的框架，上述的 Transformer 模型作为 Actor 获取动作，可以看到，Actor 策略收敛依赖于 $\hat{X}$ 优势函数的估计，而优势函数基于 Critic 对于值函数的估计。本章采用的是蒙特卡洛估计方法，这种方法必须等当前 episode 结束才能进行学习，具有无偏估计但高方差的特点，因此，希望通过控制变量减法（也称为基线技法），来降低价值函数估计的方差。同时，置信分配的问题也可以通过加入基线的方式使得不同智能体不同动作的价值估计表现出差异性，回报分配更加合理。

基线 b 是一个函数，但并不依赖于动作 a，故

$$E[b\cdot\nabla_{\theta}log\pi_{\theta}(a\,|\,s)]=0 \tag{6-6}$$

而在多智能体策略梯度的计算中，

中国知网 https://www.cnki.net

$$\nabla_{\theta^i}\mathcal{J}(\theta)=\mathbb{E}_{s_{0:\infty}\sim d_\theta^{0:\infty},a_{0:\infty}^{-i}\sim\pi_\theta^{-i},a_{0:\infty}^{i}\sim\pi_\theta^{i}}\left[\sum_{t=0}^{\infty}\gamma^t Q_\theta\left(s_t,\boldsymbol{a}_t^{-i},a_t^i\right)\nabla_{\theta^i}\log\pi_\theta^i\left(a_t^i|\ s_t\right)\right] \tag{6-7}$$

当带入 baseline 时，

$$\begin{aligned}\nabla_{\theta^i}\mathcal{J}(\theta,b)&=\mathbb{E}_{s\sim d_\theta,\boldsymbol{a}_{0:\infty}^{-i}\sim\pi_\theta^{-i},a_{0:\infty}^{i}\sim\pi_\theta^{i}}\left[\sum_{t=0}^{\infty}\gamma^t(Q_\theta\left(s_t,\boldsymbol{a}_t^{-i},a_t^i\right)-b(s_t,\boldsymbol{a}_t^{-i}))\nabla_{\theta^i}\log\pi_\theta^i\left(a_t^i|\ s_t\right)\right]\\&=\mathbb{E}_{s_{0:\infty}\sim d_\theta^{0:\infty},\boldsymbol{a}_{0:\infty}^{-i}\sim\pi_\theta^{-i},a_{0:\infty}^{i}\sim\pi_\theta^{i}}\left[\sum_{t=0}^{\infty}\gamma^t Q_\theta\left(s_t,\boldsymbol{a}_t^{-i},a_t^i\right)\nabla_{\theta^i}\log\pi_\theta^i\left(a_t^i|\ s_t\right)\right]\\&-\sum_{t=0}^{\infty}\mathbb{E}_{s_{0:\infty}\sim d_\theta^{0:\infty},\boldsymbol{a}_{0:\infty}^{-i}\sim\pi_\theta^{-i},}\left[b(s,\boldsymbol{a}^{-i})\nabla_{\theta^i}\log\pi_\theta^i\left(a_t^i|\ s_t\right)\right]\\&=\nabla_{\theta^i}\mathcal{J}(\theta)\end{aligned} \tag{6-8}$$

可以看到，无论函数 b 是什么，梯度的期望是不变的，所以基线的加入并不会影响策略梯度的正确性。

本章共选择了三种基线供不同任务选择：状态价值函数、反事实基线以及通过数学计算得出的最优基线。

第一，以状态价值函数作为基线，这种情况下计算得出的优势函数即为一般的优势函数。

$$\hat{X}=Q(s,a)-V(s) \tag{6-9}$$

第二，使用反事实基线。反事实基线的思想是，评价一个智能体所执行动作的贡献有多少，可以把该智能体的动作替换为一个默认的动作，然后观察替换前与替换后，团队回报的变化，当回报上升，则说明当前动作的效果不如基线动作。为了降低计算的复杂性，采用当前策略的平均效果作为默认动作的效果[114]。

$$\hat{X}=Q(s,a^i)-\sum\pi^i(a_t\mid s_t)Q(s_t,a_t) \tag{6-10}$$

第三，使用数学计算得出的最优基线。在最优基线理论中，对多智能体环境中方差产生的原因进行了分析，估计方差主要来源于状态的方差、当前智能体动作的方差以及其余智能体的方差三个部分[84]。

中国知网 https://www.cnki.net

$$
\begin{aligned}
&Var_{s_t\sim d_\theta^t,a_t\sim\pi_\theta}\left[Q_t^i(b)\right]=Var_{s_t\sim d_\theta^t}\left[Q_t^i(b)\mid \boldsymbol{a}_t\sim\boldsymbol{\pi}_\theta\right]\\
&=Var_{s_t\sim d_\theta^t}\left[\mathbb{E}_{\boldsymbol{a}_t\sim\boldsymbol{\pi}_\theta}\left[Q_t^i(b)\right]\right]+\mathbb{E}_{s_t\sim d_\theta^t}\left[Var_{\boldsymbol{a}_t\sim\boldsymbol{\pi}_\theta}\left[Q_t^i(b)\right]\right]\\
&=Var_{s_t\sim d_\theta^t}\left[\mathbb{E}_{\boldsymbol{a}_t\sim\boldsymbol{\pi}_\theta}\left[Q_t^i(b)\right]\right]+\mathbb{E}_{s_t\sim d_\theta^t}\left[Var_{\boldsymbol{a}_t^{-i}\sim\boldsymbol{\pi}_\theta^{-i}}\left[\mathbb{E}_{a_t^i\sim\pi_\theta^i}\left[Q_t^i(b)\right]\right]+\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\boldsymbol{\pi}_\theta^{-i}}\left[Var_{a_t^i\sim\pi_\theta^i}\left[Q_t^i(b)\right]\right]\right]\\
&=\underbrace{Var_{s_t\sim d_\theta^t}\left[\mathbb{E}_{\boldsymbol{a}_t\sim\boldsymbol{\pi}_\theta}\left[Q_t^i(b)\right]\right]}_{\text{Variance from state}}+\underbrace{\mathbb{E}_{s_t\sim d_\theta^t}\left[Var_{\boldsymbol{a}_t^{-i}\sim\boldsymbol{\pi}_\theta^{-i}}\left[\mathbb{E}_{a_t^i\sim\pi_\theta^i}\left[Q_t^i(b)\right]\right]\right]}_{\text{Variance from other agents' actions}}+\underbrace{\mathbb{E}_{\boldsymbol{a}_t^{-i}\sim\boldsymbol{\pi}_\theta^{-i}}\left[Var_{a_t^i\sim\pi_\theta^i}\left[Q_t^i(b)\right]\right]}_{\text{Variance from agent i's action}}]
\end{aligned}
\tag{6-11}
$$

由于基线函数受状态以及其余智能体动作的影响，但不依赖于当前智能体的动作，故最小化第三项方差即为最小化整个方差。

最优基线为：

$$
b^{\text{optimal}}\left(s,\boldsymbol{a}^{-i}\right)=\frac{\mathbb{E}_{a^i\sim\pi_\theta^i}\left[\hat{Q}\left(s,\boldsymbol{a}^{-i},a^i\right)\left\|\nabla_{\theta^i}\log\pi_\theta^i\left(a^i\mid s\right)\right\|^2\right]}{\mathbb{E}_{a^i\sim\pi_\theta^i}\left[\left\|\nabla_{\theta^i}\log\pi_\theta^i\left(a^i\mid s\right)\right\|^2\right]} \tag{6-12}
$$

上述三种基线中，需要用到动作价值函数$Q(s,a)$以及状态价值函数$V(s)$，故本章算法建立了相应的评估网络，一方面借用 Transformer 中 encoder 的编码网络，以智能体状态编码结果为输入，状态价值为输出，计算得出状态价值$V(s)$，用于进行第一种基线。另一方面，后两项基线的计算需要依赖于某一时刻某一智能体所有可能动作的动作价值，为了便于计算，算法在决策模型的顶端扩展建立了一个 COMA 价值网络，见图 6-1，用于估计某状态下某个智能体下一步所有可能动作的价值。该网络以某一智能体的状态嵌入信息以及其余智能体所有动作的嵌入信息（以a^{-i_m}表示除智能体i_m之外的智能体动作信息）为输入，输出该智能体在此状态下的所有下一步动作的价值Q，用于后续基线的计算，并且在训练中加入了 Target 网络，保证估计的稳定性。

训练该网络所采用的损失函数为：

$$
L_{Critic}(\phi)=r+\gamma\sum_{i=0}^{n}\omega_i\hat{Q}_i-\sum_{i=0}^{n}\omega_iQ_i \tag{6-13}
$$

其中，$\hat{Q}$是 Target 网络的结果，ω是某状态下智能体的所有动作选择概率。

6.2.4 算法流程

依据上述的三个部分，就组成了本章所提出了 MA2DBT 算法。算法采用经典的 AC 架构，Actor 进行动作的决策，而 Critic 进行当前状态及动作的价值评估，评估结果用以指导 Actor 的更新。其中 Actor 网络采用 Encoder-Decoder 的

中国知网 https://www.cnki.net

结构，由 Encoder 对观测信息进行编码，Decoder 根据观测信息进行序列决策直至决策出完整的动作序列，而 Critic 部分则采用 COMA 网络的结构，根据状态信息和其余智能体的动作信息计算得到动作价值函数，并采用基线方法来减小价值估计的方差，从而更平稳地指导 Actor 的更新。算法详细流程伪代码见下方。

表 6-1 MA2DBT 算法的伪代码

Table 6-1 Pseudocode of MA2DBT

算法：MA2DBT 算法
输入：智能体个数 n，智能体动作个数 d，总训练 episode 个数 K，每个 episode 的样本训练次数 ep，episode 长度 T；
初始化：DeLighT Transformer 的参数 θ，经验池 $\mathcal{D}$；
循环 $k=0\to K-1$：
循环 $t=0\to T-1$：
收集来自环境的观测向量序列 $o_t^{i_1},\ldots,o_t^{i_n}$；
通过 Embedding 模块对观测向量信息进行嵌入编码；
通过模型中的编码器（Encoder）模块对观测嵌入向量进行进一步表征，得到序列 $\hat{o}_t^{i_1},\ldots,\hat{o}_t^{i_n}$；
将序列 $\hat{o}_t^{i_1},\ldots,\hat{o}_t^{i_n}$ 输入到解码器中；
循环 $m=0\to n-1$:
循环 $c=0\to d-1$:
将动作序列 $a_t^{i_0,j_0},\ldots,a_t^{i_0,j_c},\ldots,a_t^{i_m,j_c}$ 通过 Embedding 模块进行嵌入编码；
嵌入结果给到解码器(Decoder)进行下一动作决策；
在环境中执行联合复合动作 $a_t^{i_0,j_0},\ldots,a_t^{i_0,j_d},\ldots,a_t^{i_n,j_d}$；
将得到的样本 $(\boldsymbol{o}_t,\boldsymbol{a}_t,R)$ 存入经验池 $\mathcal{D}$ 中；
循环 $p=1\to ep$：
从经验池 $\mathcal{D}$ 中随机采集 batch size 的样本 $(\boldsymbol{o}_t,\boldsymbol{a}_t,R,\boldsymbol{o}_{t+1})$；
通过决策模型中 Critic 的部分计算得到
$Q_\phi(o^{i_0},a^{i_0}),\ldots,Q_\phi(o^{i_0},a^{i_d}),\ldots,Q_\phi(o^{i_n},a^{i_d})$；
根据公式(6-13)计算 $L_{critic}(\phi)$；
根据 $Q_\phi(o^{i_0},a^{i_0}),\ldots,Q_\phi(o^{i_0},a^{i_d}),\ldots,Q_\phi(o^{i_n},a^{i_d})$ 和每个动作的概率计算基线 b；
计算价值函数 $X=Q-b$；
根据公式(6-4)计算 $L_{actor}(\theta)$；
通过最小化 $L=L_{critic}+L_{actor}$ 的策略梯度更新 Transformer 决

中国知网 https://www.cnki.net

策网络

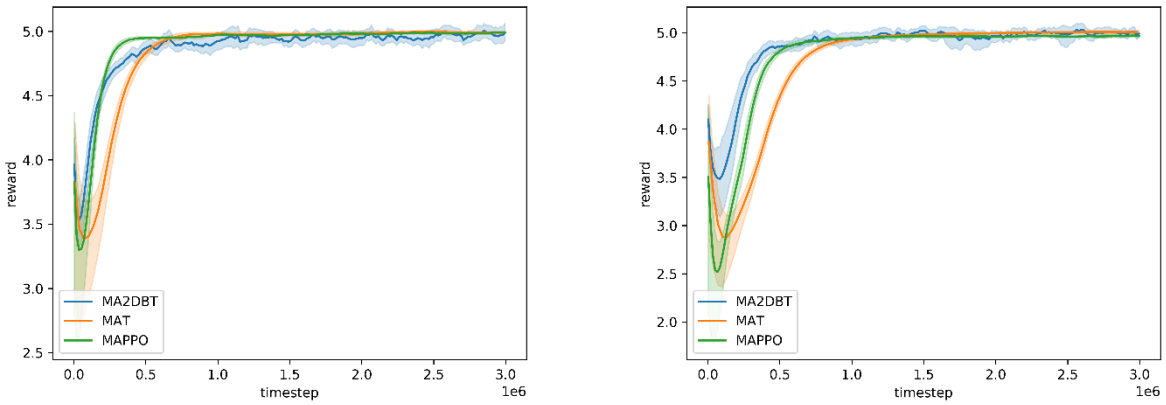
6.3 实验结果与分析

在这一部分，为了评估 MA2DBT 算法是否有效地提高了多智能体算法在复合动作空间的多无人机护航任务场景下的决策能力，将其与当前的一些算法进行了比较实验。为了获得更具有普适意义的结论，同样在一些开源的环境上进行了比较实验。另外，针对本章的算法进行一些消融实验以探索算法各部分所起到的效果。

6.3.1 单智能体任务决策比较实验

首先，在复合动作单智能体任务中，与几种现行的强化学习算法进行了比较，包括 MAT、PPO 算法。这里选用谷歌足球（Google Research Football）中的足球单智能体场景 academy_empty_goal_close（近距离空门射门），并且为了贴近复合动作的情况，对动作空间进行了修改，从每次智能体只需决策一步的动作变为了一次决策未来的多步，一次决策的步数越多，则动作空间越复杂。

本节设置总的仿真训练步数为 3000000，episode 长度为 1000，并行线程数为 16，每次训练神经网络时，从容量为一个 episode 长度的经验池中多次随机抽取 256 组样本进行批量训练。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。

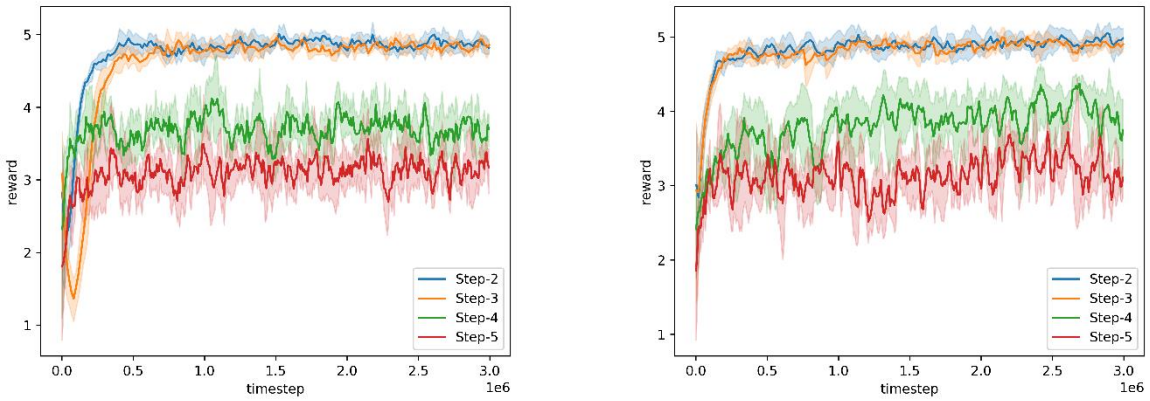


(a) 2 步版 academy_empty_goal_close　　**(b) 3 步版 academy_empty_goal_close**

图 6-3 不同算法在 academy_empty_goal_close 场景中的收敛特性

Figure 6-3 Convergence of different algorithms in the academy_empty_goal_close

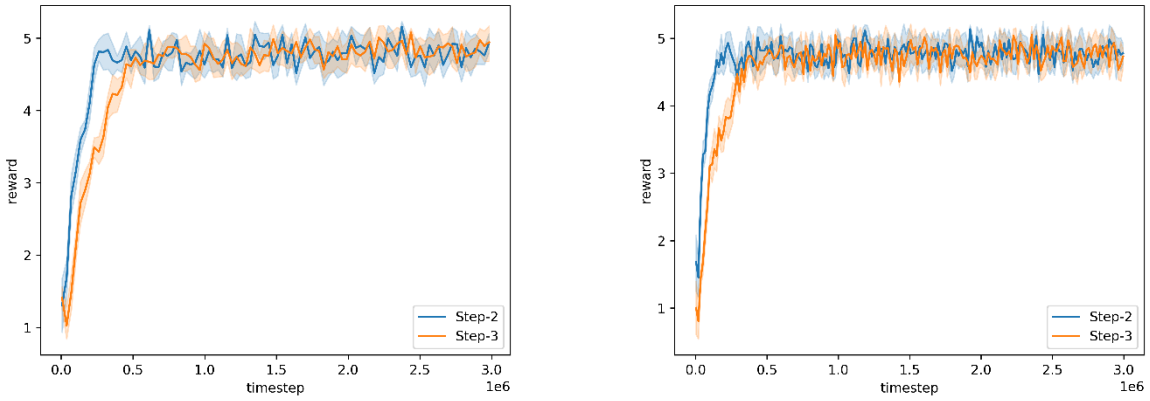
中国知网 https://www.cnki.net



(a) DeLighT 网络结构的 MA2DBT **(b) 朴素 MA2DBT**

图 6-4 MA2DBT 算法在同一场景下不同决策复杂度下的收敛特性

Figure 6-4 Convergence of MA2DBT algorithm in the same scenario and different decision complexity



(a) MAT **(b) MAPPO**

图 6-5 MAT 和 MAPPO 算法在同一场景下不同决策复杂度下的收敛特性

Figure 6-5 Convergence of MAT and MAPPO under different decision complexity in the same scene

可以看到，在每次决策未来两步、三步的情况下，MA2DBT 算法的收敛速率要优于 MAT、MAPPO。而随着动作空间的复杂度上升，算法收敛速率在降低，由于智能体每次决策未来四步时，每个智能体的复合动作空间已经达到了 13 万，现有算法下本文所使用的设备无法支持该体量的计算，相比之下，虽然 MA2DBT 算法没有在最优策略的探索中有突出表现，但通过对复合动作空间进行了分解，在复杂场景下有更好的收敛稳定性，有效的缓解了复合动作所带来的动作空间维度爆炸问题。此外，图 6-4 为 MA2DBT 算法采用带 DeLighT 模块的 Transformer 网络与朴素 Transformer 网络之间的比较，DeLighT 模块的加入虽

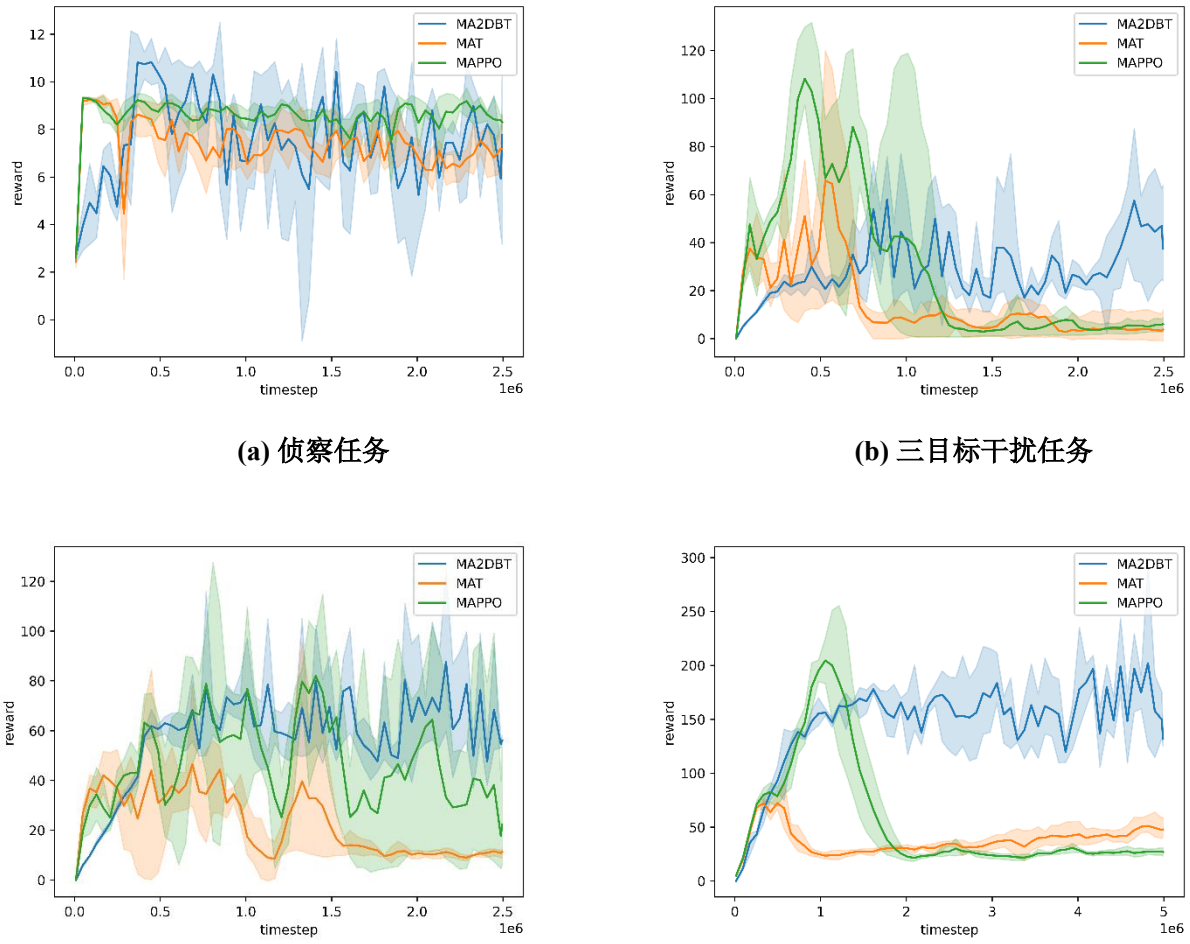
中国知网 https://www.cnki.net

然降低了收敛的速率，但其最终的效果要略强于朴素 Transformer。

6.3.2 多智能体任务决策比较实验

单智能体场景的结果并不能完全体现算法的优势，所以本文拟在更为复杂的具备复合动作的多智能体环境下进行比较实验，在本实验中，主要目的是探索算法在不同决策复杂度，即不同复合动作空间大小的决策任务中的表现。多智能体实验环境仍采用本文所设计的多无人机协作护航任务，将多无人机协作护航任务分解为四级复杂度任务，分别为侦察任务、针对三个雷达的干扰任务、针对六个雷达的干扰任务以及完整的多无人机协作护航任务，其单个智能体的复合动作空间大小分别为 45、150、330 和 990。

设置总的仿真训练步数为 2500000，episode 长度为 1000，并行线程数为 16，每次训练神经网络时，批量样本数为 256。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。



(a) 侦察任务　　**(b) 三目标干扰任务**

(c) 六目标干扰任务　　**(d) 多无人机协作护航任务**

图 6-6 不同复杂度任务下的算法收敛特性

Figure 6-6 Algorithm convergence characteristics under different complexity tasks

中国知网 https://www.cnki.net

实验结果显示，在复杂度较低的侦察任务中，三种算法最终的收敛效果是相近的，甚至相比之下，MA2DBT 算法的震荡范围要更大一些，因为 MA2DBT 算法中对多种类动作进行分解并依次决策，在一个时间步进行了多次决策，决策次数的增多造成了随机性的增大。在三目标干扰任务中，三个目标并不能充分反映当前的战场环境情况，在认知不足的情况下，智能体无人机为寻求更高的奖励而不断接近已知的目标雷达，但过度的接近也导致了无人机被发现和打击的风险不断上升，尤其是未知雷达带来的风险，这就导致累计奖励曲线会产生一个较大的峰值而又因被打击的负奖励而急速下降，造成最终的收敛效果不理想，而 MA2DBT 算法由于一个时间步要进行多次决策，且用基线方法降低了更新的方差，更新较为谨慎，最终收敛的效果要更优。在信息完备的六目标干扰任务中，MAPPO 和 MAT 算法的曲线中仍存在较大的起伏，但信息的完备性使其有跳出局部最优的可能性。在完整的多无人机协作护航任务中，由于需要平衡侦察与干扰两个任务，不完全的侦察结果会面临三目标干扰任务类似的情况，易出现断崖式下跌，进而陷入局部最优，而 MA2DBT 算法整体收敛过程较为平滑，没有过大的起伏，没有陷入局部最优的陷阱，取得了优于其他算法的策略模型收敛结果。可以发现，在复杂度逐渐上升的任务中，MA2DBT 算法展现出了更好的适应性。

6.3.3 决策顺序实验

接下来，进行多无人机协作护航任务策略的探索学习并进行消融实验以探索本章算法各部分对策略探索整体的贡献。

针对多无人机护航任务，首先探索决策顺序对决策结果的影响，本文提出两种多智能体复合动作决策顺序，其一，是先根据动作种类顺序决策一个智能体所有种类的动作，再依次决策下一个智能体所有种类的动作，直至完成所有决策，称为智能体优先；其二，是先根据智能体顺序决策所有智能体某一种类动作，再依次决策所有智能体的下一种类动作，直至完成所有决策，称为动作优先。

本文对 MA2DBT 算法的不同决策顺序进行了比较实验，设置总的仿真训练步数为 5000000，episode 长度为 1000，并行线程数为 16，每次训练神经网络时，从容量为一个 episode 长度的经验池中多次随机抽取 256 组样本进行批量训练。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。结果见图 6-7。

中国知网 https://www.cnki.net

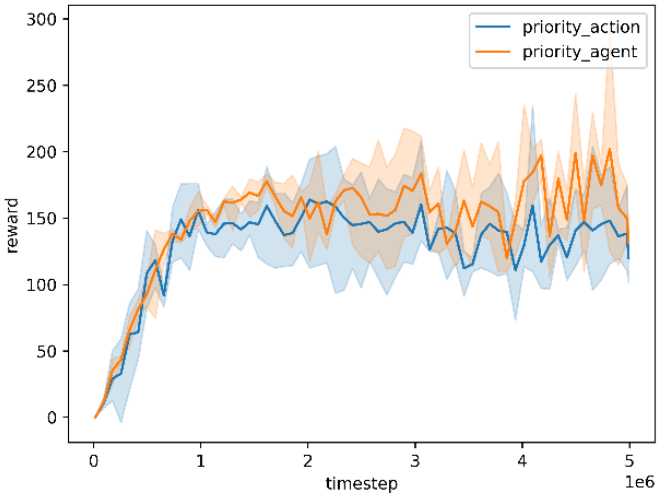


图 6-7 不同序列决策顺序的收敛特性

Figure 6-7 Convergence characteristics of different sequence decision order

结果显示，在本文建立的多无人机协作护航任务场景中，无论采取哪种决策树顺序，都取得了良好的收敛结果，但相比之下，动作优先的决策顺序在训练前期的标准差更大，说明该顺序下训练的稳定性较弱，而采用智能体优先决策顺序时获得了略胜一筹的决策模型收敛结果，并且后期有很强的向上波动的趋势，是一个向更优缓慢发展的趋势。

6.3.4 算法消融实验

在本章的算法中，策略更新依赖于 Critic 所得出的动作价值函数进行指导，而蒙特卡洛方法计算的价值函数具有无偏差但高方差的特点，所以基线技法的使用就显得尤为重要。在这一节中，我们将在多无人机协作护航任务中进行实验，进行不同基线的比较实验探索最适配该任务的基线以及对 MA2DBT 算法的各个环节进行消融比较实验。其中，朴素版的 MA2DBT 算法，在 Transformer 的损失函数中 X 采用的是 GAE 优势估计算法得到的优势值。

本节设置总的仿真训练步数为 5000000，episode 长度为 1000，并行线程数为 16，每次训练神经网络时，从容量为一个 episode 长度的经验池中多次随机抽取 256 组样本进行批量训练。本节所有算法将 Actor 网络策略梯度更新的学习率分别设置为 0.001，Critic 网络的学习率为 0.0001。任务重复执行实验次数为 1000。

中国知网 https://www.cnki.net

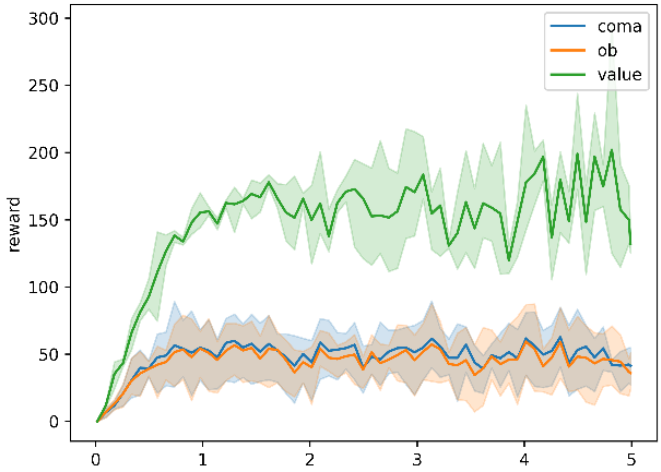


图 6-8 不同基线收敛特性的比较

Figure 6-8 Comparison of convergence characteristics of different baselines

图 6-8 为 MA2DBT 算法在 coma、ob、adv 三种基线方法下进行仿真实验的收敛结果比较，可以看到，在本文所建立的多无人机协同作战场景中，以状态价值作为基线取得了更好的收敛效果。

表 6-2 是对各个算法的决策模型进行同样次数的评估而得出的平均数据，其中包括单步决策时间以及平均每次任务中红蓝方的一些任务相关数据。

表 6-2 决策模型任务执行情况

Table 6-2 Performance of Decision model in mission

算法	单次决策时间	侦察结果个数	侦察协作率	干扰效能	安全区域开辟时间	干扰协作率	任务成功率
MA2DBT	0.0216	6.91(98.7%)	75.5%	45.4	2.8	31.8%	34.8%
MA2DT	0.021	5.03(71.8%)	71.2%	35.7	1.2	25.2%	13.2%
MA2T	0.017	2.43(34.7%)	34.2%	22.1	0.1	8.6%	0%
MAT	0.007	4.85(69.2%)	44.3%	38.2	1.9	26.6%	23.6%
MAPPO	0.0014	3.34(47.7%)	33.2%	21.6	1.4	20.9%	19.8%
vs MAT		29.5%	31.2%	7.2	0.9	5.2%	11.2%

其中，干扰效能是指任务执行过程中无人机累计干扰引起的雷达探测能力衰减所占的比例。多无人机协作护航任务的最终目标是为了护送轰炸机对蓝方重要目标完成打击，轰炸打击是环境的脚本动作，该动作的成功率与安全区域开辟时间长短息息相关。根据设定，每多一个时间步即可提升 10%的成功率，理想情况下达到 10 个时间步即可 100%完成打击，任务成功率即为完成轰炸动作的比例。MA2T 表示 MA2DBT 算法采用朴素 Transformer 网络模型且不加入基

中国知网 https://www.cnki.net

线的版本，MA2DT 表示 MA2DBT 算法不加入基线的版本。“vs MAT”是本章算法 MA2DBT 在多无人机协作护航任务中，与表现较好的 MAT 算法相比，任务执行结果评判指标的增量。

可以看到，MA2DBT 算法在单次决策时间上做出了一定的牺牲，但依旧保持在 0.1 秒之内，与 MAT 算法相比，协同侦察任务的完成度提高了 29.5%，侦察效能提升了 1.8%，干扰效能提升了 7.2，安全区域开辟时间提升了 0.9，在协作性方面，侦察协作率提升了 31.2%，干扰协作率提高了 5.2%，任务成功率提升了 11.2%。

在 MA2DBT 算法的各个环节中，编码嵌入（Embedding）模块和 DeLighT 结构的使用进一步延长了单步决策时间，而基线的使用并没有单步决策时间产出较大影响，此外，DeLighT 结构以及基线的加入，使得多无人机协同作战任务策略模型的各项任务数据均有不同程度的提升。

多无人机协作护航任务的理想结果是开辟出安全区域并维持 10 个时间步以上，显然，所有的策略模型都没有达到理想结果。在该任务中，飞行、侦察以及干扰是三个并不独立的子任务，相互之间是会影响的，比如，一般情况下，覆盖目标位置的蓝方雷达在三到四个左右，MA2DBT 算法虽然很好的完成了侦察任务，但有一部分侦察定位结果对后续的安全区域的打开并没有太多帮助，故 MAT 算法以较少的侦察结果实现了和 MA2DLBT 算法相近的干扰效能，可见某一子任务完成度高并不代表着主任务的完成度也高，这里面就涉及到了多目标任务中多个目标之间的协调问题。如何更合理的分配多目标之间的奖励机制、如何选取更合适的基线、如何使得多智能体更高效合理地在多目标任务中进行决策是我们下一阶段的研究重点。

6.4 本章小结

本章在前述基于动作依赖的强化学习算法的基础上，将动作依赖的范围从单个智能体的多种类动作扩展到多智能体，将同时决策多个智能体的多种类动作的马尔可夫决策过程序列扩展为一次决策某一智能体的某一动作的决策序列，显式地建立多种类动作之间、多智能体之间的协作关系。考虑到需要高效处理决策序列，采用自然语言处理中的改进 Transformer 序列模型来进行序列决策，并且以冗余嵌入信息的形式为决策模型提供智能体观测信息、多智能体之间的关系信息以及已决策的动作信息。此外，为了降低蒙特卡洛方法在梯度更新时的高方差问题，结合基线的思想，提出了一种面向复合离散动作空间的多智能体序列决策算法。MA2DBT 可以在不详细分析任务的情况下应用于其他的任务场景，仿真实验结果表明，算法取得了优于其他算法的策略收敛效果，且在任务执行数据中，展现出了更好的任务执行结果评估数据。在本文重点探索的多

中国知网 https://www.cnki.net

无人机协作护航任务中，MA2DBT 算法在同样分解高维动作空间的同时进一步提高了智能体之间的协作性。

中国知网 https://www.cnki.net

第7章 总结与展望

7.1 总结

随着信息技术的不断发展，战场电磁环境日益复杂，给电子对抗决策带来了严峻挑战。无人作战是电子对抗的重要作战样式，无人作战系统通过自主控制技术和人工智能的深度融合，使得系统具备环境态势感知、自主行为决策等能力，具备独立执行复杂任务的能力。但是，面对瞬息万变的战场态势，如何科学高效地进行作战决策，使得己方的 OODA 环在整个作战过程中占据主导性优势地位，越来越成为制胜的关键所在。

本文以典型电子对抗任务场景——多无人机协作护航任务为例，构建了多无人机协作护航任务仿真系统模型，针对任务决策中无人机需要同时决策多种类动作以及多智能体之间协作性不强的问题，提出了基于深度强化学习的智能决策算法。

本文的研究工作总结如下：

（1）针对典型电子对抗场景中的多无人机协作护航任务，分析其任务流程以及任务细节，以红蓝对抗为背景，设计了多无人机协作护航任务仿真系统模型。将多无人机协作护航任务建模为部分可观测马尔可夫决策过程，以红方无人机作为智能体，设计了智能体的观测空间、动作空间，并且根据任务目标设计了任务环境的奖励函数。此外，对多无人机协作护航任务中实体的功能模型进行了构建，包括红方无人机侦察模型、蓝方雷达探测模型、蓝方雷达受干扰模型等。

（2）针对多无人机协作护航任务中，由于智能体无人机需要在多种类动作组成的高维动作空间中进行决策，而任务包含侦察、干扰等多个任务目标，不同任务目标对多种类动作的选择有不同需求，容易产生互相干扰的情况，进而影响智能体的整体决策的问题，基于分层决策的思想，提出了基于子任务分解的多智能体强化学习决策算法。将完整的多无人机协作护航任务分解为侦察和干扰两个子任务，分别进行策略学习，并根据任务的进展情况综合两个策略，得到完整的智能体决策结果。本文为两个子任务分别设计了不同的奖励函数，解决了策略学习中的任务冲突问题，加强其对子任务的针对性学习，仿真实验证明了子策略相比于MADDPG算法在完整任务中的策略在各自子任务上取得了更好的任务完成度和协作率。算法的综合策略将子策略的决策结果根据任务进展进行了结合，仿真实验结果表明，该算法取得了更好的任务完成度。

（3）在前述的基于子任务分解的多智能体强化学习决策算法中，两个子策略对飞行动作进行了重复决策，存在决策冗余，而且综合策略结果中的飞行动

中国知网 https://www.cnki.net

作对某一任务目标存在倾向性，容易造成针对另一任务目标的动作孤立甚至失效。针对该问题，提出了一种基于动作依赖的多智能体强化学习决策算法（AD-PPO 算法），对于决策模型的神经网络结构进行了特定性的优化，将前置动作的决策结果与观测向量一起作为后置动作的决策输入，建立动作与动作之间的依赖关系不孤立任何一类动作，同时降低智能体决策的难度。仿真实验结果表明，AD-PPO 算法有效降低了干扰动作的失效率，并且与 MADDPG 算法、MAPPO 算法相比，获得了更好的任务执行结果，同时也对智能体之间的协作率有一定的提升，可见智能体自身决策的稳定性对于智能体之间寻求协作是有益的。此外，进行了动作决策顺序的比较仿真实验，实验结果显示，决策模型可以适应多种的决策顺序，为后续序列处理模型的引入奠定了基础。

（4）在前述的基于动作依赖的多智能体强化学习决策算法中，虽然加强了动作与动作之间的协作性，并通过决策模型网络结构的优化降低了单个决策模型决策的动作维度，但面对不同的任务场景时，需要对任务实际进行深入分析并根据实际情况重新设计网络结构，不具备通用性，且智能体之间的协作性没有受到重点关注。针对上述问题，本文提出了一种面向复合动作空间的多智能体强化学习序列决策算法，将序列模型引入智能决策，每一步的决策过程拆分为一个决策序列，每次决策只选择一个智能体的一种动作，且决策的输入包括当前观测信息以及已决策出的动作信息，增强动作之间协作性的同时降低单次决策的维度，并且考虑到每个时间步决策序列的延长，通过优化网络模型、优化价值估计模型的方法提升决策模型的性能。最后，通过仿真实验验证了 MA2DBT 算法在复合动作空间场景下面对单智能体任务和多智能体任务均展现了更好的性能。并且针对多无人机协作护航任务，进行了 MA2DBT 算法的消融实验，结果表明，在该任务中，采用智能体优先的决策顺序得到了略好的收敛效果，并且最适合该任务的是采用状态价值函数作为基线，而收敛特性以及策略评估结果也体现出了 MA2DBT 算法在多无人机协作护航任务上的优越性和适应性。

7.2 创新性工作

本文主要的创新工作如下：

（1）针对多无人机协作护航任务中，智能体无人机需要在多种类动作组成的高维动作空间中进行决策，而多个任务目标对于多种类动作的不同需求容易互相干扰，进而影响智能体整体决策的问题，基于分层决策的思想，提出了基于子任务分解的 TDPA 决策算法，底层策略针对相应的任务目标进行决策，顶层策略负责子策略决策结果的综合，解决了不同任务目标需求冲突下策略难以收敛的问题，且降低了单个策略决策的维度，有益于提升策略的收敛效果。

中国知网 https://www.cnki.net

（2）针对基于子任务分解的多智能体强化学习决策算法中，两个子策略对飞行动作进行了重复决策，存在决策冗余，而且综合策略结果中的飞行动作对某一任务目标存在倾向性，容易造成针对另一任务目标的动作孤立甚至失效的问题，提出了一种基于动作依赖的 AD-PPO 算法。通过对决策模型神经网络结构的特定性优化，将前置动作的决策结果纳入后置动作的决策考虑，建立动作与动作之间的依赖关系，不孤立任何一类动作，同时降低了每部分决策模型决策的维度，有益于提升策略的收敛效果。

（3）针对前述算法面对不同任务场景的通用性问题以及没有重点关注智能体之间的协作性问题，提出了一种面向复合动作空间的 MA2DBT 算法，将序列模型引入智能决策，将单步多智能体多种类动作的决策以序列的形式依次进行决策，降低单次决策的维度，且动作的决策输入包含上一步的决策，显式地建立了动作之间、智能体之间的协作相关性，并且不需要进行特定的结构设计，加强了算法的通用性。

7.3 工作展望

未来新兴电子对抗攻防技术发展，特别是认知电子技术的发展，正在加速新兴认知电子战能力的形成，而无人作战逐渐成为电子战的重要作战样式。决策博弈是智能化战争对抗的核心和中枢，以深度学习和强化学习为代表的人工智能技术取得了巨大突破，军事智能化成为了人工智能的重要应用方向。多智能体强化学习是当前机器学习研究领域的前沿技术，采用深度学习与强化学习相结合的方法研究多智能体间的完全合作任务，为解决无人装备协同对抗问题提供了有效途径。然而，智能体如何在更加复杂多变的作战场景进行决策是一项极具挑战性的难题。结合本文的研究成果，针对典型电子对抗任务场景中可能出现的问题，我们将从以下几个方面对电子对抗智能决策问题展开更深入的研究。

（1）在作战中，指挥员通常需要在了解当下战场态势的基础上，结合过去发生的一些信息来进行决策，下一步我们将研究如何将时序信息纳入智能体的决策考虑，使得智能体在整个任务执行过程中的决策有环环相扣的性质；

（2）在作战任务中，通常有多个任务目标，这些任务目标有主次，有不同的重要性等级，如何在更多更复杂的任务目标中进行权衡决策，是未来的研究难点之一；

（3）随着先进网络信息体系、多学科交叉融合等关键技术群的突破，未来战争将向跨军种作战、跨功能域作战、联合多域作战发展，如何将各种异构作战力量协同起来是需要重点关注的研究方向之一。

中国知网 https://www.cnki.net

参考文献

[1] Zhen L U, Huang Y. Electronic warfare confrontation between the United States and Russia[J]. Journal of Beijing University of Posts and Telecommunications, 2020, 43(5): 1-8.

[2] 何梅昕,刘恩凯. 信息化战争中的电子对抗技术[J].电子技术与软件工程,2018,(18):71.

[3] 王立楠,蔡楚瀚,刘国生,等. 基于效能评估的战斗机末端光电对抗仿真[J].航空学报,2021,42(08):358-370.

[4] 石荣. 电子对抗在中国的起源与早期发展历程回顾[J].舰船电子工程,2021,41(02):14-19.

[5] 邱志明,李恒,周玉芳,等.模拟仿真技术及其在训练领域的应用综述[J].系统仿真学报,2023,35(06):1131-1143.

[6] Power D J. Decision support systems: concepts and resources for managers[M]. Westport: Quorum Books, 2002.

[7] Pasqual G M, Mansfield J. Development of a prototype expert system for identification and control of insect pests[J]. Computers and Electronics in Agriculture, 1988, 2(4): 263-276.

[8] Wang S, Shi W. Data mining and knowledge discovery[J]. Springer Handbook of Geographic Information, 2012: 49-58.

[9] Russell S J, Norvig P. Artificial intelligence: a modern approach[M]. Pearson, 2016.

[10] Cutler A, Cutler D R, Stevens J R. Random forests[J]. Ensemble machine learning: Methods and applications, 2012: 157-175.

[11] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[12] 李理,李旭光,郭凯杰,等. 国产化环境下基于强化学习的地空协同作战仿真[J].兵工学报,2022,43(S1):74-81.

[13] 章胜,周攀,何扬,等. 基于深度强化学习的空战机动决策试验[J].航空学报,2023,44(10):122-135.

[14] 周攀,黄江涛,章胜,等. 基于深度强化学习的智能空战决策与仿真[J].航空学报,2023,44(04):99-112.

[15] Chebotar Y, Vuong Q, Hausman K, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions[C]//Conference on Robot Learning. PMLR, 2023: 3909-3928.

[16] 张耀中,赵雪芳,丰文成. 基于改进蛙跳算法的多无人机协同任务分配研究[J].

中国知网 https://www.cnki.net

火力与指挥控制,2023,48(04):52-58+64.
[17] 左家亮,杨任农,张滢,等.基于启发式强化学习的空战机动智能决策[J].航空学报,2017,38(10):217-230.
[18] Wang S, Bao Y, Li Y. The architecture and technology of cognitive electronic warfare[J]. Science in China (Information Sciences), 2018, 48(12): 1603-1613..
[19] Liu S, Lei Z, Wen Z, et al. A development review on cognitive electronic warfare[J]. Journal of Detection & Control, 2020, 42(5): 1-15.
[20] Wang Z, Zhang H, Zhao H, et al. Intelligent electromagnetic metasurface camera: system design and experimental results[J]. Nanophotonics, 2022, 11(9): 2011-2024.
[21] Ya S, Li L I, Zhuo W, et al. Metasurface-assisted intelligent electromagnetic sensing: theory, design and experiment[J]. Chinese Journal of Radio Science, 2021, 36(6): 858-866.
[22] Jia S, Yang F. Research on intelligent detection method of weak sensing signal based on artificial intelligence[C]//Advanced Hybrid Information Processing: Third EAI International Conference, ADHIP 2019, Nanjing, China, September 21–22, 2019, Proceedings, Part II. Springer International Publishing, 2019: 90-98.
[23] Chi W, Wang H, Xie W, et al. Research on distributed cooperative intelligent spectrum sensing of UAV cluster[J]. Wireless Communications & Mobile Computing, 2022.
[24] Ning W, Zhe L, Xiaolong L, et al. Cooperative region search of UAV swarm with limited communication distance[J]. Systems Engineering & Electronics, 2022, 44(5): 1615-1625.
[25] Wei R X, Zhou K, Ru C J, et al. Study on fuzzy cognitive decision-making method for multiple UAVs cooperative search[J]. Scientia Sinica (Technologica), 2015, 45(6): 595-601.
[26] Rong S, Jiang L. Application of intelligent optimization methods in jamming resource allocation: a review[J]. Electronics Optics and Control, 2019, 26(10): 54-61.
[27] Xing H, Wu H, Chen Y, et al. A cooperative interference resource allocation method based on improved firefly algorithm[J]. Defence Technology, 2021, 17(4): 1352-1360.
[28] Tang M N, Xiong W L, Xu B G. Improved TOPSIS Method in the Distribution of Jamming Resources of Radar[J]. Fire Control and Command, 2012, 37(1): 91-94.
[29] Xing H, Wu H, Chen Y, et al. Multi-efficiency optimization method of jamming resource based on multi-objective grey wolf optimizer[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(10): 1990-1998.
[30] Ye F, Che F, Gao L. Multi-objective cognitive cooperative jamming decision-making method based on Tabu search-artificial bee colony algorithm[J]. International Journal of Aerospace Engineering, 2018, 2018: 1-10.

中国知网 https://www.cnki.net

[31] Han G X, He J, Mao X Q, et al. Research on optimal distribution of radar jamming resource based on improved genetic algorithm[J]. Fire Control and Command, 2013, 38(3): 99-102.
[32] Huang X, Li Y. The allocation of jamming resources based on Double Q-learning algorithm[J]. Journal of System Simulation, 2021, 33(8): 1801-1808.
[33] Conway M D, Du Russel D, Morris A, et al. Multifunction phased array radar advanced technology demonstrator nearfield test results[C]//2018 IEEE Radar Conference (RadarConf18). IEEE, 2018: 1412-1415.
[34] Ye F, Li X, Li Y, et al. Research on jamming decision making based on feedback Iterative-Brown algorithm[C]//2020 IEEE USNC-CNC-URSI North American Radio Science Meeting (Joint with AP-S Symposium). IEEE, 2020: 3-4.
[35] Zhang B, Zhu W. Research on decision-making system of cognitive jamming against multifunctional radar[C]//2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). IEEE, 2019: 1-6.
[36] Song H, Xiao M, Xiao J, et al. A POMDP approach for scheduling the usage of airborne electronic countermeasures in air operations[J]. Aerospace Science and Technology, 2016, 48: 86-93.
[37] Lyu Rui, Wu Da, Zhao Yan. A summary of researches on key technologies of cognitive interference decision-making[J]. Electronics Optics & Control, 2021, 28(11): 60-64.
[38] Li H, Li Y, He C, et al. Cognitive electronic jamming decision-making method based on improved Q-learning algorithm[J]. International journal of aerospace engineering, 2021, 2021: 1-12.
[39] Li Y, Huang D, Xing S, et al. A review of synthetic aperture radar jamming technique[J]. Journal of Radars, 2020, 9(5): 753-764.
[40] Zhu B, Zhu W, Li W, et al. A review on reinforcement learning based radar jamming decision-making technology[J]. Electronics Optics And Control, 2022, 29(4): 52-58.
[41] Ye F, Che F, Tian H. Cognitive cooperative-jamming decision method based on bee colony algorithm[C]//2017 Progress in Electromagnetics Research Symposium-Fall (PIERS-FALL). IEEE, 2017: 531-537.
[42] Zhang B, Zhu W. DQN based decision-making method of cognitive jamming against multifunctional radar[J]. Systems Engineering and Electronics, 2020, 42(4): 819-825.
[43] Feng L W, Liu S T, Xu H Z. Multifunctional radar cognitive jamming decision based on dueling double deep Q-network[J]. IEEE Access, 2022, 10: 112150-112157.
[44] Feng C, Fu X, Lang P, et al. A radar anti-jamming strategy based on game theory with temporal constraints[J]. IEEE Access, 2022, 10: 97429-97438.
[45] Yi W, Varshney P K. Adaptation of frequency hopping interval for radar anti-

中国知网 https://www.cnki.net

jamming based on reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2022, 71(12): 12434-12449.

[46] Zhu J L, Ma Y T, Liu K H. Multi-Agent joint anti-jamming decision algorithm based on LSTM and Deep Q Network[J]. Chinese Journal of Sensors and Actuators, 2021, 34(6): 811-817.

[47] Niu Y, Feng X, Kou S, et al. A novel anti-jamming decision-making algorithm based on knowledge graph technology[J]. Applied Sciences, 2022, 12(10): 4960.

[48] Ran Y, Cheng Y, Chen D, et al. Intelligent anti-jamming decision engine based on BP neural network[J]. Signal Process, 2019, 35: 1350-1357.

[49] Ye F, Zhou Z, Tian H, et al. Intelligent anti-jamming decision method based on the mutation search artificial bee colony algorithm for wireless systems[C]//2019 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium). IEEE, 2019: 27-28.

[50] Song B, Xu H, Jiang L, et al. An intelligent decision-making method for anti-jamming communication based on deep reinforcement learning[J]. Journal of Northwestern Polytechnical University, 2021, 39(3): 641-649.

[51] Wang M, Song X, Niu Y, et al. Anti-jamming decision-making in wireless communication based on rough sets theory[C]// Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC). IEEE, 2013: 2535-2539.

[52] Jia L, Qi N, Chu F, et al. Game-theoretic learning anti-jamming approaches in wireless networks[J]. IEEE Communications Magazine, 2022, 60(5): 60-66.

[53] Li W, Chen J, Liu X, et al. Intelligent dynamic spectrum anti-jamming communications: A deep reinforcement learning perspective[J]. IEEE Wireless Communications, 2022, 29(5): 60-67.

[54] Yu W, Sun Y, Wang X, et al. Modeling and analyzing of fire-control radar anti-jamming performance in the complex electromagnetic circumstances[C]//Man–Machine–Environment System Engineering: Proceedings of the 17th International Conference on MMESE 17. Springer Singapore, 2018: 611-619.

[55] Hu M, Gao L, Zhang Z. Comprehensive evaluation on ISAR anti-jamming effectiveness via ELECTRE-III[C]//2013 IEEE International Conference on Information and Automation (ICIA). IEEE, 2013: 979-984.

[56] Park S R, Nam I, Noh S. Modeling and simulation for the investigation of radar responses to electronic attacks in electronic warfare environments[J]. Security & Communication Networks, 2018.

[57] Luo D, Xu Y, Zhang J. New progresses on UAV swarm confrontation[J]. Science & Technology Review, 2017, 35(7): 26-31.

[58] MA Z, HE M, LIU Z, et al. Survey of unmanned aerial vehicle cooperative control[J]. Journal of Computer Applications, 2021, 41(5): 1477-1483.

[59] Duan H B, Zhang D F, Fan Y M, et al. From wolf pack intelligence to UAV swarm

中国知网 https://www.cnki.net

cooperative decision-making[J]. Scientia Sinica Informationis, 2019, 49(1): 112-118.

[60] Shen Y, Wei C. Multi-UAV flocking control with individual properties inspired by bird behavior[J]. Aerospace Science and Technology, 2022, 130: 107882.

[61] Gao Y, Li D. Unmanned aerial vehicle swarm distributed cooperation method based on situation awareness consensus and its information processing mechanism[J]. Knowledge-Based Systems, 2020, 188: 105034.

[62] Zhang Y Z, Li J W, Hu B, et al. An improved PSO algorithm for solving multi-UAV cooperative reconnaissance task decision-making problem[C]//2016 IEEE International Conference on Aircraft Utility Systems (AUS). IEEE, 2016: 434-437.

[63] Yue L, Yang R, Zhang Y, et al. Research on reinforcement learning-based safe decision-making methodology for multiple unmanned aerial vehicles[J]. Frontiers in Neurorobotics, 2023, 16: 1105480.

[64] Baek S, York G. Optimal sensor management for multiple target tracking using cooperative unmanned aerial vehicles[C]//2020 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2020: 1294-1300.

[65] Giacomossi L, Dias S S, Brancalion J F, et al. Cooperative and decentralized decision-making for loyal wingman UAVs[C]//2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE). IEEE, 2021: 78-83.

[66] Chen J, Liang J, Cheng L, et al. Cooperative attack decision modeling method of multiple UAVs based on FCM[J]. Acta Aeronaut Astronaut Sin, 2022, 43(7):377-394.

[67] Chen C, Mo L, Zheng D, et al. Cooperative attack-defense game of multiple UAVs with asymmetric maneuverability[J]. Acta Aeronaut. Astronaut. Sin, 2020, 41(12): 342-354.

[68] Rahmes M, Chester D, Clouse R, et al. Cooperative cognitive electronic warfare UAV game modeling for frequency hopping radar[C]//Unmanned Systems Technology XX. SPIE, 2018, 10640: 170-177.

[69] Ma Y, Wang G, Hu X, et al. Cooperative occupancy decision making of Multi-UAV in Beyond-Visual-Range air combat: A game theory approach[J]. IEEE Access, 2019, 8: 11624-11634.

[70] Xu J, Guo Q, Li Z. Dynamic selection method for cooperative decision-making center of multi-UAV system based on cloud trust model[C]//2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018: 922-926.

[71] Zhang J, Yang Q, Shi G, et al. UAV cooperative air combat maneuver decision based on multi-agent reinforcement learning[J]. Journal of Systems Engineering and Electronics, 2021, 32(6): 1421-1438.

[72] Jiang S, Yang X, Delin L U O. Cooperative combat decision-making research for

中国知网 https://www.cnki.net

multi UAVs[J]. Information and Control, 2018, 47(3): 347-354.
[73] Le N T T. Multi-agent reinforcement learning for traffic congestion on one-way multi-lane highways[J]. Journal of Information and Telecommunication, 2023, 7(3): 255-269.
[74] Yue Y, Lakshminarayanan S. Multi-agent reinforcement learning system for multiloop control of chemical processes[C]//2022 IEEE International Symposium on Advanced Control of Industrial Processes (AdCONIP). IEEE, 2022: 48-53.
[75] Liu J, Gu Y, Cheng Y, et al. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning[J]. Acta Autom. Sinica, 2022, 48: 1-13.
[76] Bhamre P, Gupta S. Constrained waveform designing for MIMO radar using Jaya optimization[J]. Wireless Personal Communications, 2020, 111(1): 331-342.
[77] Do S, Baek J, Jun S, et al. Battlefield environment design for multi-agent reinforcement learning[C]//2022 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2022: 318-319.
[78] Ahmed I H, Brewitt C, Carlucho I, et al. Deep reinforcement learning for multi-agent interaction[J]. Ai Communications, 2022, 35(4): 357-368.
[79] Chen X, Liu X, Luo C, et al. Robust multi-agent reinforcement learning for noisy environments[J]. Peer-to-Peer Networking and Applications, 2022, 15(2): 1045-1056.
[80] Riley J, Calinescu R, Paterson C, et al. Assured deep multi-agent reinforcement learning for safe robotic systems[C]//International Conference on Agents and Artificial Intelligence. Cham: Springer International Publishing, 2021: 158-180.
[81] Malysheva A, Kudenko D, Shpilman A. Magnet: Multi-agent graph network for deep multi-agent reinforcement learning[C]//2019 XVI International Symposium" Problems of Redundancy in Information and Control Systems"(REDUNDANCY). IEEE, 2019: 171-176.
[82] Liu X, Tan Y. Feudal latent space exploration for coordinated multi-agent reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.
[83] Kim H, Kim S, Lee D, et al. Avoiding collaborative paradox in multi‐agent reinforcement learning[J]. ETRI Journal, 2021, 43(6): 1004-1012.
[84] Kuba J G, Wen M, Meng L, et al. Settling the variance of multi-agent policy gradients[J]. Advances in Neural Information Processing Systems, 2021, 34: 13458-13470.
[85] Chen Y, Song G, Ye Z, et al. Scalable and transferable reinforcement learning for multi-agent mixed cooperative–competitive environments based on hierarchical graph attention[J]. Entropy, 2022, 24(4): 563.
[86] Jiang S, Amato C. Multi-agent reinforcement learning with directed exploration and selective memory reuse[C]//Proceedings of the 36th annual ACM symposium on applied computing. 2021: 777-784.

中国知网 https://www.cnki.net

[87] Elias Alonso G, Jin X. Skeleton-level control for multi-agent simulation through deep reinforcement learning[J]. Computer Animation and Virtual Worlds, 2022, 33(3-4).

[88] Selvakumar J, Bakolas E. Min–max Q-learning for multi-player pursuit-evasion games[J]. Neurocomputing, 2022, 475: 1-14.

[89] Wang S, Yue W, Chen Y, et al. Cooperative learning with difference reward in large-scale traffic signal control[C]//2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022: 2307-2312.

[90] Grosnit A, Cai D, Wynter L. Decentralized deterministic multi-agent reinforcement learning[C]//2021 60th IEEE Conference on Decision and Control (CDC). IEEE, 2021: 1548-1553.

[91] Huang S, Yang B, Chen H, et al. MA-TREX: Multi-agent trajectory-ranked reward extrapolation via inverse reinforcement learning[C]//International Conference on Knowledge Science, Engineering and Management. Cham: Springer International Publishing, 2020: 3-14.

[92] Chen T, Bu S, Liu X, et al. Peer-to-peer energy trading and energy conversion in interconnected multi-energy microgrids using multi-agent deep reinforcement learning[J]. IEEE transactions on smart grid, 2021, 13(1): 715-727.

[93] Lee H, Jeong J. Multi-agent deep reinforcement learning (MADRL) meets multi-user MIMO systems[C]//2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021: 1-6.

[94] Shi D, Tong J, Liu Y, et al. Knowledge reuse of multi-agent reinforcement learning in cooperative tasks[J]. Entropy, 2022, 24(4): 470.

[95] Ma Y, Wu L, Xu X. Cooperative targets assignment based on multi-agent reinforcement learning[J]. Systems Engineering & Electronics, 2023, 45(9): 2793-2801.

[96] Fukumoto Y, Tadokoro M, Takadama K. Cooperative multi-agent inverse reinforcement learning based on selfish expert and its behavior archives[C]//2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020: 2202-2209.

[97] Ikeda T, Shibuya T. Centralized training with decentralized execution reinforcement learning for cooperative multi-agent systems with communication delay[C]//2022 61st Annual Conference of the Society of Instrument and Control Engineers (SICE). IEEE, 2022: 135-140.

[98] Wang X, Zhao C, Huang T, et al. Cooperative learning of multi-agent systems via reinforcement learning[J]. IEEE Transactions on Signal and Information Processing over Networks, 2023, 9: 13-23.

[99] Sheikh H U, Bölöni L. Multi-agent reinforcement learning for problems with combined individual and team reward[C]//2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1-8.

中国知网 https://www.cnki.net

[100] Wang M, Xie S, Luo X, et al. HCTA: Hierarchical cooperative task allocation in multi-agent reinforcement learning[C]//2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2023: 934-941.

[101] Chen L, Lu K, Rajeswaran A, et al. Decision transformer: Reinforcement learning via sequence modeling[J]. Advances in neural information processing systems, 2021, 34: 15084-15097.

[102] Zheng Q, Zhang A, Grover A. Online decision transformer[C]//international conference on machine learning. PMLR, 2022: 27042-27059.

[103] Meng L, Wen M, Yang Y, et al. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all smac tasks[J]. arXiv preprint arXiv:2112.02845, 2021.

[104] Wen M, Kuba J, Lin R, et al. Multi-agent reinforcement learning is a sequence modeling problem[J]. Advances in Neural Information Processing Systems, 2022, 35: 16509-16521.

[105] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]//International conference on machine learning. PMLR, 2016: 1928-1937.

[106] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[107] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[108] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International conference on machine learning. PMLR, 2018: 1587-1596.

[109] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International conference on machine learning. PMLR, 2018: 1861-1870.

[110] Kulkarni T D, Narasimhan K, Saeedi A, et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation[J]. Advances in neural information processing systems, 2016, 29.

[111] Lohse O, Pütz N, Hörmann K. Implementing an online scheduling approach for production with multi agent proximal policy optimization (MAPPO)[C]//IFIP International Conference on Advances in Production Management Systems (APMS). Springer International Publishing, 2021 (Part V): 586-595.

[112] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

[113] Li C, Liu J, Zhang Y, et al. Ace: Cooperative multi-agent q-learning with bidirectional action-dependency[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(7): 8536-8544.

中国知网 https://www.cnki.net

[114] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

中国知网 https://www.cnki.net