规则引导的智能体决策框架

牟轩庭,张宏军,廖湘琳,章乐贵
(陆军工程大学 指挥控制工程学院,江苏 南京 210000)

摘 要:虽然近年来深度强化学习在决策智能中取得突破,但复杂场景中的巨大动作空间仍然是算法成功学习的一大挑战。导致这一问题的主要原因在于缺乏指导的智能体难以累积足够的成功经验,样本数据质量低下,影响模型正确收敛,而加入人类知识进行辅助是一种有效的方法。为此提出了规则引导的智能体决策框架,介绍了决策框架的总体组成;针对不同态势下存在的无效动作导致探索困难的问题,提出了规则引导的智能体决策方法,选择近端策略优化算法和注意力机制构建了简单的智能体网络,利用专家经验设计规则引导层,根据态势特征对智能体的动作空间进行动态约束。实验结果表明:该方法提高了智能体在星际争霸 II 小型任务“训练陆战队员”中的成绩,并且去掉规则引导层后仍然能够保持部分性能。

关键词:深度强化学习;专家经验;规则;动作空间;近端策略优化算法;注意力机制
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)10-0156-08
doi:10.3969/j.issn.1673-629X.2022.10.026

Rule-guided Agent Decision-Making Framework

MU Xuan-ting, ZHANG Hong-jun, LIAO Xiang-lin, ZHANG Le-gui
(School of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210000, China)

Abstract: Although deep reinforcement learning has made breakthrough in intelligent decision-making in recent years, the large action space in complex scenes is still a big challenge for successful learning. The main reason for this problem is that it is difficult for the agent without guidance to accumulate enough successful experience, which leads to the low-quality sample data and prevents the model from converging correctly. However, adding human knowledge can help solve this problem. For this reason, a rule-guided agent decision-making framework is proposed, and the overall composition of the decision-making framework is introduced. In order to solve the hard exploration problem due to invalid actions in different situations, a rule-guided agent decision-making method is proposed, which chooses proximal policy optimization algorithm and attention mechanism to build a simple agent network and uses expert experience to design a rule guidance layer thus dynamically constraining the action space of the agent according to the situational features. Experimental results show that the proposed method improves the agent's performance in StarCraft II minigame "BuildMarines", and the agent is able to maintain part of its performance after removing the rule guidance layer.

Key words: deep reinforcement learning; expert experience; rule; action space; proximal policy optimization algorithm; attention mechanism

0 引 言

现代战争的节奏不断加快,复杂性不断上升,人脑因其生理上的限制,难以快速、持续地对多维态势做出准确的分析判断,在需要反复进行的作战实验中尤为明显。而人工智能相继在 Atari[1]、围棋[2]和星际争霸[3]等复杂程度递增的环境中取得了突破,表明人工智能有望解决具有实时性、不确定性的复杂战场决策问题。越来越多的研究人员开始在作战实验中使用智能技术,特别是将以深度强化学习为代表的智能算法与特定的仿真环境相结合,利用其强大的学习探索能力从高维决策空间中发掘出可行的行动决策序列,应用于智能推演与指挥员辅助决策中。

然而单纯使用深度强化学习进行推演有其局限性,一是巨大的动作空间与状态空间增加了神经网络的训练难度,智能体很容易陷入局部最优,二是推演中的复杂任务需要多步决策完成,只采用简单算法和网络结构的智能体可能无法进行有效的探索。许多研究者从引入专家经验的角度出发尝试改进算法的表现,

收稿日期:2021-10-12 **修回日期**:2022-02-18
基金项目:国家自然科学基金(61806221)
作者简介:牟轩庭(1997-),男,硕士研究生,研究方向为军事智能决策;通讯作者:张宏军(1963-),男,博士,教授,研究方向为计算机仿真理论。

主要的思路包括学习高质量复盘数据,构建基于专家经验的奖励机制[4-6];或者将作战条令、作战规则、指挥员经验等非结构化数据进行建模,构建基于专家经验的规则库支撑智能体决策[7-8];或者将两者的优点相结合,设计基于知识和数据驱动的决策方法[9-10],在对应的领域均取得了较好的效果。

从实际应用的效果来看,专家经验辅助智能体决策的效果受它对决策空间的约束程度的影响,过度约束可能导致模型完全拟合为某种固定的决策模式,缺乏探索和发现新方案的能力,约束过少或者错误的约束又可能导致智能体重新面对巨大搜索空间的难题,增加求解难度。针对这个问题,该文提出了一种规则引导的强化学习智能体决策方法,在智能体策略产生过程中,利用专家经验对环境中的无效动作进行动态过滤,并在星际争霸 II 的小型任务场景中进行了实验验证。结果表明,与单纯使用强化学习的方法相比,规则引导机制在困难的任务中能够降低智能体的探索难度,帮助模型尽快收敛,在高质量样本数据难以获得或者计算资源不足等情况下,为强化学习算法在人不在回路的仿真、智能蓝军建设等军事领域中的快速应用提供了一个可行的解决方案。

1　智能体决策框架

为将智能决策算法与专家经验知识更好地融合,该文提出规则引导的智能体决策框架,分为数据资源层、数据处理层、决策支持层和决策输出层 4 个部分,整体结构如图 1 所示。

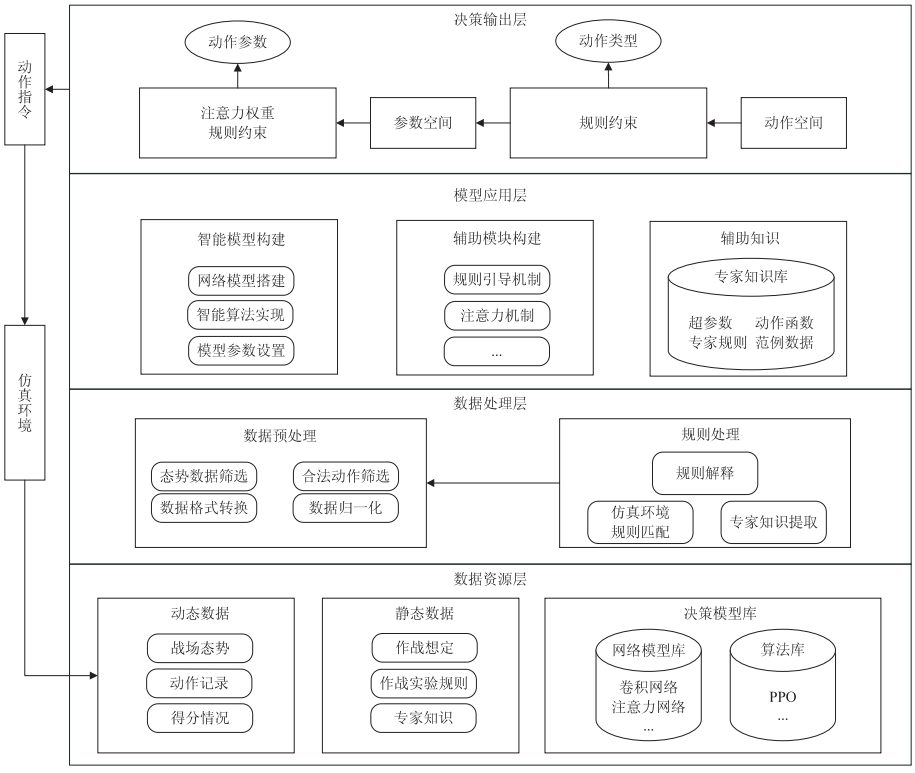

图 1　规则引导的智能体决策框架

数据资源层主要为智能体的决策提供基础的数据和模型算法支撑。数据分为动态数据和静态数据,动态数据存储智能体与仿真环境交互后产生的反馈信息,主要包含战场态势信息、智能体的历史动作以及当前得分状况,是决策模型自我迭代更新的数据基础;静态数据包含作战想定、作战实验规则及专家知识等,为下一步的数据处理以及规则引导机制的生成提供数据支撑。决策模型库存储着深度强化学习等决策算法的基本神经网络模型和更新算法的实现,方便模型具体实现时进一步的调整。

数据处理层主要实现基础数据的初步处理。数据预处理通过仿真环境的规则以及部分专家知识对态势信息进行筛选,留下必要的决策信息,同时根据仿真环境的规则初步确定智能体的合法动作空间,之后将这些数据进行格式转换、归一化等处理,转化为模型更易于利用的形式。规则处理一方面将仿真环境的规则通过解释器进行转化,用于数据预处理,另一方面将专家知识进行分类、抽象,并以特定的形式存储于专家知识库中,为模型应用层提供辅助数据。

模型应用层主要实现决策模型的具体部署与应用。智能模型构建完成智能体网络模型与强化学习等智能算法的具体结合,以及关键参数的设置。辅助模块构建根据任务需求构建特殊的数据处理模块以辅助决策的产生,如规则引导机制、注意力机制等。上述两

项活动由专家知识库提供数据支撑,包括超参数设置、动作函数、专家规则和范例数据等。超参数主要包含智能算法在不同想定情况下的模型参数设置,可根据后期结果进行不断调整;动作函数包含仿真环境中所有可调用的原子动作命令,及根据研究需要设计的多个动作的组合命令;专家规则与范例数据分别以显式和隐式的方式表达了人类的先验知识,智能体决策既可通过专家规则在线指导,又可通过范例数据离线学习。

决策输出层主要根据决策流程完成最终动作指令的输出。由于不同的动作类型具有不同的参数空间,动作指令的输出顺序为:首先使用专家规则约束动作空间后进行采样或选择最大概率的选项得到动作类型值,然后根据选择的动作类型选择对应的参数空间,规则约束与注意力权重一同作用于参数空间后采样得到动作参数值,将两者进行封装即可得到完整的动作指令。根据辅助模块的不同,动作输出方式可进行相应的调整。

该框架将专家经验和智能算法进行了有机结合,可灵活应用不同的网络结构、决策算法和不同场景下的专家知识进行决策,具备将专家知识规则化应用的功能,便于军事研究人员开展实际研究。

2 规则引导的智能体决策方法

在提出规则引导的智能体决策框架基础上,本节从规则引导机制的思路、仿真环境建模、智能体决策流程以及网络结构四个方面介绍决策框架的实现细节。

2.1 规则引导机制

深度强化学习智能体在训练过程中需要面对数据质量低、数据利用率不高、稀疏奖励、探索与利用等难题,导致模型难以收敛或训练时间过长。从数据来源分析,上述问题产生的很大一个原因在于训练初期智能体普遍采用随机探索策略。由于缺乏目的性,随机策略在复杂任务的探索中难以频繁地发现解决问题的行动序列,导致生成的数据利用价值较低,极大地影响了训练效果。针对这个问题可采用辅助强化学习思想[11],引入外部信息指导智能体的训练过程,常见的指导方式如图 2 所示。

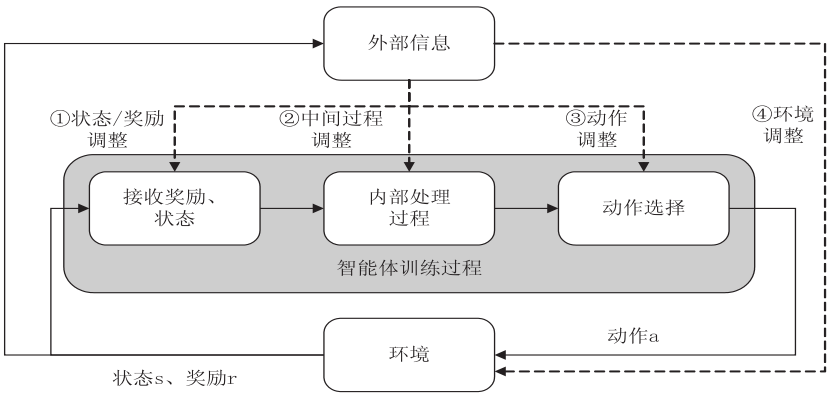

图 2 外部信息辅助强化学习途径

状态和奖励是智能体进行决策以及决策优化最重要的依据,根据需要对输入的状态进行扩充、简化或者在奖励中加入额外信号,有助于帮助智能体快速收敛;中间过程调整主要通过调整智能体训练时的超参数来改变训练效果,如智能体因陷入局部最优而反复尝试同一个动作时,可增加学习率或贪婪系数等超参数,鼓励其尝试更多的动作;动作调整影响的是智能体产生动作的过程,如根据规则推理或模仿专家范例直接选择动作等,帮助智能体直接积累正确的经验;环境调整主要是对状态空间、动作空间、智能体的初始状态等训练环境的相关设置进行调整,通过改变任务的难易度来影响智能体的训练效果。

对状态空间的充分探索有利于发现稀疏奖励,避免陷入局部最优。但从人类的角度来看,许多问题中,动作对于完成任务的必要性会根据不同的状态发生变化。一些状态下的可执行动作明显与任务目标无关,它们时常使得智能体花费时间进行探索,甚至深陷其中。最直接的方法就是对这些无效的动作施以惩罚(对应图 2 中的②),或者将多个低层次的动作组合为高层次动作(对应图 2 中的③),最近的一项研究[12]总结了常用的三种动作空间简化的方法,包括离散空间组合、连续空间离散化、和移除无效动作,并通过充分的实验发现移除无效动作和连续空间离散化这两项措施对改善智能体表现有明显的效果。而另一项研究[13]表明,策略梯度算法在探索较大的动作空间时,对无效动作进行惩罚并不能改善智能体的训练效果,但使用掩码将无效动作进行遮蔽后再进行动作采样,模型更容易收敛,并且在去除遮蔽后仍然保持了一定的性能。

受以上研究启发,军事领域的环境复杂,但基于仿真系统的作战方案推演可以结合专家经验将合法动作以规则的形式进行筛选,根据任务需求以及可能遇到

的状态对动作空间进行动态过滤而不是简单地移除,在简化环境的同时,能够在一定程度上引导智能体进行更加高效的探索,提高算法本身生成有效样本的概率,促进模型的快速收敛。虽然规则引导机制依赖专家经验的形式化表示,但它容易在具有可调整环境和规则库的仿真系统中实现,并且与其他方法相比,规则引导机制直接作用于动作空间,能够与大部分强化学习算法有效结合,具有较强的实用性。

2.2　仿真环境建模

作战仿真系统是人类对于特定军事活动建立的计算机模型,对系统中各个要素的描述必须遵循相关的作战实验规则,不同的规则决定了描述对象的不同特点。围绕实体、行为、交互这三类系统基本要素将作战实验规则分为物理规则、行为规则和裁决规则[14]。物理规则是描述仿真实体物理结构、功能、组合方式等固有属性的规则,行为规则是描述仿真实体执行作战行动时需满足的主客观条件的规则,裁决规则是描述仿真实体发生交互后产生效果的规则。强化学习智能体与仿真系统进行交互时,物理规则与裁决规则隐含于态势信息中,为智能体决策所需的重要依据,而行为规则直接影响动作空间大小,决定了动作搜索的难度。通过 2.1 节的描述可以知道,规则除了正确约束仿真系统的运行外,还可以用于提高智能体的训练效率。为了更好地描述规则引导机制,需要先对仿真环境的关键要素进行建模。

基于作战仿真系统的推演属于即时战略博弈,具有实时性、不确定性、不完全信息等特点,可以将推演活动定义为元组 $G=(S,A,P,\rho,L,W)$。其中,S、A、P 分别指特定想定下的状态空间、混合动作空间和实体所属方;$\rho:S_t\times A_1\times A_2\to S_{t+1}$ 为状态转移函数,规定了在时间 t 时,对抗双方(假设参加推演人数为两人)分别执行动作 A_1 和 A_2 后状态 S_t 向 S_{t+1} 转移的方式,包含了对动作执行效果的裁决;$L:S\times A\times P\to\{\text{True},\text{False}\}$ 为合法动作判别函数,在当前态势下,根据仿真系统中定义的规则对指挥员执行的动作的合法性进行判别;$W:S\to P\cup\{\text{ongoing},\text{draw}\}$ 为态势判别函数,根据双方态势返回博弈的进展情况,根据是否结束可输出的对抗状态包括获胜方、正在进行和平局。按照上述描述,元组 G 中的 L 对应行为规则,ρ 和 W 对应裁决规则。

与单纯的离散动作空间或连续动作空间不同,推演决策问题涉及混合动作空间[15],这类动作空间的特点是:完成动作时,在指定离散的动作类型后还需要指定该动作相关的连续动作参数,如完成实体机动需要选择执行动作的实体 u,选择对应的动作函数 a_{type},并指定机动目标的坐标 $a_{\text{arg}}=(x,y)$,上述选择构成元组 $a=(u,a_{\text{type}},a_{\text{arg}})$,并最终封装为系统可以调用的动作函数。智能体在自动推演过程中,不论最终的目标如何,在每个时刻做出的决策必定能够分解为选择实体,选择动作类型和选择动作参数这三个微观动作,选择实体视为特殊的动作,因此动作空间主要分为动作类型空间和动作参数空间,即 $A=(A_{\text{type}},A_{\text{arg}})$。

2.3　规则引导的智能体决策流程

规则引导的智能体决策流程从整体来看分为 5 个模块,包括态势感知模块、策略生成模块、强化学习模块、数据存储模块以及辅助模块,辅助模块包括注意力层和规则引导层。各模块间的交互情况如图 3 所示。

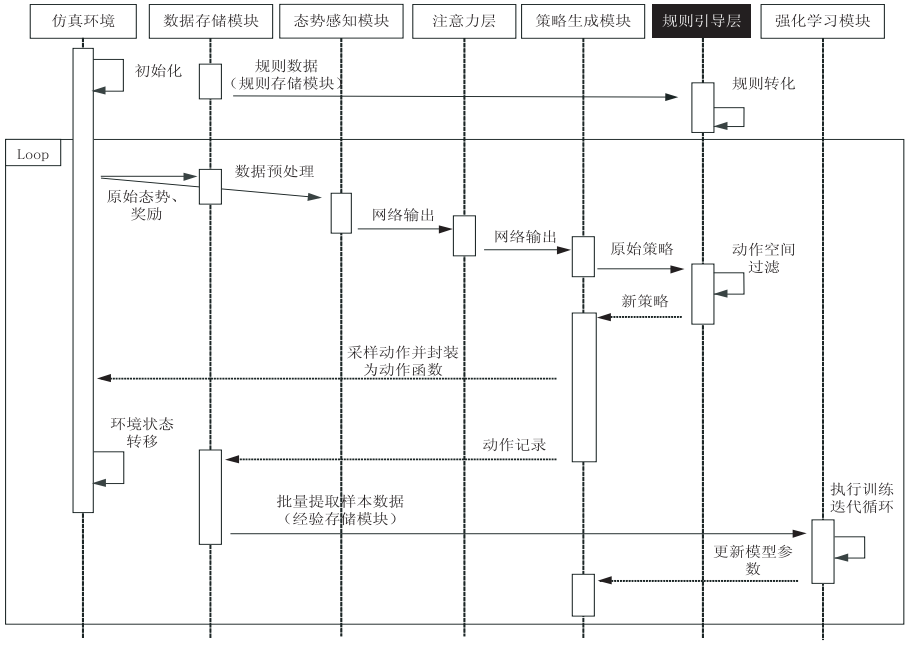

图 3　规则引导的智能体决策流程

数据存储模块包含经验存储模块和规则存储模块,前者采集环境和智能体在每个时间步产生的当前状态、动作、奖励等交互信息,存储为样本数据,在训练环节输入神经网络进行训练,后者存储了专家经验知识,在训练开始前转化为形式化的规则构建规则引导层以辅助智能体决策;态势感知模块接收环境输入的实时态势信息,包括反映地图的特征层,以及各实体的状态信息,数据输入后经过筛选、转换、归一化等预处理,输入神经网络;策略生成模块输出原始策略,由规则引导层进行动作空间的过滤后返回新的策略,各动作参数经过采样后被封装为环境可识别的指令输出;强化学习模块是整个框架的核心,可根据需求实现相应的强化学习算法,提取经验存储模块中的样本数据进行训练,并提供网络的更新参数,本流程中使用近端策略优化算法[17]进行策略梯度更新;辅助模块是设计者为优化智能体训练过程,提高智能体性能等目的而设计的数据处理模块,在本流程设计中为注意力层和规则引导层,其中注意力层为态势中不同的实体位置计算权重,具体实现参考文献[16];规则引导层简化环境的动作空间,引导智能体正确决策。

智能体网络结构如图 4 所示。为处理复杂的星际争霸环境状态输入,使用多层卷积网络接收连续 5 帧的屏幕信息和小地图信息,网络之间使用残差链接;使用编码层接收单位状态和历史动作等非空间信息,生成一维向量。将两者的输出进行维度调整、拼接后输入注意力层。注意力层直接输出的向量保留了空间特征,经过上采样层进行逆卷积后与输出动作一起作为空间参数输出层的特征输入。另外一个输出则采用平铺层进行拉伸,与非空间信息编码拼接后一同作为动作、辅助参数以及状态价值的输入。动作输出层、辅助参数输出层及价值输出层都是全连接层,输出对应参数概率分布和价值。

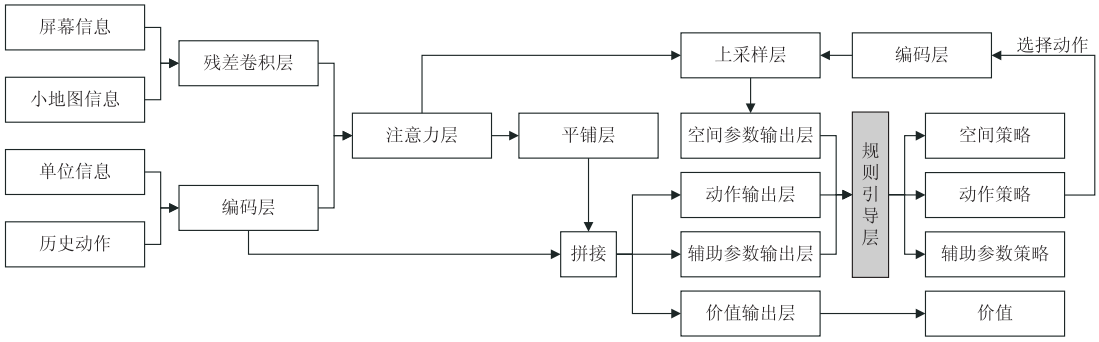

图 4 智能体网络结构

规则引导层根据知识库中专家经验的描述生成规则集 R。规则集 R 使用产生式规则描述,其中约束动作类型空间的规则表示如下:

$$r_i:\text{IF } C_1 \wedge C_2 \wedge \cdots \wedge C_n \text{ THEN } \{a_1,a_2,\cdots,a_m\} \text{ is invalid} \tag{1}$$

选定动作为 a_j 后,约束对应参数空间的规则表示如下:

$$r_i:\text{IF } C_1 \wedge C_2 \wedge \cdots \wedge C_n \text{ and } a_{\text{type}} = a_j \text{ THEN } \{a_{\text{arg1}},a_{\text{arg2}},\cdots,a_{\text{argm}}\} \text{ is invalid} \tag{2}$$

式(1)和式(2)分别对应动作类型空间和动作类型为 a_j 时动作参数空间的约束规则,式中的 $\wedge$ 也可替换为 $\vee$。C_i 为态势的限制条件,满足条件则动作空间中的相关参数集合将被设为无效参数,并生成相应的掩码。掩码操作则是在原始策略的基础上将无效动作对应的概率值设为负无穷,确保其无法被采样,而规则生成的掩码根据不同态势具有不同的值,这样就实现了动作空间的动态约束。在给定状态 s 下,神经网络输出的原策略分布 l_θ 输入规则引导层,与规则引导层产生的掩码 L_R 相乘,得到最终的策略 π'_θ,可描述如下:

$$\pi'_\theta(\cdot \mid s_t) = \text{softmax}(L_R(l_\theta(s))) \tag{3}$$

$$L_R(l_\theta(s)) = \begin{cases} l_\theta(a_i \mid s) & \text{if } a_i \text{ is valid in } R \\ -\infty & \text{otherwise} \end{cases} \tag{4}$$

因此综合上述描述,该文采用近端策略优化算法进行训练的流程如算法 1。根据演员-评论家(Actor-Critic)方法构建了共享参数的智能体网络,目的是减少训练的开销;设置参照网络 θ_{old},采用梯度裁剪的方法限制策略梯度更新的幅度,提高算法训练的稳定性,具体实现可参考文献[17]。

算法 1:规则引导的近端策略优化算法。

输入:初始化网络参数 θ,经验池容量,规则集 R

输出:最优网络参数 θ^*,最优策略 π^*

1. 载入规则集 R,初始化规则引导层;

2. 对于训练回合 1,2,…,M:

3. 初始化环境状态 s_0,

4. 对于回合中的第 n 步:

5. 获取当前环境状态 s_n,输入规则引导层,生成掩码 L_R;

6. 根据原始策略 l_θ 和 L_R,得到新策略 π'_θ;

7. 根据策略 π'_θ 执行动作 $a_n = \pi'_\theta(s_n)$,获得奖励 r_n 和下一个状态 s_{n+1};

8.　　将交互数据 (s_n, a_n, r_n, s_{n+1}) 存入经验池;

9.　　执行 L 步后,根据 PPO 算法更新网络参数 θ;

10. 更新参数 $\theta_{old} = \theta$;

11. 单个训练回合结束;

12. M 个训练回合结束。

3　实验验证

该文基于 python3.8 和 pytorch1.7.0 工具包以及星际争霸 II 机器学习接口 pysc2[18] 进行决策方法的应用与评估。

3.1　实验设置与任务描述

实验场景为星际争霸 II 的战胜蟑螂和训练陆战队员。战胜蟑螂的初始状态为 9 名陆战队员和 4 只蟑螂,击败一只蟑螂可以得到 10 分,一名陆战队员死亡扣 1 分,当 4 只蟑螂全部被消灭后地图会重置双方的位置和数量,智能体要在有限的时间中控制陆战队员击杀数量尽量多的蟑螂。蟑螂的生命力与攻击力较高,陆战队员必须集中火力攻击才能有效地取得较高的分数。训练陆战队员的初始状态为 1 个指挥中心和 12 个工人,通过训练工人、采集水晶矿、建造补给站、建造兵营等步骤在有限的动作步长内尽可能多地训练陆战队员,建造和训练动作需要消耗水晶矿,工人和陆战队员需要占用补给,每训练一个陆战队员可以得到 1 分的奖励。此外,该任务还对智能体提出了规划要求,完成任务的同时要统筹好资源和补给的平衡,各步骤关系如图 5 所示。

从总体上来看,虽然环境本身实时提供了合法动作列表,但依然存在许多无效动作,使得两个任务面临较大的动作空间和稀疏奖励的挑战。

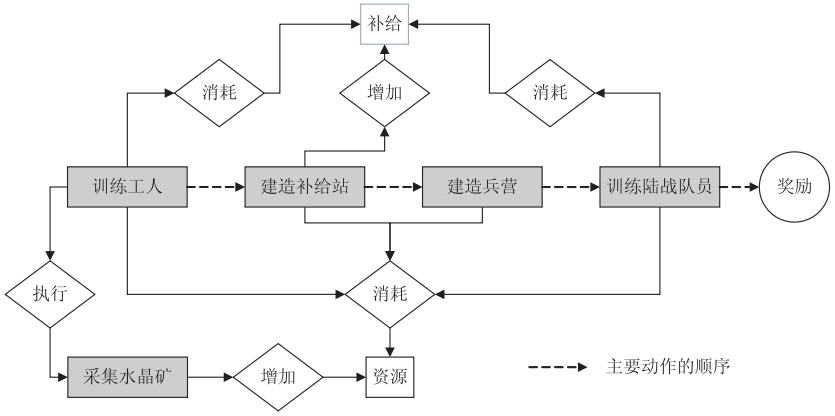

图 5　训练陆战队员中关键动作的关系

战胜蟑螂任务相对简单,只需要智能体控制陆战队员快速集火敌人即可,更多是考验智能体感知地图上敌人位置的能力,该文将规则设置为只允许选择单位与攻击这两个动作。训练陆战队员的环境稍显复杂,实验前对随机策略与无规则引导的智能体行为进行分析,发现无规则引导下,环境的合法动作平均为 10 个,最多可达 15 个。智能体初期探索的时候就容易花费大量时间操作工人四处移动而不是采集矿物,中期容易反复执行建造命令,或者操作陆战队员进行无意义移动和攻击等,这些动作对完成任务没有任何帮助。另外,任务设置中陆战队员被消灭后不会受到惩罚,还能节省补给数量,此条经验可嵌入规则中帮助提高任务成绩。因此制定以下规则对动作空间进行约束,主要规则简单描述如下:

(1)禁止小地图动作。

(2)禁止工人和陆战队员通过移动指令进行移动。

(3)禁止建筑设置生产单位集结点。

(4)陆战队员的攻击目标限定为陆战队员。

(5)限定建造补给站和兵营的地图坐标范围。

(6)限制补给站个数。

为与人类操作速度相近,智能体每 1 秒执行一个动作。游戏环境的动作接口涉及到动作类型、辅助参数和空间参数三种,考虑到本任务中的辅助参数对任务影响不大,规则集中将特定动作对应的辅助参数设为定值。最终规则引导的智能体动作集被限制为:选择单位、训练工人、训练陆战队员、建造补给站、建造兵营、采集水晶矿和陆战队员攻击。

实验将有规则引导层的智能体(rule-guided-agent)、无规则引导层的智能体(attention-agent)以及无注意力层的智能体(no-attention-agent)进行对比,无注意力智能体的注意力层被替换为相同层数的卷积层。智能体输入状态包括屏幕特征层中的单位类型、已选单位信息等空间信息,以及当前矿物数量、当前补给上限、当前占用的补给上限、场上的建筑与单位数量等标量信息。

3.2　实验结果分析

将每 30 回合的平均得分经过一定的平滑处理后进行分析。图 6 为战胜蟑螂任务中三类智能体收敛后的训练得分情况。为了减少智能体在图像认知方面的

训练时间,事先使用脚本产生的高质量动作状态序列对智能体进行了预训练。可以看出,任务本身存在攻击细节上的难度,因此得分波动幅度较大;规则引导的智能体收敛于一个平均得分最高的策略,能够做到按顺序集火消灭蟑螂,并且更加容易产生高得分的行动序列,相比之下,无规则引导的智能体和无注意力的智能体因为频繁执行攻击以外的动作,导致陆战队员得分的效率降低,并经常被蟑螂消灭,表现不稳定,收敛到了次优的策略。

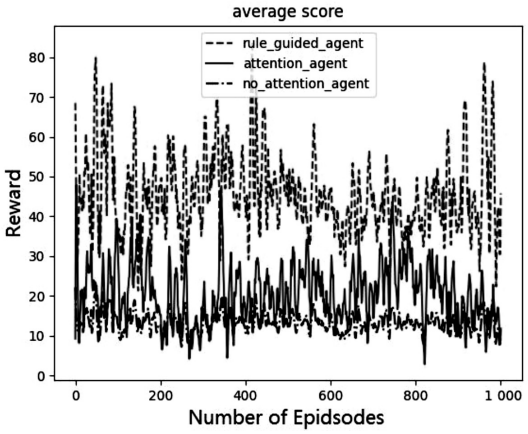

图 6 战胜蟑螂平均得分变化

训练陆战队员的平均成绩对比如图 7 所示。通过观察智能体行为可以发现,没有规则引导层和注意力机制支撑的智能体存在巨大的探索困难,除了初始随机探索获得一些奖励以外,模型最终收敛到了一个无意义的动作集中,导致无法有效地发现训练陆战队员的动作序列。而拥有注意力层的智能体经过初期探索后成功发现目标动作序列,但仍然存在操作陆战队员四处游走的现象;而且训练后期存在攻击已经建造完成的补给站和兵营等行为,导致平均得分反而有所下降。规则引导的智能体因为专家经验的指导,能够快速找到目标动作序列,比只拥有注意力层的智能体更快逼近收敛,并且通过逐步调整资源与补给的平衡,得分在后期还能稳步提升。

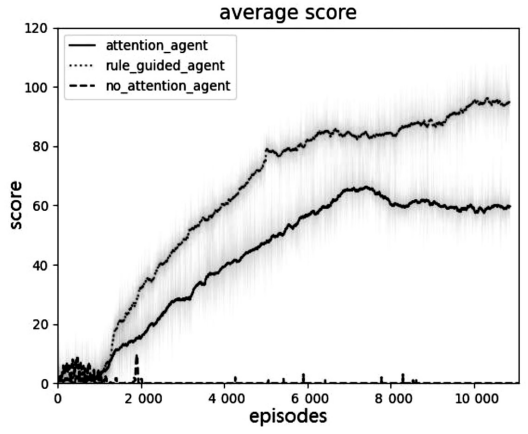

图 7 训练陆战队员平均成绩对比

为进一步评估智能体策略的稳健性,在模型收敛后撤掉规则引导层,其他设置不变,继续训练,得到的训练成绩如图 8 所示。由于失去了规则引导层的帮助,智能体的表现有所下降,但很快又恢复到原来的水平,这一结果证明规则引导机制能够帮助智能体快速地收敛到行之有效的策略中,并且更好地发挥其原有的性能。

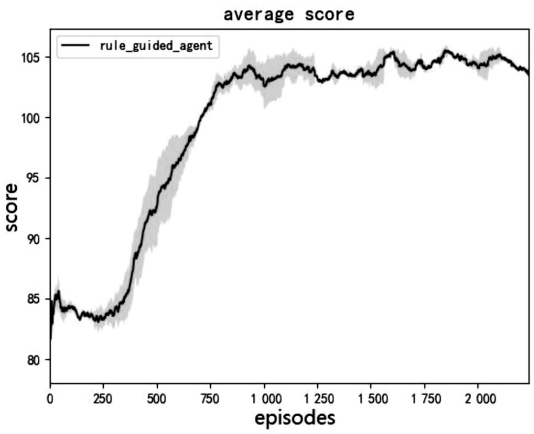

图 8 去掉规则引导层后平均得分变化

最后将本实验的结果与 DeepMind 的基准 AI[18] 和同样使用了注意力机制的关系强化学习模型[16]进行对比,见表 1。可以看到,提出的两种机制在简单任务上接近基准 AI,在规划性强的复杂任务上甚至能超过基准 AI,并且和关系强化学习模型的得分相近。关系强化学习模型使用了大规模分布式强化学习算法 IMPALA[19],在此任务上进行了约一千万局的训练,对计算资源和时间的要求较高,因此拥有普通的网络和小规模的训练难以达到的性能。相比之下,规则引导机制对算力的要求不高,并且在数据样本和计算资源有限的情况下,使用合理的网络结构和相对简单的算法同样能够达到不错的效果。另外,作为对照,在随机策略的智能体中应用规则引导机制,平均得分也能够接近基准 AI,进一步说明对动作空间的合理约束能够排除干扰,增强智能体的探索能力。

表 1 最佳平均得分对比

玩家	战胜蟑螂	训练陆战队员
DeepMind 业余玩家	41	138
星际争霸职业玩家	215	133
DeepMind 随机策略	1	1
DeepMind FullyConvLSTM 网络	98	6
DeepMind Atari 网络	101	1
Relational Agent	303	123
Control Agent	295	120
规则引导的随机策略	2	4
无引导的智能体	38	65
规则引导的智能体	81	110

4　结束语

该文提出了一种规则引导的智能体决策生成框架,利用专家经验生成的掩码对智能体的动作空间进行动态过滤,起到了简化环境作用,在复杂环境中帮助智能体进行更多的有效探索,加快了模型的收敛,并且在去掉规则引导层后仍然能够保持对于环境的适应性。结合课程学习的思想,规则引导机制还可以通过一步步减少规则约束来实现环境简单到复杂的变化,具有进一步应用的价值。它也存在一定的局限性,比如机制的实现依赖人工设计,在没有规则库或无法调整环境的场景中应用相对困难,复杂的专家经验转化为形式化规则存在描述困难问题等。对于依靠仿真系统进行的智能化作战推演来说,如何从实际想定出发,针对不同的态势制定引导规则是下一步研究的方向。

参考文献:

[1] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[2] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[3] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[4] 施　伟,冯旸赫,程光权,等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.

[5] PIAO H, SUN Z, MENG G, et al. Beyond-visual-range air combat tactics auto-generation by reinforcement learning[C]//2020 international joint conference on neural networks (IJCNN). Glasgow: IEEE, 2020: 1-8.

[6] 陈希亮,张永亮. 基于深度强化学习的陆军分队战术决策问题研究[J]. 军事运筹与系统工程, 2017, 31(3): 20-27.

[7] 李　琛,黄炎焱,张永亮,等. Actor-Critic 框架下的多智能体决策方法及其在兵棋上的应用[J]. 系统工程与电子技术, 2021, 43(3): 755-762.

[8] 谭玉玺,王洪军,侯　俊. 陆军作战指挥决策仿真模型构建[J]. 指挥控制与仿真, 2021, 43(2): 57-60.

[9] 刘　满,张宏军,郝文宁,等. 战术级兵棋实体作战行动智能决策方法[J]. 控制与决策, 2020, 35(12): 2977-2985.

[10] 张　可,郝文宁,余晓晗,等. 基于遗传模糊系统的兵棋推演关键点推理方法[J]. 系统工程与电子技术, 2020, 42(10): 2303-2311.

[11] BIGNOLD A, CRUZ F, TAYLOR M E, et al. A conceptual framework for externally-influenced agents: an assisted reinforcement learning review[J]. arXiv: 2007.01544, 2020.

[12] KANERVISTO A, SCHELLER C, HAUTAMÄKI V. Action space shaping in deep reinforcement learning[C]//2020 IEEE conference on games (CoG). Osaka: IEEE, 2020: 479-486.

[13] HUANG S, ONTAÑÓN S. A closer look at invalid action masking in policy gradient algorithms[J]. arXiv: 2006.14171, 2020.

[14] 王佳胤,张宏军,程　恺,等. 作战实验规则形式化表达建模研究[J]. 火力与指挥控制, 2020, 45(10): 54-62.

[15] XIONG J, WANG Q, YANG Z, et al. Parametrized deep q-networks learning: reinforcement learning with discrete-continuous hybrid action space[J]. arXiv: 1810.06394, 2018.

[16] ZAMBALDI V, RAPOSO D, SANTORO A, et al. Relational deep reinforcement learning[J]. arXiv: 1806.01830, 2018.

[17] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv: 1707.06347, 2017.

[18] VINYALS O, EWALDS T, BARTUNOV S, et al. Starcraft ii: a new challenge for reinforcement learning[J]. arXiv: 1708.04782, 2017.

[19] ESPEHOLT L, SOYER H, MUNOS R, et al. Importance weighted actor-learner architectures: scalable distributed deep-rl with importance weighted actor-learner architectures[J]. arXiv: 1802.01561, 2018.