软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(8):3821−3835 [doi: 10.13328/j.cnki.jos.006593]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

显式知识推理和深度强化学习结合的动态决策*

张昊迪[1], 陈振浩[1], 陈俊扬[1], 周 熠[2], 连德富[3], 伍楷舜[1], 林方真[4]

[1](深圳大学 计算机与软件学院, 广东 深圳 518052)

[2](上海脑科学与类脑研究中心, 上海 200031)

[3](中国科学技术大学 计算机科学与技术学院, 安徽 合肥 230026)

[4](香港科技大学 计算机科学与工程系, 香港 999077)

通信作者: 伍楷舜, E-mail: wu@szu.edu.cn; 林方真, E-mail: flin@cse.ust.hk

摘 要: 近年来, 深度强化学习在序列决策领域被广泛应用并且效果良好, 尤其在具有高维输入、大规模状态空间的应用场景中优势明显. 然而, 深度强化学习相关方法也存在一些局限, 如缺乏可解释性、初期训练低效与冷启动等问题. 针对这些问题, 提出了一种基于显式知识推理和深度强化学习的动态决策框架, 将显式的知识推理与深度强化学习结合. 该框架通过显式知识表示将人类先验知识嵌入智能体训练中, 让智能体在强化学习中获得知识推理结果的干预, 以提高智能体的训练效率, 并增加模型的可解释性. 将显式知识分为两种, 即启发式加速知识与规避式安全知识. 前者在训练初期干预智能体决策, 加快训练速度; 而后者将避免智能体作出灾难性决策, 使其训练过程更为稳定. 实验表明, 该决策框架在不同强化学习算法上、不同应用场景中明显提高了模型训练效率, 并增加了模型的可解释性.

关键词: 知识表示与推理; 可解释性; 深度强化学习; 动态序列决策

中图法分类号: TP18

中文引用格式: 张昊迪, 陈振浩, 陈俊扬, 周熠, 连德富, 伍楷舜, 林方真. 显式知识推理和深度强化学习结合的动态决策. 软件学报, 2023, 34(8): 3821–3835. http://www.jos.org.cn/1000-9825/6593.htm

英文引用格式: Zhang HD, Chen ZH, Chen JY, Zhou Y, Lian DF, Wu KS, Lin FZ. Dynamic Decision Making Based on Explicit Knowledge Reasoning and Deep Reinforcement Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(8): 3821–3835 (in Chinese). http://www.jos.org.cn/1000-9825/6593.htm

Dynamic Decision Making Based on Explicit Knowledge Reasoning and Deep Reinforcement Learning

ZHANG Hao-Di[1], CHEN Zhen-Hao[1], CHEN Jun-Yang[1], ZHOU Yi[2], LIAN De-Fu[3], WU Kai-Shun[1], LIN Fang-Zhen[4]

[1](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518052, China)

[2](Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 200031, China)

[3](School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China)

[4](Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China)

Abstract: In recent years, deep reinforcement learning has been widely used in sequential decisions with positive effects, and it has outstanding advantages in application scenarios with high-dimensional input and large state spaces. However, deep reinforcement learning faces some limitations such as a lack of interpretability, inefficient initial training, and a cold start. To address these issues, this study proposes a dynamic decision framework combing explicit knowledge reasoning with deep reinforcement learning. The framework

* 基金项目: 国家自然科学基金 (61806132, U2001207, 61872248); 广东省自然科学基金 (2017A030312008); 深圳市自然科学基金 (ZDSYS20190902092853047, R2020A045); 珠江人才计划 (2019ZT08X603) ; 广东省普通高校创新团队项目 (2019KCXTD005)
本文由“智能系统的分析和验证”专题特约编辑明仲教授、张立军教授和秦胜潮教授推荐.
收稿时间: 2021-09-05; 修改时间: 2021-10-14; 采用时间: 2022-01-10; jos 在线出版时间: 2022-01-28
CNKI 网络首发时间: 2023-01-19

successfully embeds the priori knowledge in intelligent agent training via explicit knowledge representation and gets the agent intervened by the knowledge reasoning results during the reinforcement learning, so as to improve the training efficiency and the model's interpretability. The explicit knowledge in this study is categorized into two kinds, namely, heuristic acceleration knowledge and evasive safety knowledge. The heuristic acceleration knowledge intervenes in the decision of the agent in the initial training to speed up the training, while the evasive safety knowledge keeps the agent from making catastrophic decisions to keep the training process stable. The experimental results show that the proposed framework significantly improves the training efficiency and the model's interpretability under different application scenarios and reinforcement learning algorithms.

Key words: knowledge representation and reasoning; interpretability; deep reinforcement learning (DRL); sequential decision making

深度强化学习 (deep reinforcement learning, DRL)[1,2]将深度神经网络和强化学习结合, 近年来已被成功应用在诸多序列决策领域, 尤其在如 AlphaGo、OpenAIFIve、Atari 等大规模状态空间的决策问题中优势明显. 自 2013 年深度 Q 网络 (deep Q-network, DQN) 被提出后, 大量的深度强化学习算法与模型被相继提出, 其中基于价值的强化学习算法 (例如 DQN[1,2]、Double DQN[3]和 Dueling DQN[4]) 和基于策略的强化学习算法 (例如 A3C[5]、PPO[6]和 SAC[7]等) 都在决策领域表现出良好效果. 然而, 完全由数据驱动的深度强化学习方法也存在一些问题. 首先, 作为黑盒模型的深度强化学习方法缺乏可解释性, 其方法的基本假设是领域任务相关的智能或知识可以被深度神经网络隐式地、分布式地表示. 然而人类进行表示与推理的知识通常以显式方式呈现, 如一阶逻辑、非单调推理、动作语言等. 可解释性的缺失导致深度强化学习难以利用人类显式知识解决模型本身存在的问题, 例如训练初期的冷启动与低效问题. 深度神经网络往往需要大量的训练才能达到良好效果, 而在智能体与环境中的交互过程中, 所获奖励驱动下的更新往往存在低效的问题. 由于深度神经网络相关方法的黑盒特性, 强化学习过程中在状态空间中的探索缺乏理论依据. 尤其是训练初期, 模型对于状态-动作奖励函数估计不准时, 所做出的动作决策具有较大的随机性. 因此, 在训练稳定之前, 智能体的这种随机探索很有可能作出不良动作, 甚至灾难性的决策. 此问题在机器学习领域被称为"冷启动". 对于在真实场景或仿真成本高昂的环境下训练的智能体, 冷启动严重影响了深度强化学习方法的实用性. 产生该问题的原因在于智能体缺乏对环境或任务的基本认知, 而深度神经网络又不支持显式的知识表示, 无法便捷地将这些基本认知嵌入网络模型中. 为了解决这一问题, 一些模型与方法被相继提出. 例如, 有部分学者认为, 人类学习一项新的技能或操作时最直接有效的途径就是模仿他人从而学习其中的知识. 基于此观察而被提出的模仿学习[8]和演示学习[9]通过学习人类专业的演示动作, 使智能体习得给定领域中的决策方法. 然而在许多情况下, 通过模仿学习和演示学习来学得知识的方式受限于以下 3 方面因素: (1) 模型对于包含人类知识的训练数据质量要求极高; (2) 训练数据获取成本极大, 需要人类大量重复性的操作; (3) 是在特定情况下人类无法作出演示. 人们提出一些新的方法试图缓解以上问题, 例如以动作序列代替单个动作作为训练数据, 通过人类评价某一状态在所作出的决策[10]; 人类根据自己的偏好对智能体所作出的大量决策轨迹选出较优秀的一批轨迹[11,12]; 智能体逐个实现人类制定的高层级目标[13]等. 虽然以上方法一定程度上降低了训练数据的质量门槛和获取成本, 但模型训练过程中仍需要大量的人工干预.

本文针对深度强化学习中的缺乏可解释性与训练效率低的问题, 提出了一种基于显式知识推理和深度强化学习的决策框架, 以提高智能体的训练效率. 本文中的显式知识可以是启发式的加速规则, 对智能体在训练初期加以启发式的正向引导, 以避免过多的无效探索; 也可以是规避式的安全规则, 避免智能体在训练过程中做出灾难性的动作. 本文的基本假设是显式表达的知识对于模型设计者是直观的、易懂的, 且符合人类逻辑. 因此, 通过将显式知识整合到深度强化学习模型中, 不仅使得智能体能够更快更好的进行学习, 提高训练效率与效果, 而且增加了深度强化学习模型的可解释性.

本文的主要贡献包括如下 3 方面.

(1) 针对深度强化学习中的缺乏可解释性的问题, 提出了一个知识推理与深度强化学习结合的决策框架, 其中的显式知识推理增加了决策系统的可解释性.

(2) 在显式知识与深度神经网络的结合上, 综合考虑知识生效机制的一般性需求, 提出两种模式的显式知识, 即启发式的加速知识与规避式的安全知识. 前者有效地提高了训练初期模型表现, 后者为模型训练提供安全性保障, 有效地提高了模型的训练效率.

(3) 在多种场景下, 对多种深度强化学习算法进行了实证研究. 结果显示, 该动态决策框架中显式知识的结合方式及效果具有一般性, 不依赖于特定场景与特定算法.

本文第 1 节介绍关于人类知识与深度强化学习相结合的相关工作以及本工作的研究动机与意义. 第 2 节简要介绍本文工作的基础知识. 第 3 节详细介绍了基于显式知识推理和深度强化学习的动态决策框架 (KB-DRL), 包括框架的特点、训练过程和技术细节. 第 4 节讲述了所提的决策框架的实验环境、实验设计和分析结果. 第 5 节则对本框架的研究分析得出最后总结, 以及介绍未来的研究方向.

1 相关工作

自 2013 年 DQN[1]被提出后, 深度强化学习受到了广泛关注. 2015 年目标网络分离的 DQN[2]版本被提出. 其后很多变体模型相继出现, 包括 Double DQN[3]、Dueling DQN[4]、C51 DQN[14]、Bootstrapped DQN[15]和 Rainbow DQN[16]等. 除了以上这些基于价值的深度强化学习算法, 基于策略的相关算法, 包括 DPG[17]、DDPG[18]、A3C[5]、TRPO[19]、PPO[6]和 SAC[7]等, 也在不同领域与任务中表现出良好效果. 然而这些深度强化学习算法在实际应用中都存在数据依赖和训练低效等问题, 而且缺乏可解释性. 怎样更好结合并利用抽象的、可解释的领域知识, 成为近年来人工智能领域的研究热点.

部分学者关注于让模型在任务中模仿人类行为. 当人类需要学习一项新的技能或操作时, 一种最直观高效的方式就是模仿其他人的演示, 即从演示者对该任务或操作的理解与执行中进行直接模仿, 是一种利用他人知识进行引导性学习的方式. 例如, 序列决策任务中端到端学习的模仿学习算法 (imitation learning)[8]可以让智能体直接模仿人类演示专家的行为从而获提高学习效率, 即仅通过人类演示专家在应场景下的正确动作序列 $\{(s_t,a_t),(s_{t+1},a_{t+1}),\ldots\}$ 传递给智能体. 模仿学习可以主要分为行为克隆算法 (behavioral cloning)[20]和逆强化学习算法 (inverse reinforcement learning)[21]两大类. 行为克隆智能体通过监督学习直接模仿学习人类的策略; 而逆强化学习智能体则可以根据人类的演示估算出其中的奖励函数, 再通过标准的强化学习算法进行学习. 但模仿学习相关方法面临一个主要问题, 即模仿学习算法对人类演示专家的演示数据质量要求极高. 因为部分质量不达标的数据对智能体而言将可能是误导性的演示, 而这种误导性的演示对学习过程的影响又难以矫正, 因此高质量数据的制作成本非常大. 此外, 并非所有的场景都能由人类专家提供高质量的演示, 如无人机一些复杂场景、陌生环境中的高难度飞行任务. 一些工作提出的实时评价反馈的人类知识强化学习算法, 能够在一定程度上缓解该问题. 智能体根据环境做出决策, 人类观察智能体在该环境下所作出的决策进行实时评价, 给出反馈值, 以表达对该决策的满意程度, 智能体再根据反馈值优化模型. 其中对智能体的决策进行实时最优判断的相关的工作包括 Policy Shaping[9,22], 其简单直接地对智能体的决策进行评价对错, 以表明是否做出最优决策. 与 Policy Shaping 类似的 Reward Shaping[23]则可以将人类的反馈直接作为奖励函数并直接代替原有的奖励函数, 如, TAMER[24]将人类对某一状态下的动作分为 3 种评价, 即负面、中立和正面评价, 对应的人类奖励函数值为−1, 0 和 1. 又如, Deep TAMER[25]加入了深度神经网络以估计人类对某状态对应动作的反馈值函数, 相关的 TAMER 优秀衍生算法还有 TAMER+RL[26]和 DQN-TAMER[27]等.

另一种整合人类知识的方法是人为干预, 即人类专家观察智能体的整个训练过程, 当智能体决策出危险动作时, 人类专家用一个安全动作去代替这个危险动作, 以避免灾难性的后果发生, 相关的工作有 HIRL[28]等. 此类方法与 Policy Shaping 和 Reward Shaping 的不同之处在于 HIRL 只有负反馈没有中立和正面反馈. 当在智能体做出灾难性动作时, 人类会做出安全动作以保护智能体, 并传递给智能体一个新的奖励值. 模型 HIRL 的另一个特性是人类决策可以直接被作用在实际决策动作上. 但 HIRL 同样面临着高昂的人力成本问题, 即使智能体做出灾难性动作的概率较低, 但人类专家依然需要全程监测智能体的整个学习过程. 另一个相关的难题是, 在许多深度强化学习的任务中, 人类很难给出高质量的演示和精确的反馈值. 而且由于任务智能体的动作形态和人类有可能有巨大差异, 人类也无法给出高质量的演示数据. 一种可行的解决方式是根据智能体的多组决策序列, 人类专家根据各组决策序列的表现和自己的偏好选出较优的决策序列[11,12].

层次化深度学习是另一类整合人类知识的方法. 例如, 在一些过于复杂的问题中, 尤其是带有大量延时奖励的

任务中, 部分强化学习方法可能无法学得一个较为合理的策略. 而在分层指导框架[13]中, 人类专家可以提前将复杂任务划分为多个子目标. 高层决策模块可以根据当前状态, 选出一个子目标; 底层决策模块则根据当前状态与可执行的原子动作来实现当前子目标. 当前子目标完成后, 高层决策模块再选出新的子目标, 直至整体任务被完成. 这种分层学习方法在一些涉及复杂规划、延时奖励的任务中可以提高学习效率, 例如一个机器人要学会从高层楼下电梯, 则需要分为先移动到电梯旁、按下电梯按钮、进入电梯和离开电梯这几个步骤, 人类通过为智能体提前制定好所需的子目标, 可以快速提高智能体的学习效率. 此类方法基于知识表示与推理领域相关知识描述语言为基础, 如动作语言 BC[29]、BC+[30]等, 以及被广泛应用的非单调推理工具回答集编程 ASP[31]、NeurASP[32]. 然而此类方法的局限性也显而易见, 即子任务划分需要人类专家手动完成; 而在很多场景中, 任务划分带来高昂的人力成本. 今年来也有很多工作在结合深度神经网络与符号化方法上进行尝试, 包括神经产生系统 (neural production system)[33], 不确定性动作语言[34,35]等.

2 基础知识

本文所提方法主要基于深度强化学习与非单调推理, 以下介绍相关概念和基本知识.

2.1 深度强化学习

强化学习的基本思想是智能体在与环境的交互过程中迭代地学习最优决策. 强化学习智能体与环境的交互示意图如图 1. 智能体接收从环境中获得的环境状态 s_t, 再根据该环境状态决策出动作 a_t, 动作 a_t 作用于环境后获得奖励值 r_t; 在下一时刻环境发生变化, 智能体感知新的环境状态 s_{t+1}, 再做出相应决策动作 a_{t+1}. 智能体的目标是要在交互过程中学得一个最优策略, 以使期望的长期累计奖励最大化.

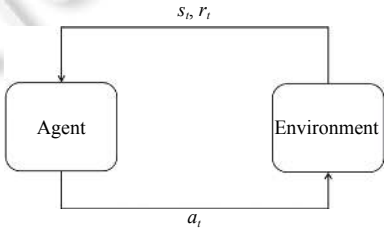


图 1 强化学习智能体交互过程

一个智能体的强化学习过程可视为一个马尔可夫决策过程 (MDP), 由五元组$<S,A,r,P,\gamma>$表示:

- 状态空间 S, 表示环境状态的集合.
- 动作空间 A, 表示智能体能够选择的所有动作的集合.
- 奖励函数 $r: S\times A\to R$, $r(s_t,a_t)$ 表示智能体根据状态 s_t 决策出动作 a_t 后, 所获得的即时奖励值, 记为 r_t.
- 状态转移概率分布 $P\sim S\times A\to S$, $P(s_t,a_t,s_{t+1})$ 表示智能体在状态 s_t 下, 执行决策动作 a_t 后, 下一时刻环境转移到状态 s_{t+1} 的概率.
- 折扣因子 $\gamma: (0\leqslant\gamma\leqslant 1)$, 未来奖励的折扣系数.

其具有马尔可夫性质, 在当前状态 s_t 下, 下一时刻的状态与之前状态 $(s_0,s_1,\ldots,s_{t-1})$ 无关, 即有:

$$P(s_t,a_t,s_{t+1}|s_0,a_0,s_1,a_1,\ldots,s_t,a_t)=P(s_t,a_t,s_{t+1}|s_t,a_t) \tag{1}$$

在大规模的状态空间中, 传统的强化学习无法计算出价值函数和策略函数, 而结合深度学习, 则可以利用神经网络来拟合强化学习中的价值函数和策略函数, 即输入是环境的状态数据, 输出是价值函数值或策略函数值, 如基于价值的 DQN[1,2]及其变体与基于策略的 A3C[5]等.

2.2 动作语言

在本文结合知识推理的框架中, 我们将以稳定模型语义下的逻辑程序作为工具、以动作语言 BC[29]为基础, 定义动作语言 BC-dynamic 来表示知识. 稳定模型语义也称回答集语义[31,36], 是表达动态领域的一种描述性逻辑方

法. 稳定模型语义下的逻辑程序非常直观, 易于解释、维护以及更新.

定义 1. 给定时变事实 (fluent) 集合 F_A, 一个逻辑程序 P 由如下形式的逻辑规则构成:

$$A_0 \leftarrow L_1,\ldots,L_m,not\ L_{m+1},\ldots,not\ L_n \tag{2}$$

其中, $A_0 \in F_A$ 为时变事实; 对于事实 $A_i \in F_A$, L_i 代表 A_i 或者 $\neg A_i$, 称为 A_i 对应的文字.

在公式 (1) 中的一元联结词 *not* 为缺省否定, 表示若没有显式证据支撑该事实为真, 则该事实为假. 一个给定的逻辑程序的稳定模型由以下不动点来定义.

定义 2. 对于逻辑程序 P 以及时变事实集合 π, P 在 π 下的归约程序 (reduction program) 记作 P_π. 对于 P 中的每一条形如公式 (1) 的逻辑规则, 若 $\{L_{m+1},\ldots,L_n\}\cap\pi=\varnothing$, 则 P_π 包含以下规则:

$$A_0 \leftarrow L_1,\ldots,L_m \tag{3}$$

归约程序 P_π 的生成可视为在原逻辑程序上进行遵循闭世界假设的操作, 即对于所有包含缺省否定 $not\ L_i$ 的规则, 如果在 π 中 L_i 被判定为真, 则将该规则整体删去; 否则将规则中的 $not\ L_i$ 部分从原规则中删去.

定义 3. 对于一个逻辑程序 P 以及时变事实集 π, 如果 π 是归约程序 P_π 的最小模型 (相对于集合包含关系), 即:

$$\pi = \arg\min_{\pi'} \pi' \models p_\pi \tag{4}$$

则 π 是一个 P 的稳定模型.

动作语言 BC 是一种用于指定状态转换系统的语言, 其语义由相应逻辑程序的稳定模型所定义. 语法上, 一个用 BC 编写的动作理论 R 由两个子集组成, 即静态规则集 R_S 以及动态规则集 R_D. 其中, 静态规则形式如下:

$$consequence \text{ caused } premise \text{ incons } justification \tag{5}$$

而动态规则形式如下:

$$consequence \text{ after } premise \text{ ifcons } justification \tag{6}$$

其中, *consequence*, *premise* 以及 *justification* 分别为规则的结论、前提与缺省条件, 形式为时变事实对应文字的合取. 静态规则描述了同一时刻不同事实间的因果关系; 而动态规则用以描述动作在给定的 *justification* 下对下一时刻状态的直接影响. 一个 BC 动作理论的语义是由一个稳定模型下的逻辑程序定义的. 对于一个以 BC 语言编写的动作理论 $R = R_S \cup R_D$, 其对应的逻辑程序 $P(R)$ 由下面规则组成.

- 对于每个在 R_S 中形式如下的静态规则:

$$A_0^t \text{ caused } \wedge_{i+1}^m A_i^t \text{ ifcons } \wedge_{j=m+1}^n A_j^t \tag{7}$$

$P(R)$ 都包含以下逻辑程序规则:

$$A_0^t \leftarrow A_1^t,\ldots,A_m^t,not\ \neg A_{m+1}^t,\ldots,not\ \neg A_n^t \tag{8}$$

- 对于每个在 R_S 中形式如下的动态规则:

$$A_0^{t+1} \text{ after } \wedge_{i+1}^m A_i^t \text{ ifcons } \wedge_{i=m+1}^n A_i^{t+1} \tag{9}$$

$P(R)$ 都包含以下逻辑程序规则:

$$A_0^{t+1} \leftarrow A_1^t,\ldots,A_m^t,not\ \neg A_{m+1}^{t+1},\ldots,not\ \neg A_n^{t+1} \tag{10}$$

- 对于每个时变事实 A^t, $P(R)$ 都包含以下选择规则:

$$A^t \leftarrow not\ \neg A^t \tag{11}$$

$$\neg A^t \leftarrow not\ A^t \tag{12}$$

- 对于每个时变事实 A^t, $P(R)$ 都包含以下存在性和唯一性规则:

$$\leftarrow not\ A^t, not\ \neg A^t \tag{13}$$

$$\leftarrow A^t, \neg A^t \tag{14}$$

以上的逻辑程序 $P(R)$ 的稳定模型即为动作理论 R 的模型. 与人类演示数据等其他形式的知识相比, 这些动作推理中的规则高度抽象并易于描述.

3 显式知识推理和深度强化学习结合的动态决策

针对现有强化学习方法的不可解释性与训练效率低的问题, 本文提出了一种基于显式知识推理和深度强化学

习的动态决策框架. 与相关工作中所提到的大部分算法不同的是, 本框架不需要人类专家在训练过程中频繁交互演示, 不需要对智能体决策进行评价和干预, 也不需要任何预先人工定义的子任务划分.

如上文所述, 本框架中的显式知识可以分为两类.

(1) 启发式的加速知识: 作为加速器以加快模型的训练, 在规则生效时, 若智能体做出非正确的决策, 加速器则用该状态下对应的正确决策代替原决策, 让智能体在训练初期做出更多有效探索.

(2) 规避式的安全知识: 作为保护器保护智能体的安全, 安全规则会在智能体内始终生效, 当智能体面临危险场景且可能做出灾难性决策时, 保护器会排除掉所有的灾难性决策, 保证智能体在训练以及应用过程中的安全性.

在本文中, 我们对以上两种知识进行统一的知识表示.

3.1 显式知识表示

我们定义 BC 动作语言的变体, 记作 BC-dynamic, 对以上两种知识进行统一表示.

定义 4. 语法上, 给定动作集 ACT, 时变事实集 F, 由 BC-dynamic 动作语言定义的知识集 R 是形式如下的规则的集合:

$$action(params, t+1)\ \text{if} \wedge_i condition_i(t)\ \text{default} \wedge_j context_j(t) \tag{15}$$

其中, $action \in ACT$, $condition_i \in F$, $context_j \in F$. 语义上, 对于给定动作集 $X \subseteq ACT$, X 是 R 的模型当且仅当 X^{t+1} 是对应逻辑程序 $P(R)$ 的稳定模型. 其中 $P(R)$ 包含如下规则.

- 对于每一形如公式 (15) 的规则与当前时刻 t, $P(R)$ 都包含以下逻辑程序规则:

$$action^{t+1} \leftarrow condition_1^t, \ldots, condition_m^t, not\ context_1^t, \ldots, not\ context_n^t \tag{16}$$

- 对于每个时变事实 $f \in F$ 与当前时刻 t, $P(R)$ 都包含以下选择规则:

$$f^t \leftarrow not\ \neg f^t \tag{17}$$

$$\neg f^t \leftarrow not\ f^t \tag{18}$$

- 对于每个时变事实 $f \in F$ 与当前时刻 t, $P(R)$ 都包含以下存在性与唯一性规则:

$$\leftarrow not\ f^t, not\ \neg f^t \tag{19}$$

$$\leftarrow f^t, \neg f^t \tag{20}$$

- 对于每个动作 $action \in ACT$ 与当前时刻 t, $P(R)$ 都包含以下选择规则:

$$action^t \leftarrow not\ \neg action^t \tag{21}$$

$$\neg action^t \leftarrow not\ action^t \tag{22}$$

- 对于每个动作 $action \in ACT$ 与当前时刻 t, $P(R)$ 都包含以下存在性与唯一性规则:

$$\leftarrow not\ action^t, not\ \neg action^t \tag{23}$$

$$\leftarrow action^t, \neg action^t \tag{24}$$

在以上逻辑程序刻画中, 时变事实的时间戳用于区分事实在不同时刻的真值. 在一个仅用逻辑程序刻画的完整动态决策系统中, 形如以上的逻辑规则需要被实例化在时间序列空间 (0, 1, …, T) 上. 而在本文中, 动作语言与逻辑程序仅被用于在神经网络一步迭代中进行推理, 因此涉及的时间戳只包含当前时刻 t 与下一时刻 t+1. 故以上时间戳可简单用常量带入, 如 t=0. 例如, 在 breakout 游戏中, 我们采用以下简单的启发式规则:

$$move(left, t+1)\ \text{if}\ \ at(ball, left, t)\ \text{default}\ \ \neg at(ball, leftest, t) \tag{25}$$

即当默认条件 (小球不在屏幕最左边界) 成立时, 当小球在挡板左侧, 则下一时刻建议动作为向屏幕左侧方向移动挡板.

3.2 动态决策框架 KR-DRL

本研究中所提出的基于显式知识和深度强化学习的动态决策框架如图 2 所示, 我们在智能体内维持一个显式知识规则库 (即上文所述 BC-dynamic 知识集), 用于表示先验知识. 当智能体与环境进行交互并根据当前状态作出决策动作时, 如果满足规则生效的条件, 该决策将传递给规则库中进行判断. 若不符合规则逻辑, 则用规则库中对应的规则决策以一定的条件代替原决策. 框架包括以下主要模块.

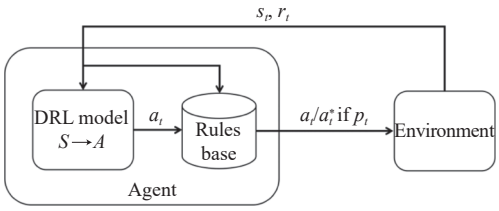


图 2 显式知识推理和深度强化学习结合的动态决策

● 深度强化学习模块: 与标准深度强化学习相同, 智能体感知来自环境的状态 s_t 后进行决策, 在环境执行决策动作 a_t, 该决策动作作用于环境中, 环境状态变化为 s_{t+1} 并产生即时奖励 r_{t+1}; 智能体追踪下一时刻的环境状态 s_{t+1} 并根据奖励进行参数学习.

● 知识规则库: 在特定的环境中, 知识规则库将维护一系列的规则, 规则将在一定的环境下生效, 用于取代智能体由模型所决策的动作.

● 知识干预机制模型: 决策框架中所维护的知识干预机制以 $P(t) = p_0 \cdot \gamma^{\lambda t}$ 的形式刻画. 对于启发式加速规则集, 系统维护一个单调递减的函数 $P(t)$, 其中 $0 < \gamma < 1$, $0 < p_0 \leqslant 1$; 而对于规避式的安全规则, 系统维护一个大小恒定为 1 的 $P(t)$, 即 $\gamma = 1$, $p_0 = 1$.

由于知识表示与推理模块相对独立, 与深度神经网络通过推理结果生效机制进行协同, 因此该框架具有广泛的适用性. 显式知识推理可以作用于基于价值的深度强化学习算法, 如 DQN、Double DQN 和 Dueling DQN 等, 还可以作用于基于策略的深度强化学习算法, 如 A2C、PPO 和 Discrete-SAC 等, 该框架的算法流程如算法 1 所示.

算法 1. 显式知识推理和深度强化学习结合的动态决策的算法流程.

输入: 知识集 R, 规则初始生效概率 p_0, 衰减系数 γ, 步长 λ, 以及所用深度强化学习设置.

1. 游戏从第 1 局开始, 直至局数上限 M 或模型收敛:
2. 初始化环境状态 s_0, 并预处理为 $\eta(s_0)$
3. 循环: 从第 1 帧开始直至第 T 帧或至该局游戏结束
4. 如果 DRL 算法为基于价值的深度强化学习, 则
5. 以 ε 概率随机选取一个动作 a_t
6. 以 $1-\varepsilon$ 概率选取动作 $a_t \leftarrow \max_a Q^*(\eta(s_t), a; \theta)$
7. 如果 DRL 算法为基于策略的深度强化学习算法, 则
8. 选取 $a_t \leftarrow \pi(a_t|s_t)$
9. 以概率 $P(t) = p_0 \cdot \gamma^{\lambda t}$ 进行知识推理与干预:
10. 计算逻辑程序 $P(R)$ 的稳定模型集合 X
11. 如果当前 $a_t \notin X$, 则在 X 中随机选取动作替换 a_t
12. 执行动作 a_t, 获得即时奖励 r_t
13. 更新状态 s_{t+1}
14. 根据 s_t, a_t, r_t 更新模型参数 θ
15. 循环结束

需要说明的是, 知识规则集产生的推理结果并不要求是最优策略.

4 实验分析

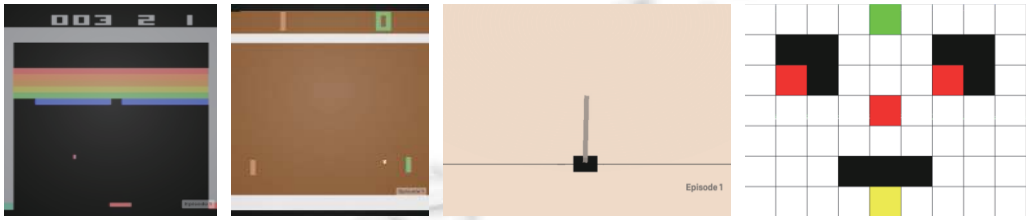
4.1 实验数据

为了验证基于显式知识和深度强化学习的动态决策框架的有效性, 我们在 Breakout、Pong、CartPole 和

GridWorld 游戏中根据环境的特性实现了我们的框架, 这些游戏分为两组, 用于分别演示启发式加速规则和规避式安全规则, 在以下的实验中, 每种规则在其对应的实验环境中独立工作, 以证明该决策框架的有效性. 需要特别说明的是, 由于实验中我们采用的知识规则不要求最优, 而只是两条以内形式统一的启发式规则. 因此我们没有按照 BC 的逻辑程序翻译方式将其翻译为回答集逻辑程序并求解, 而是直接编译为过程式模块, 在深度神经网络训练时调用.

4.1.1 Breakout

在该游戏中, 如图 3, 玩家控制位于游戏画面下方的砖块, 通过左右移动接由上方反弹下来的球, 球与球拍碰撞后, 球会向上反弹, 游戏中一共有 6 层不同颜色的砖块, 每层共有 18 块砖块, 不用颜色的砖块对应的分数不同, 由下往上数, 第 1 层和第 2 层的砖块每块 1 分, 第 3 层和第 4 层的砖块每块 4 分, 第 5 层和第 6 层的砖块每块 7 分, 当游戏中所有砖块被打完后, 会重新刷新 6 层砖块但不会结束游戏, 继续击打可以继续得分, 因此理论上该游戏得分没有上限, 当球拍接不到球时游戏才会结束.



(a) Breakout (b) Pong (c) CartPole (d) GridWorld

图 3 实验环境

我们在该游戏中使用启发式加速规则: 如果球在球拍的左边, 则球拍向左移动, 如果球在球拍的右边, 则球拍向右移动.

$$move(left,t+1)\ \text{if}\ at(ball,left,t)\ \text{default}\ \neg at(ball,leftest,t) \tag{26}$$

$$move(right,t+1)\ \text{if}\ \text{at(ball,right,t)}\ \text{default}\ \neg at(ball,rightest,t) \tag{27}$$

4.1.2 Pong

在该游戏中, 玩家控制右方球拍, 左方为敌方球拍, 双方进行乒乓球游戏, 乒乓球碰到球拍后会进行反弹, 当敌方未能接住我方的球时, 我方获得 1 分, 当我方未能接住敌方的球时, 敌方获得 1 分, 当其中一方达到 21 分时游戏结束, 总得分为我放得分减去敌方得分, 既当我方 21 分, 敌方 0 分是, 游戏得分 21, 当敌方 21 分, 我方 0 分时, 游戏得分−21.

由于该游戏和 Breakout 游戏的任务场景非常相似, 都是根据迎面而来的球与我方球拍的相对位置进行移动, 因此可以通过简单地修改 Breakout 中的规则就能在 Pong 中使用, 由于规则是高度抽象且符合逻辑的, 因此相似的任务场景能共享相同的规则集:

$$move(up,t+1)\ \text{if}\ at(ball,above,t)\ \text{default}\ \neg at(ball,top,t) \tag{28}$$

$$move(down,t+1)\ \text{if}\ at(ball,below,t)\ \text{default}\ \neg at(ball,bottom,t) \tag{29}$$

4.1.3 CartPole

在该游戏中, 有一个可水平移动的小车与一只竖杆, 竖杆一端连接着车体且可旋转. 玩家需要控制小车以维持受重力影响的竖杆的初值状态. 当竖杆与垂直方向的倾斜角度大于 15 度时、车的水平位移超过中心位置 2.4 个单位长度或达到游戏的最高奖励值 500 分时, 游戏结束. 在游戏结束前每坚持 1 个时间单位, 则获得 1 分的奖励.

在该游戏中, 根据杆的角度以及角速度进行启发式决策: 如果杆偏向左边且角速度为正, 则手推车应该向左移动, 如果杆偏向右边且角速度为负, 则手推车应该向右移动.

$$move(left,t+1)\ \text{if}\ leaning(left,t)\wedge angular\_velocity(pos)\ \text{default}\ \neg at(leftest,t) \tag{30}$$

$$move(right,t+1)\ \text{if}\ leaning(right,t)\wedge angular\_velocity(neg)\ \text{default}\ \neg at(rightest,t) \tag{31}$$

4.1.4 GridWorld

在该游戏中有 4 种不同的格子, 其中黄色格子为玩家控制的角色, 红色格子是陷阱, 黑色格子是墙体, 绿色格子是目的地, 白色格子是道路. 该游戏的目的是被玩家控制的黄色格子要达到目的地绿色格子, 每走一步扣 1 分, 如果 500 步内还没有走到目的地, 则游戏结束, 如果踩到红色陷阱则扣 600 分, 游戏结束, 如果走到了目的地绿色格子, 则获得 600 分, 游戏结束.

在该游戏中, 我们的规则是保证智能体在探索过程中不会落入任何陷阱以发生灾难性的后果, 即排除掉灾难性决策后, 在多个安全决策中随机选取其中一个.

$$\neg walk(dir, t+1) \text{ if } neighbor(trap_i, t) \wedge at(trap_i, dir, t) \text{ default } \top \tag{32}$$

我们在 Breakout, Pong 和 CartPole 中使用了启发式加速规则, 在 GridWorld 中使用了规避式安全规则, 并在每个游戏中采用了多个基准算法以验证方法的有效性与一般性, 实验数据集如表 1.

表 1 实验数据集

分组	项目	基准算法数量	知识规则数量
Acceleration KB	Breakout	3	2
	Pong	3	2
	CartPole	5	2
Safety KB	GridWorld	5	1

4.2 评价指标及基准模型

在实验中, 我们以重复实验的局内累计奖励的平均值与方差作为评价指标.

Breakout: 在该游戏中, 一局游戏的所有砖块打完后, 刷新砖块且游戏继续, 分数累计; 直至球拍接不到球时游戏结束. 因此如果没有时间限制, 得分并无上限. 我们设定评价指标为游戏进行到指定局数后所得分数.

Pong: 在该游戏中, 当一方得分达到 21 分时游戏结束, 即该游戏的最高分为我方获得 21 分, 对手 0 分, 总分 21 分. 评价指标为实验组和对照组在指定局数中的平均得分.

CartPole: 在该游戏中, 竖杆每竖持 1 个时间单位获得 1 分, 当竖杆与垂直方向的倾斜角度大于 15 度时、车的水平位移超过中心位置 2.4 个单位或达到游戏的最高奖励值 500 分时, 游戏结束. 评价指标为实验组和对照组在训练到指定步数中的平均得分.

GridWorld: 在该游戏中, 其得分具有上限和下限, 对于实验组和对照组会收敛在同一分数的算法, 评价指标为实验组和对照组的得分稳定时所需要的游戏局数; 对于实验组和对照组最后稳定在不同分数的算法, 评价指标为实验组和对照组在训练到一定步数结束后所达到的对比效果.

我们将显式知识推理应用在两类基准模型上进行实验测试, 即基于价值的深度强化学习算法, 包括 DQN, Double DQN, Dueling DQN, 与基于策略的深度强化学习算法 A2C, Discrete SAC, PPO 等.

4.3 实验方法

我们使用 Breakout、Pong、CartPole 和 GridWorld 作为实验测试环境. 对于 Breakout 和 Pong 游戏, 我们在基于价值的 DQN、Double DQN 和 Dueling DQN 上加入了启发式加速规则; 对于 CartPole 游戏, 我们在基于价值的 DQN、Double DQN 和 Dueling DQN, 以及基于策略的 A2C 和 Discrete SAC 上加入了启发式加速规则; 对于 GridWorld 游戏, 我们在基于价值的 DQN、Double DQN 和 Dueling DQN, 以及基于策略的 A2C 和 PPO 上加入了规避式安全规则. 启发式加速规则生效函数:

$$P(t) = p_0 \cdot \gamma^{\lambda t} \tag{33}$$

其中, 折扣因子 $0 < \gamma < 1$, 初始生效概率 $0 < p_0 \leqslant 1$, γ 为步长. 规避式安全规则使用相同的生效函数, 生效参数为 $\gamma = 1$, $p_0 = 1$.

4.4 实验结果与分析

为了评估基于知识推理与深度强化学习的动态决策框架 KB-DRL 的有效性, 我们研究了以下 4 个问题.

• RQ1: KB-DRL 是否可以有效提高训练效率?

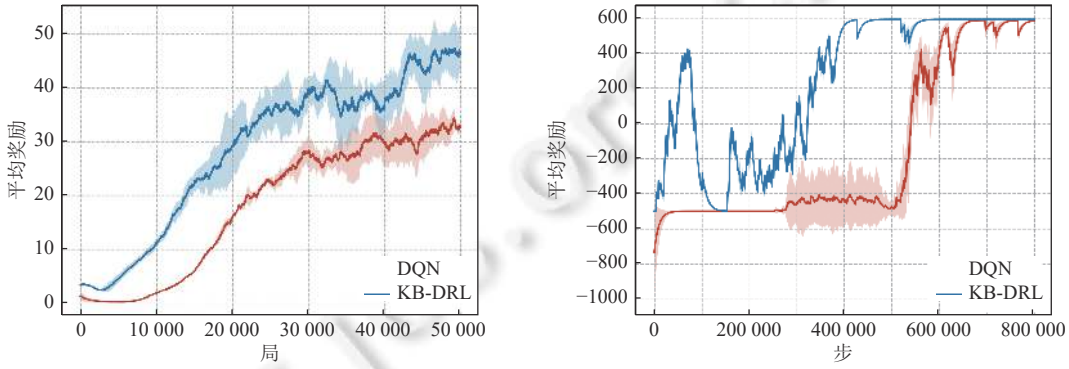
• RQ2: KB-DRL 是否提高了模型的可解释性?

• RQ3: KB-DRL 的性能优势是否具有一般性?

• RQ4: KB-DRL 调用知识推理带来的额外时间代价是否可接受?

4.4.1 RQ1: KB-DRL 是否可以有效提高训练效率?

在 Breakout 游戏中, 我们在 DQN 算法中引入了启发式加速规则, 横轴为训练过程中的游戏局数, 纵轴为得分, 图中绘制重复实验下的平均奖励与方差. 由实验结果图 4(a) 可以发现, 在训练到第 10 000 局时, 基准算法由于冷启动仅获得 2 分, 最初的 10 000 局模型的学习效率非常的低, 而加入显式知识后奖励可以达到 10 分, 且最初的 10 000 局均有奖励, 在训练到上限局数时, 基准算法奖励约为 33, 而加入显式知识的奖励约为 47. 加入显式知识的模型有效地提高了训练效率与模型表现.



(a) Breakout (b) GridWorld

图 4 加速规则 (在 Breakout 中) 与安全规则 (在 GridWorld 中) 实验结果

在 GridWorld 游戏中, 我们在 DQN 算法中引入了规避式安全规则, 实验结果如图 4(b) 横轴为训练的步数, 纵轴为每训练到一定的步数, 保存模型且保证加规则生效地玩 10 局游戏的平均奖励, 浅色为 10 局游戏的方差. 从实验结果图可见, 在 DQN 算法中, 基准算法在 60 万步时智能体能够学会走到目的地, 但是效果并不稳定, 而加入显式知识后在 40 万步学会走到目的地, 且效果更为稳定. 可以得出结论: 加入显式知识的模型可以有效提高训练效率. 综上, KB-DRL 可以有效提高训练效率.

4.4.2 RQ2: KB-DRL 是否提高了模型的可解释性?

显式知识的引入明显提高了模型的可解释性. 首先, 以 BC-dynamic 表示的知识符合人类直观认知, 本身具有较高的抽象与可解释性. 其次, 由于显式知识的高度抽象性, 知识可以较为直观的横向比较. 例如, 我们在 Breakout 和 Pong 采用了极其相似的知识库.

在 Breakout 中使用的启发式加速规则: 如果球在球拍的左边, 则球拍应该往左, 如果球在球拍的右边, 则球拍应该往右.

$$move(left,t+1)\ \text{if}\ at(ball,left,t)\ \text{default}\ \neg at(ball,leftest,t) \tag{34}$$

$$move(right,t+1)\ \text{if}\ at(ball,right,t)\ \text{default}\ \neg at(ball,rightest,t) \tag{35}$$

而在 Pong 游戏中, 使用的启发式加速规则为:

$$move(up,t+1)\ \text{if}\ at(ball,above,t)\ \text{default}\ \neg at(ball,top,t) \tag{36}$$

$$move(down,t+1)\ \text{if}\ at(ball,below,t)\ \text{default}\ \neg at(ball,bottom,t) \tag{37}$$

由于该游戏和 Breakout 游戏的任务场景非常相似, 都是根据迎面而来的球与我方球拍的相对位置进行移动, 因此可以通过简单地修改 Breakout 中的规则就能在 Pong 中被使用.

4.4.3 RQ3: KB-DRL 为动态决策带来的优势是否具有一般性?

为了验证这个问题的结论, 我们在 Breakout、Pong、CartPole 和 GridWorld 游戏中采用了多个算法用于验证.

在 Breakout 游戏中, 我们引入了启发式加速规则, 为验证该方法的一般性, 除了 RQ1 中 Breakout 的 DQN 算法, 我们还在 Double DQN 和 Dueling DQN 中运用了此方法, 如图 5, 横轴为训练过程中的游戏局数, 纵轴为重复实验下的平均奖励, 阴影为重复实验下的奖励方差, 从 3 个算法的实验中, 均可以发现, 在训练到第 10 000 局时, 基准算法由于冷启动仅获得 2 分, 最初的 10 000 局模型的学习效率非常低, 而加入显式知识后奖励可以达到 10 分, 且最初的 10 000 局均有奖励, 在训练到上限局数时, 3 个算法的 Basedline 奖励分别约为 33、36 和 41, 而加入显式知识的奖励分别约为 47、48 和 50.

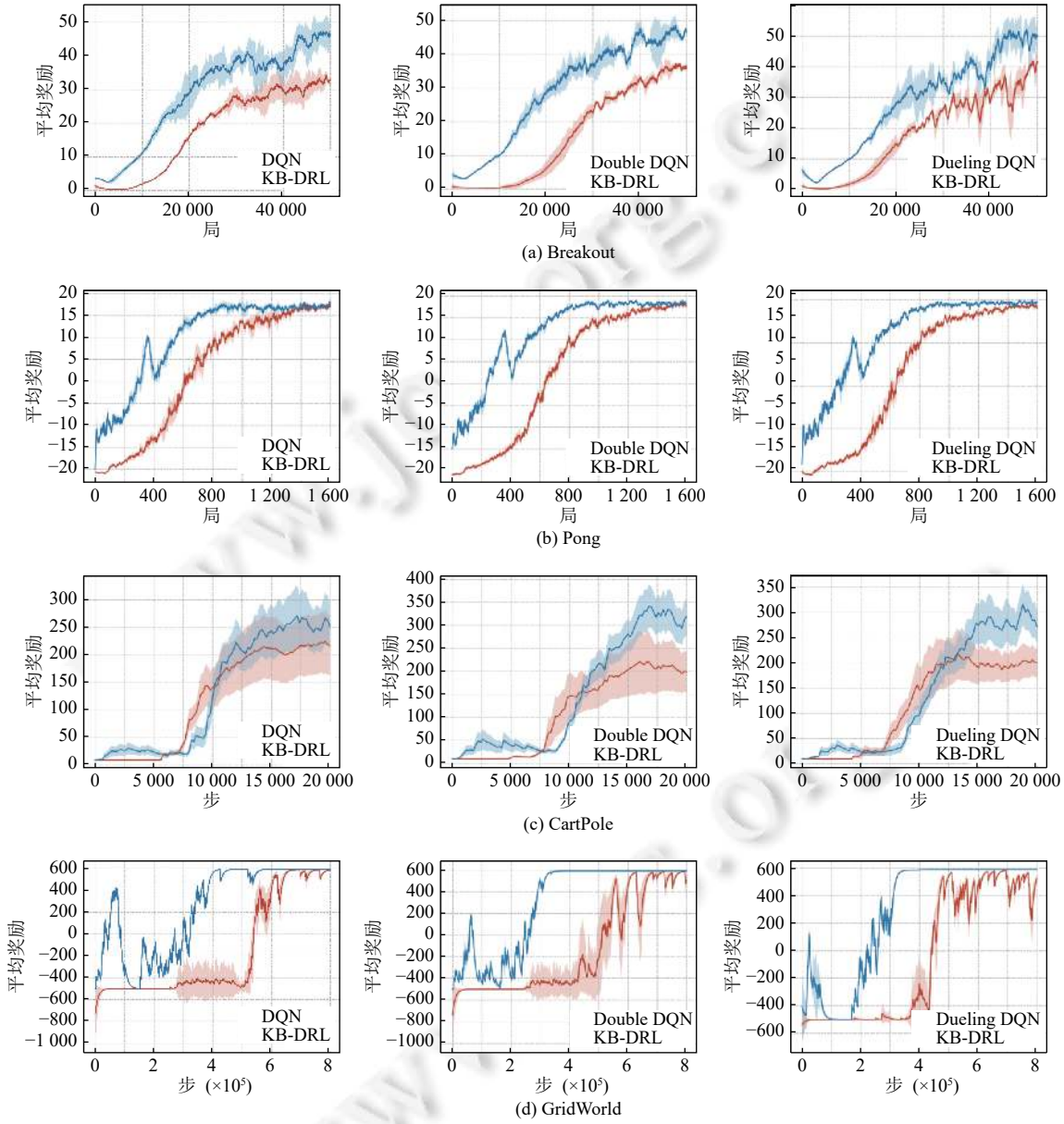
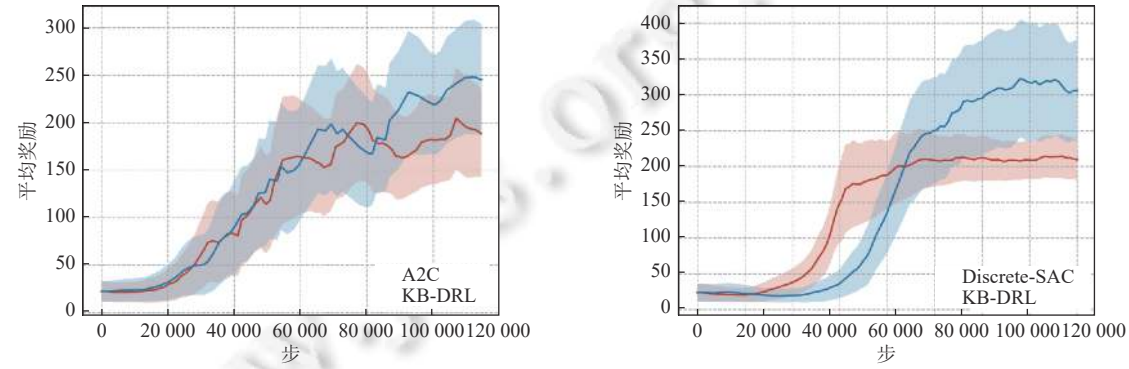


图 5 基于价值的深度强化学习算法在不同游戏上的实验结果

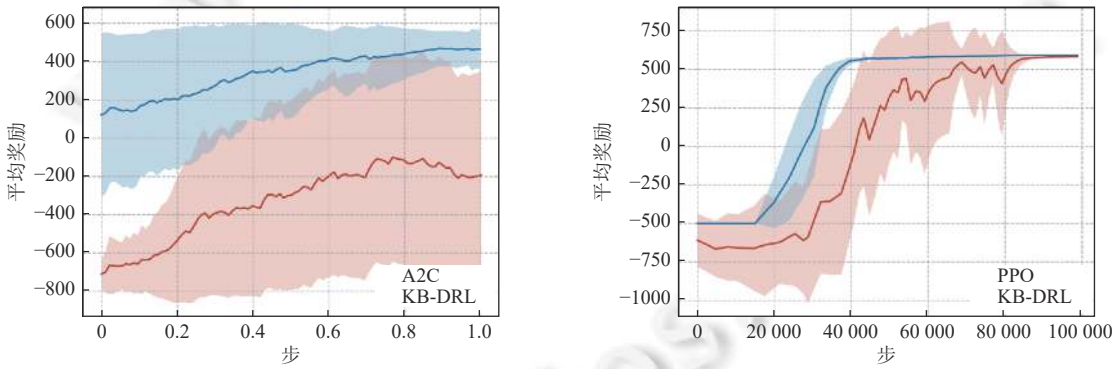
在 Pong 游戏中, 与 Breakout 类似, 我们引入了启发式加速规则, 我们在 DQN、Double DQN 和 Dueling DQN 中运用了此方法, 如图 5, 横轴为训练过程中的游戏局数, 纵轴为重复实验下的平均奖励, 阴影为重复实验下的奖

励方差, 从 3 个算法的实验中, 均可以发现, 基准算法在前约 100 局前由于模型还不会接球, 智能体一直 0 分而敌方 21 分, 基本处于最低分−21 分的状态, 在约 100 局后, 智能体逐渐能够获得一定奖励. 而在加入了显式知识之后, 智能体在前 100 局的奖励在−10 分震荡, 比最低分高出 11 分左右, 此结论可以充分说明在显式知识的引导下, 智能体在训练初期的探索变得更加有意义, 由于显式知识中的启发式加速规则是在生效函数的作用下生效, 启发式加速规则的生效概率会随着时间的推移而下降, 在训练到约 350 局时, 规则逐渐不起作用, 但由于前期知识的引导做了更多的有效探索, 模型能够更快地学得更优的效果. 与 Breakout 游戏奖励无上限不同, Pong 游戏在任一方达到 21 分时游戏结束, 我们的框架在约 900 局时便能稳定在 21 分的效果.

在 CartPole 游戏中, 我们引入了启发式加速规则, 我们不仅在基于价值的深度强化学习算法 DQN、Double DQN 和 Dueling DQN 加入了显式知识, 实验结果如图 5, 而且还在基于策略的深度强化学习算法 A2C 和 Discrete-SAC 加入了显式知识, 实验结果如图 6, 横轴为训练的步数, 纵轴为每训练到一定的步数, 保存模型不加规则地玩 100 局游戏的平均奖励, 浅色为 100 局游戏的方差. 加入显式知识的模型效果都有大幅提升.



(a) CartPole



(b) GridWorld

图 6 基于策略的深度强化学习算法在不同游戏上的实验结果

在 GridWorld 游戏中, 我们引入了规避式安全规则, 除了 RQ1 中 GridWorld 的 DQN 算法外, 我们还在基于价值的深度强化学习算法 Double DQN 和 Dueling DQN 中加入了安全规则, 实验结果如图 5, 以及基于策略的深度强化学习算法 A2C 和 PPO 中加入了安全规则, 实验结果如图 6, 横轴为训练的步数, 纵轴为每训练到一定的步数, 保存模型且规则生效下统计 10 局游戏的平均奖励与方差. 实验结果显示, 在 DQN、DDQN 和 Dueling DQN 算法中, 基准算法分别在 60 万、60 万和 50 万步时智能体能够学会走到目的地, 但是效果并不稳定, 而加入显式知识后分别在 40 万、35 万和 35 万步学会走到目的地, 且效果更为稳定. 而在 A2C 中, 基准算法较难学会走到目的地, 加入显式规则后能够逐渐学会且趋于稳定. 在 PPO 中, 基准算法约在 8.5 万步时学会走到目的地, 而加入显式规则

后模型在 4 万步时就能够学会走到目的地, 且效果稳定.

综上, KB-DRL 不仅可以在不同的环境下使用, 并与多种主流算法正交, 在以上实验中均表现出明显优势. 因此可以得出结论: KB-DRL 为动态决策带来的优势具有一般性.

4.4.4 RQ4: KB-DRL 调用知识推理带来的额外时间代价是否可接受?

为了考查加入显式知识带来的额外时间代价, 我们在启发式加速规则和规避式安全规则中分别进行了实验对比. 图 7 为训练耗时统计分布, 横轴表示模型在单批训练样本上的训练耗时, 纵轴表示频数, 组距为 0.05 s.

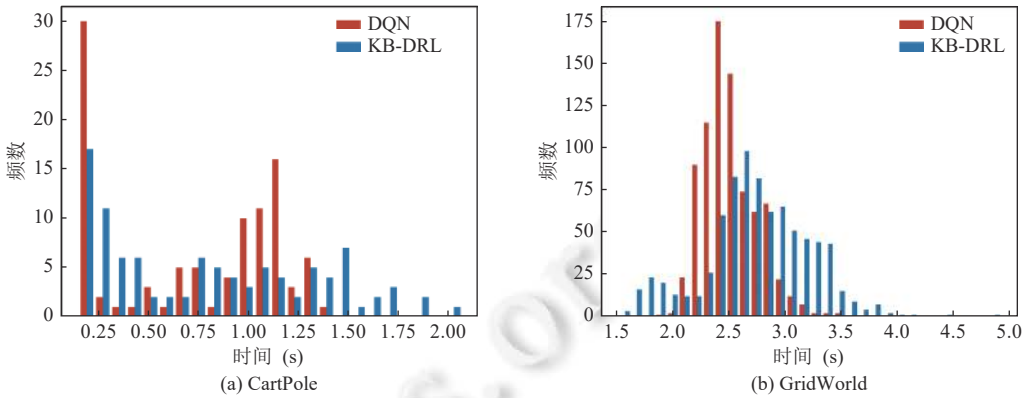


图 7 加速规则 (在 CartPole 中) 与安全规则 (在 GridWorld 中) 单位数据训练时间分布

在 CartPole 游戏中, 每 200 步训练一次模型并记录所需时间, KB-DRL 在大于 1.40 s 的部分均多于 DQN. 在 GirdWorld 游戏中, 每 1 000 步训练一次模型并记录所需时间, KB-DRL 在大于 2.95 s 的部分均多于 DQN. 由图 7 及表 2 中统计数据可见, 加入显式知识的智能体的部分训练轮次需要花费更多的时间, 但其频数均在可以接受的范围之内.

表 2 单批数据训练耗时统计 (s)

知识	算法	最大	最小	平均	标准差
加速规则 (CartPole)	DQN	1.36	0.15	0.72	0.43
	KB-DRL	2.06	0.16	0.79	0.52
安全规则 (GridWorld)	DQN	3.50	1.88	2.50	0.23
	KB-DRL	4.91	1.52	2.74	0.46

5 总 结

本文针对深度强化学习相关算法与模型缺乏可解释性, 初期训练低效等问题, 提出了一种基于显式知识推理和深度强化学习的动态决策框架. 与模仿学习等其他知识整合方法相比, 该框架在深度强化学习中加入显式知识时并不需要大量人力成本. 该决策框架提供了在深度强化学习中整合显式知识的方法, 可以有效提高训练效率, 提高模型的可解释性, 且不依赖特定场景与算法. 正如我们在多个基准算法上的实验所证实, 在不同场景下, 启发式的加速规则能够引导智能体去做更有价值的探索, 而规避式的安全规则能够帮助智能体避开灾难性的后果.

在未来的工作中, 我们将考虑: (1) 自动化的显式知识获取; (2) 规则类显式知识的高维空间表示.

References:

[1] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing atari with deep reinforcement learning. arXiv:1312.5602, 2013.

[2] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep

reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

[3] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proc. of the 2016 AAAI Conf. on Artificial Intelligence, 2016, 30(1): 2094–2100. [doi: 10.1609/aaai.v30i1.10295]

[4] Wang ZY, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N. Dueling network architectures for deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR, 2016. 1995–2003.

[5] Mnih V, Badia AP, Mirza M, Graves A, Harley T, Lillicrap TP, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR, 2016. 1928–1937.

[6] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

[7] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: ICML, 2018. 1861–1870.

[8] Osa T, Pajarinen J, Neumann G, Bagnell JA, Abbeel P, Peters J. An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics, 2018, 7(1–2): 1–2. [doi: 10.1561/2300000053]

[9] Liu YX, Gupta A, Abbeel P, Levine S. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 1118–1125. [doi: 10.1109/ICRA.2018.8462901]

[10] Griffith S, Subramanian K, Scholz J, Isbell CL, Thomaz A. Policy shaping: Integrating human feedback with reinforcement learning. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2625–2633.

[11] Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4302–4310.

[12] Ibarz B, Leike J, Pohlen T, Irving G, Legg S, Amodei D. Reward learning from human preferences and demonstrations in Atari. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 8022–8034.

[13] Le HM, Jiang N, Agarwal A, Dudík M, Yue YS, Daumé III H. Hierarchical imitation and reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2923–2932.

[14] Bellemare MG, Dabney W, Munos R. A distributional perspective on reinforcement learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR, 2017. 449–458.

[15] Osband I, Blundell C, Pritzel A, Van Roy B. Deep exploration via bootstrapped DQN. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 4033–4041.

[16] Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, Dabney W, Horgan D, Piot B, Azar MG, Silver D. Rainbow: Combining improvements in deep reinforcement learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 3215–3222.

[17] Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proc. of the 31st Int'l Conf. on Machine Learning. Beijing: JMLR, 2014. 387–395.

[18] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: ICLR, 2016.

[19] Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR, 2015. 1889–1897.

[20] Bain M, Sammut C. A Framework for Behavioural Cloning. Technical Report, Oxford: St. Catherine's College, 1995. 103–129.

[21] Abbeel P, Ng AY. Apprenticeship learning via inverse reinforcement learning. In: Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM, 2004. 1. [doi: 10.1145/1015330.1015430]

[22] Cederborg T, Grover I, Isbell CL, Thomaz AL. Policy shaping with human teachers. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. Buenos: Buenos Aires Argentina, 2015. 3366–3372.

[23] Tenorio-Gonzalez AC, Morales EF, Villaseñor-Pineda L. Dynamic reward shaping: Training a robot by voice. In: Proc. of the 12th Ibero-American Conf. on Artificial Intelligence. Bahía Blanca: Springer, 2010. 483–492. [doi: 10.1007/978-3-642-16952-6_49]

[24] Knox WB, Stone P. Interactively shaping agents via human reinforcement: The TAMER framework. In: Proc. of the 5th Int'l Conf. on Knowledge Capture. California: ACM, 2009. 9–16. [doi: 10.1145/1597735.1597738]

[25] Warnell G, Waytowich NR, Lawhern V, Stone P. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 1545–1554.

[26] Knox WB, Stone P. Reinforcement learning from simultaneous human and MDP reward. In: Proc. of the 11th Int'l Conf. on Autonomous Agents and Multiagent Systems. Valencia: AAMAS, 2012. 475–482.

[27] Arakawa R, Kobayashi S, Unno Y, Tsuboi Y, Maeda SI. DQN-TAMER: Human-in-the-loop reinforcement learning with intractable

feedback. arXiv:1810.11748, 2018.
[28] Saunders W, Sastry G, Stuhlmüller A, Evans O. Trial without error: Towards safe reinforcement learning via human intervention. In: Proc. of the 17th Int'l Conf. on Autonomous Agents and MultiAgent Systems. Stockholm: AAMAS, 2018. 2067–2069.
[29] Lee J, Lifschitz V, Yang FK. Action language BC: Preliminary report. In: Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence. Beijing: AAAI Press, 2013. 983–989.
[30] Babb J, Lee J. Action language BC+. Journal of Logic and Computation, 2020, 30(4): 899–922. [doi: 10.1093/logcom/exv062]
[31] Lifschitz V. What is answer set programming? In: Proc. of the 23rd National Conf. on Artificial Intelligence. Chicago: AAAI Press, 2008. 1594–1597.
[32] Yang Z, Ishay A, Lee J. NeurASP: Embracing neural networks into answer set programming. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI, 2020. 1755–1762.
[33] Didolkar A, Goyal A, Ke NR, Blundell C, Beaudoin P, Heess H, Mozer M, Bengio Y. Neural production systems. In: Proc. of the 34th Neural Information Processing Systems. 2021. 25673–25687.
[34] Scheck S, Niveau A, Zanuttini B. Knowledge compilation for nondeterministic action languages. In: Proc. of the 31st Int'l Conf. on Automated Planning and Scheduling. Guangzhou: AAAI Press, 2021. 308–316.
[35] Wang Y, Lee J. Elaboration tolerant representation of Markov decision process via decision-theoretic extension of probabilistic action language pBC+. In: Proc. of the 15th Int'l Conf. on Logic Programming and Nonmonotonic Reasoning. Philadelphia: Springer, 2019. 224–238. [doi: 10.1007/978-3-030-20528-7_17]
[36] Niemelä I. Logic programs with stable model semantics as a constraint programming paradigm. Annals of Mathematics and Artificial Intelligence, 1999, 25(3–4): 241–273. [doi: 10.1023/A:1018930122475]

张昊迪(1986－), 男, 博士, 助理教授, CCF 专业会员, 主要研究领域为人工智能, 知识表示与推理, 深度学习, 人工智能在自然语言处理、游戏、医疗等领域的应用.

连德富(1985－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为人工智能, 数据挖掘, 深度学习, 推荐系统.

陈振浩(1995－), 男, 硕士生, 主要研究领域为深度学习, 知识表示与推理, 游戏智能.

伍楷舜(1985－), 男, 博士, 特聘教授, 博士生导师, CCF 专业会员, 主要研究领域为物联网, 智能可穿戴计算, 无线传感网络, 无线干扰管理, 无线室内定位, 认知无线电, 普适计算.

陈俊扬(1990－), 男, 博士, 助理教授, CCF 专业会员, 主要研究领域为数据挖掘, 人工智能, 推荐系统, 自然语言处理, 图神经网络的理论研究, 时序推荐模型.

林方真(1963－), 男, 博士, 教授, 博士生导师, 主要研究领域为人工智能, 知识表示与推理, 逻辑程序语言, 机器人, 多智能体, 博弈论与社会选择理论.

周熠(1981－), 男, 博士, 研究员, 博士生导师, 主要研究领域为认知人工智能基础, 真实动态复杂环境中知识的表示、推理和学习, 显式的符号知识与隐式的神经网络深度融合的理论模型.