华为-清华团队提出一种数据高效的分层强化学习算法，在AI顶会NeurIPS MineRL竞赛中夺冠

原创 诺亚方舟实验室 诺亚实验室 2021-11-22 18:54

华为诺亚方舟实验室的决策推理研究团队联合清华大学交叉信息研究院，提出了一种面向数据高效的分层强化学习解决方案SEIHAI，在参与NeurIPS2020-MineRL比赛的90+队伍中脱颖而出，斩获初赛和决赛冠军；该方案仅需与环境交互300万次，远远低于官方规定的800万次，并且最终得分大幅领先于第二名（39.55-vs-13.29分），相关成果近日以论文形式发表在国际会议Distributed AI 2021上。

近年来，人工智能疾速发展，强化学习作为最具备潜力的人工智能技术之一，已经在诸多领域取得了重要成果。然而强化学习仍然面临着一些技术困境，其中之一是数据低效问题。例如，能够轻而易举地打败人类围棋高手李世石的围棋智能体AlphaGo采用的核心技术就是强化学习，但是AlphaGo的成功离不开人类3000万的棋谱作为训练数据来指导其下棋，3000万的棋谱相当于人类上百年的下棋经验；之后的AlphaGo Zero等变体也采用了上千万次的自我博弈来学习下棋。对于现实世界的应用场景，比如自动驾驶、推荐系统、机器人等，和真实环境大量交互的代价会非常高，现有强化学习算法数据效率低的问题成为了其在真实场景落地的关键瓶颈之一。因此，发展数据高效的强化学习已经成为整个研究领域的共识。

MineRL比赛就是在上述背景下产生的。该比赛是由CMU, OpenAI, DeepMind, Microsoft Research等机构联合举办的比赛，是强化学习方向最负盛名的比赛之一。该比赛旨在引导学术界和工业界基于有限人类玩家数据和有限环境交互数据，发展数据高效的大规模强化学习和模仿学习算法。比赛自2019年起，每年在NeurIPS会议上举办，吸引了工业界和学术界大量的关注。此次比赛有超过90支队伍参赛，成员来自斯坦福大学，清华大学，香港中文大学，南洋理工大学等国际知名高校。

MineRL比赛

MineRL比赛需要参赛人员训练一个能够在我的世界（MineCraft）游戏中挖到钻石的智能体。比赛存在诸多难点，例如：

- 物品之间有强依赖：获取钻石的前提是获得一系列的先决物品和工具（参考图1），并且先决物品以及工具之间也存在相互依赖的关系。智能体必须按照如下图所示的顺序依次获得足够多的物品和工具，并最终获得钻石。

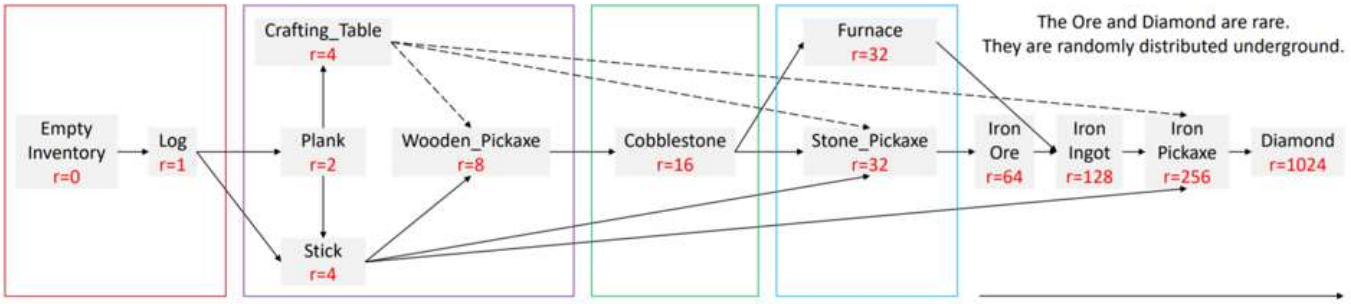

图1 物品之间的依赖关系

- 奖励稀疏：智能体只在第一次获得相应的物品或工具的时候获得环境给予的奖励，例如智能体砍了两块木头，但只会在得到第一块木头的时候获得1分的奖励，此外中间的探索过程不会得到任何奖励。为了获得钻石，智能体需要学习获取各种不同物品或者工具的能力。例如为了制作出木镐，智能体必须至少收集3块木头，而后续的工具制作也会消耗木头，智能体必须能够积累足够的木头才能完成后续的任务。

- 状态和动作空间巨大、并且语义信息未知：状态空间包含智能体所处的MineCraft环境的视觉信息以及智能体背包中物品的信息（如当前木头数量，木板数量等等），动作空间包含智能体能够执行的所有动作集合（如前进，后退，制作木镐，制作石镐等等）。为了防止采用规则，举办方使用了参赛者未知的编码器将状态空间的背包信息和整个动作进行了模糊化处理（如图2），模糊化之后的状态和动作空间巨大、并且语义信息未知。

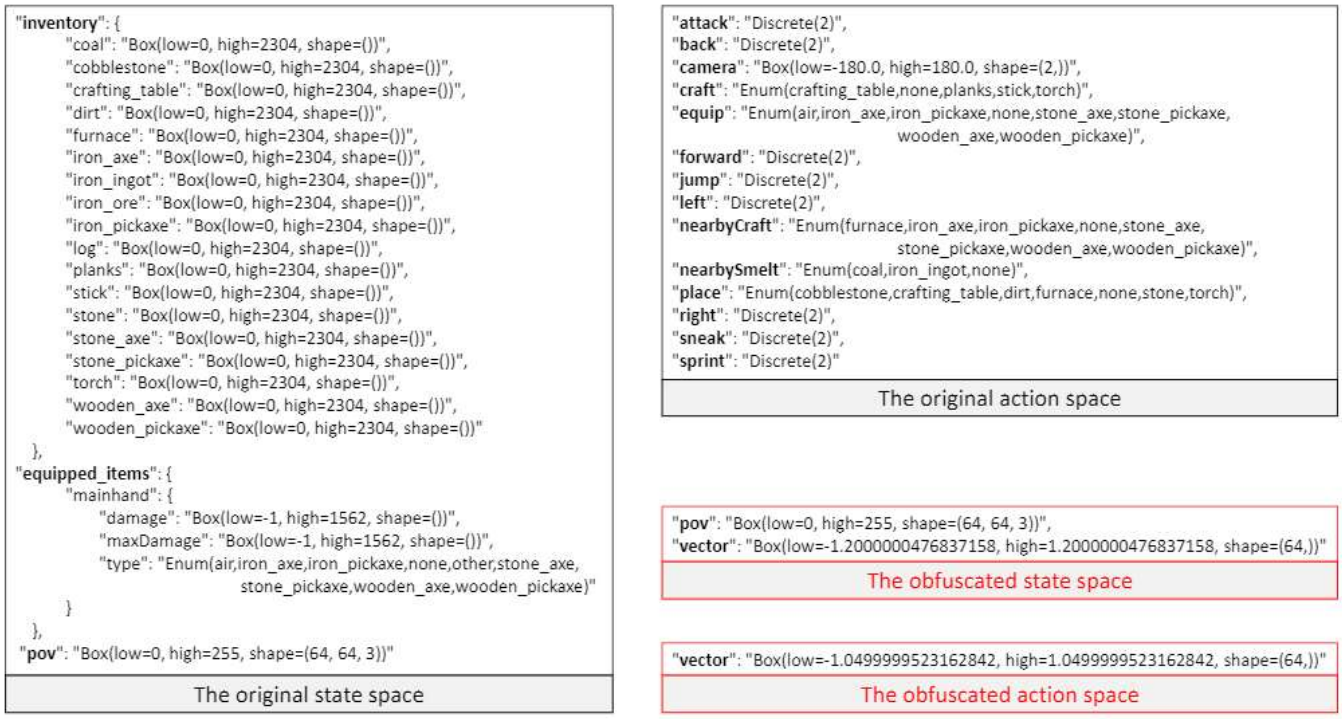

图2 原始状态和动作空间，以及模糊化之后的状态和动作空间

- 有限人类玩家数据：除了MineRL游戏环境本身，主办方提供了有限数量的经过编码的人类玩家数据（211条轨迹片段），这些轨迹中仅有不到200个获得了钻石。

- 数据高效：一个高效的玩家也需要上千步才能成功挖到钻石，但举办方要求参赛队伍的算法至多与MineRL环境交互800万步。

目前没有任何一个强化学习算法能够在不违反规则的条件下直接学习有效的控制策略，例如举办方提供的flat baseline最多只能得到2.94分。来自华为诺亚方舟实验室的决策推理研究团队（参赛队名HelloWorld）提出了一种面向数据高效的分层强化学习解决方案SEIHAI，仅需要与环境交互300万步就可以将分数提高到39.55分，并且大幅领先第二名的13.29分（如图3，https://www.aicrowd.com/challenges/neurips-2020-minerl-competition/leaderboards）

图3　官方baseline、初赛队伍和决赛队伍的得分情况

强化学习方案SEIHAI

智能体需要在不同的游戏阶段获取不同的技能才能够在MineCraft环境中到达最终目标——挖到钻石。例如，游戏开始时需要寻找树木并且砍伐足够数量的木头；之后再用木头制作木板，木棍以及木镐；之后智能体利用木镐去收集石头和铁矿石；然后制作石镐和铁镐以及之后的一系列工具，收集制作完所有的先决物品和工具之后，才可能挖到钻石。显然，单个策略很难实现上面所有的这些技能（而不影响各个技能的效果）。

考虑到上面的游戏特性，尤其是物品以及工具间存在的强依赖的特性，华为诺亚方舟实验室团队提出的数据高效的分层强化学习方案SEIHAI（图4所示）将挖钻石的任务拆分成了5个子任务，为每个子任务设计一个智能体（具体包括砍树智能体、制作木镐智能体、挖石头智能体、制作石镐智能体和随机搜索智能体），并设计一个上层调度策略调度不同的智能体，实现挖钻石的任务。该方案的优点在于：

- 数据效率高：上层调度策略和下层子策略均利用数据驱动的学习算法得到，并且不同子任务的智能体可以使用不同的算法进行实现，有效利用人类玩家数据和环境交互数据，解决数据高效问题和稀疏奖励问题。例如：砍树智能体需要和环境大量交互才能学习到比较稳定的策略，SEIHAI方案采用SQIL算法来实现；制作木镐智能体需要按照物品依赖关系执行一系列制作动作，SEIHAI方案利用人类玩家数据，采用BC算法来实现。

- 不依赖规则：上层调度策略会根据智能体背包中不同物品的数量状态决定调用哪一个子策略，不需要人为指定调度规则。例如游戏开局背包中空无一物，上层调度策略就会激活砍树智能体去收集木头，收集一定数量的木头之后，上层策略就会激活制作木镐的智能体。而在制作木镐阶段，若木头不够使用，上层策略就会重新调用砍树智能体来砍更多的木头。

- 可扩展性好：下层子策略可以进一步细化来实现更好的性能，同时也可以训练更多的下层智能体来提高方案能够掌握的技能数量。例如在初赛中SEIHAI方案仅仅包含了前三个子策略，达到了19.84分的成绩；而在决赛时SEIHAI方案将子策略扩展到五个，达到了39.55分的成绩。如果将随机搜索智能体进一步细化，SEIHAI方案将能够掌握更多技能、得到更多的分数。

除了分层方案本身之外，SEIHAI方案还采用自适应的动作聚类方案将高维连续空间的动作聚类为少量的离散动作，由于这些离散动作都是关键动作（能够造成环境状态发生大变化或者直接获取奖励），这样进一步增加了SEIHAI方案的数据高效性。方案的具体细节可参考最近发表在DAI2021的论文（SEIHAI: A Sample-efficient Hierarchical AI for the MineRL Competition，https://arxiv.org/pdf/2111.08857.pdf）。（点击底部左下角“阅读原文”可直接进入）

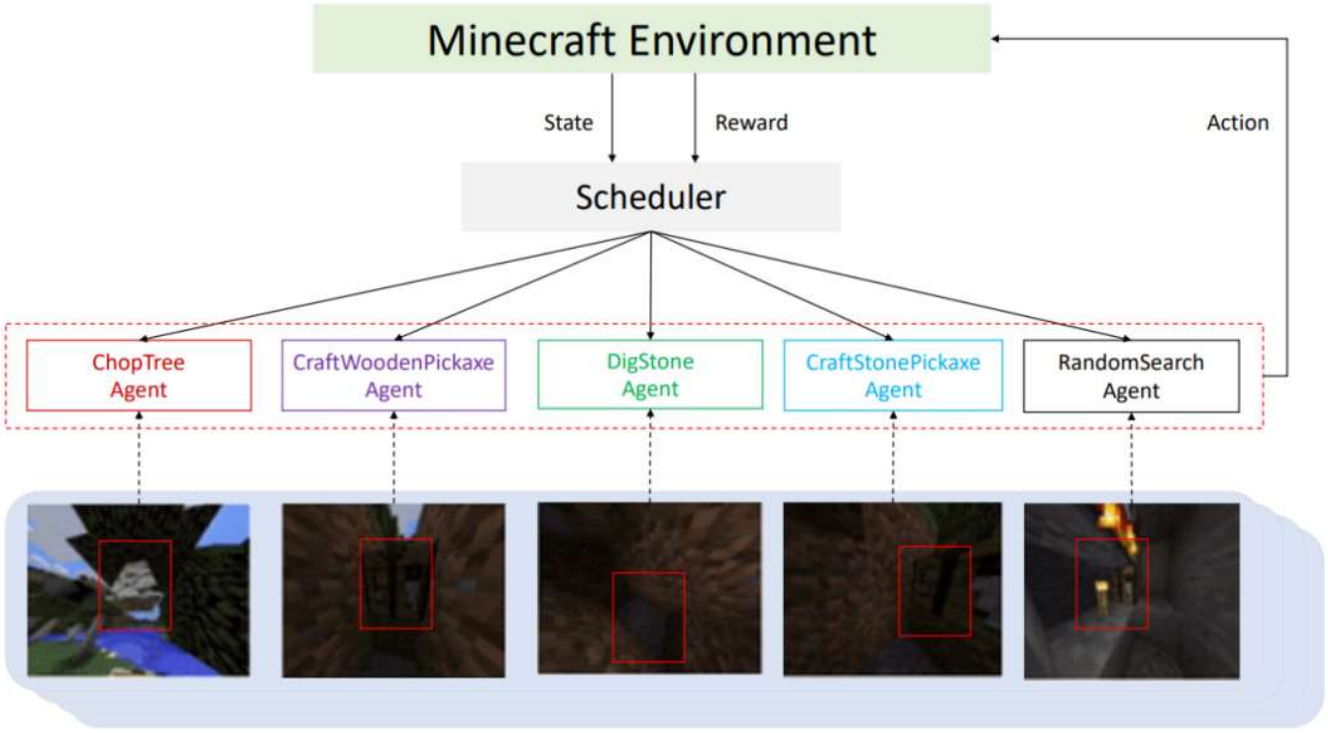

图4　华为-清华联合团队提出的数据高效的分层强化学习方案SEIHAI

参赛队伍HelloWorld成员

团队成员

指导老师

【免责声明】

华为在本公众号所载的材料和信息，包括但不限于文本、图片、数据、观点、建议、网页或链接，虽然华为力图在网站平台上提供准确的材料和信息，但华为并不保证这些材料和内容的准确、完整、充分和可靠性，并且明确声明不对这些材料和内容的错误或遗漏承担责任，也不对这些材料和内容作出任何明示或默示的、包括但不限于有关所有权担保、没有侵犯第三方权利、质量和没有计算机病毒的保证。

华为可以在没有任何通知或提示的情况下随时对网站上的内容进行修改，为了得到最新版本的信息，请您定时访问本网站。华为（含其关联公司）在本网站上所提及的非华为产品或服务仅仅是为了提供相关信息，并不构成对这些产品、服务的认可或推荐。华为并不就网址上提供的任何产品、服务或信息做出任何声明、保证或认可，所有销售的产品和服务应受华为的销售合同和条款的约束。

— 完 —

欢迎有才华的您与我们一起，迎接人工智能的大航海时代！

诺亚方舟实验室（Noah's Ark Lab） 是华为公司从事人工智能基础研究的实验室，致力于推动人工智能领域的技术创新和发展，并为华为公司的产品和服务提供支撑。

简历投递： 请发送简历至**Noahlab@huawei.com**，邮件标题“应聘职位+姓名”

— Welcome Aboard! —

阅读原文