研究提案：基于期望回报预测的子任务切换策略

背景介绍

在机器人长视野操作(Long-horizon Manipulation)的任务中，常采用强化学习生成高自由度机械臂-灵巧手的动作策略。因为长视野操作任务动作空间状态空间维数高、执行过程长造成轨迹复杂，各阶段任务目标复杂不能通过统一的奖励函数进行描述，在强化学习中难以通过完整任务的强化学习生成可靠的策略。此前已有研究[1-4]通过将一个长程任务拆分为一个或多个子任务并分别进行预训练后进行微调，以提高强化学习训练的可行性。

如图 1 所示，作者构建了一种包含“前向传播训练”和“后向传播微调”的训练框架。“前向传播训练”指根据前一子任务终止状态分布构造后一子任务的初始状态分布，并依此进行子任务学习。“后向传播微调”指通过神经网络预测后一子任务期望回报作为前一子任务终止奖励来传递微调目标。文章在提出使用神经网络基于前一子任务的结束状态预测后一子任务的期望回报后，又将后一子任务的期望回报与经验阈值得比较并以此参量指导相邻子任务之间的切换。在上述过程中，后续子任务期望回报阈值的设置具有较强主观性、不能通过学习策略进行调整，且在阈值附近会发生不合理的策略突变。

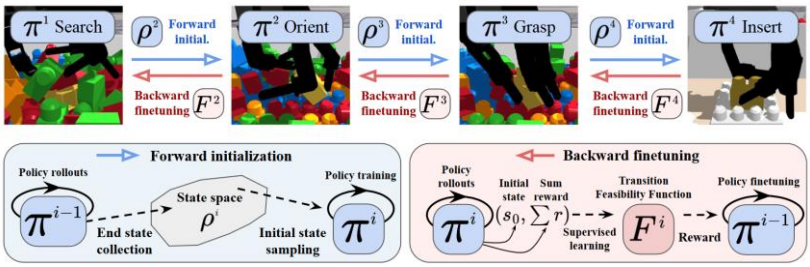

图 1 机器人长程操作中的子任务训练[4]

研究内容

本项研究旨在基于原始论文[4]所完成的工作，通过层级强化学习(Hierarchical RL)训练一个高级决策 Agent，其可以通过观测各个子任务的预期回报，指导各子任务策略切换和回滚。其具体过程应当包括以下基础内容：

0. 研读原始论文[4]，并根据其开源代码在仿真环境中复现其主要工作；
1. 提出以上高级决策 Agent 强化学习的 MDP 问题描述；
2. 采用 1-3 种主流 RL 方法训练上述 Agent，并对比训练过程和部署效果；
3. 研究报告撰写。

若以上任务顺利完成，希望基于个人研究兴趣完成以下工作：

0. 为防止原始问题过于复杂难以训练，基于以上 MDP 问题描述，设计并搭建一个简化问题的环境模型；
1. 为提高 RL 决策的可解释性和人类指导经验，引入基于 RL 修改隶属度函数的模糊推理(Fuzzy Reasoning)替代神经网络分类器生成策略，指导子任务之间的切换和回滚。这一过程需要重新设计模糊推理的反向传播；
2. 在简化环境中通过 1-3 种强化学习方法训练以上模糊推理 Agent，并对比训练过程和部署效果；
3. 对训练得到的模糊推理隶属度函数进行解释；

提出的科学问题

0. 子任务切换、回滚的高层决策 Agent 强化学习 MDP 问题；
1. 可通过反向传播修改模糊推理隶属度函数问题。

预期的技术成果

0. 针对机器人长视野操作问题的层级强化学习训练架构；
1. 可应用于原问题场景的高级 Agent；
2. 新提出的基于强化学习调整隶属度函数的模糊推断方法。

时间安排

研究时间	研究内容
校历 第 10-11 周	论文[4]研读与复现
校历 第 11-14 周	代码编写、强化学习训练
校历 第 14-15 周	兴趣工作
校历 第 15-16 周	报告撰写、准备汇报

参考资料

[1] Y. Lee, J. J. Lim, A. Anandkumar, and Y. Zhu, “Adversarial Skill Chaining for Long-Horizon Robot Manipulation via Terminal State Regularization.” arXiv, Nov. 15, 2021. doi: 10.48550/arXiv.2111.07999.

[2] Y. Li, Y. Wu, H. Xu, X. Wang, and Y. Wu, “Solving Compositional Reinforcement Learning Problems via Task Reduction.” arXiv, Mar. 18, 2021. doi: 10.48550/arXiv.2103.07607.

[3] U. A. Mishra, S. Xue, Y. Chen, and D. Xu, “Generative Skill Chaining: Long-Horizon Skill Planning with Diffusion Models,” in *Proceedings of The 7th Conference on Robot Learning*, PMLR, Dec. 2023, pp. 2905–2925. Accessed: Mar. 21, 2024. [Online]. Available: https://proceedings.mlr.press/v229/mishra23a.html

[4] Y. Chen, C. Wang, L. Fei-Fei, and C. K. Liu, “Sequential Dexterity: Chaining Dexterous Policies for Long-Horizon Manipulation.” arXiv, Oct. 16, 2023. doi: 10.48550/arXiv.2309.00987.