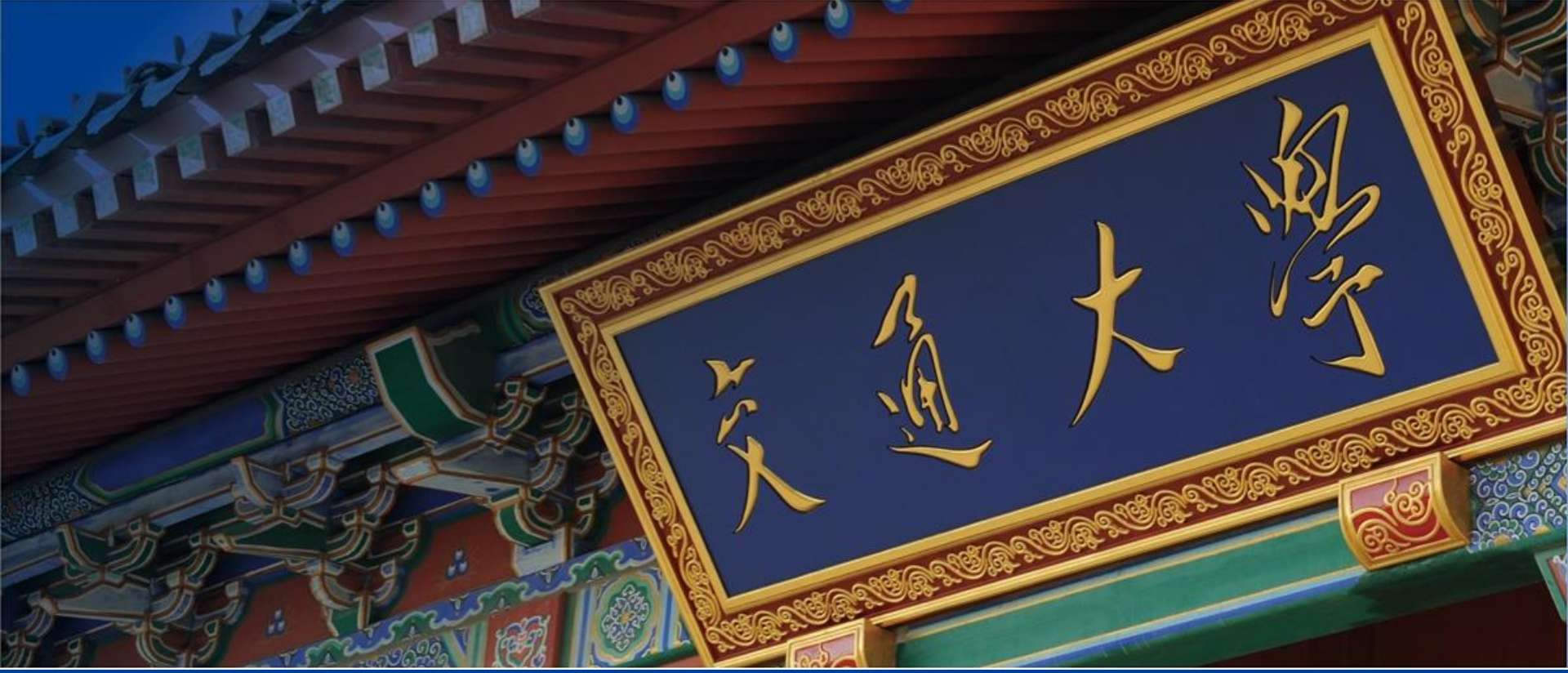


2024 AI3601 强化学习大作业

2024年5月

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

作业形式

- 分组完成，每组3人，完成multi task imitation learning 算法或者multi task offline RL算法。该课题列出了一些参考算法，也可以调研其他相关算法。
  - 在 Mujoco Walker2D 上实现
    - Multi task imitation learning算法 MH-AIRL [1], MAML-IL [2]或者
    - Multi task offline RL 算法 CDS [3]，Multi head CQL [4]

    并且获得在各个task都表现良好的智能体

课题一：Multi task Imitation learning

- 模仿学习（Imitation learning）旨在只使用记录的专家数据 (s, a, s') 来学习策略。

- 多任务模仿学习旨在通过模仿多任务专家演示，以在不同任务上达到优秀的表现。

课题二：Multi task Offline RL

- 离线强化学习（Offline RL）旨在只使用记录的数据 (s,a,s',r) 来学习行为，例如预先记录的实验过程或人类演示的数据，而不需要进一步的环境交互。

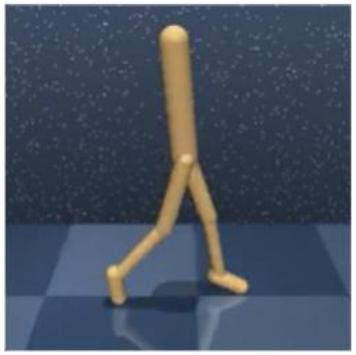
- 多任务离线强化学习旨在不与真实环境进一步交互的情况下，汇集各种场景下的大量数据来提高样本利用效率和策略的鲁棒性，以在不同任务上达到优秀的表现。

实验环境介绍

- 实验环境是 OpenAI Gymnasium Walker2D 环境下的两个任务:

(1) 向前跑；（2）向前走。这两个任务的动作空间，状态空间和转移函数，仅有奖励函数不同。

(a) 向前跑



(b) 向前走

实验环境-动作空间

- OpenAI Gymnasium Walker2D 任务简介:
- Walker是一个二维的双腿人形，由七个主要身体部分组成——顶部的单个躯干（躯干后两条腿分开），躯干下方中间的两条大腿，底部的两条腿大腿下方，还有两只脚连在腿上，整个身体靠在腿上。目标是通过在连接七个身体部位的六个铰链上施加扭矩来向前（右）行走或者奔跑。
- 动作空间(Action Space):

Num	Action	Control Min	Control Max	Name (in corresponding XML file)	Joint	Unit
0	Torque applied on the thigh rotor	-1	1	thigh_joint	hinge	torque (N m)
1	Torque applied on the leg rotor	-1	1	leg_joint	hinge	torque (N m)
2	Torque applied on the foot rotor	-1	1	foot_joint	hinge	torque (N m)
3	Torque applied on the left thigh rotor	-1	1	thigh_left_joint	hinge	torque (N m)
4	Torque applied on the left leg rotor	-1	1	leg_left_joint	hinge	torque (N m)
5	Torque applied on the left foot rotor	-1	1	foot_left_joint	hinge	torque (N m)

实验环境-状态空间

- OpenAI Gymnasium Walker2D 任务简介:
  - 状态空间(Observation Space):

Num	Observation	Min	Max	Name (in corresponding XML file)	Joint	Unit
excluded	x-coordinate of the torso	-Inf	Inf	rootx	slide	position (m)
0	z-coordinate of the torso (height of Walker2d)	-Inf	Inf	rootz	slide	position (m)
1	angle of the torso	-Inf	Inf	rooty	hinge	angle (rad)
2	angle of the thigh joint	-Inf	Inf	thigh_joint	hinge	angle (rad)
3	angle of the leg joint	-Inf	Inf	leg_joint	hinge	angle (rad)
4	angle of the foot joint	-Inf	Inf	foot_joint	hinge	angle (rad)
5	angle of the left thigh joint	-Inf	Inf	thigh_left_joint	hinge	angle (rad)
6	angle of the left leg joint	-Inf	Inf	leg_left_joint	hinge	angle (rad)
7	angle of the left foot joint	-Inf	Inf	foot_left_joint	hinge	angle (rad)
8	velocity of the x-coordinate of the torso	-Inf	Inf	rootx	slide	velocity (m/s)
9	velocity of the z-coordinate (height) of the torso	-Inf	Inf	rootz	slide	velocity (m/s)
10	angular velocity of the angle of the torso	-Inf	Inf	rooty	hinge	angular velocity (rad/s)
11	angular velocity of the thigh hinge	-Inf	Inf	thigh_joint	hinge	angular velocity (rad/s)
12	angular velocity of the leg hinge	-Inf	Inf	leg_joint	hinge	angular velocity (rad/s)
13	angular velocity of the foot hinge	-Inf	Inf	foot_joint	hinge	angular velocity (rad/s)
14	angular velocity of the thigh hinge	-Inf	Inf	thigh_left_joint	hinge	angular velocity (rad/s)
15	angular velocity of the leg hinge	-Inf	Inf	leg_left_joint	hinge	angular velocity (rad/s)
16	angular velocity of the foot hinge	-Inf	Inf	foot_left_joint	hinge	angular velocity (rad/s)

课题一：Multi task imitation learning

- 本实验中，我们提供了对每个任务提供了两个数据集，共4个数据集，每个数据集包含5w个样本 (50 trajectory，每个trajectory 长度为1000)的数据集。
  - 初级专家数据集：该数据集由TD3 算法经过少量训练得到的智能体收集得来的。
  - 专家数据集：该数据集由TD3 算法经过充分训练得到的智能体收集得来的。
- 本课题需将奖励信息从上述4个数据集中移除。
- 可参考基线算法:
  - MH-AIRL [1]：常用模仿学习算法AIRL的multi task 版本。
  - MAML-IL [2]: 集成了MAML (一种常见的元学习算法)和行为克隆算法 (BC)的算法。

课题二：Multi task offline RL

- 本实验中，我们使用课题一中提供的4个数据集 (包括奖励信息)
- 为了模拟离线强化学习的环境，此任务中本地训练智能体的过程中**只能**在我们提供的数据集上进行，而不能使用额外的数据集或直接通过与Walker2D环境交互直接进行在线强化学习(Online RL)训练。
- 可参考基线算法:
  - CDS [3]: 使用保守数据共享策略实现多任务离线强化学习高效数据共享，以在各任务中有良好表现。
  - Multi-head CQL [4]: 采用多头构架拓展离线强化学习算法CQL以适应多任务。

课题参考资料

- Multi task imitation learning
  - 参考资料：https://cs330.stanford.edu/fall2021/slides/cs330_2021_mtrl.pdf
  - 参考视频：https://www.youtube.com/watch?v=_ND7muYS9qY&list=PLoROMvodv4rMIJ-TvblAIkw28Wxi27B36&index=9
- Multi task offline RL
  - 参考资料：https://cs330.stanford.edu/fall2021/slides/cs330_2021_offline_rl.pdf
  - 参考视频：https://www.youtube.com/watch?v=VGLqzbsOSJY&list=PLoROMvodv4rMIJ-TvblAIkw28Wxi27B36&index=13

课程补充文件以及提交要求

- 课程补充文件 Project.zip 已上传至canvas
- Project.zip 包括
  - 数据集文件夹 collected_data
  - 环境构建文件夹 custom_dmc_tasks
  - 智能体实例文件 agent_example.py
- **提交要求**
  - 智能体接口要求:

    本项目提供智能体参考文件：agent_example.py 文件
  - 提交测评文件：网络结构及其对应的检查点

组队

- 请同学们在5.19（十三周周末） 23:59之前于共享文档中完成组队注册

【腾讯文档】2024春-AI3601-Project

https://docs.qq.com/sheet/DSkpKSUh5UlNNUWpV?tab=BB08J2

评分标准及时间安排

根据提交的 report 和最终的 presentation 进行打分：

Report占总成绩的30分 (包括model, results, novelty, discussion)，presentation 占10分。

时间安排如下：

- 第15周周末： Canvas 提交presentation slides以及模型评估文件 (agent_example.py、网络结构、相应的检查点)。
- 第 16 周：答辩，展示大作业的研究问题，采用的模型，实验结果与自己的思考。
- 第16周末：提交所有材料，包括report, 代码和附件。

注：本次大作业不强调模型性能，而是专注项目设计本身的创新性。

材料提交及答辩要求

Presentation slides：

- Presentation slides
  - 格式为.ppt或.pdf
  - 该文件将在答辩环节被使用
  - 在第一页，请注明小组编号、小组成员（角色和相应的贡献百分比）和演讲者姓名
  - 所有团队成员都应该在场，可以自行决定是由一个成员还是多个成员完成答辩

材料提交及答辩要求 2

- Report及源码：
  - 格式为.zip文件，其中包含一个.pdf的report和.zip的源码
  - Report使用NeurIPS 2024 Style Files，正文部分不超过9页（包含图表），附录部分不作限制

    https://media.neurips.cc/Conferences/NeurIPS2024/Styles.zip
  - 请在report中明确写出每个成员在小组中的角色和相应的贡献百分比
  - 最终材料不允许迟交

Reference

[1] Chen, J., Tamboli, D., Lan, T., & Aggarwal, V. (2023, July). Multi-task hierarchical adversarial inverse reinforcement learning. In International Conference on Machine Learning (pp. 4895-4920). PMLR.

[2] Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017, October). One-shot visual imitation learning via meta-learning. In Conference on robot learning (pp. 357-368). PMLR.

[3] Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., & Finn, C. (2021). Conservative data sharing for multi-task offline reinforcement learning. Advances in Neural Information Processing Systems, 34, 11501-11516.

[4] Offline Q-learning on Diverse Multi-Task Data Both Scales And Generalizes. ICLR 2023 Oral