2023 AI3601 强化学习大作业

2023年4月

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

作业形式

- 分组完成，每组3人，在给出的三个课题中选择一个并完成相关任务。每个课题列出了一些参考算法，也可以调研其他相关算法。
  - 在Gym Acrobot上实现 Imitation Learning 算法

    GAIL[1], SQIL[2], AIRL[3]
  - 在及第 3v3 贪吃蛇上实现 MARL 算法

    QMIX[4], MAPPO[5], COMA[6]
  - 在 Mujoco Hopper 上实现 offline RL 算法

    BCQ [7], BEAR [8]

课题一：Imitation Learning

在 Gym Acrobot数据集上，利用单智能体的模仿学习方法得到比较好的策略。

Acrobot机器人系统包括两个关节和两个连杆，其中两个连杆之间的关节可以被致动。 最初，连杆是向下悬挂的，目标是将下部连杆的末端摆动到给定的高度。

Acrobot - Gymnasium Documentation (farama.org)

连续状态空间、离散动作空间。

课题一：Imitation Learning

本实验中，我们提供了一个包含500个 (s, a, s') 样本的专家数据集。（该数据集由我们使用一个预先训练好的DDPG agent和Acrobot-v1环境连续交互采集并保存样本而来，专家策略可以达到平均奖励-76）。

为了模拟模仿学习的环境，此任务中本地训练智能体的过程中不能从环境中获得奖励 (reward) 信息。

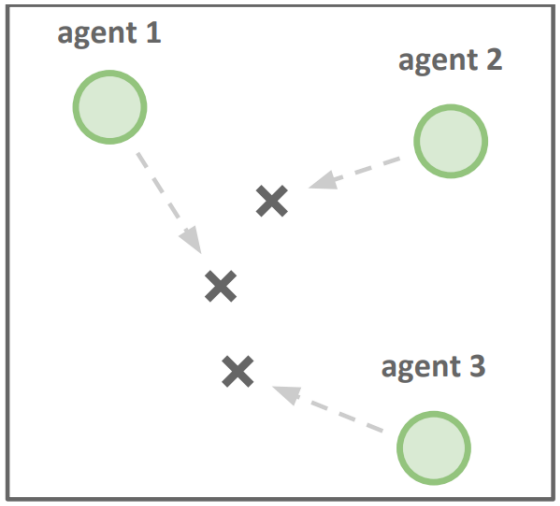
课题一：Imitation Learning

以下三个算法可进行参考：

- GAIL: 采用生成对抗的思路，将策略作为 GAN 中的生成器，去逼近策略采样得到的数据和专家数据之间的距离；
- AIRL: 整体思路类似于 GAIL，将策略加入到辨别器的结构中，从而能恢复出专家的奖励函数；
- SQIL: 采用固定的奖励函数来学习策略：专家数据对应的奖励函数是 1，采样得到的数据对应的奖励函数是 0。

课题二：Multi-agent Reinforcement Learning

多智能体强化学习（Multi-agent Reinforcement Learning）旨在处理环境中具有多个智能体的情况。

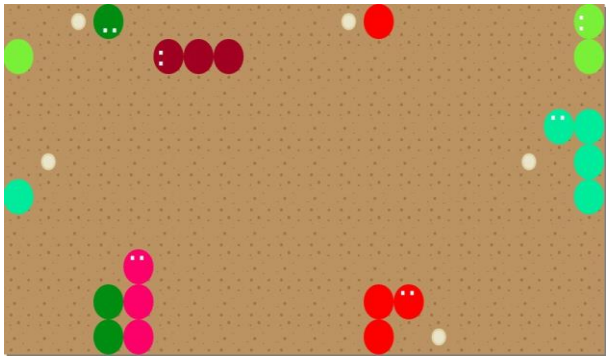


Particle

StarCraft II

课题二：Multi-agent Reinforcement Learning

3v3 贪吃蛇环境：
模拟环境代码库



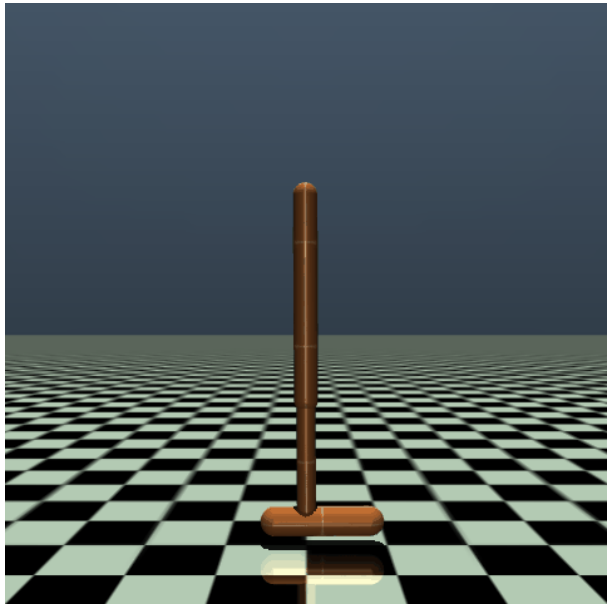
- 在 10x20 的网格中，对战双方各控制三条蛇；
- 初始时每条蛇长度为 3，位置随机；每一步，玩家可控制每条蛇向上、下、左、右四个方向移动，移动超出边界时可穿越到对侧位置；若蛇头撞击己方或对方某条蛇的身体，该蛇死亡，并在随机空位以长度 3 重生；
- 在地图空位上会随机产生 5 个食物，游戏过程中每当有一个食物被蛇吃掉，该蛇长度 +1，且会立即随机在某个空位上产生新的食物；
- 最终经过 200 步后，本方三条蛇长度之和更大的玩家获胜。

课题二：MARL

- 以下三个算法都适用于中心化训练，分布式执行的设定：
- QMIX：使用Hyper-Network处理多智能体的 Credit Assignment 和 Centralized Q 的维度爆炸问题，应用于合作型多智能体环境。
- MAPPO：使用Centralized Q，将单智能体算法引入多智能体环境中。
- COMA：采用反事实基线（Counterfactual Baseline）来解决 Credit Assignment 的问题。

课题三：Offline RL

- 离线强化学习（Offline RL）旨在只使用记录的数据来学习行为，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。
- 实验环境是 OpenAI Gymnasium Hopper 任务。



课题三：Offline RL

- OpenAI Gymnasium Hopper 任务简介:
- Hopper 是一个二维的单腿人形，由四个主要身体部位组成——躯干在顶部，大腿在中间，腿在底部，还有一只脚，整个身体都靠它休息。目标是通过在连接四个身体部位的三个铰链上施加扭矩，使跳跃向前移动。
  - 动作空间(Action Space):

The action space is a `Box(-1, 1, (3,), float32)`. An action represents the torques applied between *links*

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)

课题三：Offline RL

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

Num	Observation	Min	Max	Name (in corresponding XML file)	Joint	Unit
0	z-coordinate of the top (height of hopper)	-Inf	Inf	rootz	slide	position (m)
1	angle of the top	-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	velocity of the x-coordinate of the top	-Inf	Inf	rootx	slide	velocity (m/s)
6	velocity of the z-coordinate (height) of the top	-Inf	Inf	rootz	slide	velocity (m/s)
7	angular velocity of the angle of the top	-Inf	Inf	rooty	hinge	angular velocity (rad/s)
8	angular velocity of the thigh hinge	-Inf	Inf	thigh_joint	hinge	angular velocity (rad/s)
9	angular velocity of the leg hinge	-Inf	Inf	leg_joint	hinge	angular velocity (rad/s)
10	angular velocity of the foot hinge	-Inf	Inf	foot_joint	hinge	angular velocity (rad/s)

课题三：Offline RL

- 本实验中，我们提供了一个包含100,000个样本的数据集(该数据集由我们使用一个预先训练好的agent和Hopper环境交互采集并保存样本而来)。
- 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Hopper环境交互直接进行在线强化学习(Online RL)的训练。
- 可参考基线算法:
  - BCQ [7]: 通过限制动作空间来让智能体在给定的批数据集上和on-policy算法表现接近。
  - BEAR [8]: 限制分布外(out-of-distribution)动作来控制自举误差(bootstrapping error)。

课题补充文件

- 课题1文件：Imitation Learning Acrobot-v1 expert data.zip
- 课题3文件：Offline RL Hopper-v4 dataset.zip
- 所需数据文件已上传至canvas/文件/作业/

组队及注册

- 请同学们在4.9（周日） 23:59之前于共享文档中完成组队注册

【腾讯文档】AI3601-Project

- https://docs.qq.com/sheet/DTWRsQ0tMSkRGaVBB?tab=BB08J2

- 本次大作业将统一使用JIDI平台提交测评

JIDI平台使用

- 三个课题擂台链接如下：
  - Imitation Learning: http://www.jidiai.cn/compete_detail?compete=38
  - Multi-agent RL: http://www.jidiai.cn/compete_detail?compete=37
  - Offline RL: http://www.jidiai.cn/compete_detail?compete=39
- 注册事项：
  - 所有参赛成员均需使用邮箱/手机号注册及第账号，并填入真实姓名和组织
  - 参赛队伍队长登录自己的及第账号，在擂台的“立即报名”中填写队名(Group#) (组队腾讯文档中的队伍编号)、组织机构 (SJTU-AI3601) 、队员邮箱/手机号，提交成功后即完成报名
  - 在每场热身赛和正赛提交截止时间前，登录队长的及第账号，在“智能体提交”中提交智能体代码、相关参数文件等，并在“个人页-提交列表”查看智能体是否通过验证（可多次提交，但需在上一次验证完成后再进行新的提交，否则可能产生错误的验证结果）；
  - 各阶段结束后，在“赛果”中查看成绩排名、对局回放等比赛结果。
  - 每位参赛选手只可加入一支队伍，不可使用小号提交，也不可将代码分享给他人提交，一经发现将取消参赛资格。

JIDI平台使用 2

- 赛程包括两场热身赛（不强制参加）和一场正赛。正赛得分为最终总分的评定。
- 每一轮比赛均有一个提交截止时间，到达该时间节点后，以当前选手提交的最新测试通过的智能体参与此轮比赛的评测。
- 具体赛程
  - 开放报名：4月10日
  - 第一轮热身赛（提交截止：2023年4月30日 10:30）：采用瑞士轮复式赛制，成绩不计入总分
  - 第二轮热身赛（提交截止：2023年5月14日 10:30）：采用瑞士轮复式赛制，成绩不计入总分
  - 正赛（提交截止：2023年5月21日 10:30）：采用瑞士轮复式赛制，成绩为最终得分
- 更多细节请参考擂台链接

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14周周末：提交presentation slides, JIDI平台截止提交。
- 第 15/16 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 2023 Style Files，正文部分不超过9页（包含图表），附录部分不作限制

    https://neurips.cc/Conferences/2023/PaperInformation/StyleFiles
  - 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] Ho J, Ermon S. Generative adversarial imitation learning[J]. Advances in neural information processing systems, 2016, 29: 4565-4573.

[2] Reddy S, Dragan A D, Levine S. Sqil: Imitation learning via reinforcement learning with sparse rewards[J]. arXiv preprint arXiv:1905.11108, 2019.

[3] Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning[J]. arXiv preprint arXiv:1710.11248, 2017.

[4] Rashid T, Samvelyan M, Schroeder C, et al. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2018: 4295-4304.

Reference

[5] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of mappo in cooperative, multi-agent games[J]. arXiv preprint arXiv:2103.01955, 2021.

[6] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[7] Fujimoto, S., Meger, D., & Precup, D. (2019, May). Off-policy deep reinforcement learning without exploration. In International conference on machine learning (pp. 2052-2062). PMLR.

[8] Kumar, A., Fu, J., Soh, M., Tucker, G., & Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. Advances in Neural Information Processing Systems, 32.