Assignment 5：A3C & DDPG

叶增渝 123033910090

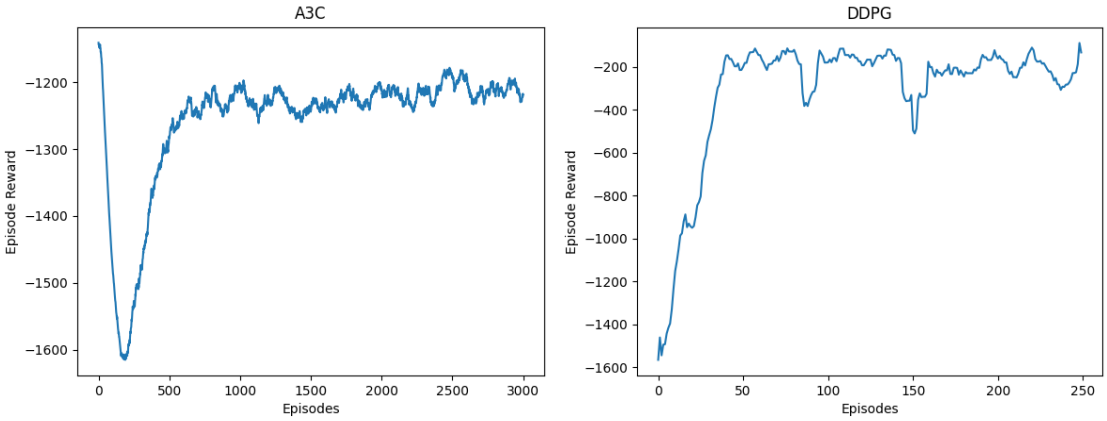
由于 Pendulum 的 v0 版本已经不适用，所以这里采用 v1 版本，使用最新的 gym0.26 来完成任务。

由于 v1 有完整的关于速度与角度的 Reward 函数计算公式，因此直接使用该 Reward 作为 buffer Reward 即可。

1）A3C 为异步 AC 算法。在传统 AC 算法中，Actor 进行动作选择，Critic 进行该动作的价值评估，两者互相联动，实现任务的学习。A3C 在 AC 算法的基础上引入了多个 local worker 同时进行 AC 算法更新参数，然后定期地提交参数到 global net 上进行参数更新。这里我们直接使用 torch.multiprocessing 进行多线程实现

2）DDPG 是一种可用于连续动作空间的强化学习算法，它除了基本的 AC 算法结构以外，使用了确定性策略、target net 与 replay buffer，是一个在连续空间上的优秀算法

以下为两个网络实验所得的 Episode Reward 随 Episode 的变化曲线图：



我们不难发现，A3C 网络虽然在逐步收敛，但是速度缓慢，在近 3000 的 episodes 下依然离收敛有很长的路要走；而 DDPG 网络则是很快能够收敛，在后续训练中虽然有波动但是基本维持在收敛状态。

Tips：源代码在本文件对应的文件夹下，A3C.py 与 DDPG.py 直接执行便可进行训练并保存 reward-episode 图像。