Homework 2

Task 1 Implementing DQN（第二小题选做）

在Atari中实现最基本的DQN算法。Atari环境会返回大小为$(height \times width \times channels)$图片作为智能体的观测。而强化学习的观测一般是采用一维向量的形式，通常只需要利用全连接神经网络，算法就可以收敛。由于Atari环境返回的是图片，因此需要用CNN对图像信息预处理，并将处理完的信息reshape成一维向量的形式作为强化学习神经网络的输入。强化学习算法的评判标准主要有reward收敛值大小、reward收敛速度等。

1. **(coding)** 在Atari `PongNoFrameskip-v4` 环境中实现DQN、Double DQN、Dueling DQN算法。环境最终的reward至少收敛至17.0。
2. **(coding)**在Atari `BreakoutNoFrameskip-v4` 环境中实现DQN、Double DQN、Dueling DQN算法。环境最终的reward至少收敛至200.0。

Task 2 Implementing Policy Gradient

在`cartpole`中实现policy gradient及其改进算法。策略梯度更新公式如下：

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta log \pi_\theta(a|s) Q^{\pi_\theta(s,a)}]$$

1. **(coding)** 在`cartpole`环境中实现基本的REINFORCE算法，即使用蒙特卡洛采样G_t作为$Q^{\pi_\theta(s,a)}$的无偏估计，更新公式如下：

$$J(\theta) = \frac{1}{\sum T_i} \sum_{i=1}^{|D|} \sum_{t=1}^{T_i} [log \pi_\theta(a_t^i|s_t^i) G_t^i],$$

   其中，$\tau^i = (s_0^i, a_0^i, r_0^i, s_1^i, \ldots, s_{T_i}^i, a_{T_i}^i, r_{T_i}^i)$，$D$是在环境中执行策略$\pi_\theta$产生的所有轨迹的集合。最终算法性能的评判标准：环境最终的reward至少收敛至180.0。

2. **(writing & coding)** 在`CartPole-v0`环境中实现REINFORCE算法的变种算法。虽然蒙特卡洛采样得到G_t是对reward的无偏估计，但环境的不确定性以及策略的随机性将会导致蒙特卡洛采样具有较大的方差。为了降低方差，在计算策略梯度的时候可以减去一个baseline，$b_\phi(s)$，常用的baseline形式就是状态值函数$V^{\pi_\theta}(s)$，具体公式：

$$J(\theta) = \frac{1}{\sum T_i} \sum_{i=1}^{|D|} \sum_{t=1}^{T_i} [log \pi_\theta(a_t^i|s_t^i) \hat{A}_t^i],$$

$$\hat{A}_t^i = G_t^i - b_\phi(s_t^i),$$

   其中$\hat{A}_t^i$被称为优势函数（advantage function）。从理论层面来说，优势函数不会影响策略梯度，请给出理论证明。最终算法性能的评判标准：环境最终的reward至少收敛至180.0.

3. **(coding)** 在`CartPole-v0`环境中实现A2C算法.最终算法性能的评判标准：环境最终的reward至少收敛至180.0.

Task 3 Implementing DDPG（选做）

在`Box2D`中实现DDPG算法。由于DQN需要评估所有动作的Q值，故无法直接应用于连续动作的环境。DDPG可以在actor中计算连续策略的高斯分布，因此可以从高斯分布中采样得到智能体的动作。

1. **(coding)** 在`LunarLanderContinuous-v2`环境中实现DDPG算法，其中`Lunar Lander`是连续动作环境。最终算法性能的评判标准：环境最终的reward至少收敛至180.0。
2. **(coding)** 在`Bipedal Walker`环境中实现DDPG、A2C、PPO 三个算法中的一个，其中`Bipedal Walker`是连续动作环境。最终算法性能的评判标准：环境最终的reward至少收敛至250.0。

Submission

作业提交内容：需提交一个zip文件，包括代码以及实验报告PDF。实验报告除了需要写writing部分的内容，还需要给出每题的reward曲线图以及算法。如果不同的题有不同的超参数，请在代码或者实验报告中说明。作业3为选做题，如果觉得有难度的同学可以不用做，学有余力的同学可以尝试完成相应的算法。

zip文件命名格式: RL_20220421_张三_homework2；如果需提交不同版本，则命名格式：RL_20220421_张三_homework2_v2等。

作业提交方式：zhangyc8@mail2.sysu.edu.cn

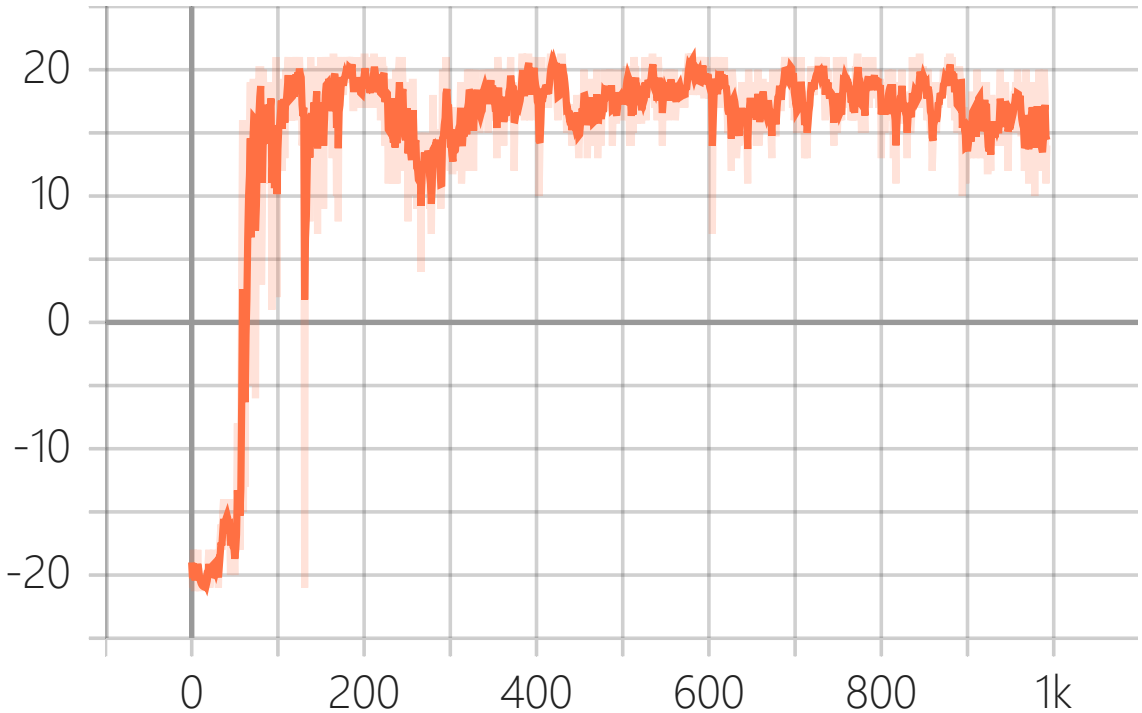
相关代码下载地址： https://github.com/ZYC9894/SYSU_RL2022

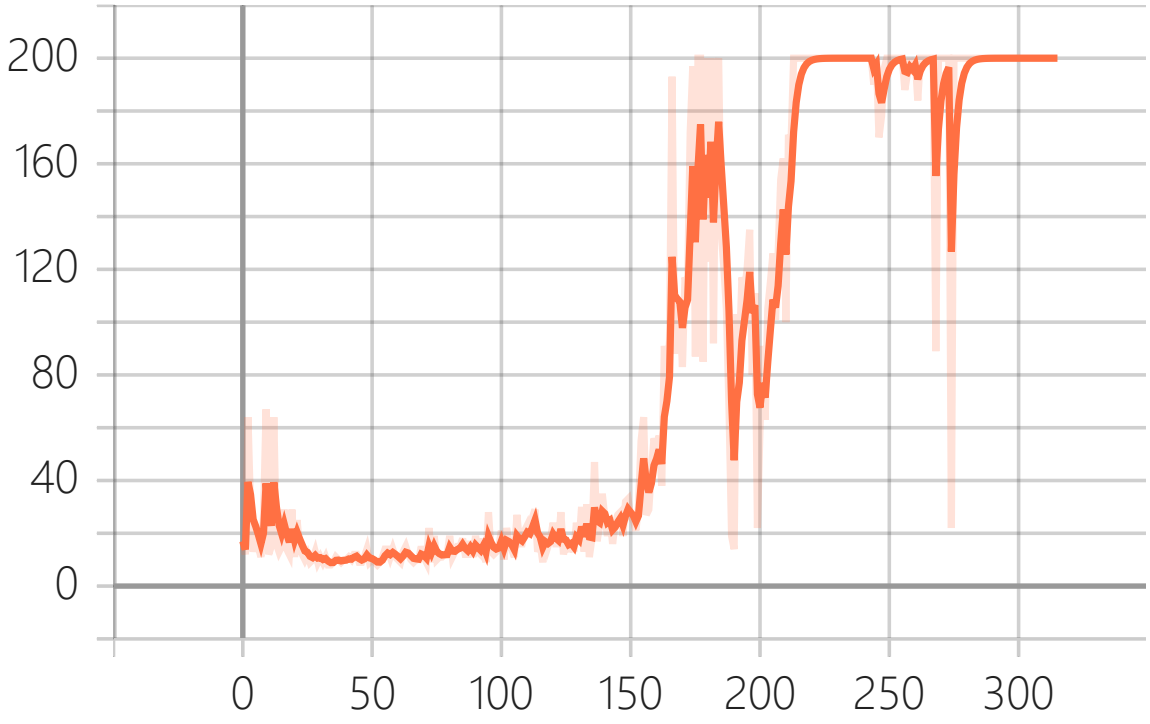
相关环境的说明文档：https://www.gymlibrary.ml/

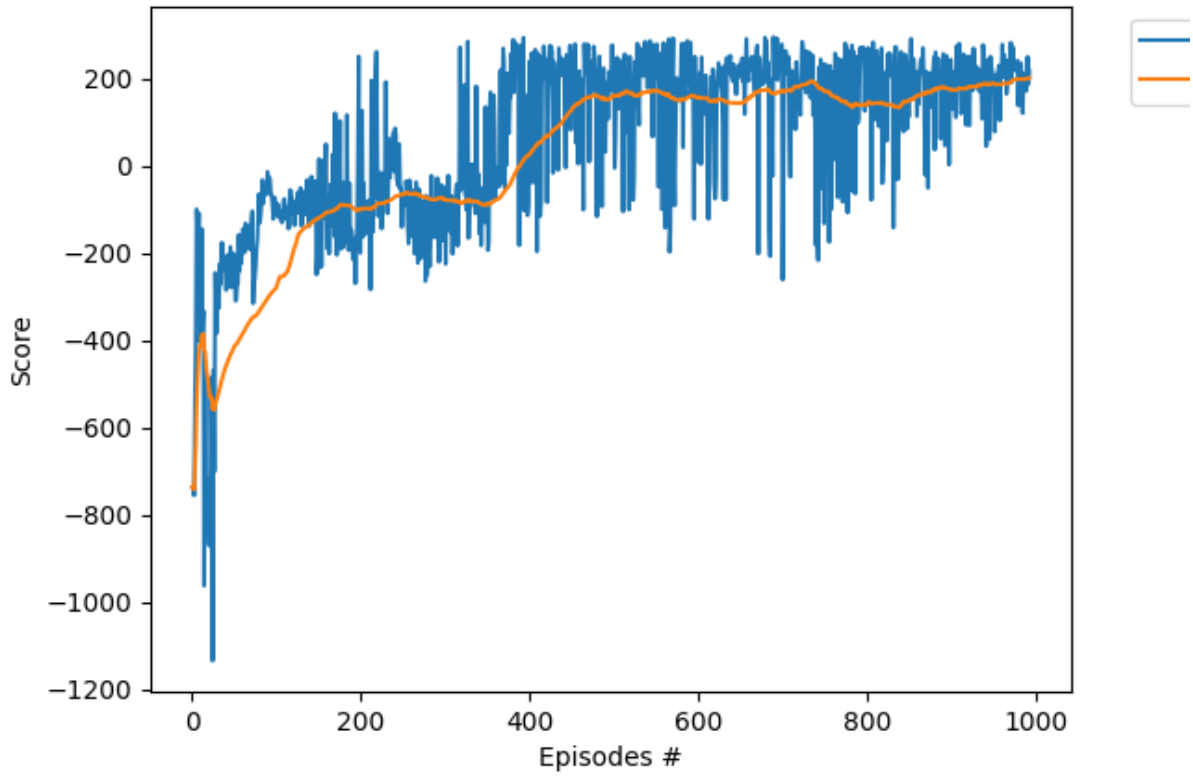
作业提交截止日期：2022年05月31日

Supplement

我们给出部分环境(`PongNoFrameskip-v4`、`cartpole`、`LunarLanderContinuous-v2`)的曲线图作为参考。







黄线是过去100个episode的平均奖励值
蓝线是当前episode的奖励值