Q1 Implement Advanced RL algorithm
2 分數

a) Choose one algorithm from REINFORCE with baseline、Q Actor-Critic、A2C, A3C or other advance RL algorithms and implement it.

b) Please explain the difference between your implementation and Policy Gradient

c) Please describe your implementation explicitly (If TAs can't understand your description, we will check your code directly.

a. Implement Advantage Actor Critic.

b. 一般的 policy gradient 使用一個 actor 的 network，且 rewards 為 discounted cumulative rewards，接著就是 minimize (-log_probs * returns) 的 loss; 而我實作的 A2C 是用 2 個 Network(分別為 actor 和 critic)，actor 要 minimize (-log_probs * advantages) 的 loss，其中 advantage 為老師在課堂中提到的 A_t=r_t+V^θ (s_(t+1) )-V^θ (s_t) ，critic 就是要 minimize critic network 產生出 predicted values 和原本的 discounted cumulative rewards。

c. 使用 2 個 Network，分別為 actor / critic Network。每次 actor 會依照當前環境 sample 出一個 action(有對應 critic network 產生的 state value)，互動完後產生出的下一個 state 會再一次經過 critic network 得到下一個 state value，然後計算 rewards 方式跟 policy gradient 一樣使用 discounted cumulative rewards，接著蒐集一定數量的 episode 資料就 update 一次 actor / critic network，actor 和 critic network 各自 optimize 目標如同 b. 所敘述，但我這個方法 train 比較久，我 train 了 10000 episodes，相較於我使用 DQN 只用了 2500 episodes 就達到了相同水平。

Q2 Below are descriptions about MuZero, which one is not correct?
2 分數

It is a tree based search + model based work

✓ Its agent doesn’t know about the real transition function

It utilizes the MCTS algorithm during training

It doesn’t need to know about the rules of those games it modeled

HW12

● 已批改

總分
2 / 4 pts

問題 1
Implement Advanced RL algorithm 2 / 2 pts

問題 2
Below are descriptions about MuZero, which one is not correct? 0 / 2 pts