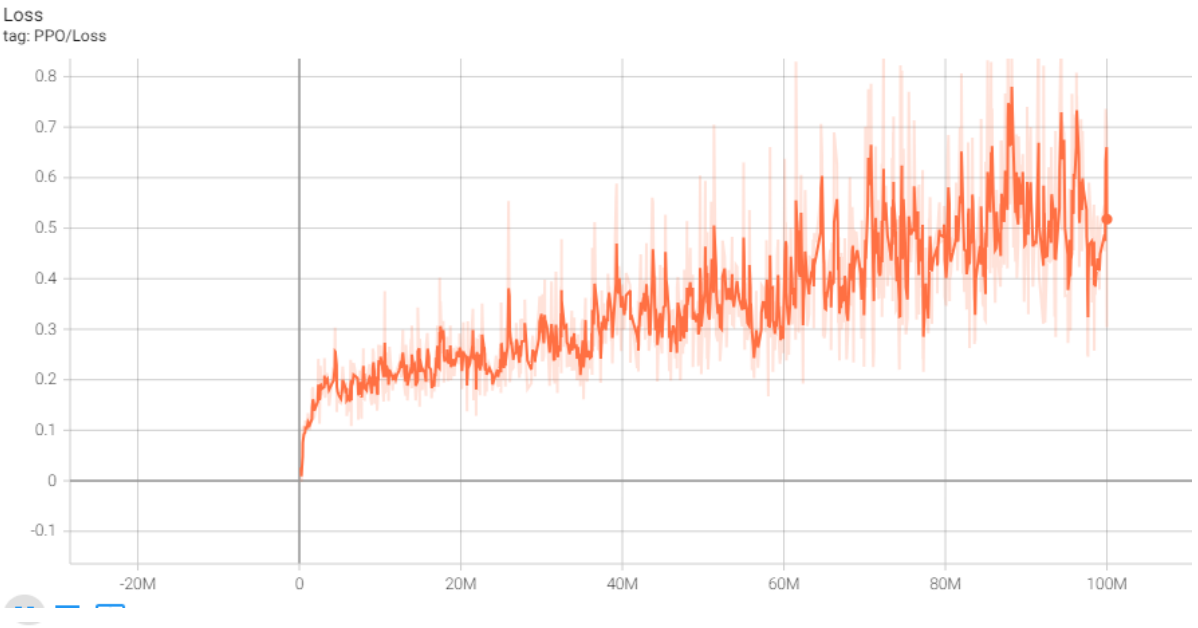
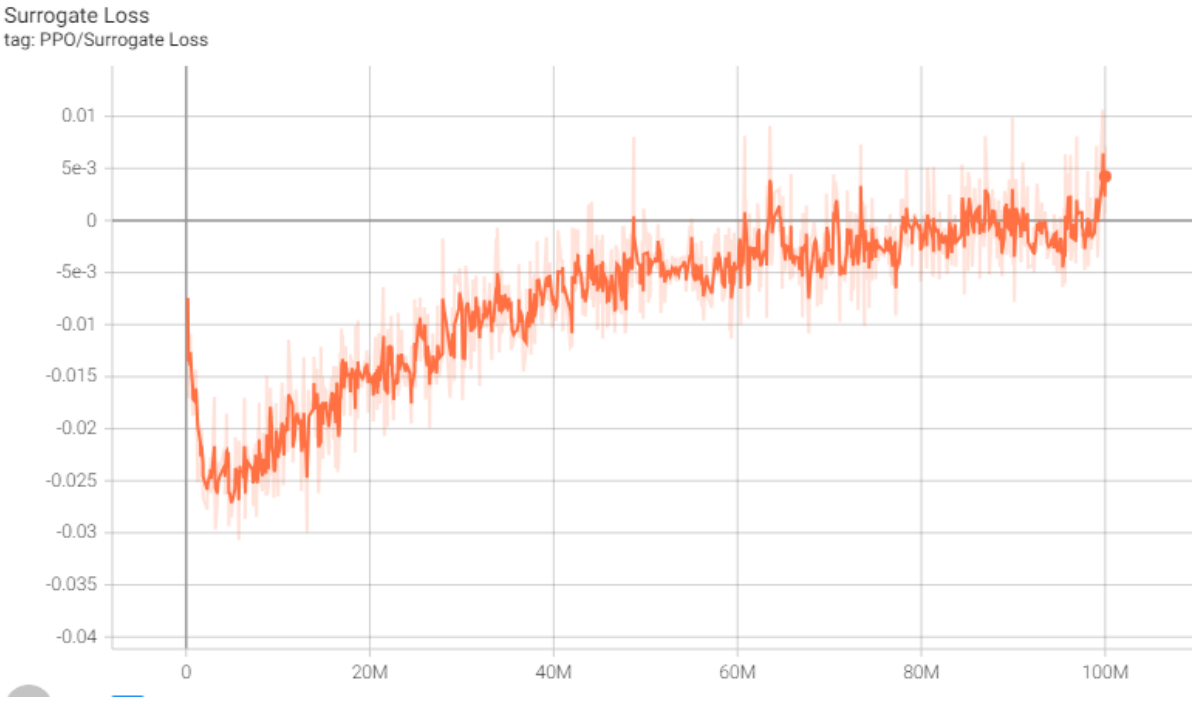
311706009 王廣和

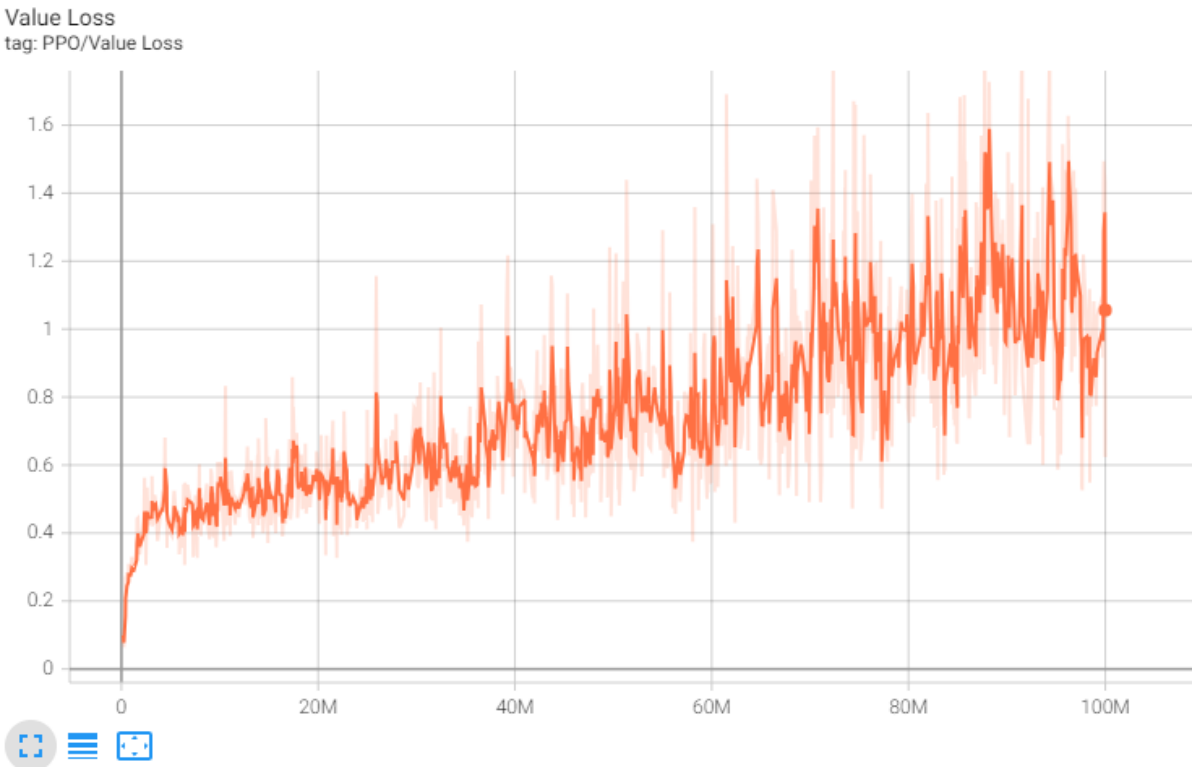
Experimental Results:

311706009 王廣和





311706009 王廣和



311706009 王廣和

Answer the questions of bonus parts (bonus) (20%)
(1) PPO is an on-policy or an off-policy algorithm? Why? (5%)
on-policy, 因為跟環境互動的network與更新的network是同一個

(2) Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization. (5%)
用clip ratio = policy/ policy_old 來確保action prob不會劇烈變動, 將ratio range keep在(1+epislon, 1-epislon)

(3) Why is GAE-lambda used to estimate advantages in PPO instead of just onestep advantages? How does it contribute to improving the policy learning process? (5%)
GAE可以透過lambda與gamma衡量未來reward對現在的重要程度, 且有效平衡 bias 與 variance

(4) Please explain what the lambda parameter represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO? (5%)
lambda指的是對未來reward的重視程度, gamma則是discount rate, 如果我比較在意遊戲後期的reward, 就應該把lambda調高一點, 如果我希望agent只專注在時間t之後幾步而已(短視近利greedy一點), 那就該調低