Machine Learning HW15

ML TAs
ntu-ml-2020spring-ta@googlegroups.com

作業內容

在本次作業當中，你們將實做並比較幾項 Deep Reinforcement Learning 方法：

- Policy Gradient
- Actor-Critic

作業的實做環境為 OpenAI 的 gym 當中的 Lunar Lander。其餘實做細節請參考助教提供的範例程式。

Policy Gradient 方法

Algorithm 1 Policy Gradient

function REINFORCE

Initialize policy parameters θ

for each episode $\{s_1, a_1, r_1, \ldots, s_T, a_T, r_T\} \sim \pi_\theta$ **do**

for $t = 1$ to T **do**

Calculate discounted reward $R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$

$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) R_t$

end for

end for

return θ

end function

Actor-Critic 方法

Algorithm 2 Actor-Critic

function REINFORCE WITH BASELINE

Initialize policy parameters θ

Initialize baseline function parameters ϕ

for each episode $\{s_1, a_1, r_1, \ldots, s_T, a_T, r_T\} \sim \pi_\theta$ **do**

for $t = 1$ to T **do**

Calculate discounted reward $R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$

Estimate advantage $A_t = R_t - b_\phi(s_t)$

Re-fit the baseline by minimizing $\|b_\phi(s_t) - R_t\|^2$

$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) A_t$

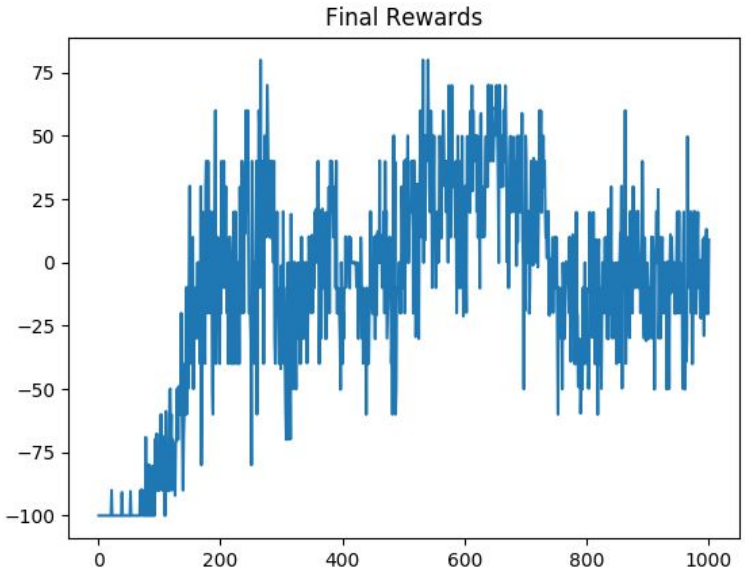
end for

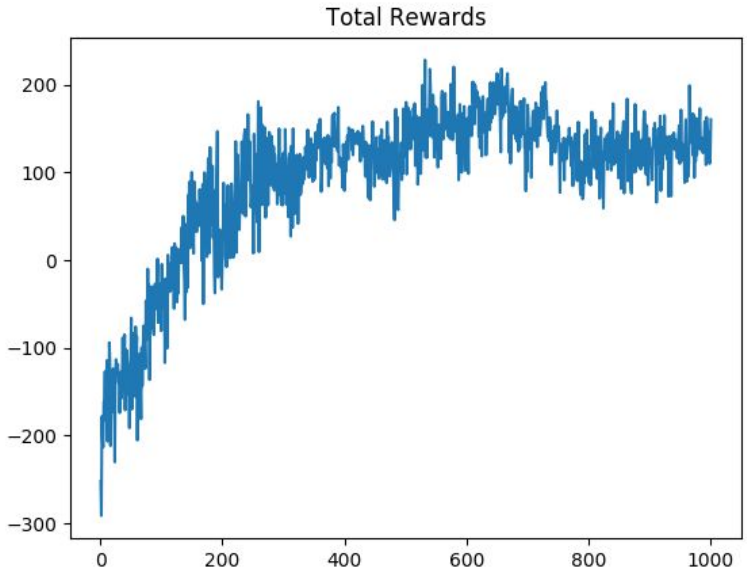
end for

return θ

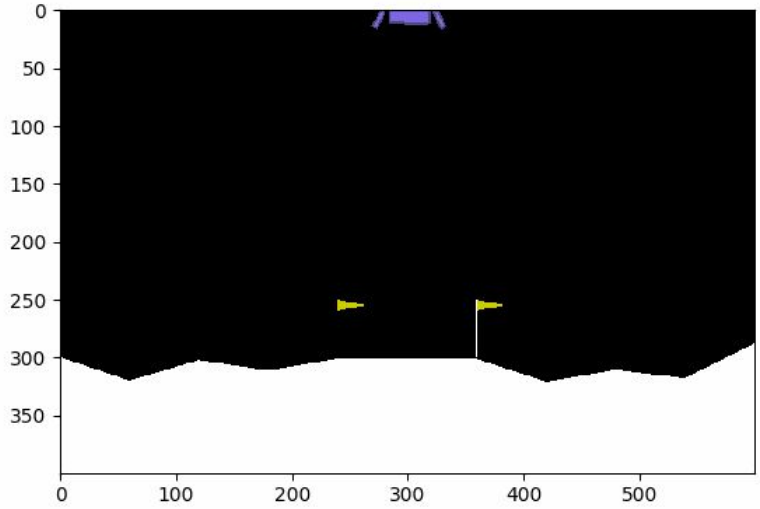
end function

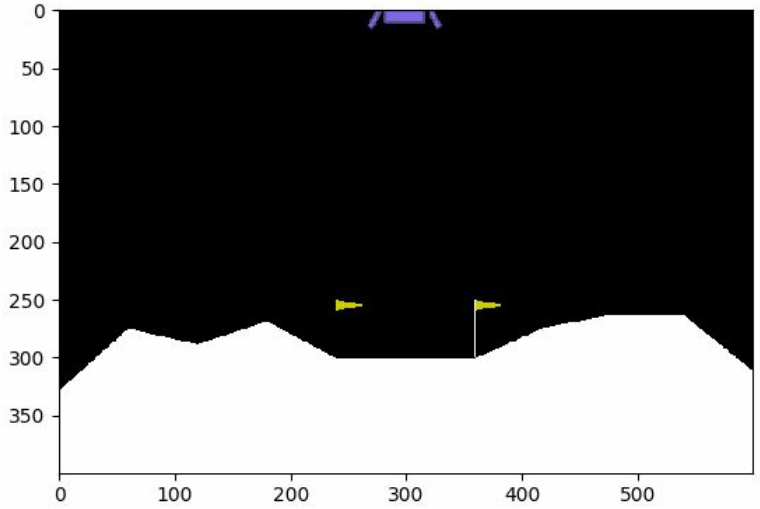
範例結果


Final Rewards
75
50
25
0
−25
−50
−75
−100
0
200
400
600
800
1000


Total Rewards
200
100
0
−100
−200
−300
0
200
400
600
800
1000

範例展示


0
50
100
150
200
250
300
350
0
100
200
300
400
500


0
50
100
150
200
250
300
350
0
100
200
300
400
500

繳交項目

GitHub 上的 hw15-<account> 裡必須包含以下檔案:

1. Python 程式碼
2. 報告
    - 報告內容請參見報告範本
    - 以 PDF 格式繳交
    - 請命名為 **report.pdf**

注意事項

- 本次作業以報告為評分標準，所有圖表、表格等請貼在報告當中以利評分
- 所有作業相關問題請在 FB 社團貼文或是寄信至助教信箱，並於信件主題處註明：**[HW15]**