Machine Learning HW12

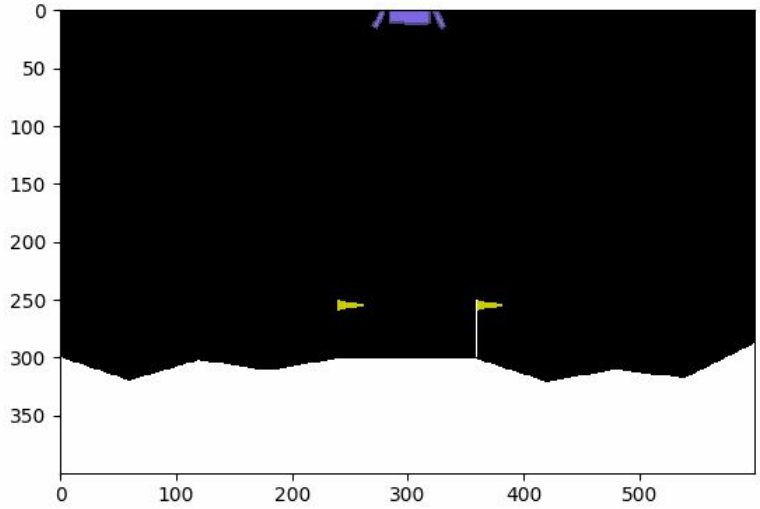
ML TAs
ntu-ml-2021spring-ta@googlegroups.com

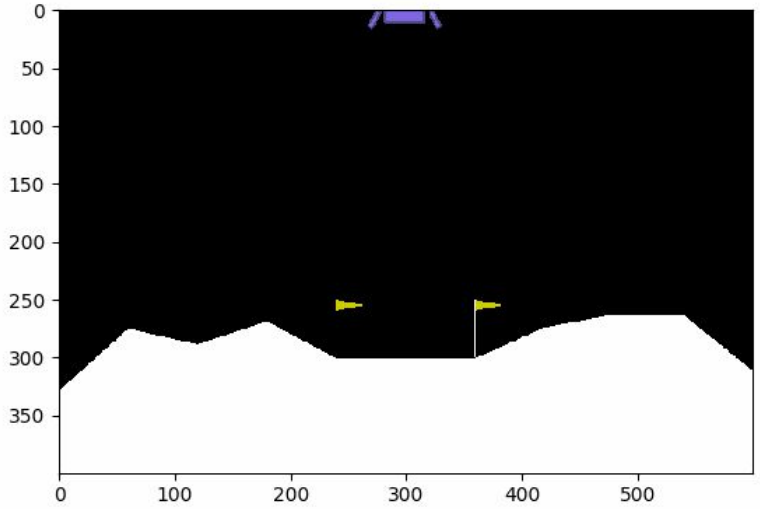
作業內容

在本次作業當中，你們將可以實做幾項 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。

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

Policy Gradient 方法(to get 8 points)

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(to get 10 points)

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 $\|b_\phi(s_t) - R_t\|^2$

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

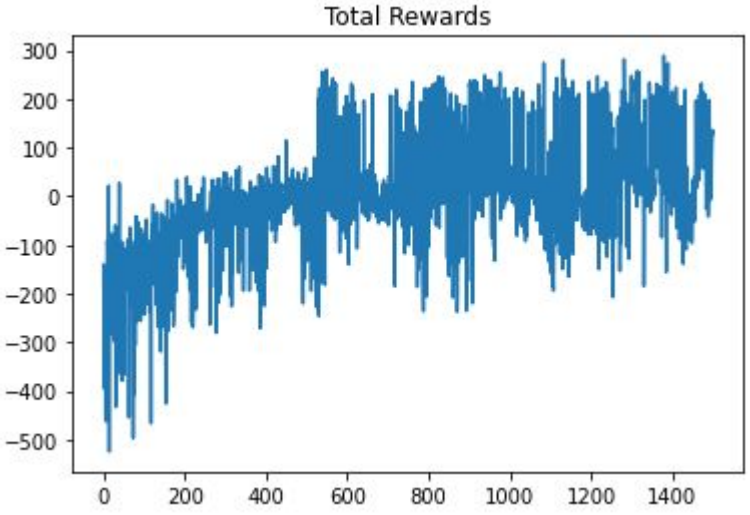
end for

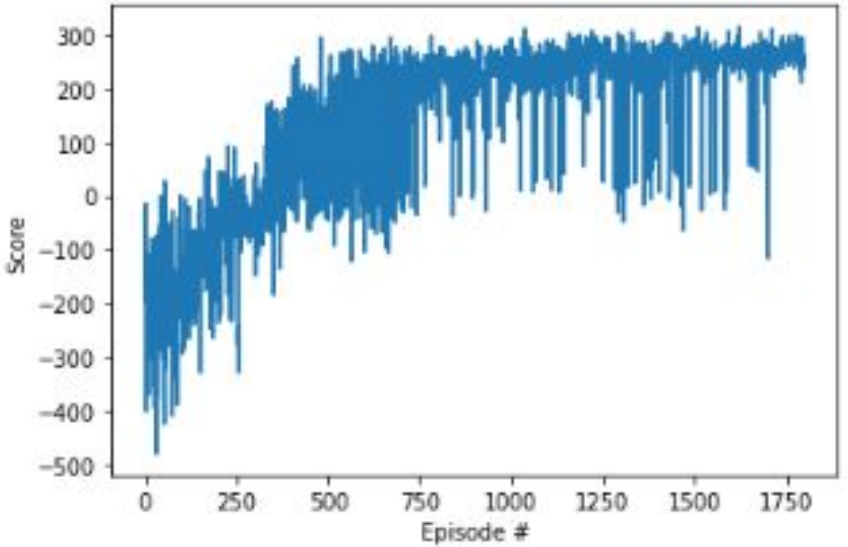
end for

return θ

end function

範例結果


Total Rewards
300
200
100
0
−100
−200
−300
−400
−500
0
200
400
600
800
1000
1200
1400


Score
300
200
100
0
−100
−200
−300
−400
−500
0
250
500
750
1000
1250
1500
1750
Episode #

繳交項目及評分標準

1. Python 程式碼 (Submit on NTU COOL) **佔4分**
2. Action List (Submit on JudgeBoi, **沒有private set, 自動選擇最高分**)
3. 給分標準: (Your submission must be valid)

15min~20min

30min

Avg_Reward	
< 0	2
0~99	3
100~199	4
200~240	5
241~	6

繳交項目及評分標準

More on a "valid submission “:

agent在action list最後一個動作輸入之後，應該要輸出done。長度過長或過短的action list都會被系統reject。

Action list 的長相

```
1 print("Action list looks like ", action_list)
2 print("Action list's shape looks like ", np.shape(action_list))
```

```
Action list looks like [[3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 2, 3, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3
Action list's shape looks like (5,)
```

Bonus

- If you successfully get 10 pts:
  - Your code will be made public to students.
  - You can submit a report in **PDF** format briefly describing what you have done (in English, less than 100 words) for **extra 0.5 pts**.
  - Reports will also be made public to students.
  - Notice, we do not have private score, so omit it in the report.

- Report template

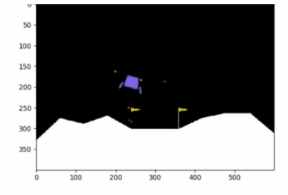
注意事項

- You should finish your homework on your own.
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- **Do NOT search or use additional data or pre-trained models.**
- Your **final grade x 0.9** if you violate any of the above rules.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

注意事項

- 所有作業相關問題請在 NTU COOL詢問(推薦)或是寄信至助教信箱，並於信件主題處註明：**[HW12]**

當你試著想讓
火箭降落在旗子內
好完成HW12



Submit Deadline: 6/04 - 6/25 (23:59)

你家的狗：
那是什麼？
可以吃嗎？