Machine Learning HW12

ML TAs
ntu-ml-2022spring-ta@googlegroups.com

HW Content

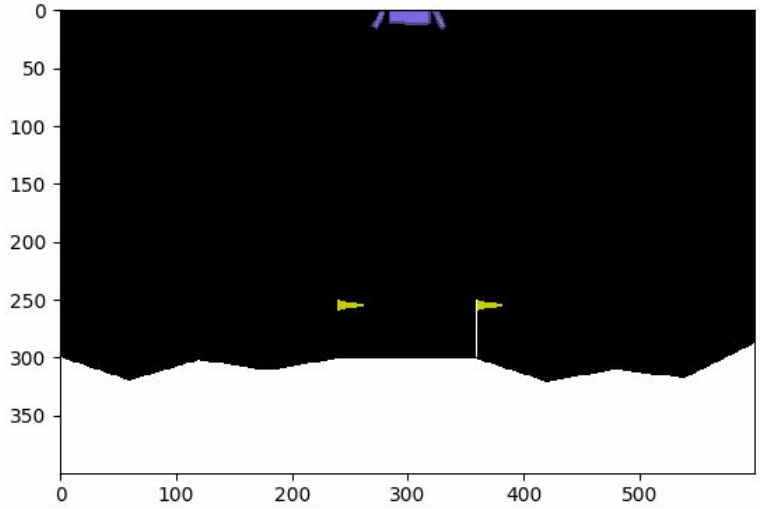
In this HomeWork, you can implement some Deep Reinforcement Learning methods by yourself :

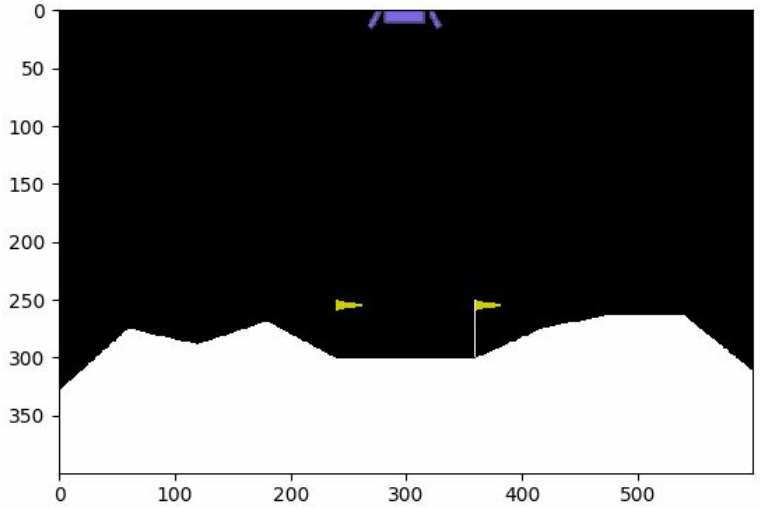
- Policy Gradient
- Actor-Critic (Implement by yourself to get high score !)

The environment of this HW is **Lunar Lander** in gym of OpenAI.

Other details can be found in the sample code.

Illustraion





Policy Gradient(to get 3 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(to get 4 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 $\|b_\phi(s_t) - R_t\|^2$

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

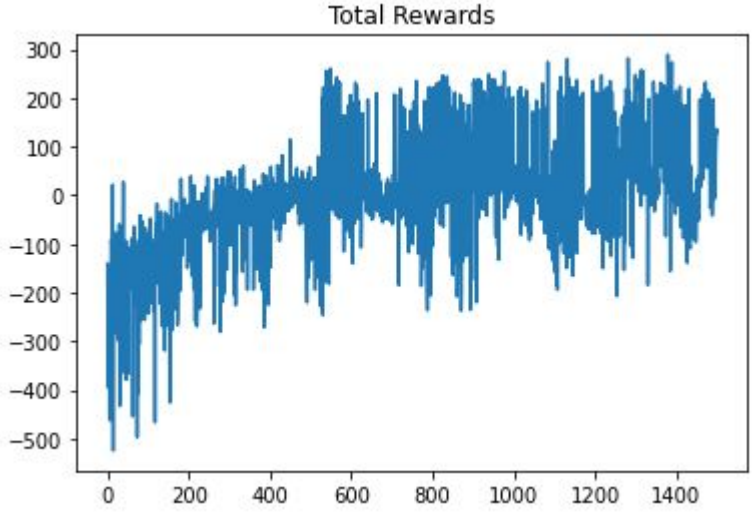
end for

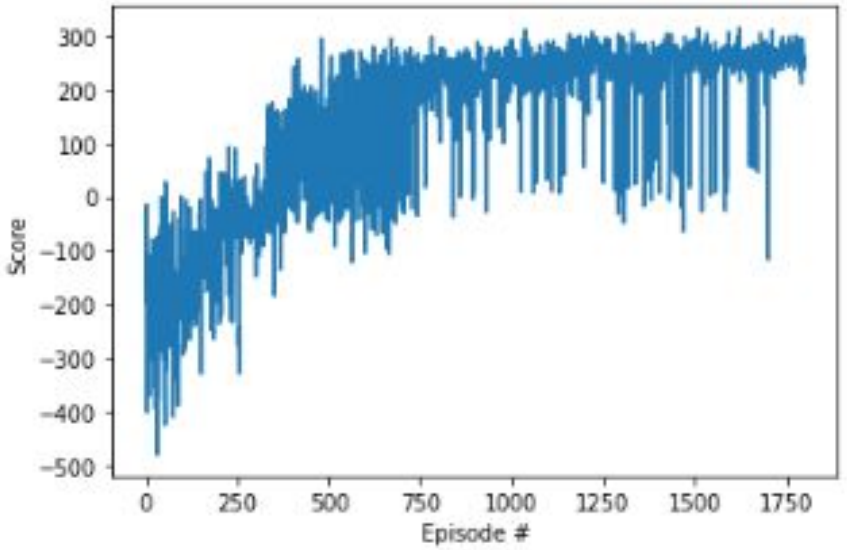
end for

return θ

end function

Sample Result





What you need to submit & Grading

1. Python file (**2 points**) (Submit on NTU COOL)
2. Action List (**4 points**) (On JudgeBoi, no private set, **the highest one is automatically selected**)
3. Report (**4 points**) (The questions are on **gradescope)**

Points	Interval
0	No valid submission or < 0
1	0 - 99
2	100 - 169
3	170 - 269
4	> 269

JudgeBoi General Rules

- 5 submission quota per day, reset at midnight.
  - Users not in the whitelist will have no quota.
- The countdown timer on the homepage is for reference only.
- We do limit the number of connections and request rate for each IP.
  - If you cannot access the website temporarily, please wait a moment.
- The system can be very busy as the deadline approaches
  - If this prevents uploads, we do not offer additional opportunities for remediation
- Please do not attempt to attack JudgeBoi.
- Every Friday from 6:00 to 9:00 is our system maintenance time.
- For any JudgeBoi issues, please post on NTUCOOL discussion
  - Discussion Link: https://cool.ntu.edu.tw/courses/11666/discussion_topics/91777

JudgeBoi HW12-Specific Rules

- Only *.npy file is allowed, file size should be smaller than **2MB**.
- You do not have to select submission since there is no private score
- JudgeBoi should complete the evaluation within one minute.
  - You do not need to wait for the progress bar to finish

What you need to submit & Grading

More on a "valid submission “:

Your agent should output done after the last input of your action list, action list with mismatched length will be rejected。

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,)
```

Report

1. (2分) Implement Advanced RL algorithm
    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.

Report

2. (2分) Below are descriptions about MuZero, **which one is not correct?**

MuZero, a new approach to model-based RL that achieves state-of-the-art performance in Atari 2600, a visually complex set of domains, while maintaining superhuman performance in pre- cision planning tasks such as chess, shogi and Go.

You have to answer according to MuZero Paper

(a) It is a tree based search + model based work

(b) Its agent doesn't know about the real transition function

(c) It utilize the MCTS algorithm during training

(d) It didn't need to know about the rules of those games it modeled

Note

- HW12 won’t use GPU by default.
- We recommend to use Colab in HW12.
- If anyone intend to use environments other than Colab, please fix reproducibility issues by yourself. TA won’t help you to fix any environment issue.
- The training of HW12 should be able to finish within 30 min.

Submission

- Submit your code to NTU COOL
  - We can only see your last submission
  - Do not submit the model or dataset
  - If your codes are not reasonable, your final grade will be x 0.9
  - You should compress your files into one single file. Wrong format may cause penalty
    - <student_id_lower_case>_hw12.zip

Grading -- Bonus

If your **ranking is in top 3**, you can choose to share a report to

NTU COOL and get extra 0.5 pts.

About the report

Report template

- Your name and student_ID
- Methods you used in code
- Reference
- in 200 words
- Deadline is same as code submission
- Please upload to NTU COOL’s discussion of HW12

If any questions, you can ask us via...

- NTU COOL (recommended)
  - https://cool.ntu.edu.tw/courses/11666
- Email
  - mlta-2022-spring@googlegroups.com
  - The title **must** begin with “[hw12]”
- TA hours
  - Each Tuesday 20:00~21:00 @ Online
  - Each Friday 16:30~17:20 @ Online
  - Each Friday 22:00~23:00 (English) @ Online

Announcement

- 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.

Announcement

- Any questions or concerns about HW can be post on NTU COOL(Recommend) or send email to ntu-ml-2022spring-ta@googlegroups.com . Please denote the subject of email by **[HW12]**

Submit Deadline: 5/20 - 6/10 (23:59)