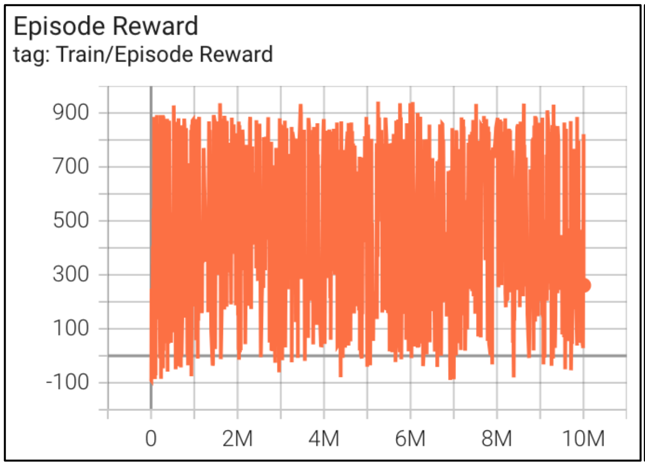
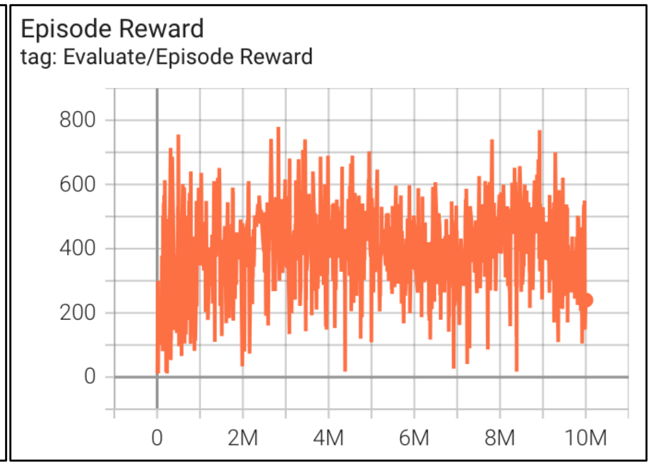
Twin Delayed DDPG
Lab Report # 4

By
312581020
許瀚丰

Selected Topics in Reinforcement Learning
Winter 2023
Date Submitted: December 3, 2023

- Screenshot of Tensorboard training curve and testing results on TD3. (30%)
  - Training curve

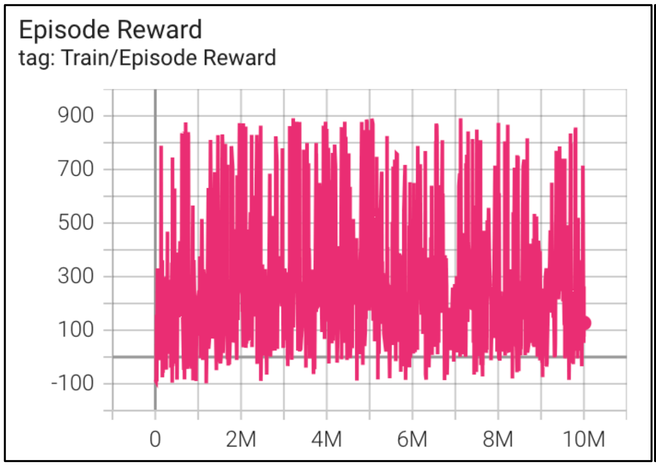


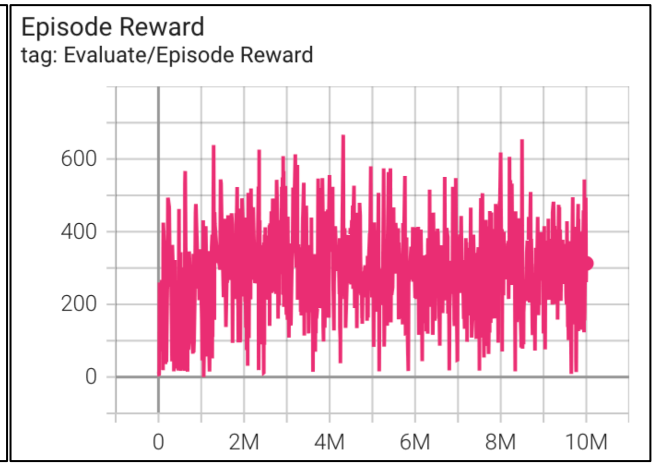


  - Testing results (10 games)

```
Episode: 1      Length: 999     Total reward: 864.05
Episode: 2      Length: 999     Total reward: 818.50
Episode: 3      Length: 252     Total reward: 246.34
Episode: 4      Length: 250     Total reward: 305.93
Episode: 5      Length: 999     Total reward: 808.47
Episode: 6      Length: 573     Total reward: 681.59
Episode: 7      Length: 999     Total reward: 767.51
Episode: 8      Length: 999     Total reward: 769.84
Episode: 9      Length: 510     Total reward: 631.75
Episode: 10     Length: 999     Total reward: 896.08
average score: 679.0067630242895
```

- Screenshot of Tensorboard training curve and compare the performance of using twin Q-networks and single Q-networks in TD3, and explain (5%)
  - Training curve

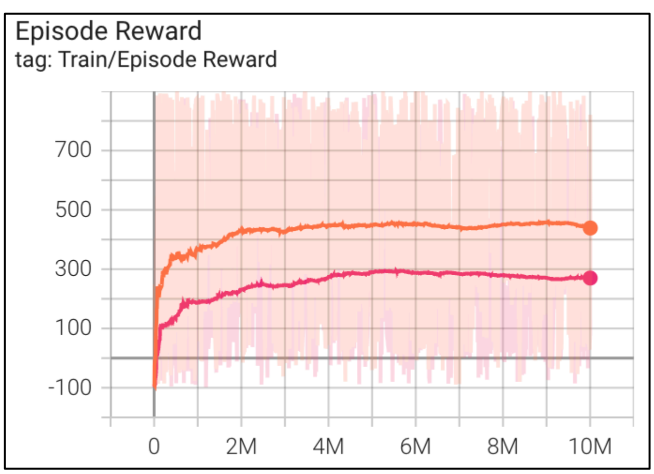


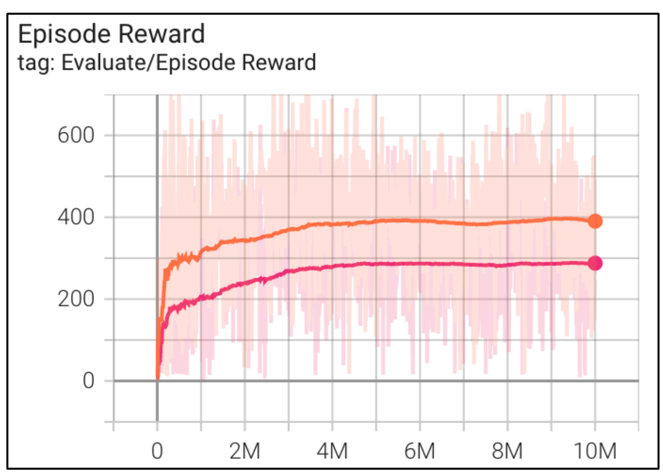


  - Testing results (10 games)

```
Episode: 1      Length: 577     Total reward: 732.17
Episode: 2      Length: 254     Total reward: 289.14
Episode: 3      Length: 312     Total reward: 272.78
Episode: 4      Length: 999     Total reward: 832.20
Episode: 5      Length: 274     Total reward: 333.09
Episode: 6      Length: 295     Total reward: 350.46
Episode: 7      Length: 137     Total reward: 90.93
Episode: 8      Length: 998     Total reward: 837.38
Episode: 9      Length: 289     Total reward: 335.52
Episode: 10     Length: 999     Total reward: 879.59
average score: 495.3266047822893
```

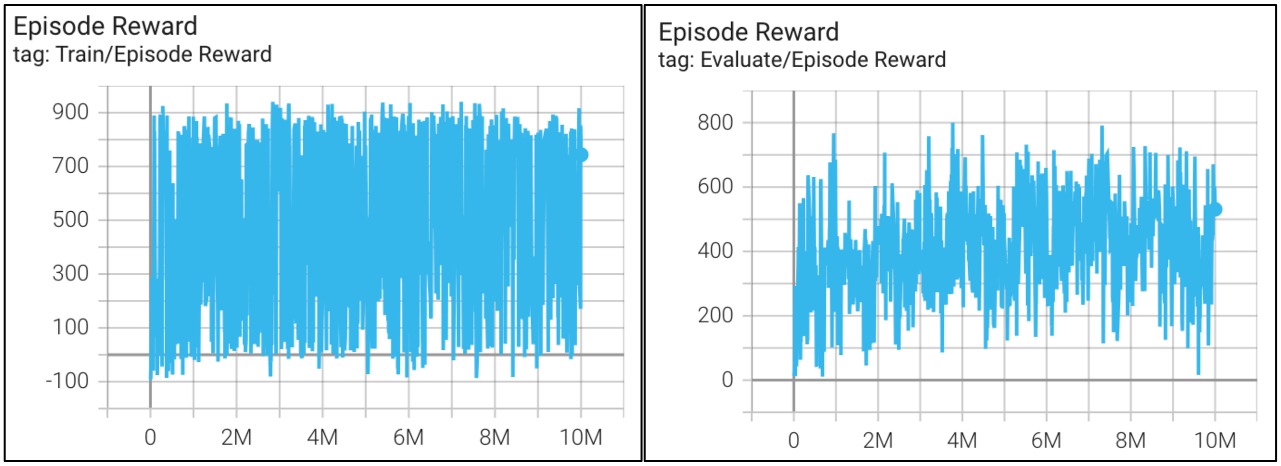
  - Training curve (comparison)
    - TD3 (Orange) and TD3 without twin Q-networks (magenta), with 0.999 smoothing





- Discuss between TD3 and TD3 without twin Q-networks
  - 我認為使用 twin Q-networks 的做法其實與 Double DQN 的想法相當類似，其目標都是希望讓預測出的 Q value 不要有高估的問題。
  - 而在實作上，TD3 的 twin Q-networks 是直接在兩個 critic 中預測出來的 Q value 取 min，相比於後者在實作上簡單許多。而在結果中可以觀察到，TD3 without twin Q-networks 在結果上相比於 TD3 差了非常多，可見 twin Q-networks 在 TD3 中是非常重要的一部分。

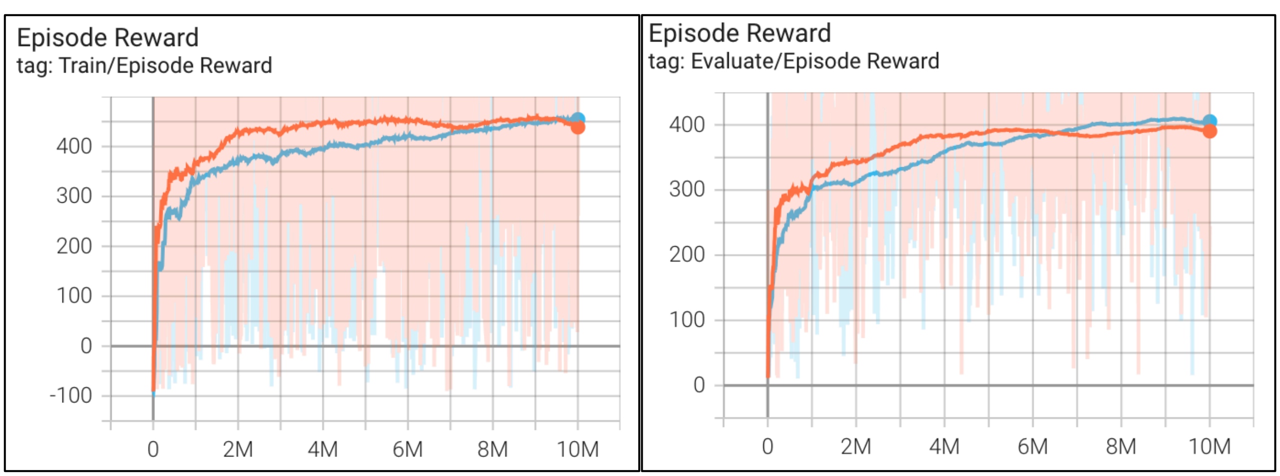
- Screenshot of Tensorboard training curve and compare the impact of enabling and disabling target policy smoothing in TD3, and explain (5%).
  - Training curve



  - Testing results (10 games)

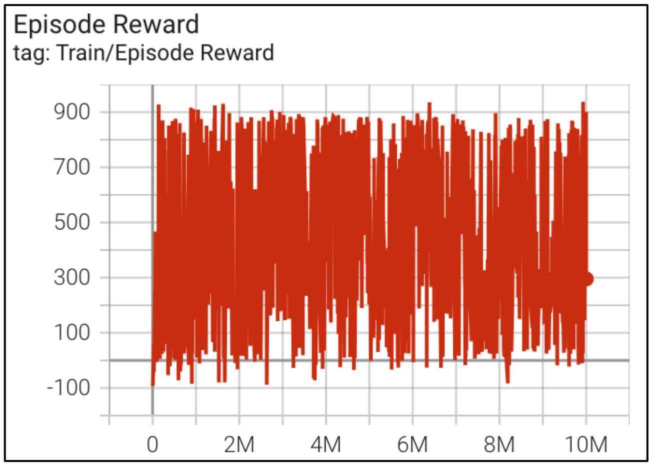
```
Episode: 1      Length: 487     Total reward: 716.35
Episode: 2      Length: 999     Total reward: 892.25
Episode: 3      Length: 999     Total reward: 838.46
Episode: 4      Length: 660     Total reward: 933.90
Episode: 5      Length: 405     Total reward: 419.33
Episode: 6      Length: 999     Total reward: 838.98
Episode: 7      Length: 999     Total reward: 864.29
Episode: 8      Length: 344     Total reward: 366.84
Episode: 9      Length: 551     Total reward: 646.44
Episode: 10     Length: 999     Total reward: 857.60
average score: 737.4434504399774
```

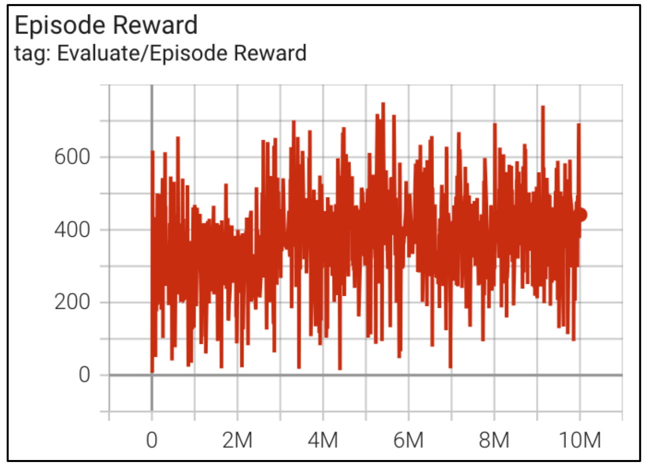
  - Training curve (comparison)
    - TD3 (Orange) and TD3 without target policy smoothing (cyan), with 0.999 smoothing



- Discuss between TD3 and TD3 without target policy smoothing
  - 在我的實驗中，由於此任務的三個 action 的值域不太一樣(-1~1, 0~1, 0~1)，對於第一個 action 的部分，我仿照了原始論文的參數設計，使用了 Normal(mean=0, std=0.2)來加上 target policy smoothing，另外兩者則是使用 Normal(mean=0, std=0.1)。
  - 而在結果中可以觀察到，使用 target policy smoothing 在訓練前期確實可以使模型的訓練上提升不少，然而在訓練後期，沒有加上 target policy smoothing 的模型卻有反超的趨勢，我認為是因為在訓練後期時，critic 對於 Q value 的預測已足夠準確，再加上 noise 可能反而會有負面影響。因此我認為若要將結果提升，也應該將此 noise 加上一個 scheduler，讓其在訓練後期逐漸減少其 std，或許就能夠讓結果更好。

- Screenshot of Tensorboard training curve and compare the impact of delayed update steps and compare the results, and explain (5%).
  - Training curve

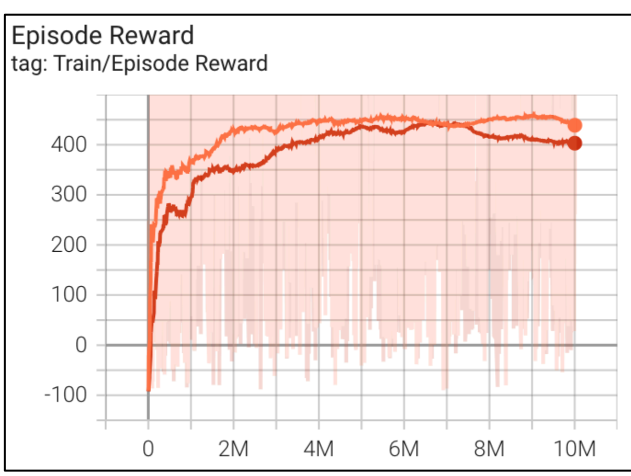


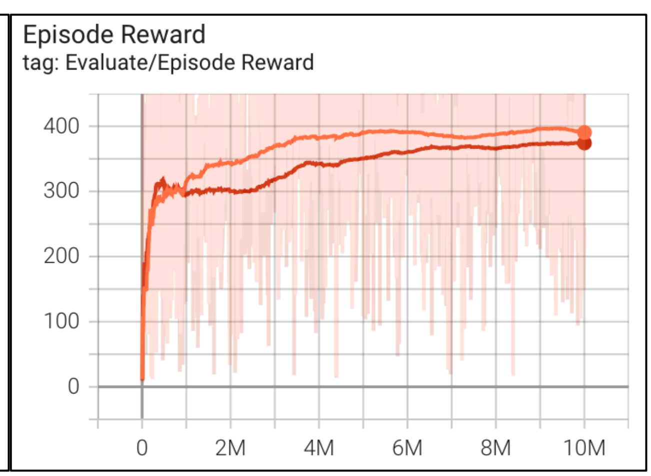


  - Testing results (10 games)

```
Episode: 1      Length: 628     Total reward: 621.62
Episode: 2      Length: 150     Total reward: 133.45
Episode: 3      Length: 868     Total reward: 891.44
Episode: 4      Length: 332     Total reward: 415.98
Episode: 5      Length: 320     Total reward: 341.28
Episode: 6      Length: 655     Total reward: 905.73
Episode: 7      Length: 480     Total reward: 607.07
Episode: 8      Length: 999     Total reward: 861.94
Episode: 9      Length: 267     Total reward: 285.49
Episode: 10     Length: 398     Total reward: 479.84
average score: 554.3831365877292
```

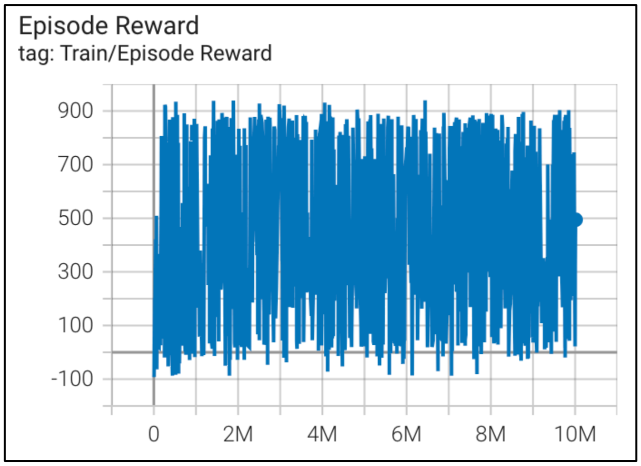
  - Training curve (comparison)
    - TD3 (Orange) and TD3 without delayed updates(red), with 0.999 smoothing

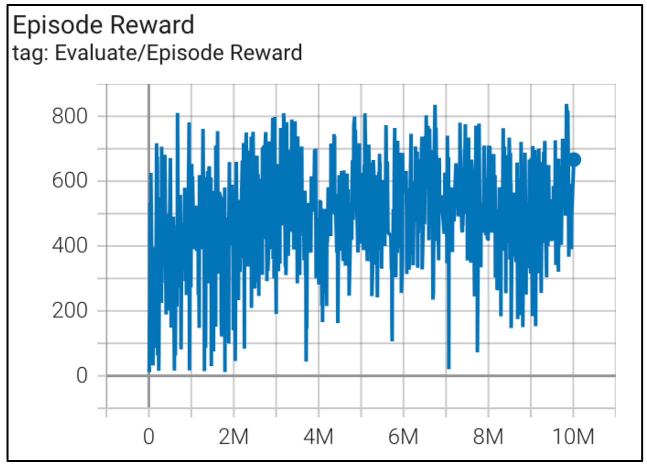




- Discuss between TD3 and TD3 without delayed updates
  - 在 TD3 中，使用 delayed policy updates 的原因我認為與 GAN 中有些類似，在 GAN 中由於 Generator 生成的結果好壞取決於 Discriminator 的輸出，因此需要 Discriminator 訓練的比 Generator 好才能夠引導 Generator。而在此可以將 Actor 看作 Generator，而 Critic 就是 Discriminator。因此就讓 Critic 的更新頻率比 Actor 高，讓 Critic 預測得更準確時再更新 Actor，以此就能讓 Actor 學得更好。
  - 而在實驗中可以看到，不管是在訓練的前期或後期，TD3 的結果在 training 與 evaluation 上相比於 TD3 without delayed updates 皆好上一些，以此也證實了此方法的有效性。

- Screenshot of Tensorboard training curve and compare the effects of adding different levels of action noise (exploration noise) in TD3, and explain (5%).
  - Training curve

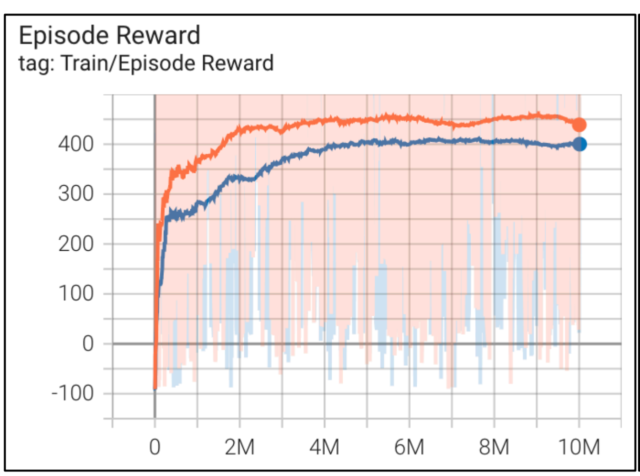


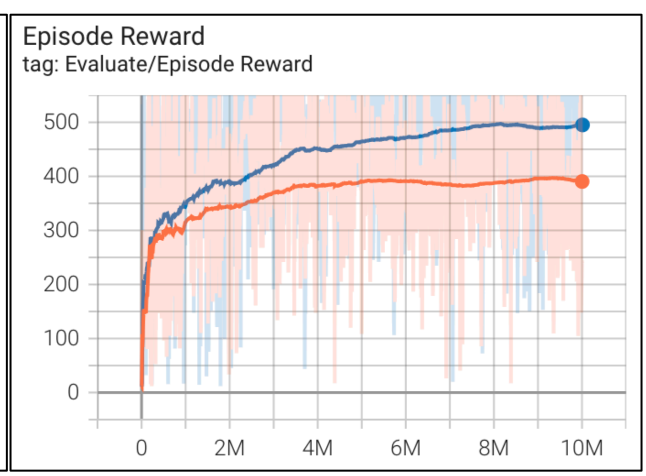


  - Testing results (10 games)

```
Episode: 1      Length: 999     Total reward: 836.24
Episode: 2      Length: 999     Total reward: 874.26
Episode: 3      Length: 892     Total reward: 910.70
Episode: 4      Length: 999     Total reward: 824.66
Episode: 5      Length: 999     Total reward: 785.99
Episode: 6      Length: 999     Total reward: 817.36
Episode: 7      Length: 995     Total reward: 824.54
Episode: 8      Length: 999     Total reward: 852.03
Episode: 9      Length: 999     Total reward: 823.32
Episode: 10     Length: 999     Total reward: 845.21
average score: 839.4308322597651
```

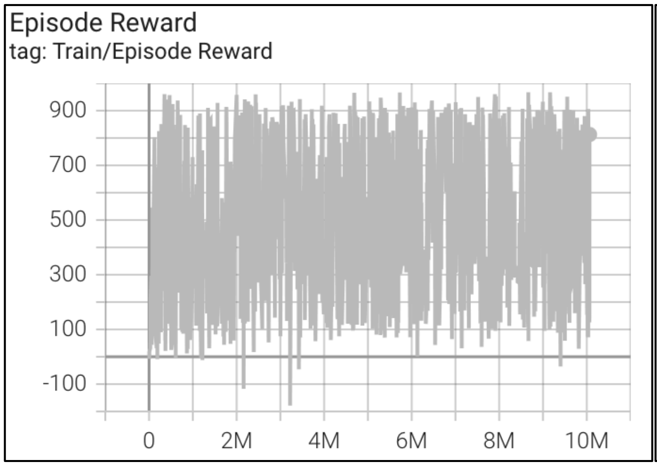
  - Training curve (comparison)
    - TD3 (Orange) and TD3 with adding different levels of action noise(blue), with 0.999 smoothing

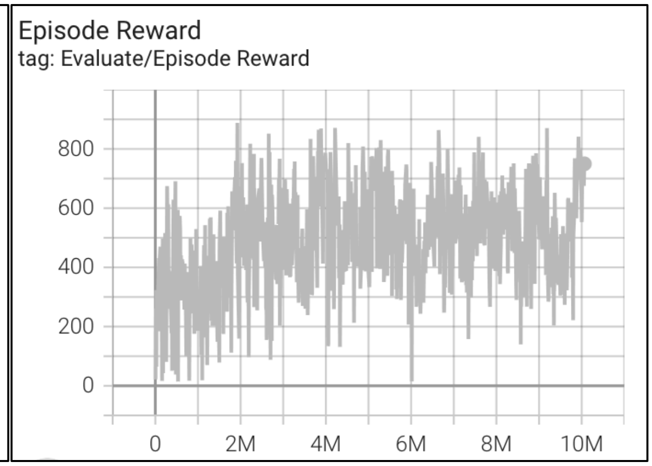




- Discuss between TD3 and TD3 with adding different levels of action noise
  - 在我的實驗中，我比較了在 action 中加上一般的 Gaussian noise 與使用 Ornstein-Uhlenbeck noise (OU noise)的差別，而在結果上，在 training 時，Gaussian noise 的結果較 OU noise 好上不少，而在 evaluate 階段則相反。我認為是因為 OU noise 每次 sample 時，資料會有一定程度的連續性，而 Gaussian noise 每次 sample 的結果皆為獨立的。讓模型在訓練時使用 OU noise 時連續幾個 frame sample 的 noise 差別相比於 Gaussian noise 不要太大，以此讓 agent 的行為間更有相關性(例如連續幾好幾個 frame 間 OU noise 都更傾向於使 agent 向左)，在 training 時可能會因為此導致結果變差，但同時也讓 agent 的 exploration 能力更好，因此在 evaluation 時由於 action 沒有加上 noise，結果才會相比於使用 Gaussian noise 更好。

- Screenshot of Tensorboard training curve and compare your reward function with the original one and explain why your reward function works better (10%).
  - Training curve

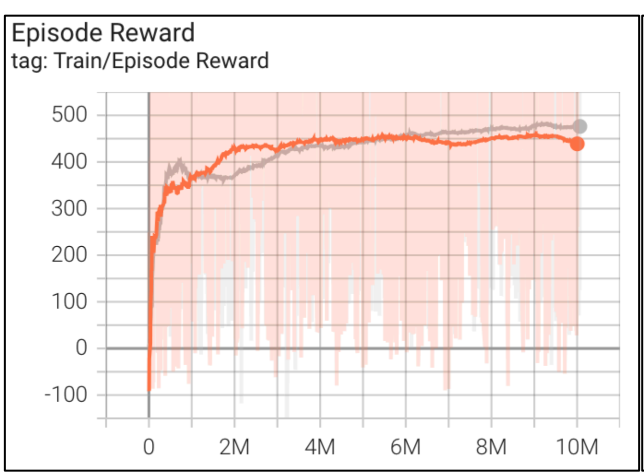


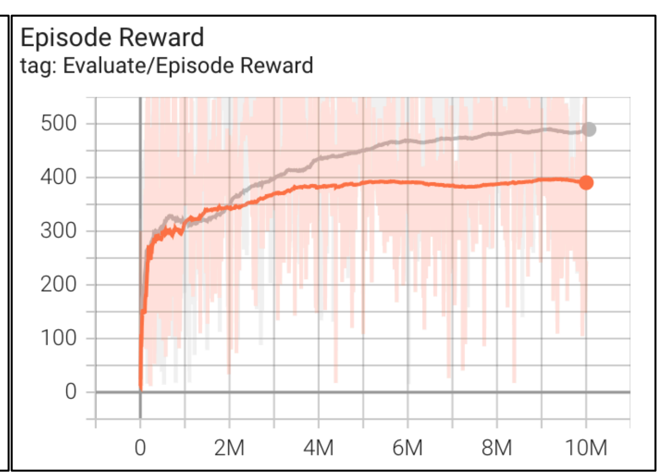


  - Testing results (10 games)

```
Episode: 1      Length: 625     Total reward: 872.18
Episode: 2      Length: 639     Total reward: 849.67
Episode: 3      Length: 634     Total reward: 855.51
Episode: 4      Length: 762     Total reward: 881.78
Episode: 5      Length: 679     Total reward: 844.73
Episode: 6      Length: 614     Total reward: 670.13
Episode: 7      Length: 772     Total reward: 823.90
Episode: 8      Length: 661     Total reward: 853.10
Episode: 9      Length: 718     Total reward: 860.58
Episode: 10     Length: 999     Total reward: 797.76
average score: 830.9341568908216
```

  - Training curve (comparison)
    - TD3 (Orange) and TD3 with with my own reward function (gray), with 0. 999 smoothing





- Discuss between TD3 and TD3 with my own reward function
  - 由於我觀察到若使用原始的 reward function，agent 會有以下行為
    - 轉彎時較為不穩定，有時會直接走草地，導致無法完整獲得所有 track tile 的分數
    - 在直線時有時不太穩定，會胡亂調整方向
  - 因此我在設計上，當 en_len＞ 40 時(遊戲剛開始時會從最大的畫面 zoom in 到只有賽道，約在 en_len = 40 時車的大小才是固定的)，而我希望讓整個 clip 過的畫面 img_size(20 * 24)賽道比例越高越好，因此我先計算了車的大小 car_size 為 20 * 24 – road_pixel – grass_pixel，而賽道佔整個畫面的比例就是 road_pixel / (img_size – car_size)。並將其-0.5 後在乘以 5，最後與 0.05 取 min，最後其範圍會是是在-2.5~0.05 間，主要是希望讓其是有獎勵項與懲罰項，希望模型能夠同時學到增加 road_pixel 的比例與減少 grass_pixel 比例，讓車子在行進時更加穩定。
  - 而從實驗上可以看到，在 evaluation 上，利用上述方式可以讓 agent 獲得更好的 return，因此我認為我的作法是有效的。