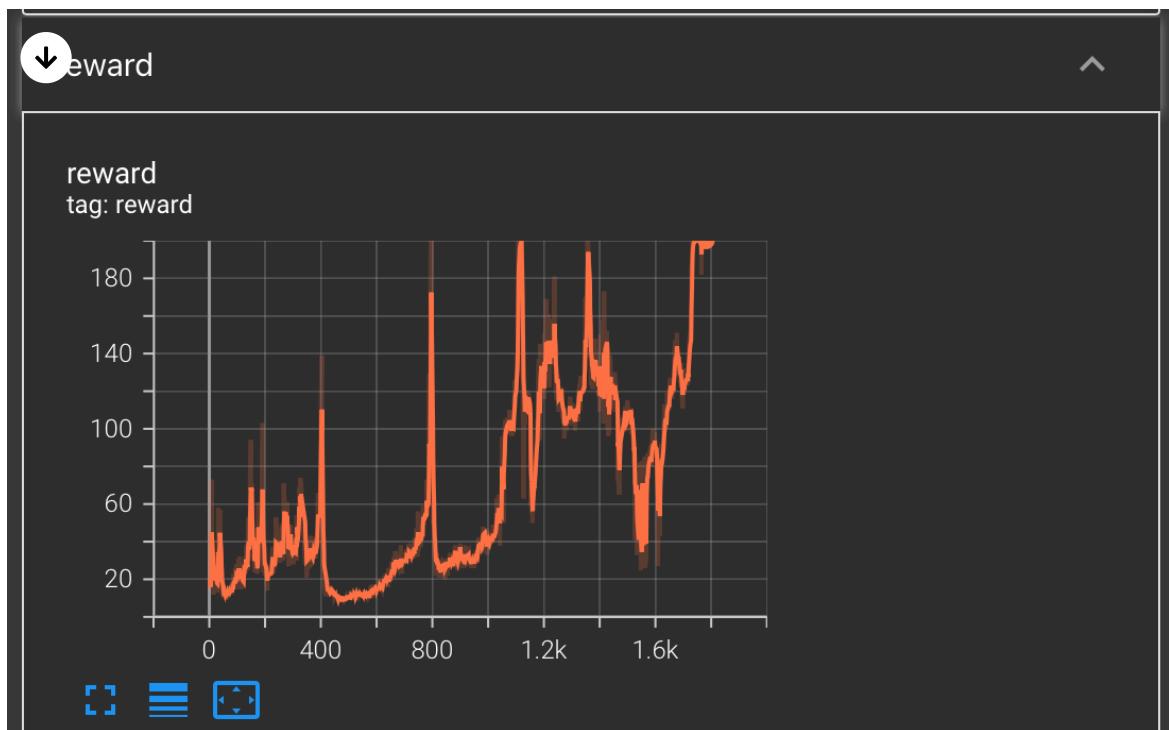


tags: 2024 年 下學期讀書計畫 Reinforcement Learning

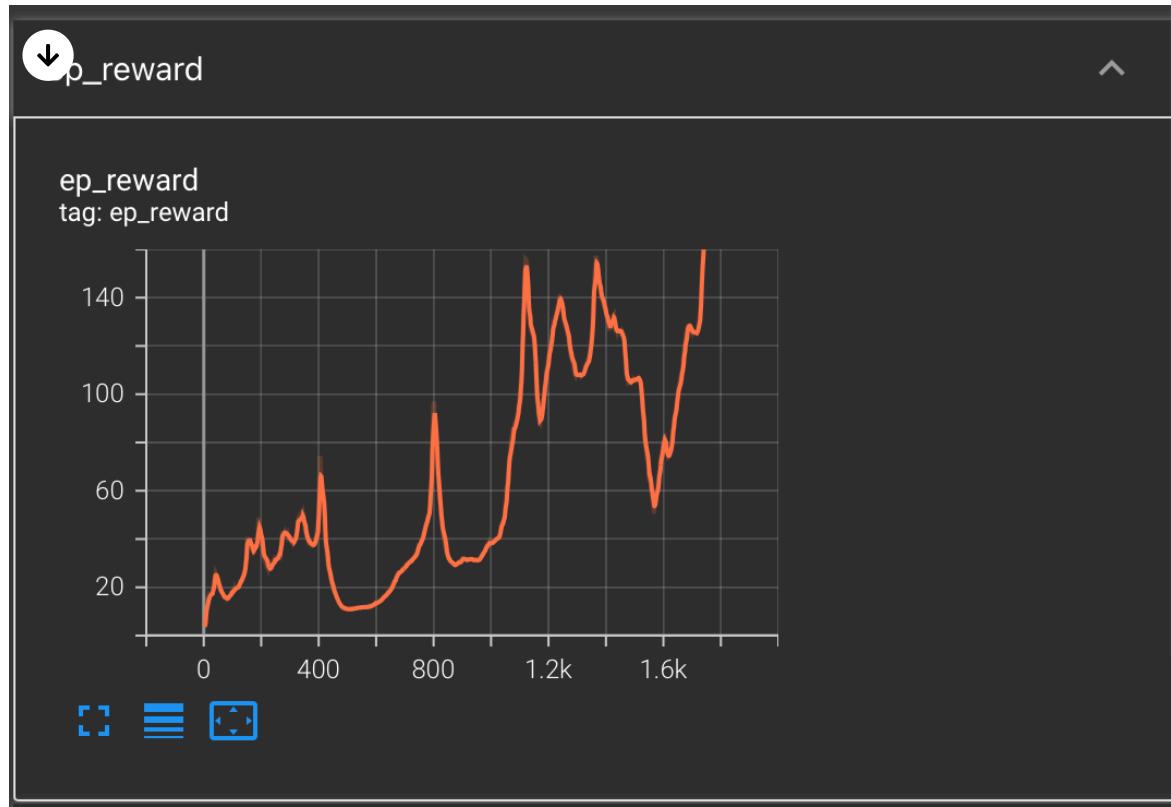
RL Homework 2: Policy Gradient and Model-Free Prediction Implements

a. reinforce

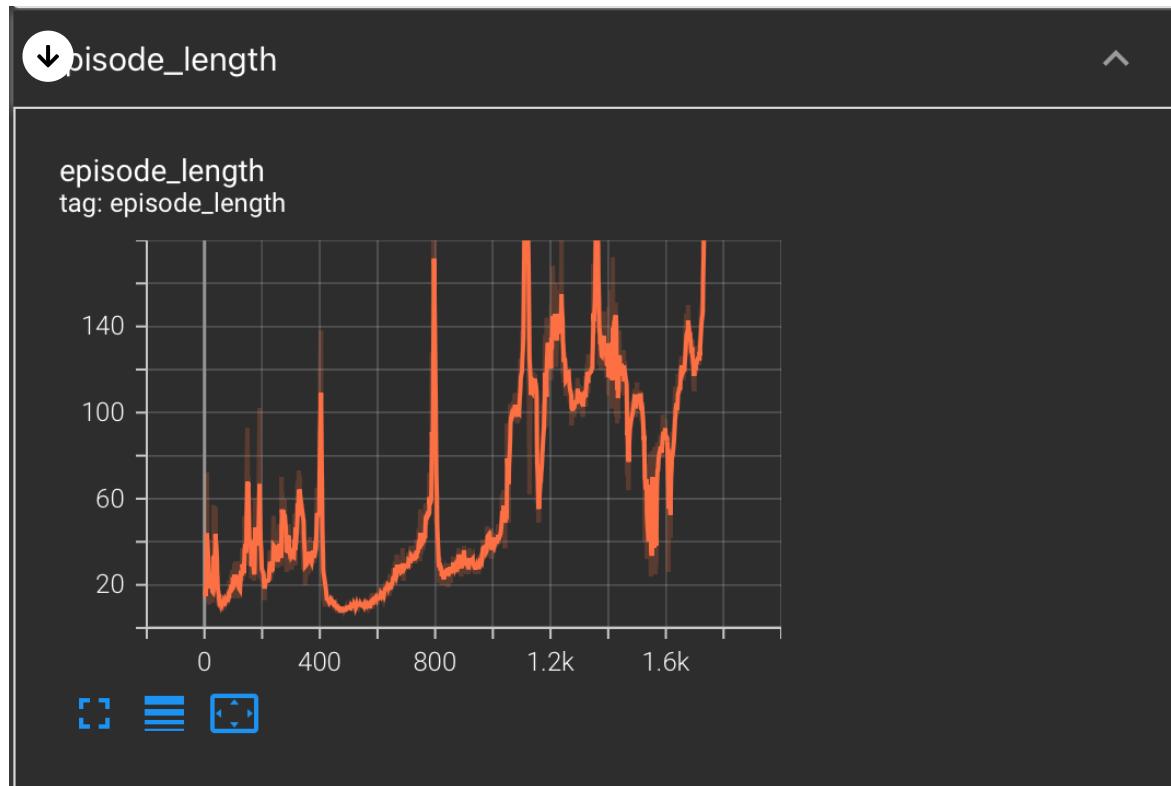
Reward



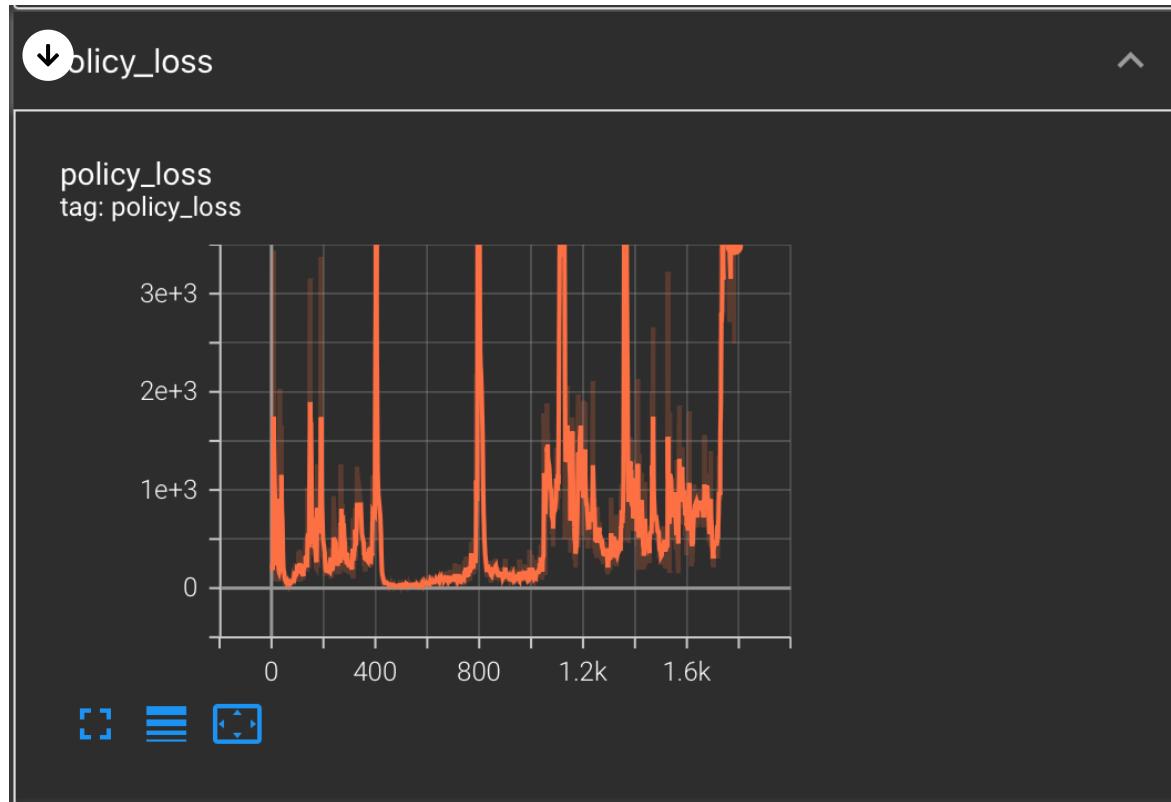
EWMA_Reward



Episode_Length



Policy_Loss



Training

Episode	length	reward	ewma reward
Episode 1200	122	125.0	127.5
Episode 1280	110	111.0	114.7 ↑
↓	113	114.0	108.04299/8545420/
Episode 1300	111	112.0	108.38507274343799
Episode 1320	110	111.0	115.46871298604162
Episode 1340	199	200.0	150.15650219078668
Episode 1360	123	124.0	142.53687328152634
Episode 1380	115	116.0	133.70775063689908
Episode 1400	148	149.0	130.68581763377045
Episode 1420	123	124.0	126.38510579481752
Episode 1440	113	114.0	123.57180326822677
Episode 1460	105	106.0	104.74062851805668
Episode 1480	114	115.0	106.19831233649252
Episode 1500	79	80.0	103.92600956409947
Episode 1520	65	66.0	75.46988866584486
Episode 1540	27	28.0	56.50736797011812
Episode 1560	79	80.0	64.79774758218277
Episode 1580	97	98.0	80.18196910618765
Episode 1600	76	77.0	74.58474108708093
Episode 1620	105	106.0	91.73164984057057
Episode 1640	115	116.0	107.20063017517944
Episode 1660	142	143.0	127.17076174241835
Episode 1680	121	122.0	125.929479784132
Episode 1700	134	135.0	126.74746693761534
Episode 1720	199	200.0	163.1986590602493
Episode 1740	199	200.0	186.80723734734212
Episode 1760	199	200.0	194.58183998630864
Episode 1780	199	200.0	Solved! Running reward is now 195.11011058764353 and the last episode runs to 199 time steps!
Episode 1	Reward: 200.0		
Episode 2	Reward: 200.0		
Episode 3	Reward: 200.0		
Episode 4	Reward: 200.0		
Episode 5	Reward: 200.0		
Episode 6	Reward: 200.0		
Episode 7	Reward: 200.0		
Episode 8	Reward: 200.0		
Episode 9	Reward: 200.0		
Episode 10	Reward: 200.0		

Hyperparameter

hyperparameter_1	value
learning_rate	0.01
hidden_size	128

b. reinforce baseline

我們利用以下兩張圖片的結論 $B(s) \approx V^{\theta_\pi}(s)$ 挑選我們的 baseline function

↓ The variance of REINFORCE PG with baseline can be written as:

$$\begin{aligned}
 & \mathbb{V}[(G_0 - \cancel{B(s_0)}) \frac{\partial}{\partial \theta_i} \log \pi_\theta(a_0 | s_0)] \quad (\text{With baseline}) \\
 &= \sum_s P(s_0 = s) \left(\mathbb{E}[(G_0 - B(s))^2 (\frac{\partial}{\partial \theta_i} \log \pi_\theta(a_0 | s))^2 | s] \right) \\
 &\quad - \left(\sum_s P(s_0 = s) \mathbb{E}[(G_0 - B(s)) \frac{\partial}{\partial \theta_i} \log \pi_\theta(a_0 | s_0) | s] \right)^2 \\
 &= \sum_s P(s_0 = s) \left(\sum_a \pi_\theta(a | s) (\mathbb{E}[(G_0 - B(s))^2 (\frac{\partial}{\partial \theta_i} \log \pi_\theta(a | s))^2 | s, a]) \right) \\
 &\quad - \left(\sum_s P(s_0 = s) \sum_a \pi_\theta(a | s) \mathbb{E}[(G_0 - B(s)) \frac{\partial}{\partial \theta_i} \log \pi_\theta(a | s) | s, a] \right)^2 \\
 &= \sum_s P(s_0 = s) \left(\sum_a \pi_\theta(a | s) \left(\frac{\partial}{\partial \theta_i} \log \pi_\theta(a | s) \right)^2 \mathbb{E}[(G_0 - B(s))^2 | s, a] \right) \\
 &\quad - \left(\sum_s P(s_0 = s) \sum_a \pi_\theta(a | s) \frac{\partial}{\partial \theta_i} \log \pi_\theta(a | s) \mathbb{E}[G_0 | s, a] \right)^2
 \end{aligned}$$

↓ Quantifying Variance Reduction By $B(s)$

$$\begin{aligned}
 & \checkmark \mathbb{V}\left[G_0 \frac{\partial}{\partial \theta_i} \log \pi_\theta(a_0 | s_0)\right] - \checkmark \mathbb{V}\left[(G_0 - B(s_0)) \frac{\partial}{\partial \theta_i} \log \pi_\theta(a_0 | s_0)\right] \\
 &= \sum_s P(s_0 = s) \left(\sum_a \pi_\theta(a | s) \underbrace{\left(\frac{\partial}{\partial \theta_i} \log \pi_\theta(a | s) \right)^2}_{:= c_a} (\mathbb{E}[G_0^2 | s, a] - \mathbb{E}[(G_0 - B(s))^2 | s, a]) \right)
 \end{aligned}$$

► Suppose $\mathbb{E}[G_0 | s, a] \equiv Q^{\pi_\theta}(s, a) \approx V^{\pi_\theta}(s)$, then we may choose $B(s) = V^{\pi_\theta}(s)$

► In practice, $B(s) = V^{\pi_\theta}(s)$ is a popular choice

To Maximize this:

Choose $B(s) = \mathbb{E}[G_0 | s, a]$
 $(Q^{\pi_\theta}(s, a))$

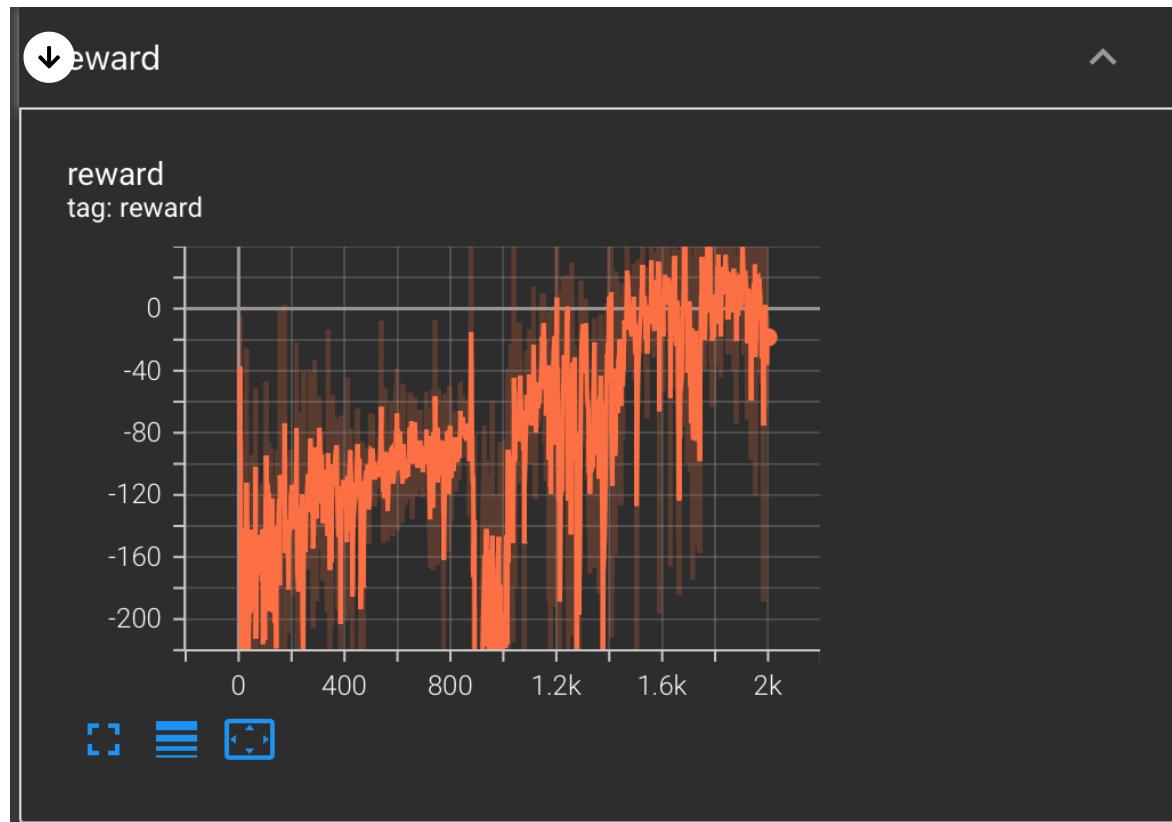
```

1 def calculate_loss(self, gamma=0.999):
2     """
3         Calculate the loss (= policy loss + value loss) to perform backprop
4         TODO:
5             1. Calculate rewards-to-go required by REINFORCE with the help
6                 of the baseline
7             2. Calculate the policy loss using the policy gradient
8             3. Calculate the value loss using either MSE loss or smooth L1
9     """
10    # Initialize the lists and variables
11    R = 0
12    saved_actions = self.saved_actions
13    loss = []
14    value_losses = []
15    returns = []
16
17    ##### YOUR CODE HERE (8-15 lines) #####
18    # Calculate baseline
19    baselines = self.get_v_values(gamma)
20
21    # Calculate advantages and returns
22    returns = []
23    for reward, baseline in zip(reversed(self.rewards), reversed(baselines)):
24        R = reward + gamma * R
25        advantage = R - baseline
26        returns.insert(0, advantage)
27
28    returns = torch.tensor(returns)
29
30    # Get log probabilities of actions
31    log_probs = [action.log_prob for action in saved_actions]
32    action_log_probs = torch.stack(log_probs, dim=0)
33
34    # Calculate policy loss with V-value baseline
35    policy_loss = -(returns * action_log_probs).sum()
36
37    # Calculate value loss using MSE loss
38    values = [action.value for action in saved_actions]
39    value_predictions = torch.stack(values, dim=0).squeeze(1)
40    value_targets = returns.detach()
41    value_loss = F.mse_loss(value_predictions, value_targets)
42
43    # Total loss is the sum of policy loss and value loss
44    loss = policy_loss + value_loss
45
46    ##### END OF YOUR CODE #####
47
48    return loss
49
50 def get_v_values(self, gamma):
51     state_values = []
52     R = 0
53
54     for action in reversed(self.saved_actions):
55         # Assuming action.value contains the value of the next state
56         R = action.value + gamma * R # V^pi_theta(s) = r + gamma * V^pi_t
57         state_values.insert(0, R)
58
59     return state_values

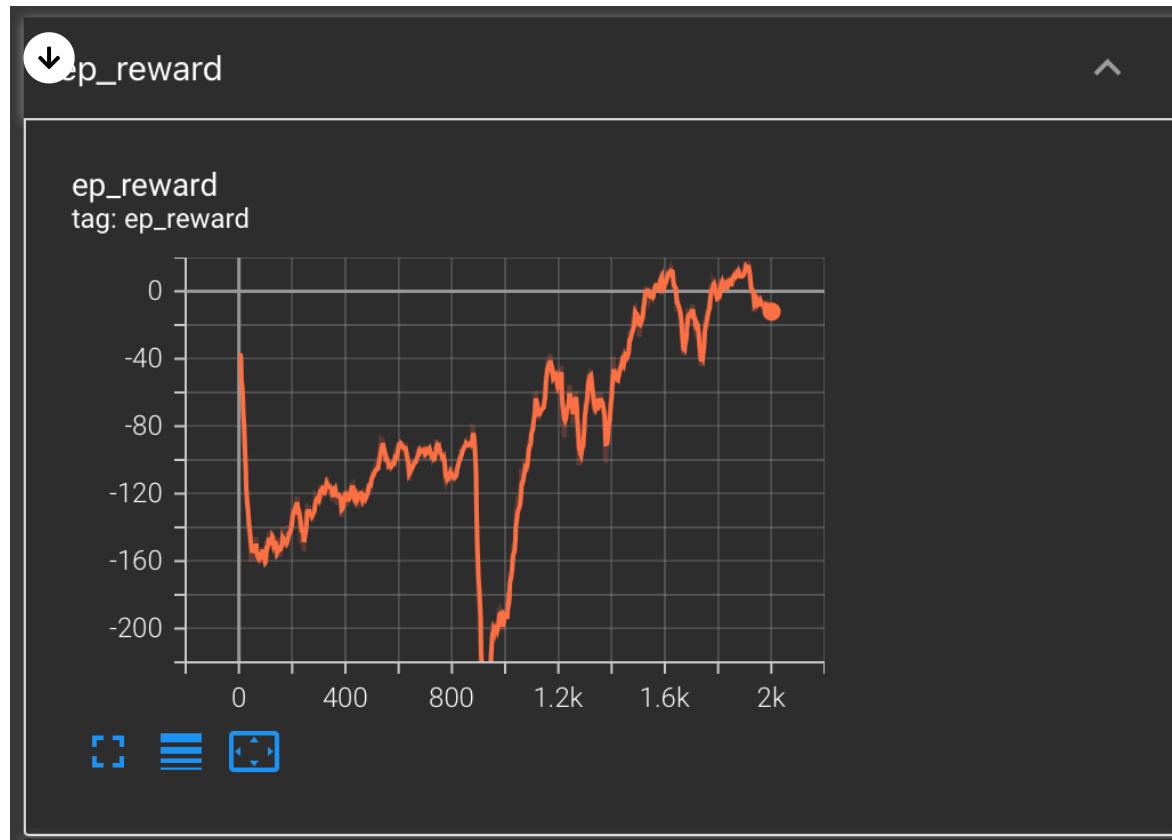
```

- 這邊有加上判斷 episode 次數超過 2000 就卡掉
- 所以不是真的跑到要求

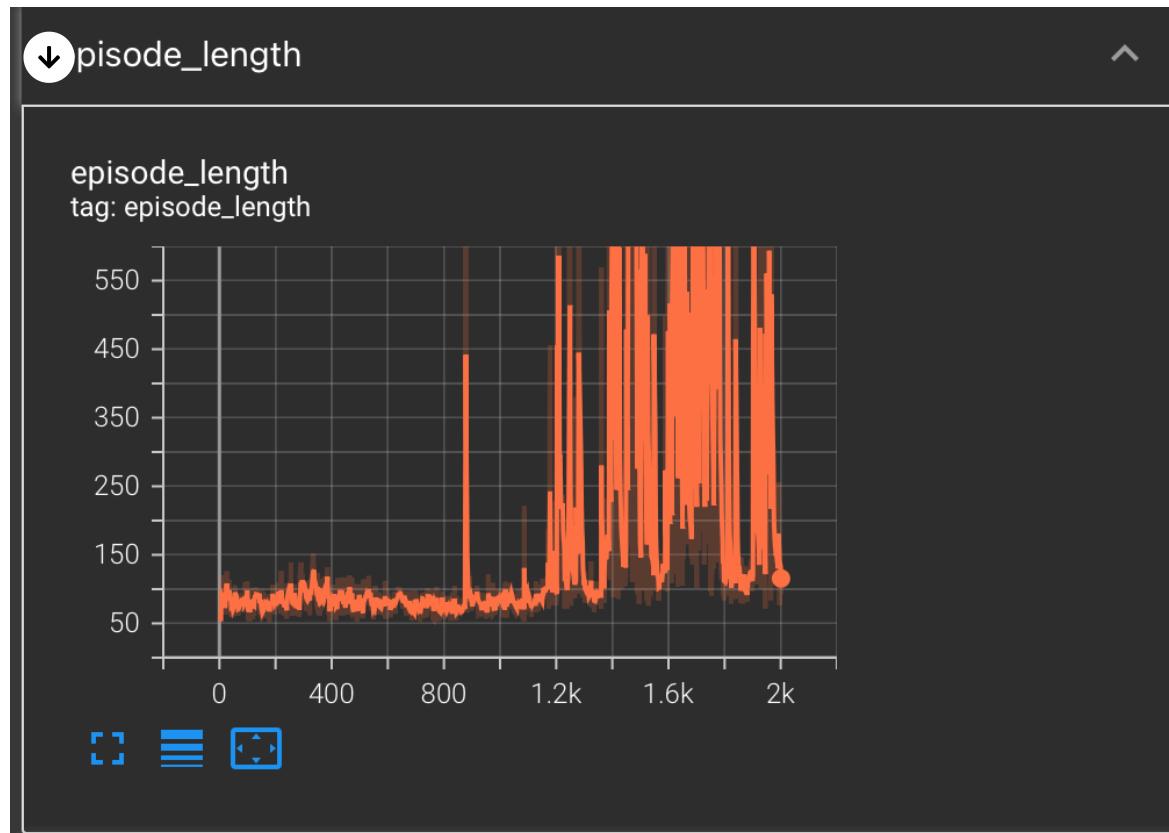
Reward



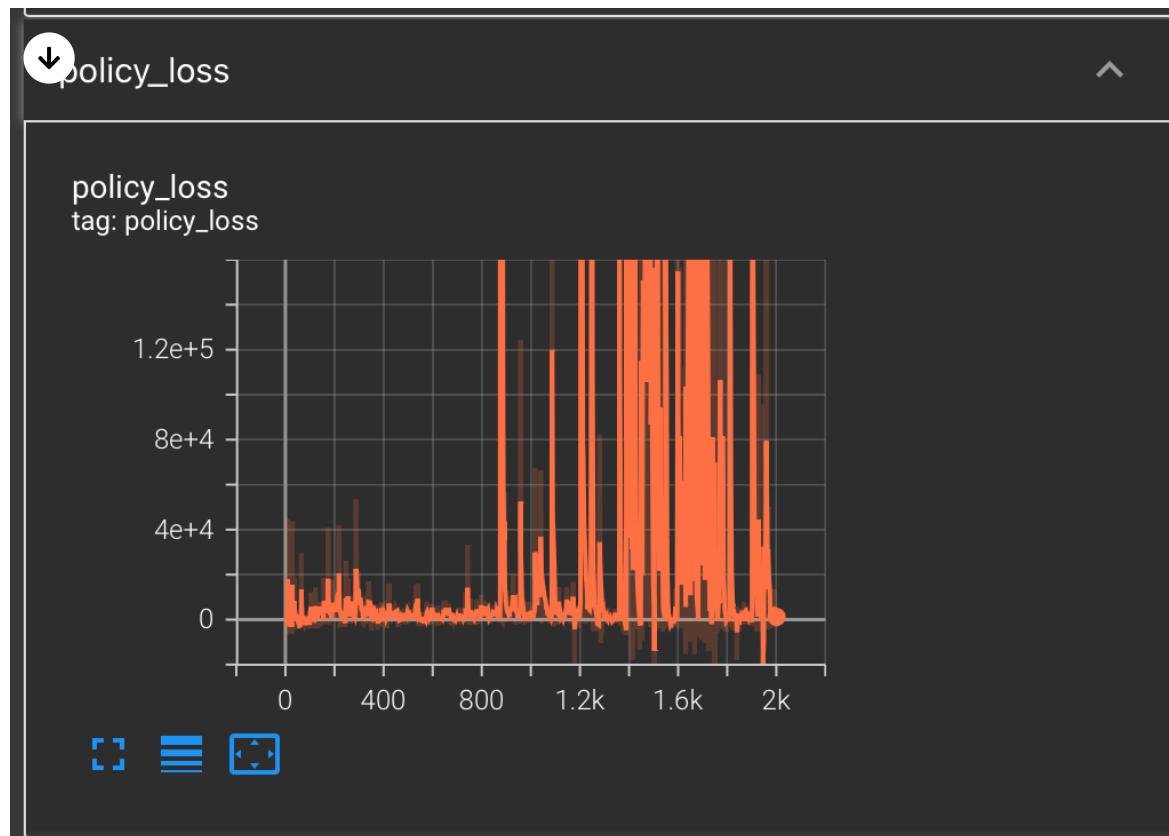
EWMA_Reward



Episode_Length



Policy_Loss



Training

Episode 1500	length: 140	reward: -2.995201004995003	ewma reward: -14.755405240103
↓ Episode 1520	length: 142	reward: -6.838910653244909	ewma reward: -6.7273257922688
Episode 1540	length: 120	reward: -5.380564289640461	ewma reward: -0.5377983278301
Episode 1560	length: 119	reward: 1.461623725587586	ewma reward: 3.14260703776590
Episode 1580	length: 152	reward: 34.81830601055708	ewma reward: 7.03498572026662
Episode 1600	length: 113	reward: 4.046959958339045	ewma reward: 6.19124600066677
Episode 1620	length: 156	reward: -26.276192906992023	ewma reward: 9.12964970160801
Episode 1640	length: 173	reward: -59.309993787762835	ewma reward: -1.2306483767513
Episode 1660	length: 999	reward: 53.78162365949972	ewma reward: -13.087592451205
Episode 1680	length: 176	reward: 15.898343308786352	ewma reward: -25.429994724678
Episode 1700	length: 999	reward: -14.013114693577428	ewma reward: -13.561396711515
Episode 1720	length: 138	reward: -64.84797235301987	ewma reward: -24.709567128956
Episode 1740	length: 999	reward: -34.666218840992336	ewma reward: -35.273647751098
Episode 1760	length: 130	reward: -38.7680188869668	ewma reward: -11.611972595533
Episode 1780	length: 163	reward: 39.69725116444994	ewma reward: 2.93883525578899
Episode 1800	length: 132	reward: 57.43289997864076	ewma reward: -0.7321955244822
Episode 1820	length: 113	reward: 6.134986596934951	ewma reward: 4.50008307479650
Episode 1840	length: 82	reward: 10.224368049741017	ewma reward: 8.73637864352512
Episode 1860	length: 127	reward: 14.50527759694232	ewma reward: 10.1521124442174
Episode 1880	length: 82	reward: -73.95997941492504	ewma reward: 8.56753085637269
Episode 1900	length: 115	reward: -20.350003301917013	ewma reward: 12.9510024211470
Episode 1920	length: 999	reward: 24.081872661693758	ewma reward: 5.99018289389370
Episode 1940	length: 105	reward: -18.18721871699148	ewma reward: -5.5770317245936
Episode 1960	length: 130	reward: -89.16052269383081	ewma reward: -7.0442197591596
Episode 1980	length: 93	reward: 30.461035001325797	ewma reward: -3.6499338153366
Episode 2000	length: 120	reward: -15.051329955840473	ewma reward: -12.365464661017
Solved! Running reward is now -12.565417942675053 and the last episode runs to 92 time steps!			
Episode 1	Reward: -81.29795442491276		
Episode 2	Reward: -43.534461142386206		
Episode 3	Reward: -166.38640362046084		
Episode 4	Reward: -2.9735699551213344		
Episode 5	Reward: -8.62361110330744		
Episode 6	Reward: 131.92094295153407		
Episode 7	Reward: 99.14846399644294		
Episode 8	Reward: -5.15241338220433		
Episode 9	Reward: -48.309868228121815		
Episode 10	Reward: -19.70331522102245		

Hyperparameter

hyperparameter_1	value
learning_rate	0.0015

```
↓ f.discrete = isinstance(env.action_space, gym.spaces.Discrete)
self.observation_dim = env.observation_space.shape[0]
self.action_dim = env.action_space.n if self.discrete else env.action_space.shape[0]
self.hidden_size = 128
self.double()
```

```
↓ if __name__ == '__main__':
    # For reproducibility, fix the random seed
    random_seed = 10
    lr = 0.001
    env = gym.make("LunarLander-v2")
    env.seed(random_seed)
    torch.manual_seed(random_seed)
    train(lr)
    test(f'LunarLander_reinforce_baseline={lr}.pth')
```

c. reinforce gae lambda=0.99

實做概念

Algorithm: REINFORCE With GAE

Recall: (P5) REINFORCE with advantage

$$\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$$

REINFORCE with GAE

Step 1: Initialize θ_0 and step size η

Step 2: Sample a trajectory $\tau \sim P_{\mu}^{\pi_{\theta}}$ and make the update as

$$\theta_{k+1} = \theta_k + \eta \left(\sum_{t=0}^{\infty} \gamma^t \hat{A}_t^{GAE(\gamma, \lambda)} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right)$$

where $\hat{A}_t^{GAE(\gamma, \lambda)}$ is constructed from $V(s)$ learned by TD

(Repeat Step 2 until termination)

我們利用這張投影片的 gradient 公式計算 gradient

```

1 # get probability and value
2 log_probs = [action.log_prob for action in saved_actions]
3 values = [action.value for action in saved_actions]
4
5 # calculate (P4) advantages = E[advantages * log_probs(a|s)] / (1 - gamma)
6 advantages = GAE(gamma, lambda_, None)(self.rewards, values)
7 advantages = advantages.detach()
8
9 action_log_probs = torch.stack(log_probs, dim=0)
10 values = torch.stack(values, dim=0)[:, 0]
11
12 policy_losses = -(advantages * action_log_probs).sum() / (1 - gamma)

```

接著按流程計算 GAE function

Come Discussions on GAE

1. Do we need to wait until the end of a trajectory to construct GAE?

Yes!

2. How to efficiently calculate GAE for different t of the same trajectory?

```
def calculate_advantages(rewards, values, discount_factor, trace_decay, normalize = True):
    advantages = []
    advantage = 0
    next_value = 0
    for r, v in zip(reversed(rewards), reversed(values)):
        td_error = r + next_value * discount_factor - v
        advantage = td_error + advantage * discount_factor * trace_decay
        next_value = v
        advantages.insert(0, advantage)
    advantages = torch.tensor(advantages)
```

Estimate $A^\pi(s_0, a_0), A^\pi(s_1, a_1), \dots, A^\pi(s_t, a_t), \dots$

3. Where does $V(s)$ in GAE come from?

Any model-free prediction methods!

$$\delta_k = r_{k+1} + \gamma V(s_{k+1}) - V(s_k)$$

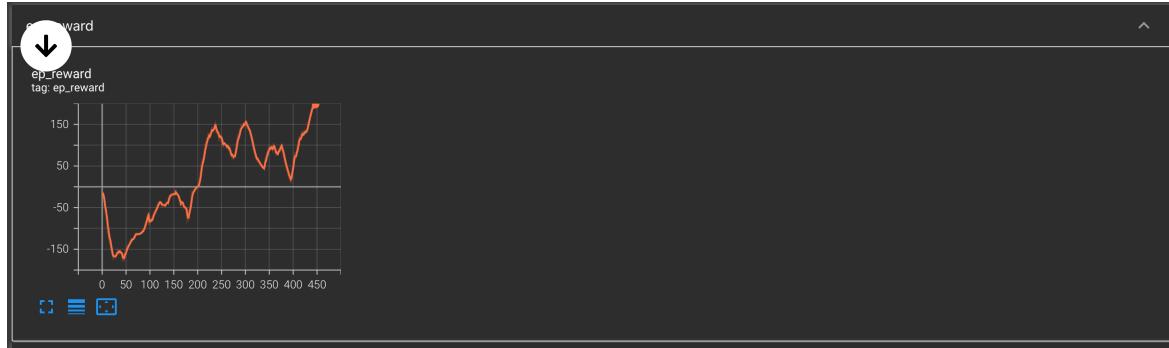
$$\delta_{k-1} = r_k + \gamma V(s_k) - V(s_{k-1})$$

```
1 # refer to slide Lec11 P.6, use defition of GAE
2 advantages = []
3 advantage = 0
4 next_value = 0
5
6 stepCount = 0
7 for r, v in zip(reversed(rewards), reversed(values)):
8     stepCount += 1
9     td_error = r + next_value * self.gamma - v
10    advantage = td_error + advantage * self.gamma * self.lambda_
11    next_value = v
12    advantages.insert(0, advantage)
13    if self.num_steps is not None and stepCount > self.num_steps:
14        break
15 advantages = torch.Tensor(advantages)
16 advantages = (advantages - advantages.mean()) / (advantages.std())
17 return advantages
```

Reward



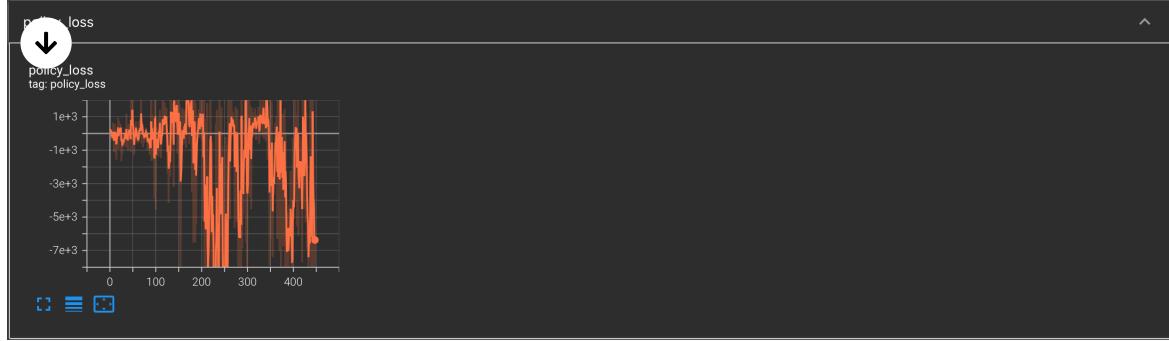
EWMA_Reward



Episode_Length



Policy_Loss



Training

```

Episode 20      length: 119      reward: -306.08468716706574      ewma reward: -158.67930241665272
↓ Episode 40      length: 188      reward: -180.30257182147074      ewma reward: -158.26239144468232
↓ Episode 60      length: 89       reward: -78.80511240438142      ewma reward: -129.34892286326019
Episode 80      length: 79       reward: -69.40044386237153      ewma reward: -110.30496278907867
Episode 100     length: 999      reward: 4.291255190501703      ewma reward: -86.11791006917521
Episode 120      length: 123      reward: 36.892970659997104      ewma reward: -34.95159069287097
Episode 140      length: 999      reward: 117.63315393843953      ewma reward: -22.275185306241937
Episode 160      length: 109      reward: -109.41942622016546      ewma reward: -26.0440866106612
Episode 180      length: 630      reward: -90.99088491636368      ewma reward: -79.66463514301721
Episode 200     length: 150       reward: 33.703125767635925      ewma reward: 1.9953777851598913
Episode 220     length: 220       reward: 16.879081604054846      ewma reward: 111.86629041416319
Episode 240      length: 390       reward: 242.02772113380448      ewma reward: 134.95395684242428
Episode 260     length: 133       reward: 55.224814187120444      ewma reward: 96.396850408255
Episode 280     length: 417       reward: 259.204169971016      ewma reward: 92.12937752124816
Episode 300     length: 283       reward: 226.3311918915353      ewma reward: 156.6171296453797
Episode 320      length: 79       reward: -47.42472689646631      ewma reward: 75.71128335227651
Episode 340      length: 212       reward: 259.9323656759929      ewma reward: 53.459291019332596
Episode 360      length: 175       reward: -10.251174622791126      ewma reward: 98.37663080842466
Episode 380      length: 279       reward: -89.16486454551442      ewma reward: 73.1449249886289
Episode 400      length: 410       reward: 257.9009331691898      ewma reward: 54.9980567810101
Episode 420      length: 166       reward: 36.805004003878395      ewma reward: 129.19434503519935
Episode 440      length: 398       reward: 250.49130871299369      ewma reward: 193.25982079258833
Solved! Running reward is now 201.27898424727456 and the last episode runs to 299 time steps!
Episode 1      Reward: 191.96337246926416
Episode 2      Reward: 208.83916730249626
Episode 3      Reward: 285.0393582553071
Episode 4      Reward: 46.45398832435623
Episode 5      Reward: 185.97919906280947
Episode 6      Reward: 236.04690554735154
Episode 7      Reward: 257.02887779900055
Episode 8      Reward: 231.50008373562733
Episode 9      Reward: 233.25313947264533
Episode 10     Reward: 17.090291902061267

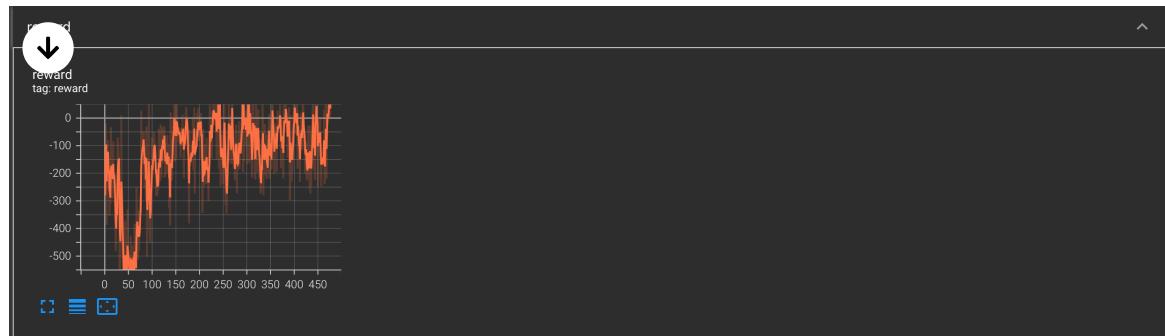
```

hyperparameter_1	value
random_seed	10
learning_rate	0.01
learning_rate_decay	0.9
discount_factor	0.99
trace_decay	0.99
hidden_size	128

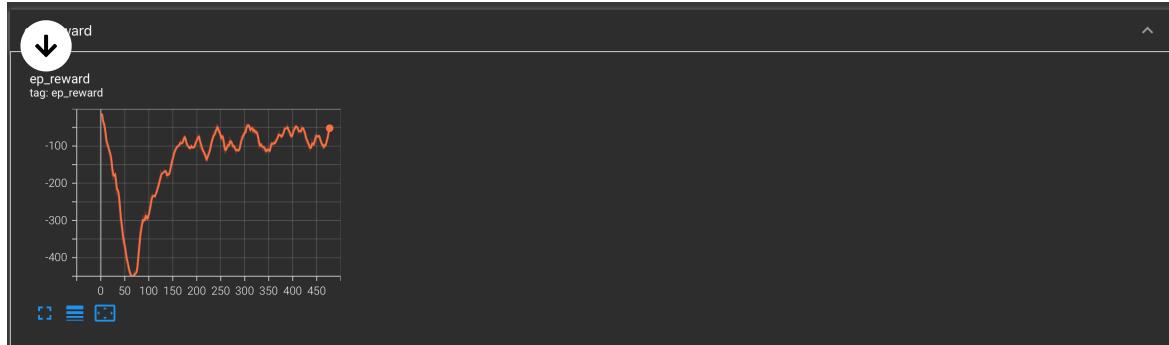
c. reinforce gae lambda=0.8

- 這邊有加上判斷 episode 次數超過 500 就卡掉
- 所以不是真的跑到要求

Reward



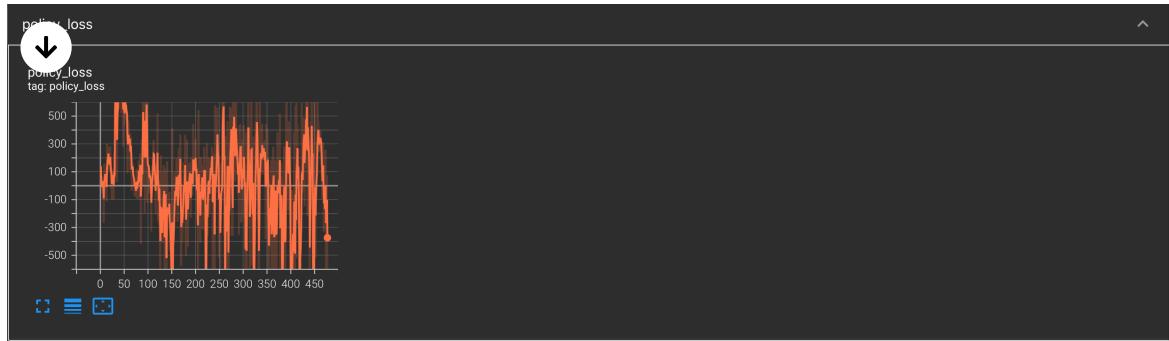
EWMA_Reward



Episode_Length



Policy_Loss



Training

```

Episode 20      length: 138      reward: -196.68395535916      ewma reward: -127.8989481220854
↓ Node 40      length: 86       reward: -657.1144955690355      ewma reward: -283.97530656088145
↓ Episode 60     length: 72       reward: -461.5938068995624      ewma reward: -441.9484943512645
Episode 80      length: 79       reward: -88.63682919181348      ewma reward: -355.68696695737
Episode 100     length: 245      reward: -71.44587811276388      ewma reward: -275.739623168452
Episode 120     length: 217      reward: -215.3588885784885      ewma reward: -204.6927809383527
Episode 140     length: 190      reward: -208.00886419172872      ewma reward: -177.7223399011013
Episode 160     length: 106      reward: -144.75595338116477      ewma reward: -101.27799046382759
Episode 180     length: 230      reward: -135.44186633383555      ewma reward: -99.2623104246206
Episode 200     length: 287      reward: 9.44845789225077      ewma reward: -80.78846716914282
Episode 220     length: 799      reward: -160.69720054000632      ewma reward: -139.9168063739193
Episode 240     length: 589      reward: 144.6672608203757      ewma reward: -48.96311581287912
Episode 260     length: 168      reward: -45.196670576694615      ewma reward: -114.27868072956166
Episode 280     length: 140      reward: -135.76438656280752      ewma reward: -109.70297320244065
Episode 300     length: 168      reward: 28.27833743404048      ewma reward: -58.962672553619726
Episode 320     length: 380      reward: -57.154336258243035      ewma reward: -56.150496674930785
Episode 340     length: 213      reward: -210.4366690754444      ewma reward: -107.38334366327985
Episode 360     length: 526      reward: -94.10642278638817      ewma reward: -93.19439482565035
Episode 380     length: 157      reward: -11.774687476388522      ewma reward: -63.27483784859901
Episode 400     length: 999      reward: -3.687993391008379      ewma reward: -64.56363392883219
Episode 420     length: 273      reward: -90.2734550177167      ewma reward: -49.88012707228572
Episode 440     length: 999      reward: 85.81342541431884      ewma reward: -86.94100605479544
Episode 460     length: 131      reward: -173.09286987965518      ewma reward: -97.71752326295282
Episode 480     length: 166      reward: -137.97600115194558      ewma reward: -50.26533224300745
Episode 500     length: 167      reward: -23.57962281439822      ewma reward: -39.15757480175209
Solved! Running reward is now -40.73127567194943 and the last episode runs to 245 time steps!
Episode 1      Reward: -31.148852063941476
Episode 2      Reward: -14.561764138150991
Episode 3      Reward: -205.53713590814056
Episode 4      Reward: -249.6750200572577
Episode 5      Reward: -133.810794843626
Episode 6      Reward: -195.03977668006928
Episode 7      Reward: -216.64175400254206
Episode 8      Reward: -196.3886072499917
Episode 9      Reward: -50.79953493323336
Episode 10     Reward: -81.4995434350619

```

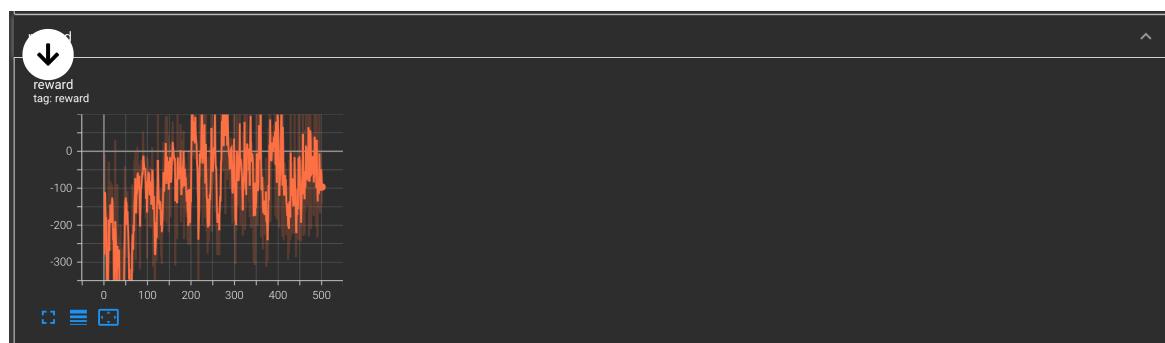
Hyperparameter

hyperparameter_1	value
random_seed	10
learning_rate	0.01
learning_rate_decay	1
discount_factor	0.99
trace_decay	0.8
NN_hidden_size	128

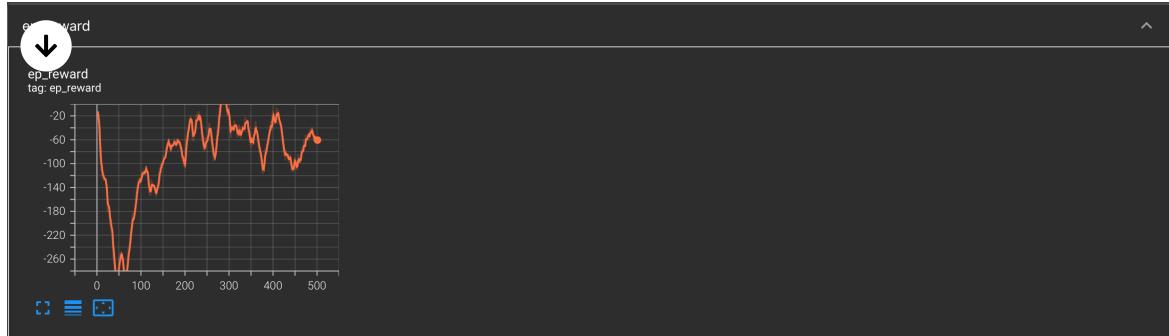
c. reinforce gae lambda=0.98

- 這邊有加上判斷 episode 次數超過 500 就卡掉
- 所以不是真的跑到要求

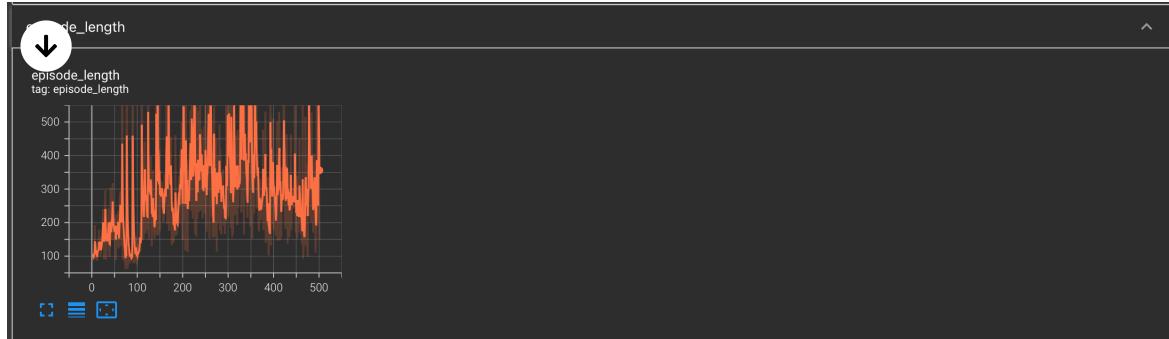
Reward



EWMA_Reward



Episode_Length



Policy_Loss



Training

```

↓
Episode 20      length: 105      reward: -155.1448111304258      ewma reward: -127.58317675422255
Episode 40      length: 133      reward: -533.9967866406203      ewma reward: -278.5537657179088
Episode 60      length: 123      reward: -525.8615380587273      ewma reward: -278.1300997749183
Episode 80      length: 115      reward: -60.40937074179726      ewma reward: -194.0818796316476
Episode 100     length: 95       reward: -31.72703397201505      ewma reward: -127.43847555289132
Episode 120     length: 308      reward: -104.25330233315677      ewma reward: -145.89463024882204
Episode 140     length: 293      reward: -20.496540048273744      ewma reward: -128.81034980416092
Episode 160     length: 341      reward: -19.61271447940777      ewma reward: -64.86602239931372
Episode 180     length: 141      reward: -141.6994060947078      ewma reward: -70.72337786547656
Episode 200     length: 148      reward: -145.76080612577562      ewma reward: -104.60061430106369
Episode 220     length: 785       reward: 221.36785077738267      ewma reward: -48.84978386688815
Episode 240     length: 235      reward: -14.594209880938294      ewma reward: -59.48205376107391
Episode 260     length: 362      reward: -280.61411555928504      ewma reward: -61.533042388909564
Episode 280     length: 347      reward: 249.6335038786512      ewma reward: -1.2077738305343608
Episode 300     length: 511      reward: -239.35530775028226      ewma reward: -20.382717816683808
Episode 320     length: 193       reward: 48.970291771271775      ewma reward: -50.90527933613881
Episode 340     length: 266      reward: 27.455870141322023      ewma reward: -28.272830482934236
Episode 360     length: 326      reward: 300.6639363704692      ewma reward: -29.0321984170468
Episode 380     length: 375      reward: 241.77634282376823      ewma reward: -93.47938949678414
Episode 400     length: 179       reward: 28.15463365386836      ewma reward: -13.359156514736656
Episode 420     length: 255      reward: -163.7926895184731      ewma reward: -53.64577291464728
Episode 440     length: 217       reward: -199.949045450901      ewma reward: -94.4456836313767
Episode 460     length: 163       reward: 10.13609102738225      ewma reward: -91.4064119213387
Episode 480     length: 218       reward: -1.9943937429893595      ewma reward: -50.33914077554006
Episode 500     length: 407       reward: -93.77025095285617      ewma reward: -59.337642399641446
Solved! Running reward is now -63.446761668640406 and the last episode runs to 150 time steps!
Episode 1      Reward: -273.7847319281313
Episode 2      Reward: 193.55078191255785
Episode 3      Reward: -18.81383811362494
Episode 4      Reward: 17.878630862263194
Episode 5      Reward: 78.74008079248344
Episode 6      Reward: 245.43378993596656
Episode 7      Reward: 215.97091089248696
Episode 8      Reward: 145.25511775083243
Episode 9      Reward: 269.1973945634143
Episode 10     Reward: -156.66545925565907

```

Hyperparameter

hyperparameter_1	value
random_seed	10
learning_rate	0.01
learning_rate_decay	0.9
discount_factor	0.9999
trace_decay	0.98
NN_hidden_size	128

summarize

- 不同的 λ 可能會影響收斂速度
- 甚至影響收斂結果
- Reward, EWMA_Reward, Policy_Loss, Episode_Length, Policy_Loss 也會不同