Reinforcement Learning
HW2 Report

309505018郭俊廷

REINFORCE with value function as baseline:

1.Network architecture & Hyperparameters:

(1) Initial:

actor跟value 都是兩個Linear layers, 並共用第一層 layer
actor第一層是(observation_dim,hidden_size),第二層是(hidden_size, action_dim)
value第一層是(observation_dim,64), 第二層是(64, 1)
hidden_size是128

(2) Forward pass:

actor第一層出來取ReLU,第二層出來取Softmax, value第一層出來取ReLU, value第二層直接output, 不加activate function

(3) Select action:

把state轉成tensor後丟到forward pass, 會得到一個機率, 再丟到Categorical裡面, 然後做sample就可以得到action

(4) Train:

把state丟到select_action裡會得到一個action, 再把action丟到env.step(action)裡會得到next_state,reward,done, 每個step的reward都是1(CartPole Env的設定), 設一個變數ep_reward把每個step的reward加起來,然後計算ewma_reward

ewma_reward = 0.05 x ep_reward + 0.95 x ewma_reward

(5) Calculate Loss: $\theta_{t+1} = \theta_t + \alpha(G_t - b(S_t)) \bigtriangledown log_{\pi}(A_t|S_t, \theta)$

policy loss :

定義 advantage = G - state_value, policy loss=(-1)*log_prob*advantage

value loss:

G和state value取 mean_square_error, G減state value後平方加總

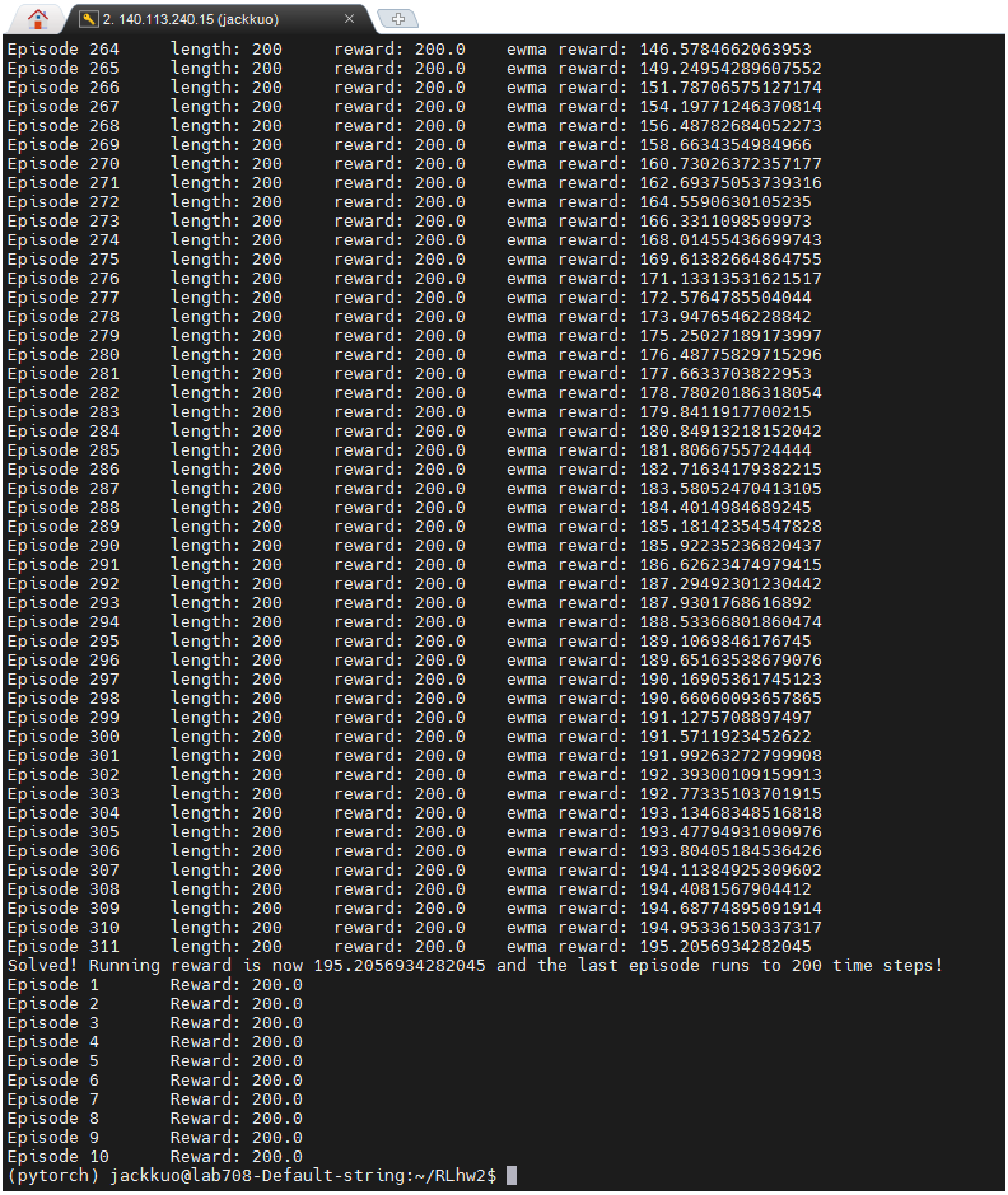
Loss = policy loss + value loss

(6) Parameters:

value的forward layer我試過64跟128, 看起來結果差不多, learning rate我試過0.01, 0.005, 0.001三種, 對結果差異不大, Loss我試過 MSE loss 和smooth L1 loss,對結果也差異不大

2.Result:

training大概在episode 311時 ewma_reward會超過 195, testing的時候每個episode都有200的reward



```
Episode 264     length: 200     reward: 200.0    ewma reward: 146.5784662063953
Episode 265     length: 200     reward: 200.0    ewma reward: 149.24954289607552
Episode 266     length: 200     reward: 200.0    ewma reward: 151.78706575127174
Episode 267     length: 200     reward: 200.0    ewma reward: 154.19771246370814
Episode 268     length: 200     reward: 200.0    ewma reward: 156.48782684052273
Episode 269     length: 200     reward: 200.0    ewma reward: 158.6634354984966
Episode 270     length: 200     reward: 200.0    ewma reward: 160.73026372357177
Episode 271     length: 200     reward: 200.0    ewma reward: 162.69375053739316
Episode 272     length: 200     reward: 200.0    ewma reward: 164.5590630105235
Episode 273     length: 200     reward: 200.0    ewma reward: 166.3311098599973
Episode 274     length: 200     reward: 200.0    ewma reward: 168.01455436699743
Episode 275     length: 200     reward: 200.0    ewma reward: 169.61382664864755
Episode 276     length: 200     reward: 200.0    ewma reward: 171.13313531621517
Episode 277     length: 200     reward: 200.0    ewma reward: 172.5764785504044
Episode 278     length: 200     reward: 200.0    ewma reward: 173.9476546228842
Episode 279     length: 200     reward: 200.0    ewma reward: 175.25027189173997
Episode 280     length: 200     reward: 200.0    ewma reward: 176.48775829715296
Episode 281     length: 200     reward: 200.0    ewma reward: 177.6633703822953
Episode 282     length: 200     reward: 200.0    ewma reward: 178.78020186318054
Episode 283     length: 200     reward: 200.0    ewma reward: 179.8411917700215
Episode 284     length: 200     reward: 200.0    ewma reward: 180.84913218152042
Episode 285     length: 200     reward: 200.0    ewma reward: 181.8066755724444
Episode 286     length: 200     reward: 200.0    ewma reward: 182.71634179382215
Episode 287     length: 200     reward: 200.0    ewma reward: 183.58052470413105
Episode 288     length: 200     reward: 200.0    ewma reward: 184.4014984689245
Episode 289     length: 200     reward: 200.0    ewma reward: 185.18142354547828
Episode 290     length: 200     reward: 200.0    ewma reward: 185.92235236820437
Episode 291     length: 200     reward: 200.0    ewma reward: 186.62623474979415
Episode 292     length: 200     reward: 200.0    ewma reward: 187.29492301230442
Episode 293     length: 200     reward: 200.0    ewma reward: 187.9301768616892
Episode 294     length: 200     reward: 200.0    ewma reward: 188.53366801860474
Episode 295     length: 200     reward: 200.0    ewma reward: 189.1069846176745
Episode 296     length: 200     reward: 200.0    ewma reward: 189.65163538679076
Episode 297     length: 200     reward: 200.0    ewma reward: 190.16905361745123
Episode 298     length: 200     reward: 200.0    ewma reward: 190.66060093657865
Episode 299     length: 200     reward: 200.0    ewma reward: 191.1275708897497
Episode 300     length: 200     reward: 200.0    ewma reward: 191.5711923452622
Episode 301     length: 200     reward: 200.0    ewma reward: 191.99263272799908
Episode 302     length: 200     reward: 200.0    ewma reward: 192.39300109159913
Episode 303     length: 200     reward: 200.0    ewma reward: 192.77335103701915
Episode 304     length: 200     reward: 200.0    ewma reward: 193.13468348516818
Episode 305     length: 200     reward: 200.0    ewma reward: 193.47794931090976
Episode 306     length: 200     reward: 200.0    ewma reward: 193.80405184536426
Episode 307     length: 200     reward: 200.0    ewma reward: 194.11384925309602
Episode 308     length: 200     reward: 200.0    ewma reward: 194.4081567904412
Episode 309     length: 200     reward: 200.0    ewma reward: 194.68774895091914
Episode 310     length: 200     reward: 200.0    ewma reward: 194.95336150337317
Episode 311     length: 200     reward: 200.0    ewma reward: 195.2056934282045
Solved! Running reward is now 195.2056934282045 and the last episode runs to 200 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
(pytorch) jackkuo@lab708-Default-string:~/RLhw2$ 
```

Advantage Actor-Critic algorithm (A2C):

1.Network architecture & Hyperparameters:

(1) Initial:

Actor-Critic架構有三層layers, affine是(8,128), action_layer是(128,4), value_layer是(128,1), Actor和Critic 共用第一層layer

(2) Forward pass:

actor第一層出來取ReLU,第二層出來取Softmax, value第一層出來取ReLU, value第二層直接output, 不加activate function

(3) Select action:

把state轉成tensor後丟到forward pass, 會得到一個機率, 再丟到Categorical裡面, 然後做sample就可以得到action

(4) Train:

把state丟到select_action裡會得到一個action, 再把action丟到env.step(action)裡會得到next_state,reward,done, 設一個變數ep_reward把每個step的reward加起來,然後計算ewma_reward

ewma_reward = 0.05 x ep_reward + 0.95 x ewma_reward

(5) Calculate Loss : $(r_t + \gamma q_{t+1} - q_t)^2$

action loss :

定義 advantage = reward + gamma* nextstate_value - state_value

其中reward + gamma* nextstate_value 稱為TD target

action loss = (-1)*logprob * advantage

value loss:

reward和state value取 smooth_l1_loss,

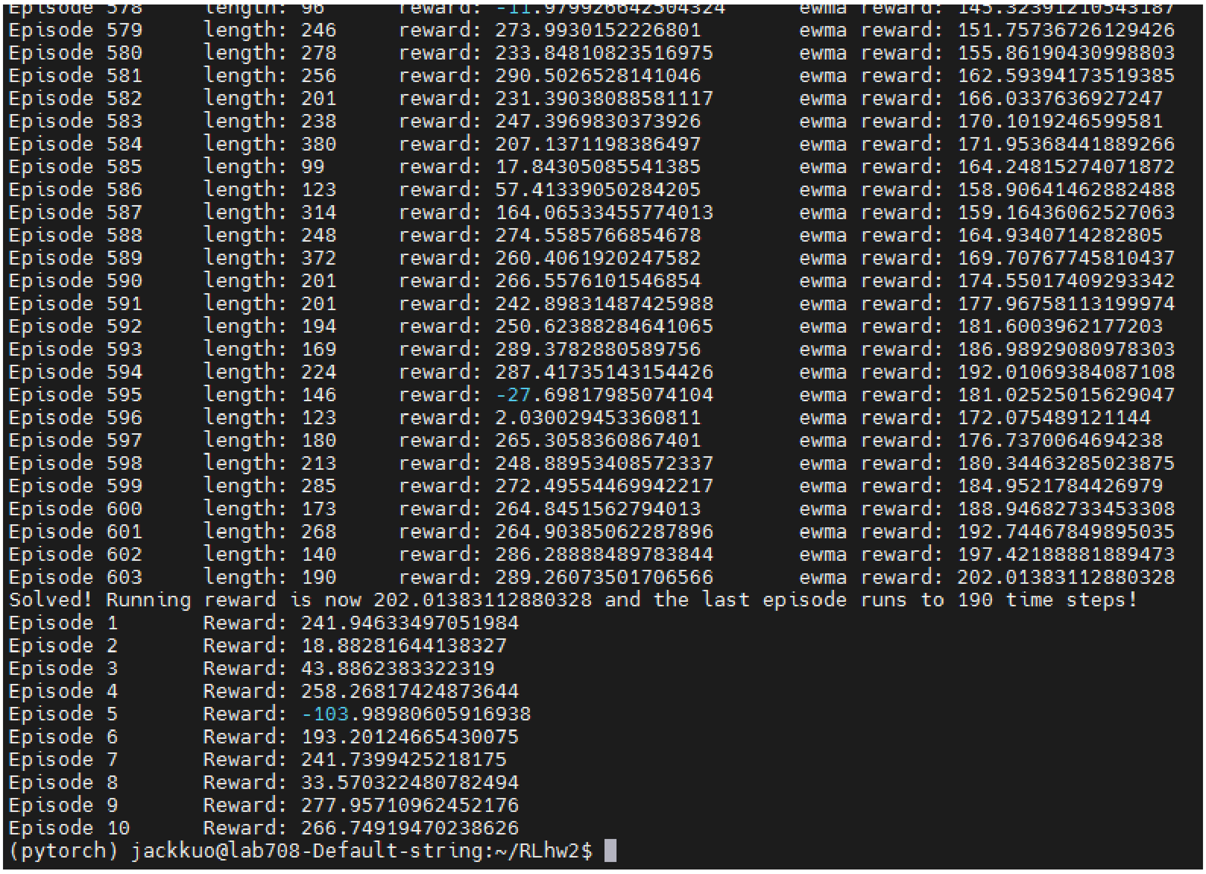
Loss = action loss + value loss

(6) Parameters:

Loss我試過 MSE loss 和smooth L1 loss,對結果差異不大, gamma我試過0.9和0.99,結果也差異不大

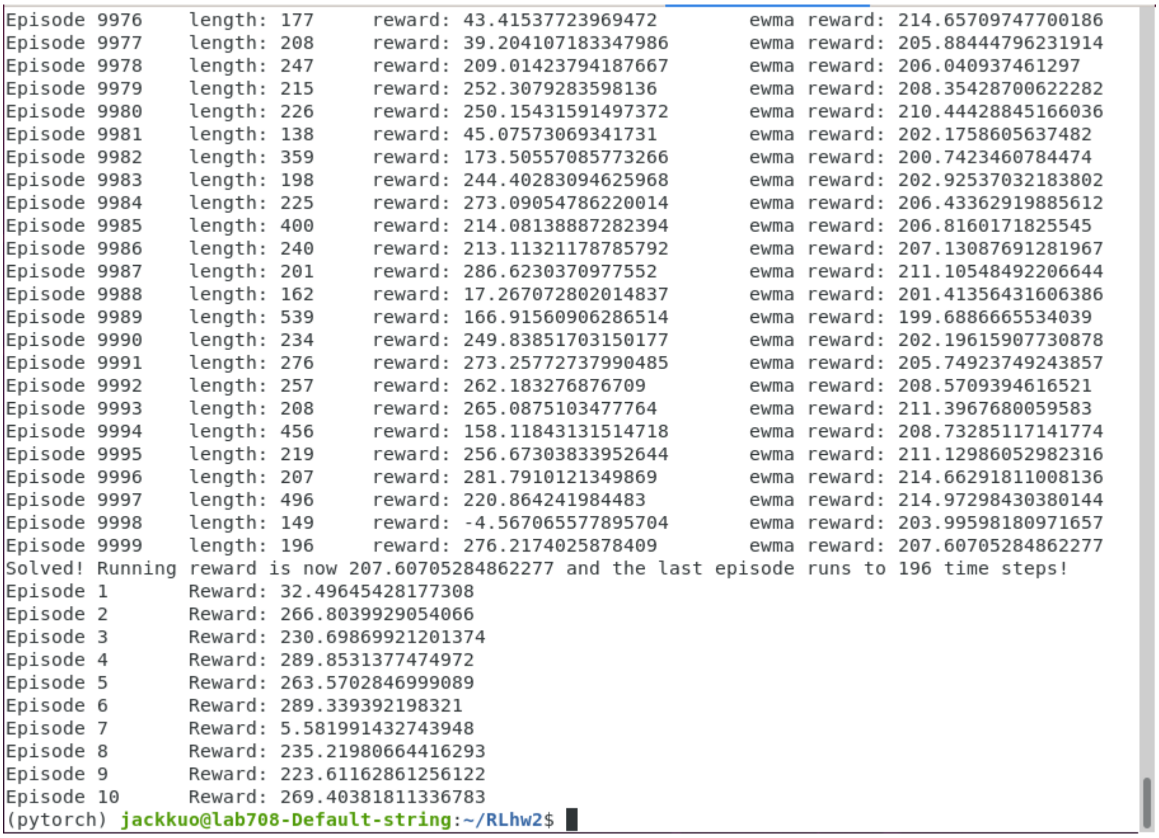
2.Result:

training大概在episode 603時 ewma_reward會超過 200, testing的時候差距比較大,我覺得可能沒有train得很好, 所以testing從-103到277都有, 但testing整體平均reward都有200以上,還可以



```
Episode 578     length: 98      reward: -11.979926642504324       ewma reward: 145.32391210543187
Episode 579     length: 246     reward: 273.9930152226801         ewma reward: 151.75736726129426
Episode 580     length: 278     reward: 233.84810823516975        ewma reward: 155.86190430998803
Episode 581     length: 256     reward: 290.5026528141046         ewma reward: 162.59394173519385
Episode 582     length: 201     reward: 231.39038088581117        ewma reward: 166.0337636927247
Episode 583     length: 238     reward: 247.3969830373926         ewma reward: 170.1019246599581
Episode 584     length: 380     reward: 207.1371198386497         ewma reward: 171.95368441889266
Episode 585     length: 99      reward: 17.84305085541385         ewma reward: 164.24815274071872
Episode 586     length: 123     reward: 57.41339050284205         ewma reward: 158.90641462882488
Episode 587     length: 314     reward: 164.06533455774013        ewma reward: 159.16436062527063
Episode 588     length: 248     reward: 274.5585766854678         ewma reward: 164.9340714282805
Episode 589     length: 372     reward: 260.4061920247582         ewma reward: 169.70767745810437
Episode 590     length: 201     reward: 266.5576101546854         ewma reward: 174.55017409293342
Episode 591     length: 201     reward: 242.89831487425988        ewma reward: 177.96758113199974
Episode 592     length: 194     reward: 250.62388284641065        ewma reward: 181.6003962177203
Episode 593     length: 169     reward: 289.3782880589756         ewma reward: 186.98929080978303
Episode 594     length: 224     reward: 287.41735143154426        ewma reward: 192.01069384087108
Episode 595     length: 146     reward: -27.69817985074104        ewma reward: 181.02525015629047
Episode 596     length: 123     reward: 2.030029453360811         ewma reward: 172.075489121144
Episode 597     length: 180     reward: 265.3058360867401         ewma reward: 176.7370064694238
Episode 598     length: 213     reward: 248.88953408572337        ewma reward: 180.34463285023875
Episode 599     length: 285     reward: 272.49554469942217        ewma reward: 184.9521784426979
Episode 600     length: 173     reward: 264.8451562794013         ewma reward: 188.94682733453308
Episode 601     length: 268     reward: 264.90385062287896        ewma reward: 192.74467849895035
Episode 602     length: 140     reward: 286.28888489783844        ewma reward: 197.42188881889473
Episode 603     length: 190     reward: 289.26073501706566        ewma reward: 202.01383112880328
Solved! Running reward is now 202.01383112880328 and the last episode runs to 190 time steps!
Episode 1       Reward: 241.94633497051984
Episode 2       Reward: 18.88281644138327
Episode 3       Reward: 43.8862383322319
Episode 4       Reward: 258.26817424873644
Episode 5       Reward: -103.98980605916938
Episode 6       Reward: 193.20124665430075
Episode 7       Reward: 241.7399425218175
Episode 8       Reward: 33.570322480782494
Episode 9       Reward: 277.95710962452176
Episode 10      Reward: 266.74919470238626
(pytorch) jackkuo@lab708-Default-string:~/RLhw2$
```

但是這樣testing不夠理想, 所以我train一個episode 9999的版本, 看起來比之前好多了



```
Episode 9976    length: 177     reward: 43.41537723969472         ewma reward: 214.65709747700186
Episode 9977    length: 208     reward: 39.204107183347986        ewma reward: 205.88444796231914
Episode 9978    length: 247     reward: 209.01423794187667        ewma reward: 206.040937461297
Episode 9979    length: 215     reward: 252.3079283598136         ewma reward: 208.35428700622282
Episode 9980    length: 226     reward: 250.15431591497372        ewma reward: 210.44428845166036
Episode 9981    length: 138     reward: 45.07573069341731         ewma reward: 202.1758605637482
Episode 9982    length: 359     reward: 173.50557085773266        ewma reward: 200.7423460784474
Episode 9983    length: 198     reward: 244.40283094625968        ewma reward: 202.92537032183802
Episode 9984    length: 225     reward: 273.09054786220014        ewma reward: 206.43362919885612
Episode 9985    length: 400     reward: 214.08138887282394        ewma reward: 206.8160171825545
Episode 9986    length: 240     reward: 213.11321178785792        ewma reward: 207.13087691281967
Episode 9987    length: 201     reward: 286.6230370977552         ewma reward: 211.10548492206644
Episode 9988    length: 162     reward: 17.267072802014837        ewma reward: 201.41356431606386
Episode 9989    length: 539     reward: 166.91560906286514        ewma reward: 199.6886665534039
Episode 9990    length: 234     reward: 249.83851703150177        ewma reward: 202.19615907730878
Episode 9991    length: 276     reward: 273.25772737990485        ewma reward: 205.74923749243857
Episode 9992    length: 257     reward: 262.183276876709          ewma reward: 208.5709394616521
Episode 9993    length: 208     reward: 265.0875103477764         ewma reward: 211.3967680059583
Episode 9994    length: 456     reward: 158.11843131514718        ewma reward: 208.73285117141774
Episode 9995    length: 219     reward: 256.67303833952644        ewma reward: 211.12986052982316
Episode 9996    length: 207     reward: 281.7910121349869         ewma reward: 214.66291811008136
Episode 9997    length: 496     reward: 220.864241984483          ewma reward: 214.97298430380144
Episode 9998    length: 149     reward: -4.567065577895704        ewma reward: 203.99598180971657
Episode 9999    length: 196     reward: 276.2174025878409         ewma reward: 207.60705284862277
Solved! Running reward is now 207.60705284862277 and the last episode runs to 196 time steps!
Episode 1       Reward: 32.49645428177308
Episode 2       Reward: 266.8039929054066
Episode 3       Reward: 230.69869921201374
Episode 4       Reward: 289.8531377474972
Episode 5       Reward: 263.5702846999089
Episode 6       Reward: 289.339392198321
Episode 7       Reward: 5.581991432743948
Episode 8       Reward: 235.21980664416293
Episode 9       Reward: 223.61162861256122
Episode 10      Reward: 269.40381811336783
(pytorch) jackkuo@lab708-Default-string:~/RLhw2$
```