工作总结报告

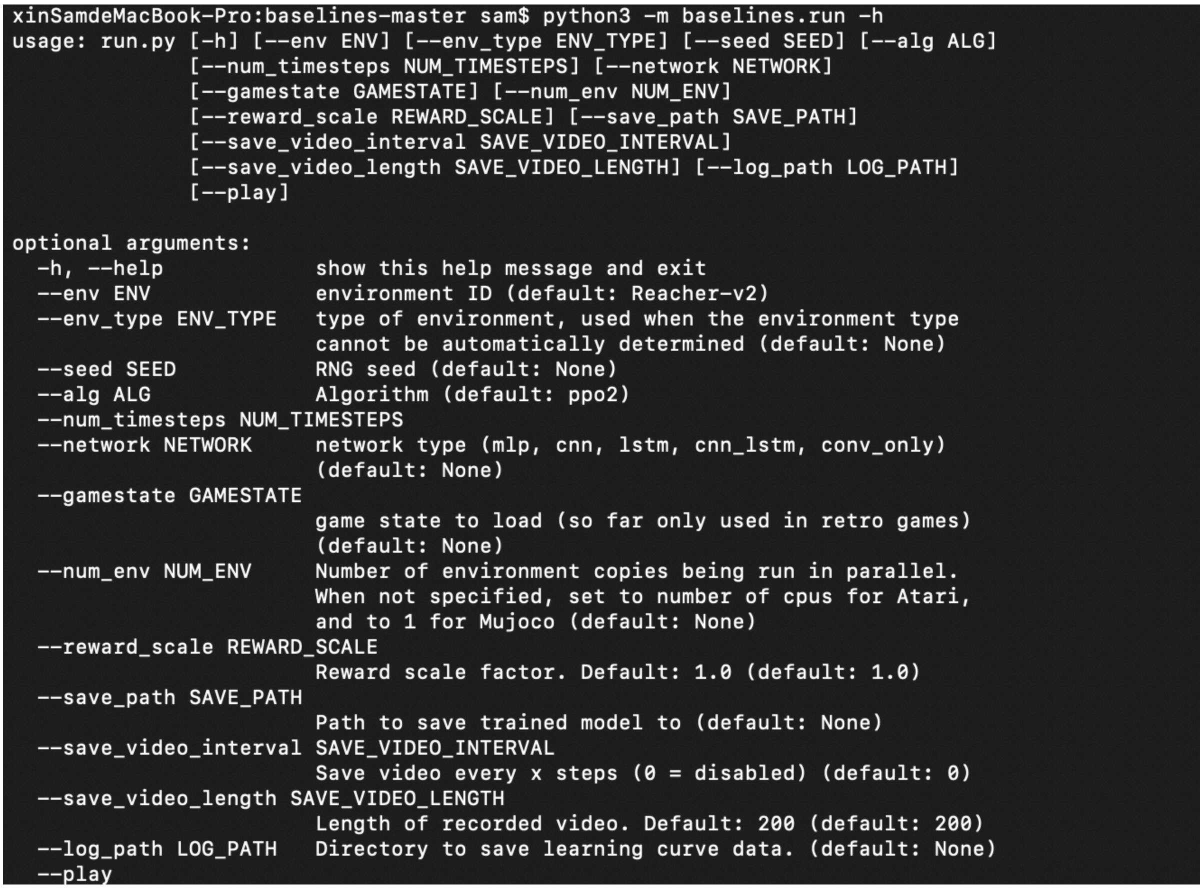
目标

使用DDPG算法对人形 双足机器人模型进行训练，使其学会直立行走

实现过程

在openAI上baselines包中有完整的强化学习的算法包，如：DQN、PPO、DDPG，通过设定相关的参数，选择使用 DDPG算法，配置Humanoid-v3环境(Mujoco)，开始训楼，训练结束后，展示当前数据量训练所得模型的结果。

- 具体使用方法：

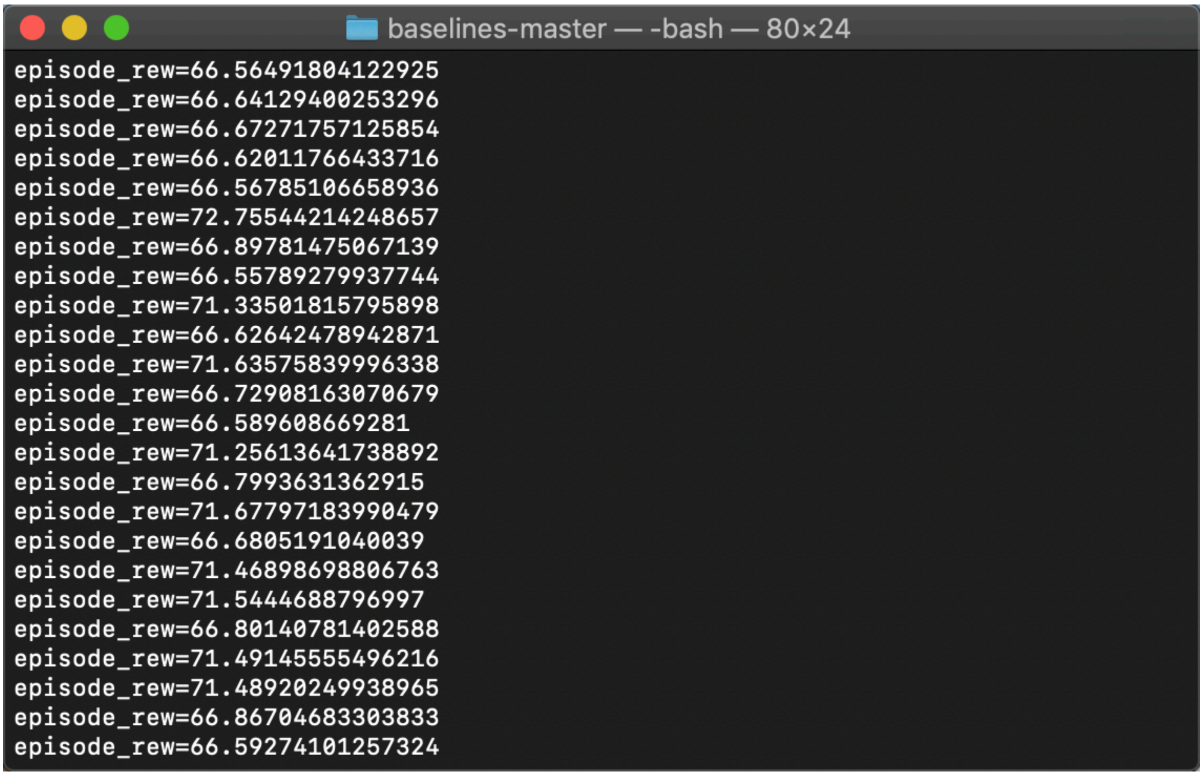


```
xinSamdeMacBook-Pro:baselines-master sam$ python3 -m baselines.run -h
usage: run.py [-h] [--env ENV] [--env_type ENV_TYPE] [--seed SEED] [--alg ALG]
              [--num_timesteps NUM_TIMESTEPS] [--network NETWORK]
              [--gamestate GAMESTATE] [--num_env NUM_ENV]
              [--reward_scale REWARD_SCALE] [--save_path SAVE_PATH]
              [--save_video_interval SAVE_VIDEO_INTERVAL]
              [--save_video_length SAVE_VIDEO_LENGTH] [--log_path LOG_PATH]
              [--play]

optional arguments:
  -h, --help            show this help message and exit
  --env ENV             environment ID (default: Reacher-v2)
  --env_type ENV_TYPE   type of environment, used when the environment type
                        cannot be automatically determined (default: None)
  --seed SEED           RNG seed (default: None)
  --alg ALG             Algorithm (default: ppo2)
  --num_timesteps NUM_TIMESTEPS
  --network NETWORK     network type (mlp, cnn, lstm, cnn_lstm, conv_only)
                        (default: None)
  --gamestate GAMESTATE
                        game state to load (so far only used in retro games)
                        (default: None)
  --num_env NUM_ENV     Number of environment copies being run in parallel.
                        When not specified, set to number of cpus for Atari,
                        and to 1 for Mujoco (default: None)
  --reward_scale REWARD_SCALE
                        Reward scale factor. Default: 1.0 (default: 1.0)
  --save_path SAVE_PATH
                        Path to save trained model to (default: None)
  --save_video_interval SAVE_VIDEO_INTERVAL
                        Save video every x steps (0 = disabled) (default: 0)
  --save_video_length SAVE_VIDEO_LENGTH
                        Length of recorded video. Default: 200 (default: 200)
  --log_path LOG_PATH   Directory to save learning curve data. (default: None)
  --play
```

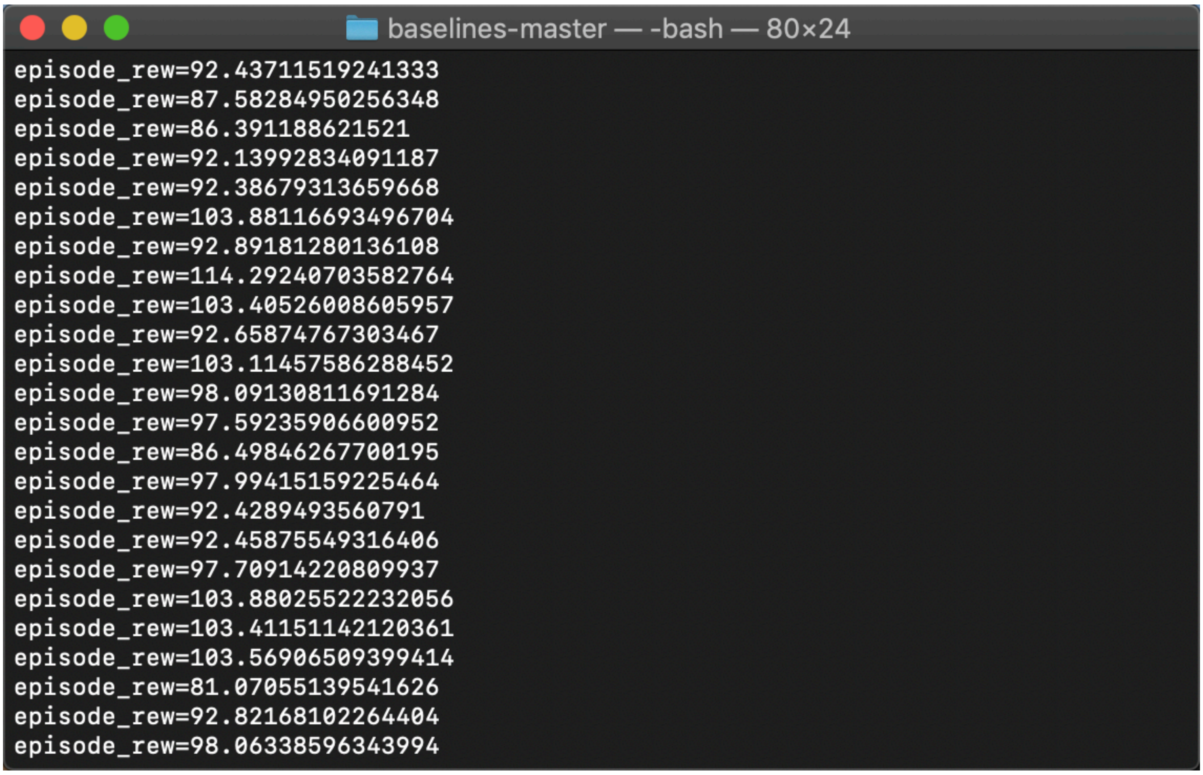
- Episodes = 1000 时的 reward

可看出，随着训练的数据量不断增加，所得到的reward的均值也在提升，达到了训练的目的。经查询资料得知，需要训练77h，平均reward达到9000分上下，才可完成对走路的学习。



```
episode_rew=66.56491804122925
episode_rew=66.64129400253296
episode_rew=66.67271757125854
episode_rew=66.62011766433716
episode_rew=66.56785106658936
episode_rew=72.75544214248657
episode_rew=66.89781475067139
episode_rew=66.55789279937744
episode_rew=71.33501815795898
episode_rew=66.62642478942871
episode_rew=71.63575839996338
episode_rew=66.72908163070679
episode_rew=66.589608669281
episode_rew=71.25613641738892
episode_rew=66.7993631362915
episode_rew=71.67797183990479
episode_rew=66.6805191040039
episode_rew=71.46898698806763
episode_rew=71.5444688796997
episode_rew=66.80140781402588
episode_rew=71.49145555496216
episode_rew=71.48920249938965
episode_rew=66.86704683303833
episode_rew=66.59274101257324
```

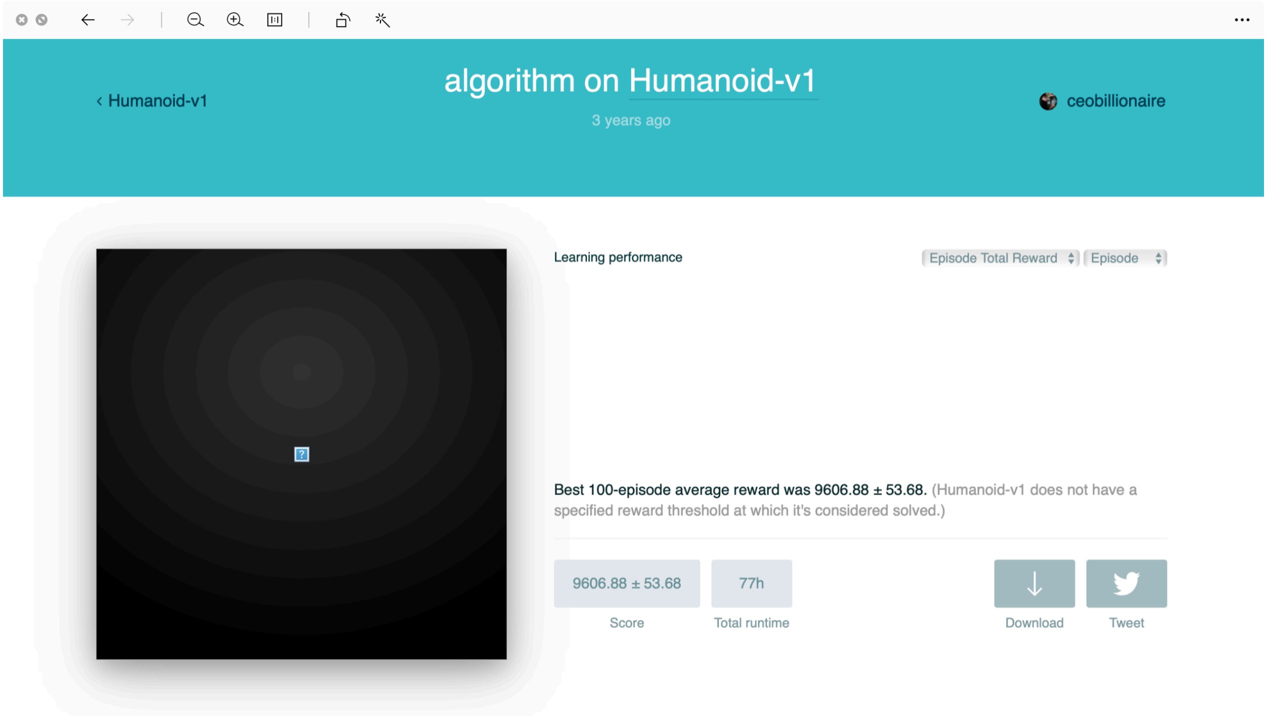
- Episodes = 10000 时的 reward



```
episode_rew=92.43711519241333
episode_rew=87.58284950256348
episode_rew=86.391188621521
episode_rew=92.13992834091187
episode_rew=92.38679313659668
episode_rew=103.88116693496704
episode_rew=92.89181280136108
episode_rew=114.29240703582764
episode_rew=103.40526008605957
episode_rew=92.65874767303467
episode_rew=103.11457586288452
episode_rew=98.09130811691284
episode_rew=97.59235906600952
episode_rew=86.49846267700195
episode_rew=97.99415159225464
episode_rew=92.4289493560791
episode_rew=92.45875549316406
episode_rew=97.70914220809937
episode_rew=103.88025522232056
episode_rew=103.41151142120361
episode_rew=103.56906509399414
episode_rew=81.07055139541626
episode_rew=92.82168102264404
episode_rew=98.06338596343994
```

- Episodes = 100000 时的 reward

可看出，随着训练的数据量不断增加，所得到的reward的均值也在提升，达到了训练的目的。经查询资料得知，需要训练77h，平均reward达到9000分上下，才可完成对走路的学习。



algorithm on Humanoid-v1

3 years ago

Best 100-episode average reward was 9606.88 ± 53.68. (Humanoid-v1 does not have a specified reward threshold at which it's considered solved.)

心得体会

经过这次实习，我们深入了解了协作完成一个项目的方法与技巧。并且对新兴的人工智能领域有了一定的认识。同时，我们还调整了自己在学校时混乱的作息，坚持早睡早起，切身体会到了通勤的感觉。可以说是收获颇丰，希望今后的学习生活中能够运用到本次实习所学到的知识和技巧，继续成长。