Deep Reinforcement Learning Nanodegree – Project 2 “Continuous Control”

Tim Wu

May 12th, 2020

**Introduction**

In this project, I built the reinforcement learning (RL) to solve the continues action spaces

1. Unity's 20 agents Reacher environment
2. Unity's 12 agents Crawler environment

In the following sessions of the report, I summarized how to build, implement, fine tune and improve the learning of DDPG and Twin Delayed DDPG agent and how I selected best agent.

Reacher

* I am able to solve the Reacher environment (version2, with 20 identical agents) in 273 episodes with average score 30.11 with DDPG agent

Crawler

* I am able to solve the Reacher environment in 273 episodes with average score 30.11 with Twin Delayed DDPG agent

**Reacher Environment (Version 2: 20 Agents)**

In the Reacher Environment has ***two*** separate versions

1. The first version contains a single agent
2. The second version contains 20 identical agents, and each of the agent has its own copy of the environment

In this project, I choose the second version to solve the environment.

**Code Location**

* The Agent Class implemented in the (ddpg\_agent.py)
* The Deep Q-Network model implemented in the file (model.py)
* The model training and etc. implemented in the (Continuous\_Control.ipynb)
* The model weights are saved in the (Reacher\_ckpt\_path) folder

**Deep Deterministic Policy Gradient (DDPG) Implementation**

The solution is based on DDPG architecture

The (Actor Critic) Network Architecture and Agent Hyperparameters

[MODEL INFO] Actor initialized with parameters:

state\_size=33

action\_size=4

seed=123

fc1\_units=128

fc2\_units=128

[MODEL INFO] CRITIC initialized with parameters:

state\_size=33

action\_size=4

seed=123

fcs1\_units=128

fc2\_units=128

The Agent Hyperparameters:

[AGENT INFO] DDPG constructor initialized parameters:

# Set parameters for training

seed = 123 # random seed number

n\_episodes\_max = 1000 # number of training episodes

max\_t = 1000 # number of timesteps per episode

actor\_fc1\_units = 128 # actor network hidden layer #1 number of unit

actor\_fc2\_units = 128 # actor network hidden layer #2 number of unit

critic\_fcs1\_units = 128 # critic network hidden layer #1 number of unit

critic\_fc2\_units = 128 # critic network hidden layer #2 number of unit

BUFFER\_SIZE = int(1e6) # replay buffer size

BATCH\_SIZE = 128 # minibatch size

GAMMA = 0.99 # discount factor

TAU = 1e-3 # for soft update of target parameters

LR\_ACTOR = 2e-4 # learning rate of the actor

LR\_CRITIC = 2e-4 # learning rate of the critic

WEIGHT\_DECAY = 0.00 # L2 weight decay

OU\_MU = 0.0 # OUNoise mu

OU\_THETA = 0.15 # OUNoise theta

OU\_SIGMA = 0.18 # OUNoise sigma

UPDATE\_EVERY\_T\_STEPS = 20 # timesteps between updates

NUM\_OF\_UPDATES = 10 # num of update passes when updating

actor\_ckpt\_path = 'Reacher\_ckpt\_path/checkpoint\_actor.pth'

critic\_ckpt\_path = 'Reacher\_ckpt\_path/checkpoint\_critic.pth'

**Results**

The best performing agent can solve the environment in 273 episodes. The file with the saved model weights of the best agent saved in the checkpoint folder Reacher\_ckpt\_path and named checkpoint\_actor.pth ,checkpoint\_critic.pth

**The Best Agent Reacher Result:**

Episode 100 (24min) Moving Average Score (over time window): 3.24

Episode 200 (50min) Moving Average Score (over time window): 16.59

Environment solved in 273 episodes! Average Score: 30.11

A screenshot of a map

Description automatically generated

**Future Improvements**

1. Extensive hyperparameter optimization, fine tune the experience replay feeding buffer size and update frequency
2. Add prioritized experience replay
3. Apply more advance model like Twin Delayed DDPG (TD3)

**Crawler Environment**

**Code Location**

* The Agent Class implemented in the (td\_ddpg\_agent.py)
* The Deep Q-Network model implemented in the file (td\_model.py)
* The model training and etc. implemented in the (Crawler.ipynb)
* The model weights are saved in the (Crawler\_ckpt\_path) folder

**Twin Delayed Deep Deterministic Policy Gradient (TD3) Implementation**

The solution is based on Twin Delayed DDPG (TD3) architecture – Actor Critic Network

Shown blow:

A close up of a map

Description automatically generated

The Network Architecture and Parameters:

[MODEL INFO] Actor initialized with parameters:

state\_size=129

action\_size=20

seed=123

fc1\_units=400

fc2\_units=300

[MODEL INFO] CRITIC initialized with parameters:

state\_size=129

action\_size=20

seed=123

fcs1\_units=400

fc2\_units=300

The Agent Hyperparameters:

[AGENT INFO] DDPG constructor initialized parameters:

num\_agents=12

state\_size=129

action\_size=20

random\_seed=123

actor\_fc1\_units=400

actor\_fc2\_units=300

critic\_fcs1\_units=400

critic\_fc2\_units=300

buffer\_size=1000000

batch\_size=128

gamma=0.99

tau=0.001

lr\_actor=0.0002

lr\_critic=0.0002

weight\_decay=0.0

ou\_mu=0.0

ou\_theta=0.15

ou\_sigma=0.1

update\_every\_t\_steps=2

**Results**

The file with the saved model weights of the best agent saved in the checkpoint folder Crawler\_ckpt\_path and named checkpoint\_actor.pth ,checkpoint\_critic.pth

**The Best Agent Reacher Result:**

**Future Improvements**

1. Extensive hyperparameter optimization, fine tune the experience replay feeding buffer size and update frequency
2. Add prioritized experience replay

**Reference**

1. <https://ai.atamai.biz/> - Reinforcement Learning with Pytorch course slides
2. <https://www.superdatascience.com/pages/drl-2-resources-page> - Deep Reinforcement Learning 2.0
3. Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., van Hasselt, H., and Silver, D. Distributed prioritized experience replay. *International Conference on Learning Representations*, 2018.
4. Scott Fujimoto, Herke van Hoof, David Meger Addressing Function Approximation Error in Actor-Critic Methods <https://arxiv.org/pdf/1802.09477.pdf> arXiv:1802.09477v3 [cs.AI] 22 Oct 2018