Navigating a Collector Agent

1 Introduction

DDPG (Deep Deterministic Policy Gradient) is a modelless policy-based learning algorithm where the agent will learn directly from raw observation spaces without knowing the dynamic information of the domain. Using policy gradient method that estimates the weights of an ideal policy through gradient ascent, similar to gradient descent used in neural network.

This algorithm is close to the DQN, but in Deep Q-Learning there is a limitation where it receives the state as an input from the network and the value function as an output, and then calculates the maximum value of the various outputs, but in continuous space it does not work very well. DDPG solves this problem with two neural networks, an actor call and a critic call. The actor is used to approach the optimal policy deterministically, that is, the action with greater certainty. As far as the critic is concerned, he evaluates the value function using the most credible action of the actors.

It uses some DQN components such as replay buffer and Soft Target Update to learn offline and Gradient Clipping, to update network weights.

2 Method

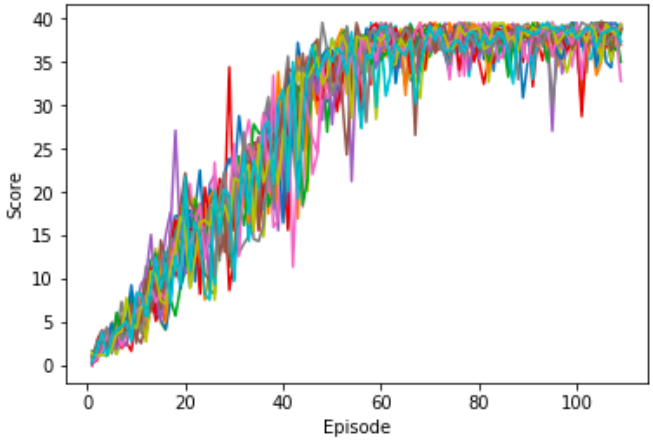
Networks of actors and critics with continuous action space are used. In the critical case, a fully connected layer produces the state value. For the actor, a fully connected layer provides the average of a normal distribution for each action. The input to these networks consists of a 33 $\times$ 20 matrix produced by the environment (the state space size times the number of agents) the first layer is fully connected linear with 400 neurons. It is followed by a hidden layer also composed of a fully connected linear layer with 300 neurons. The output layer of the critical network includes a fully connected linear layer with a single output for each agent followed by a tanh activation function that naturally constrains values between [1, 1].

The hyperparameters used for training in this environment are:

- The replay buffer size: Repeat buffer = 1e6
- The actor learning rate: Learning rate = 1e-3
- The critic learning rate: Learning rate = 1e-3
- Soft target update: 1e-3
- The discount factor: Gamma discount = 0.99
- The minibatch size: Batch Size = 128
- Weight reduction: 0
- The target soft update parameter:Tau = 1e-3
- Exploration: Epsilon = 1

3 Results

The DDPG agent takes 109 times to reach an average reward of 30 scores in about 5 hour and 35 minutes trained on the GPU instance. The stop condition is reached once, as well as the 30-point validation results as evaluation. The training results are signs in the figure.



4 Improvements

In this project, the DDPG (Deep Deterministic Policy Gradient) was implemented, to improve accuracy and with faster training, Prioritized Repetition places more important experiments with a higher sample probability.