1) Introduction

- The goal is to get an average score of +0.5 over 100 consecutive episodes.

2) Environment

- Environment yields 2 (potentially different) scores. The maximum of these 2 scores is taken

3) Learning Algorithm

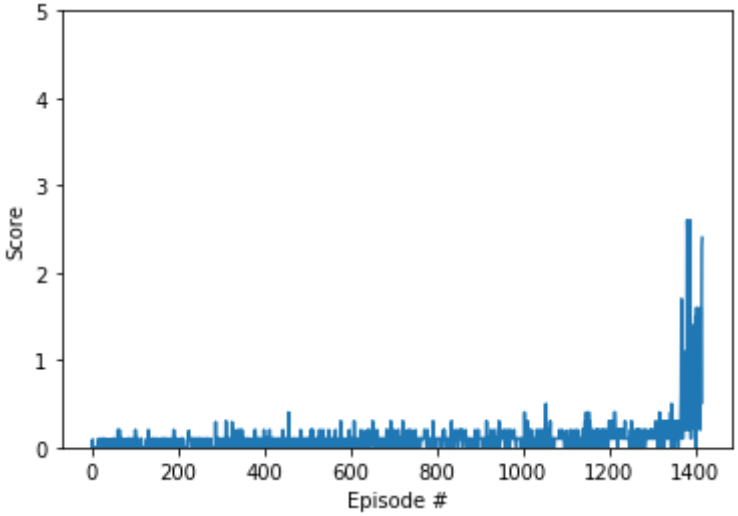
- Learning independent two DDPG agents method is adopted. Because each agent receives its own observation, each DDPG agent has its own actor_local, actor_target, critic_local, critic_target and replay memory. Both agents are trained through self-play.

4) Parameters

- Actor: 2 Hidden fully-connected layers(64, 64 units)
- Critic: 2 Hidden fully-connected layers(64, 64 units)
- Actor learning rate: 1e-3
- Critic learning rate: 1e-3
- Gamma: 0.99
- Soft update(tau): 6e-2
- Memory size: 1e6
- Batch size: 128
- Optimizer: Adam

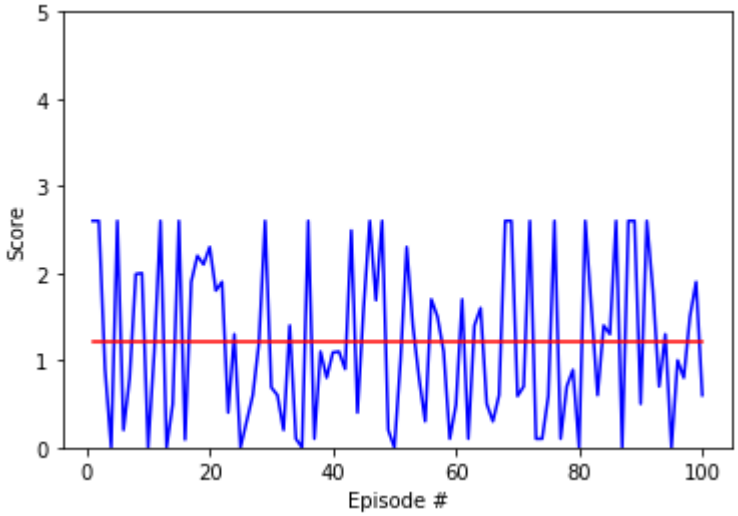
5) Training

- Training episodes: 1416



6) Results

- Average score: 1.20



7) Future work

- To improve learning method, Multi-Agent Deterministic Policy Gradients(MADDPG) will be implemented