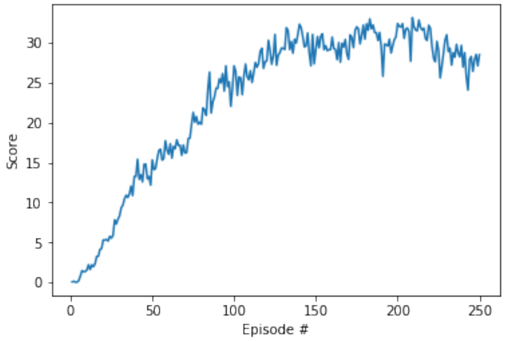
Learning Algorithm:

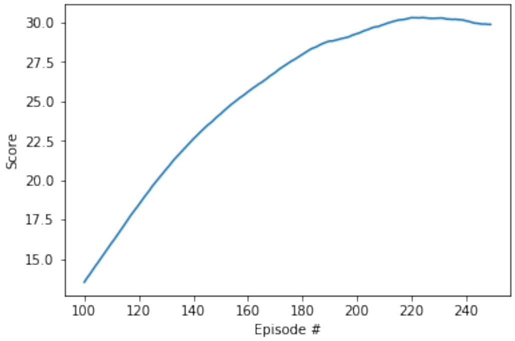
- Deep Deterministic Policy Gradient (DDPG)
- Network:
  - Actor: 2 fully connected hidden layers with 400 and 300 units. Weights were uniformly and randomly initialized between -0.003 and 0.003.
  - Critic: 2 fully connected hidden layers with 400+action size and 300 units. Weights were uniformly and randomly initialized between -0.003 and 0.003.
- Training Hyperparameters:
  - Discount (γ) = 0.99
  - Soft update ratio (τ) = 0.001
  - Learning rate for actor network = 0.0001
  - Learning rate for critic network = 0.001
  - Number of agents: 20
  - Training Episodes: 250
  - Max step in each episode: 1500

Results:

- Solved in 213 episodes.
- Reward per episode:



- Reward of averaging 100 episodes:



Future work

After roughly 180 episodes, the performance started to be unstable and declined close to the end of the training. As mentioned in the benchmark of this project, the hyperparameters for training might play important roles for the stability of the trained agents. Future work should definitely cover the exploration of different sets of hyperparameters.

Stochastic weight averaging (SWA, https://izmailovpavel.github.io/files/swa_rl/paper.pdf) could be another direction for training agents yield stable performance.

Comparison between models introduced in the actor-critic module on this Reacher task could also be another potential future work.