Report of training the agent

The model was trained for a maximum of 2,000 episodes but the agent was able to solve the environment (i.e. get atleast +30 average score over 100 adjacent episodes).

Model

The Actor Network has three dense (or fully connected layers). The first two layers have**400 and 300** nodes respectively activated with **ReLU** activation function. The final (output layer) has**4** nodes and is activated with tanh activation. This network takes in as input the **33** dimensional current state and gives as output**4** to provide the action at current state that the agent is supposed to take.

The Critic Network has three dense (or fully connected layers). The first two layers have**404 and 300** nodes respectively activated with **ReLU** activation function. The final (output layer) has**4** nodes and is activated with linear activation (no activation at all). This network takes in as input the **33** dimensional current state and**4** dimensional action and gives as output a single real number to provide the Q-value at current state and action taken in that state.

Both the neural networks used Adam optimizer and Mean Squared Error (MSE) as the loss function.

The following image provides a pictorial representation of the Actor Network model:

Pictorial representation of Q-Network

The following image provides a pictorial representation of the Critic Network model:

Pictorial representation of Q-Network

The following image provides the plot for score v/s episode number:

Plot for score v/s episode number

Performance

The model was trained on MacBook Air 2017 with 8GB RAM and Intel Core i5 Processor.

- **Number of episodes required to solve the environment** -37 episodes
- **Final score of the agent**: 30.57

Hyperparameters used

Hyperparameter	Value	Description
Buffer size	100000	Maximum size of the replay buffer
Batch size	128	Batch size for sampling from replay buffer
Gamma (**γ**)	0.99	Discount factor for calculating return

Hyperparameter	Value	Description
Tau (τ)	0.001	Hyperparameter for soft update of target parameters
Learning Rate Actor	0.0003	Learning rate for the actor neural network
Learning Rate Critic	0.001	Learning rate for the critic neural network

Future work

The following algorithms can be considered for further development of this agent:

- Proximal Policy Optimization (PPO)
- Generalized Advantage Estimation (GAE)
- Advantage Actor-Critic (A2C)
- Asynchronous Advantage Actor-Critic (A3C)