Deep Q Networks (DQN)
Lab Exercise#3

Palacode Narayana Iyer Anantharaman

04 Sep 2018

References

- Deep Q Learning with Keras and Gym, Keon’s blog: https://keon.io/deep-q-learning/
- OpenAI Gym CartPole Evaluations: https://gym.openai.com/envs/CartPole-v0/
- TensorFlow 1.10 documentation

Goal of this lab session

- Model Environment using OpenAI Gym and write the agent class as a DQN

- Today: Develop DQN agent for Cartpole example (See the references)

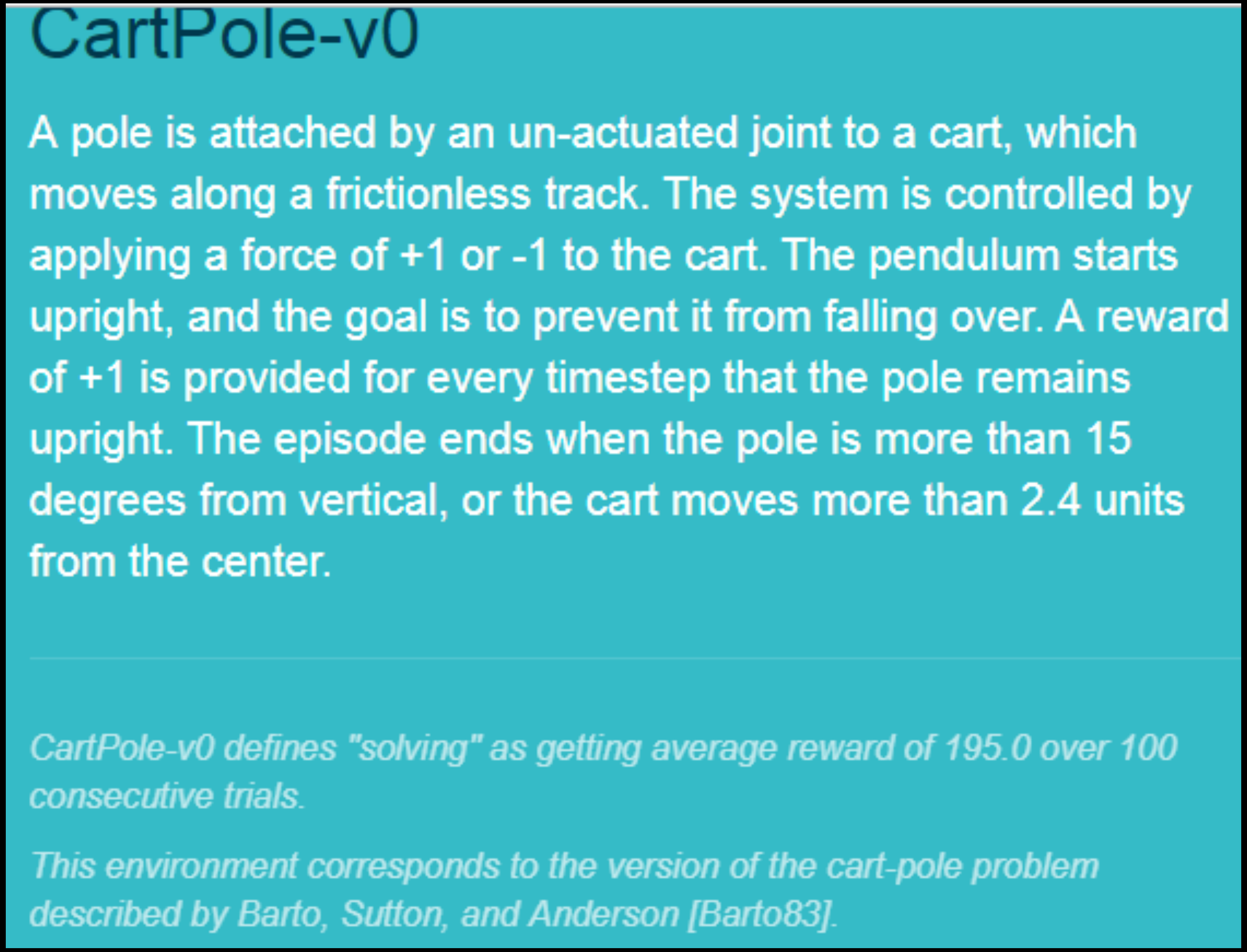
Recap: Q Learning Basics

- You are given a vector of input from CartPole environment (state) and the reward. These are inputs to the agent.
- Agent can choose an action from a list of possible actions (e.g. left, right)
- Model the agent as a DQN that takes the state as input vector, emits regression values, one unit per action, as the outputs.
- The equation below is used as the loss function

$$loss = \left(\underbrace{\overset{\text{Reward}}{\mathrm{r}} + \overset{\text{Decay Rate}}{\gamma} \max_{a^{`}} \hat{Q}(s, a^{`})}_{\text{Target}} - \underbrace{Q(s, a)}_{\text{Prediction}}\right)^2$$

Problem Statement

- In this lab, you are required to develop a DQN agent (See the references) that can balance the CartPole
- Create a cartpole environment instance using OpenAI gym.
- Create a class for the agent, use TF estimators to build the DQN model. The output of DQN is a liner unit (representing a Q Value) per action.
- Use different values of Gamma (0.2, 0.5, 0.8, 0.99) and determine the convergence and the corresponding rewards. Report your values



CartPole-v0

A pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The system is controlled by applying a force of +1 or -1 to the cart. The pendulum starts upright, and the goal is to prevent it from falling over. A reward of +1 is provided for every timestep that the pole remains upright. The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from the center.

CartPole-v0 defines "solving" as getting average reward of 195.0 over 100 consecutive trials.

This environment corresponds to the version of the cart-pole problem described by Barto, Sutton, and Anderson [Barto83].