Deep Learning Lab
Exercise 04

Rabea Turon & Megan Klaiber

January 4, 2019

1 Introduction

In this exercise a DQN agent will be implemented and its prformance willl be evaluated on the CartPole and CarRacing environment of OpenAI Gym.

2 Reinforcement Learning: Deep Q-Networks

2.1 CartPole

A simple neural network was used for the CartPole task. It consists of two fully connected layers each with 20 hidden units. The network is optimized with mean squared error and Adam optimizer. CartPole is considered solved when the average reward is greater than or equal to 195.0 over 100 consecutive episodes. So we stopped the training when this point was reached. The size of the batches sampled from the replay buffer was 64 and we used a learning rate of 0.001.

Figure 1 shows the achieved reward during the training episodes. Here CartPole was solved after 150 episodes. During training every 20th episode the agent was evaluated with deterministic actions over 5 episodes. The average reward of these 5 episodes is shown in Figure 2. The final performance where evaluated over 15 test episodes. Here the average reward was 349 with a standard deviation of 72.

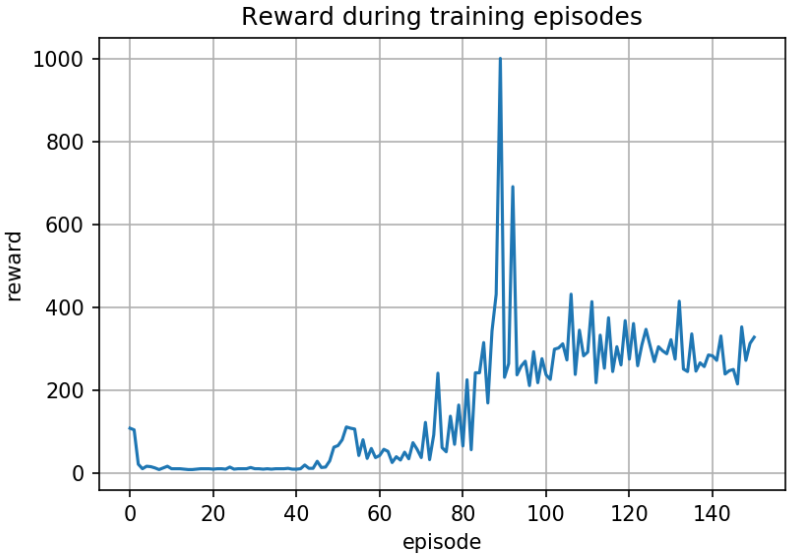

Figure 1: Achieved reward during training episodes of CartPole.

Figure 2: Evaluated CartPole agent with deterministic actions every 20th episodes (mean episode reward over 5 episodes).

2.2 CarRacing

For the CarRacing task we used a convolutional neural network. It consists of three convolutional layers with 32 8x8, 64 4x4 and 64 3x3 filters and strides of 4, 2 and 1. After that a fully connected layer with 265 hidden units is added. Like in subsection 2.1 the network is optimized with mean squared error and Adam optimizer. The size of the batches sampled from the replay buffer was 128 and we used a learning rate of 0.0003.

For training improvement we skipped four frames and repeated the agents action during this time. For the exploration we set different probabilities for the five actions so the agent prefer the actions accelerate or go straight more often.

Figure 3 shows the achieved reward during the training episodes. During training every 20th episode the agent was evaluated with deterministic actions over 5 episodes. The average reward of these 5 episodes is shown in Figure 4. The final performance where evaluated over 15 test episodes. Here the average reward was 559 with a standard deviation of 217. During testing we saw that the agent was able to follow the track but had problems with the sharp curves. For these curves the agent was to fast and didn't slow down.

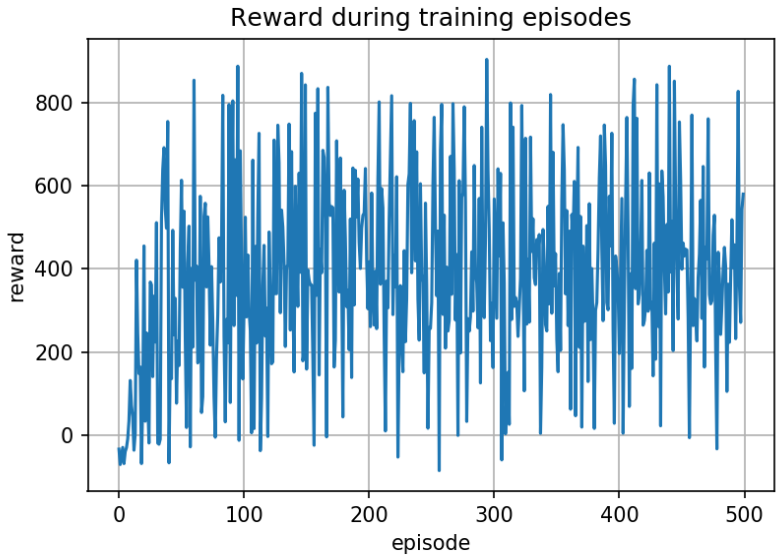

Figure 3: Achieved reward during training episodes of CarRacing.

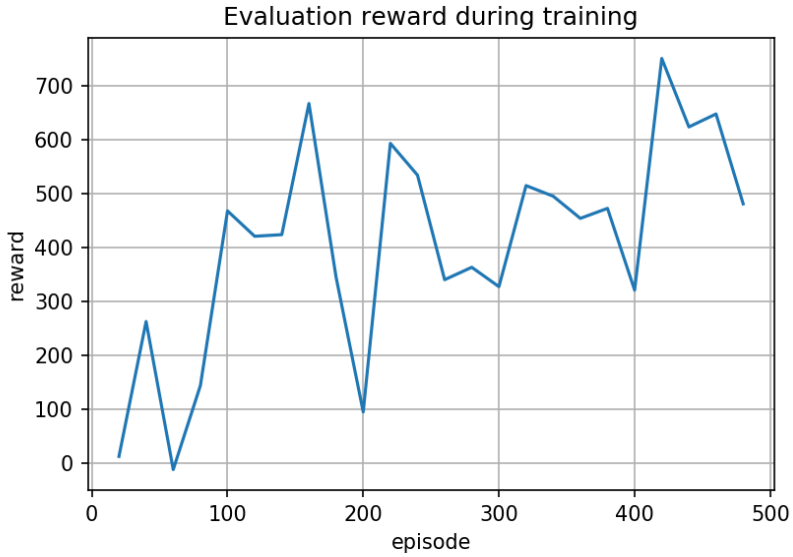

Figure 4: Evaluated CarRacing agent with deterministic actions every 20th episodes (mean episode reward over 5 episodes).

3 Exploration

4 Conclusion