Deep Reinforcement Learning

Exercise 3

Student: Arseni Pertzovskiy, ID: 317377372.

1. **[35 pts] DQN**

An implementation is inside the code.

1. **[15 pts] Replay Buffer**

An implementation is inside the code.

1. **[10 pts] Hard Target Network-Update**

An implementation is inside the code.

1. **[10 pts] Soft Target Network-Update**

An implementation is inside the code.

1. **[25 pts] Effect of Replay Buffer and Target Network**

An implementation of weights storage is inside the code.

The table with final result:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| target  update | with replay  with target | with replay  without target | without replay  with target | without replay  without target |
| hard | 195.2 | x | 9.4 | x |
| none | x | 9.8 | x | 9.4 |
| soft | 188.6 | x | 9.6 | x |

Each cell is the mean of five experiments with corresponding setting.

With replay – Without target – NoneA screenshot of a cell phone

Description automatically generated

Loss over episodes as well as mean Q error over episodes had grown with an increasing pase over the run. We are gaining inferior results.

Without replay – Without target – None

A screenshot of a cell phone

Description automatically generated

Loss over episodes as well as mean Q error over episodes had grown with an increasing pase over the run. We are gaining even more inferior and less stable results that the previous one.

**Investigation of soft and hard target updates (the difference we can observe in the loss and mean Q error)**

Without replay – With target – Hard

A screenshot of a cell phone

Description automatically generated

Last two experiments we saw loss over episodes is in between [0, 100+] range as well as mean Q error over episodes is between [0, 30+] range. While introducing target network we are already droped both ranges to [0, 1] in loss and [-1, 1] in mean Q error. The update occuring each 15 episodes therefore we can observe an cyclic behavior in the graph.

Without replay – With target – Soft

A screenshot of a social media post

Description automatically generated

In the experiments without target net we saw loss over episodes is in between [0, 100+] range as well as mean Q error over episodes is between [0, 30+] range. While introducing target network we are already droped both ranges to [0, 2] in loss and [-1, 2] in mean Q error. The update occuring in a soft manner therefore we can observe more smooth growth (and not any cyclic behaviors as with ‘hard update’ case) in the graph.

But still we are achiving highly inferior results in both cases.

With replay – With target – Hard

A screenshot of a cell phone

Description automatically generated

While combining both replay buffer and target net we are jumping in our performance and gainig the best results in the experiment. Loss over episodes as well as mean Q error over episodes decreasing over time and the score gets it’s best. The cyclic behavior proper to ‘hard update’ can be observed on the graph. The algorithm gets at some point it’s ‘ceiling’ and remains with approximatly same results (converges). Maybe some parameters tunning can help to improve perfomance.

With replay – With target – Soft

A screenshot of a cell phone

Description automatically generated

The changings of loss over episodes and mean Q error over episodes are much more smooth than with the ‘hard’ case. As we said earlier, the algorithm gets at some point it’s ‘ceiling’ and remains with approximatly same results (converges). Maybe some parameters tunning can help to improve perfomance.

Because of a stochastic nature of the problem, each approach we gain a little bit different results, therefore sometimes ‘soft’ and sometimes ‘hard’ manner of update wins in term of total score.

1. **[5 pts] Movie**

Attached to the submition.

Best performing agent: with replay, with target, hard weights update.

End