Human-level control through deep reinforcement learning

Authors: Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg & Demis Hassabis

Publisher: Nature
Year: 2015

Keywords: Reinforcement learning, DQN, Atari games

1. **Abstract:**

"The theory of reinforcement learning provides a normative account, deeply rooted in psychological and neuroscientific perspectives on animal behaviour, of how agents may optimize their control of an environment. To use reinforcement learning successfully in situations approaching real-world complexity, however, agents are confronted with a difficult task: they must derive efficient representations of the environment from high-dimensional sensory inputs, and use these to generalize past experience to new situations. Remarkably, humans and other animals seem to solve this problem through a harmonious combination of reinforcement learning and hierarchical sensory processing systems, the former evidenced by a wealth of neural data revealing notable parallels between the phasic signals emitted by dopaminergic neurons and temporal difference reinforcement learning algorithms. While reinforcement learning agents have achieved some successes in a variety of domains, their applicability has previously been limited to domains in which useful features can be handcrafted, or to domains with fully observed, low-dimensional state spaces. Here we use recent advances in training deep neural networks to develop a novel artificial agent, termed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. We tested this agent on the challenging domain of classic Atari 2600 games 12 . We demonstrate that the deep Q-network agent, receiving only the pixels and the game score as inputs, was able to surpass the performance of all previous algorithms and achieve a level comparable to that of a professional human games tester across a set of 49 games, using the same algorithm, network architecture and hyperparameters. This work bridges the divide between high-dimensional sensory inputs and actions, resulting in the first artificial agent that is capable of learning to excel at a diverse array of challenging tasks."

2. **Introduction:**

- As a process of evolution, the human beings adapted a very complex and precise sensory components to receive and process the data. And use them to handle the current situations based on the past knowledge.
- This in terms of Artificial Intelligence was referred to reinforcement learning to use the previous experience to improve the present scenario.
- So, the authors were motivated to create an architecture from the deep learning which satisfies the former statement, known as deep Q-network.
- This network is able to learn through reinforcement learning by using high dimensional sensor inputs.
- The works were also concentrated such that the individual architecture resulting in best results even in the complex domains and smaller earlier knowledge.

3. Related Works:

- Formerly the adaptation of reinforcement learning had been very successful, however limited to low dimensional inputs and fully observed environments.
- Q-networks were also implemented in prior works where the resulting scalar and the history of activities were to input again to the network as feedback. The draw back of the architecture is that the Q-value is to be computed for every action and prior action for each iteration which would affect in computational expense.
- One of the approach that authors adapted from the previous methods is frame-skipping method, where the actions were selected at every k-th step instead of every step, resulting in reducing the computations.

4. Content and Approach:

- In preprocessing, they have trained various states of the Atari game play with 2600 frames of images.
- In each image, they had reduced the dimensionality of the image and extract the maximum pixel color value.
- Secondly, they would extract the luminance of RGB and reshape the ratio to 84 x 84 pixels.
- As the previous approaches, failed in optimizing the Q-networks by feeding back all the actions done by it as input, the authors limited this to input only the current action which would reduce the computations.
- And for all the remaining possible states, certain outputs were calculated. The main pro of the network is that generating outputs for all the possible actions.
- The 84 x 84 x 4 image was input to the architecture in which the first layer convolves with stride 4 and implements rectifier non-linearity.
- The following hidden layer also convolves with stride 2 for 64 filters and the final layer would also stride 1 with rectifier nonlinearity itself.

- The training was done with 2600 different games with known results, with rewards for the accomplishments and penalties for the bad tries.
- The architecture is also robust enough to train different games even with lower prior knowledge.
- Each game is trained with a different model with three rewards if -1,0 and 1.
- To know that the game is completed for the games, which had lives, the authors used an emulator for counting number of lives.
- Adapted from previous approaches, the authors also used frame skipping technique, in which the actions were chosen for every k^{th} frame, instead of selecting at each frame resulting in the reduction of computations.
- The authors used k as 4, which would boost the run time k times (here '4') for all the games.
- With the help of informal search strategies, the parameters and hyper parameters of various games were selected, which were same for all the games.
- The training is implemented on the Q-networks by greedy policy. Instead of using all the variable histories, the model uses a fixed length of histories only backpropagating to certain limit.
- This extracting history is done by a method of storing the limited amount of history as a set and replaying all the histories through each iteration.
- However, this technique is limited as the best transitions had not being stored and cannot be used for the current situation.
- Some of the hyperparameters and corresponding values are:

○ Minibatch size	32
○ History length	4
○ Replay memory size	1000000
○ Update frequency	4
○ Discount factor	0.99

5. Results:

- The results were shown as the histograms of different games, in which the DQN networks had significantly increased the win percentage.
- More than 50% of the games played had been at or above the human level performance, which considered to be accomplishment of task.
- Although in the games like Montezuma's revenge and Double Dunk, the networks had not shown any significant improvements, it had performed excellent in majority of the games.
- In the comparisons of the combinations of Q-network and replay, the models with both the replay with target Q had outperformed all the other combinations.

6. Conclusion and Future works:

- It is obvious that the objective to create a single algorithm, which resemble biologically to the humans, and which can solve multiple challenges had been accomplished.
- Moreover, different games have different strategies which the model had successfully increased the efficiency more than the human-level performance excepting one or two cases.
- Further the works can be focused on the biasing by using the previous experiences, which the authors had not explored.