# MS-Pacman\_v0 OpenAI GYM DQN-Learning

Author: Parsa Pourzargham

e-mail: [lum2ce@inf.elte.hu](mailto:lum2ce@inf.elte.hu)

Neptun ID: LUM2CE

## Abstract

## Introduction

The problem which this article tries to solve is an Atari game called Pacman, in which, the agent must learn how to maximize its reward in the environment provided by GYM python library where agent must learn how to collect the maximum number of dots, while avoiding the moving ghosts in the closed maze.

convolutional neural network is utilized to train the agent which maps the input state to actions. The motivation of applying the DQN algorithm on this problem is to see how good this algorithm can perform in this specific task and to check it can reach a near-human level performance and beat human record in the game. Also, to find out how much benefit the model gets by utilizing the neural networks alongside with the Q-Learning, compared to more basic solutions like vanilla-Q-Learning.

The input to the algorithm is the in-game grayscale screen images, fed to the neural network that output the probability of each action also known as Q values.

## Related work

[‎1] uses a different implementation of Pacman game environment in OpenCV library, but quite interesting experiments and techniques was applied in the paper. Some experiments were carried out to see the performance of the model in an unseen environment, namely increasing the number of the ghosts which leads to a completely unseen training set for the agent. The performance of the model was expectedly decreased over the number of the ghosts. The model, however, could generalize quite fine in terms of adding new ghosts. Another experiment was also carried out to see how the model performs over an unseen game map. The results demonstrated that the performance of the model was near random in this experiment and the model failed to generalize well.

[‎1] utilized a method to determine the effect of character-centric cropping in Pacman game which leads to significantly faster model in terms of runtime. According to the paper, the model fed by cropped images centered around the agent can achieve quite same performance while decreasing the runtime duration by the factor of 6.

## Problem description

In the game, Pacman is the agent which has to collect as much seed as possible while avoiding 4 ghosts moving to catch it in the closed maze field. The agent has 9 possible actions to take in each step which include moving up, moving down, moving right, moving left, moving diagonally and staying where it is.

The environment for the problem is provided by the GYM library in python programming language. It takes the action from the agent and returns the new state of the environment after performing the action, as well as the immediate reward of the taken action as a feedback to the agent.

## Methods

As mentioned, the algorithm used to train the agent is Deep-Q-Learning, DQN in short. DQN is also a reinforcement algorithm which is derived from the classical Q-Learning algorithm. Unlike classical Q-learning algorithm which maps each state-action pair to its corresponding Q-Value using a table called Q-table, DQN maps the state-action pairs to their corresponding Q-value using a deep neural network.

The overall workflow of DQN algorithm is as following:

1. The neural network takes the in-game screen of the game as the input and calculate the tensor of probabilities of actions. To obtain a tensor of probabilities, the activation function of the las layer (fully connected layer with output number of the action space size) is set to None.
2. As the networks take the current state of the environment to give the probability tensor of actions, a grayscale image of the game is taken and fed into the network to calculate the corresponding Q-values.
3. After obtaining the probability tensor, the action with the highest probability is chosen among all possible actions. That is delivered by the function argmax ().
4. Having chosen the action in the previous step, the action is given to GYM and it provides us with the new state of the environment after taking the action, as well as the immediate reward corresponding to the taken action and whether the game is finished or not.
5. After obtaining the above-mentioned parameters from the environment, the parameters are stored in a buffer in the format of (s,a,r,s’,d), where s denotes the state before taking the action, a denote the taken action, s’ denotes the new state after taking the action and d stands for the Boolean value which decides whether the game is done or not.
6. After finishing the step 5, the algorithm needs to calculate the loss to train the network. r’ and a’ are required to calculate the loss value. a’ could be obtained by feeding s’ to the network, while r’ is discounted from r. Having r’ and a’ ,loss can be calculated.
7. Finally when the training process is finished, the network can be used by the agent to play the game and evaluate the score of the agent. This is done by feeding the current state to the network and choose the action according to the probability tensor provided by the network.

Remarks:

* Updating the model weights frequently in every step may result to low performance of the model as it updates the weights when nothing much has changed. As a result, the model weights are updated every n steps which is experimented in the experiment section to obtain the optimal n.
* Two networks are used to train the agent in this algorithm. Although both have the same architecture, they have different set of weights in the model. Two networks are called main network and target network. Main network weights are updated after each n steps in the training process, while the main network weights are copied to the target network every 100 steps. The correct frequency of copying the main network weight to the target network is obtained by trial and error to 100 steps. According to [], having these two different networks is beneficial in terms of stability in the learning process and helps the model to learn effectively.
* The environment in the game is fully observable but the gameplay of the ghosts is not deterministic and varies from episode to another and the performance of the model can not be compared with another model by first look. It is possible that an optimal model achieves low scores on some episodes, but its overall performance is better that others. Hence, apart from the rewards achieved by the agent in all the episodes, the average reward of the agent, the maximum achieved reward and the number of high scores(higher than a certain threshold) is also calculated to compare different models with different hyper parameters meaningfully. The threshold is set to 600.

## Model architecture

The code for the model is taken from [GitHub](https://github.com/EXJUSTICE/Deep_Q-learning_OpenAI_MissPacman) in the IPython file format and it is executed on Google Colab service. The original model uses 3 convolutional layers with kernel size of (8,8), (4,4) and (3,3) with strides 4,2 and 1 respectively. As the input image to the model is relatively small, no pooling layer is added to the model.

### 

## Experiments and Discussion

In the code, learning rate is set to 0.001 and the epsilon value is 0.5. The number of episodes is initially set to 800 episodes. With the default hyperparameters, the model could achieve the average reward of and of scores were higher than the threshold.

### Experiments

The following experiments were done to improve the model performance.

1. Increasing learning capacity of the network

Two additional convolutional layers were added to the model to increase the capacity of learning in the model. Additionally, the filter size was modified in a way that layers have the kernel size (5,5), (4,4), (3,3), (3,3) and (2,2) from the top layer to the last convolutional layer. The stride size was also decreased to 2 for the first two layers and 1 for the next three layers. The epsilon number is also ste to 1 to force the agent to do more explorations at the beginning which leads to more exploration and hence, learning more. Buffer size is also increased to 40000. The result from the above-mentioned modification is as following:

## Conclusion, further work

## References

Citations:

#### van der Ouderaa, Tycho. "Deep reinforcement learning in pac-man." Bachelor Thesis, University of Amsterdam (2016).

#### Spalding, Holt, Oliver Newland, and Jivko Sinapov. "Investigation Of The Effects Of Character-centric Cropping On Pacman DQN Performance." (2018).