Continuous control using Deep Q Learning with Oscillatory Activation functions

Prith Sharma

Vellore Institute of Technology, Vellore
prith.sharma2019@vitstudent.ac.in

Aditya Raj Sahoo

Vellore Institute of Technology, Vellore
adityaraj.sahoo2019@vitstudent.ac.in

Sushant Sinha

Vellore Institute of Technology, Vellore
sushant.sinha2019@vitstudent.ac.in

Abstract

The capacity of reinforcement learning (RL) to learn from the interaction between the environment and agent provides an optimal control strategy. DQN is a deep neural network structure used for estimation of Q-value of the Q-learning method. The CartPole game is essentially a game in which a stick is attached to a cart and the cart moves along a friction-less track. The goal here is to make the cart move left or right to keep the pole from falling. 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. In this paper, we explore the CartPole game and the effect oscillatory activation functions have on this.

Preprint. Under review.

1 Introduction

Reinforcement learning is a machine learning training method based on rewarding desired behaviors and/or punishing undesired ones. In general, a reinforcement learning agent is able to perceive and interpret its environment, take actions and learn through trial and error. Q-learning is a model-free, off-policy reinforcement learning that will find the best course of action, given the current state of the agent. Depending on where the agent is in the environment, it will decide the next action to be taken. The objective of the model is to find the best course of action given its current state. To do this, it may come up with rules of its own or it may operate outside the policy given to it to follow. This means that there is no actual need for a policy The OpenAI gym is an API built to make environment simulation and interaction for reinforcement learning simple. It also contains a number of built in environments (e.g. Atari games, classic control problems, etc).One such classic control problem is Cart Pole, in which a cart carrying an inverted pendulum needs to be controlled such that the pendulum stays upright. CartPole is a simple, classic reinforcement learning problem. To cope with this problem, we need something to approximate a function that takes in a state-action pair (s,a) and returns the expected reward for that pair.Suppose we want to know Q(s, a=right), we feed in the state (s) of the environment into the model. Let the neural network do the calculation, and it will return 2 values. One is the expected reward when making the left move and another is for making the right move. Since we are interested in a=right, we would just get the value from the lower node of that output. Now we have Q(s,a) by using the network. To fully train the network, loss function is essential. Intuitively, before the network outputs an approximated value Q(s,a), we will already be having an idea of what the value should be. Hence, we can punish the network for the mistake it makes, and let it learn that mistake through back-propagation.

2 Background and Related Work

A. Deep Q Networks

The underlying principle of a Deep Q Network is very similar to the Q Learning algorithm. It starts with arbitrary Q-value estimates and explores the environment using the E-greedy policy. And at its core, it uses the same notion of dual actions, a current action with a current Q-value and a target action with a target Q-value, for its update logic to improve its Q-value estimates.

B. Agent Anything that recognizes the environment through sensors and acts upon the environment initiated through actuators is called an agent. An agent performs the tasks of recognition, thinking, and acting cyclically.

C. Policies

A policy defines the way an agent acts in an environment. Typically, the goal of reinforcement learning is to train the underlying model until the policy produces the desired outcome.

D. DQN Architecture

The DQN architecture has two neural nets, the Q network and the Target networks, and a component called Experience Replay. The Q network is the agent that is trained to produce the Optimal State-Action value. Experience Replay interacts with the environment to generate data to train the Q Network.

F. Oscillatory functions:

Oscillatory activation functions comprises of multiple hyperplanes in their decision boundary. This enables the neurons to make more complex decisions than other popular sigmoidal, ReLu like Swish, and Mish activation functions. The higher representative power of networks with oscillating activation functions allows classification and regression tasks to be solved with fewer neurons. Also, the oscillations in the activation function appear to improve gradient flow and speed up back propagation learning[8]. The results presented in [8] suggest that deep networks with oscillating activation functions might potentially partially bridge the performance gap between biological and artificial neural networks. Oscillatory functions have been proved to outperform many other activation functions. The better performance can be noticed with benchmarking as well. Not only standardized, but also domain specific datasets can be used to test different use cases. For instance, testing these oscillatory activations on high entropy feature space [2] still remains a task, some of the prominent problems in this vertical include compression algorithms classification, which was earlier addressed with a reasonable accuracy across various compression classes using CNN backbone. Further experimentation using oscillatory activations could be done to better identify features in such datasets.

3 OpenAI CartPole environment

In the OpenAI CartPole environment, the status of the system is specified by an “observation” of four parameters (x, v, theta, omega), where ’x’ stand for the horizontal position of the cart (positive means to the right), ’v’ stands for the horizontal velocity of the cart (positive means moving to the right), ’theta’ stands for the angle between the pole and the vertical position (positive means clock-wise) and ’omega’ stands for the angular velocity of the pole (positive means rotating clock-wise).

Given an observation, a player can perform either one of the following two possible actions, ’0’ would imply pushing the cart to the left and ’1’ would mean pushing the cart to the right. The game is done when the pole deviates more than 15 degrees from vertical. In each time step, if the game is not done, then the cumulative reward increases by 1. The goal of the game is to have the cumulative reward as high as possible.

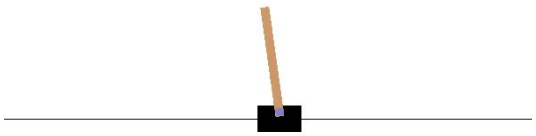

Figure 1: An image of the CartPole game [4]

4 Results and Discussion

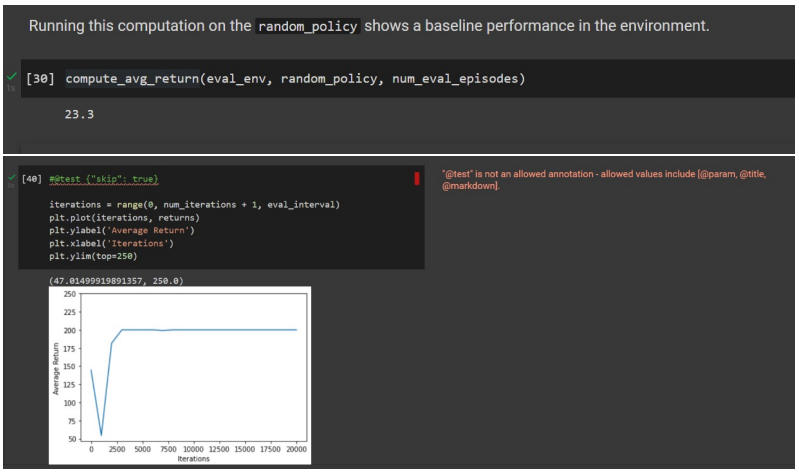

Figure 2: Activation Function: ReLU

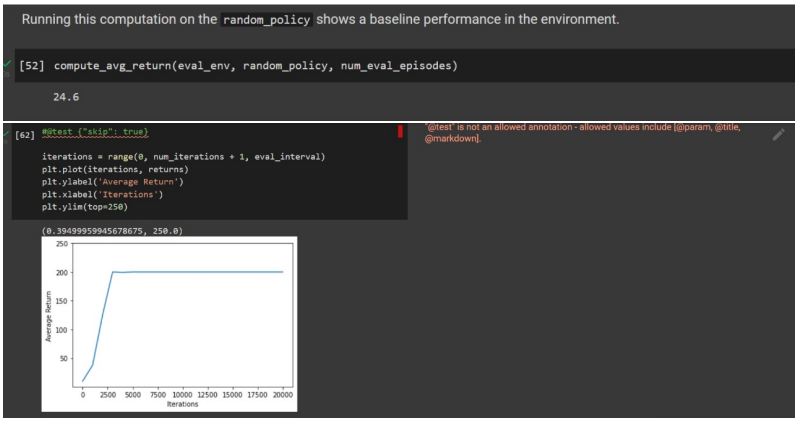

Figure 3: Activation Function: Leaky ReLU

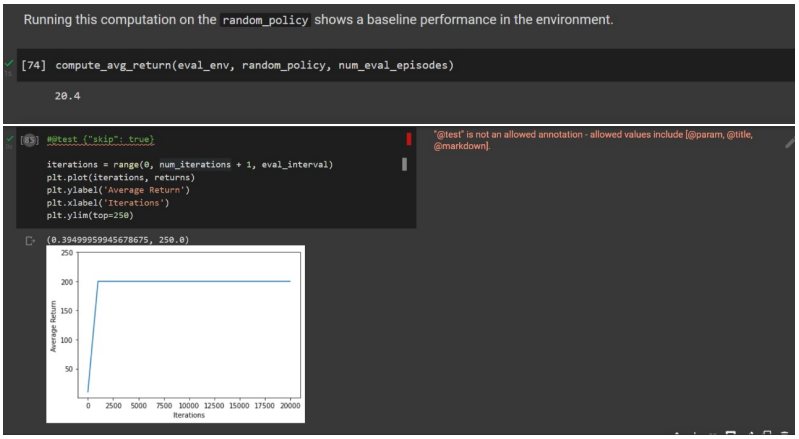

Figure 4: Activation Function: GCU

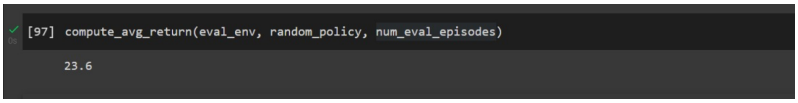

Figure 5: Activation Function: DSU (Tensor had NaN/nan values)

Figure 6: Activation Function: MC

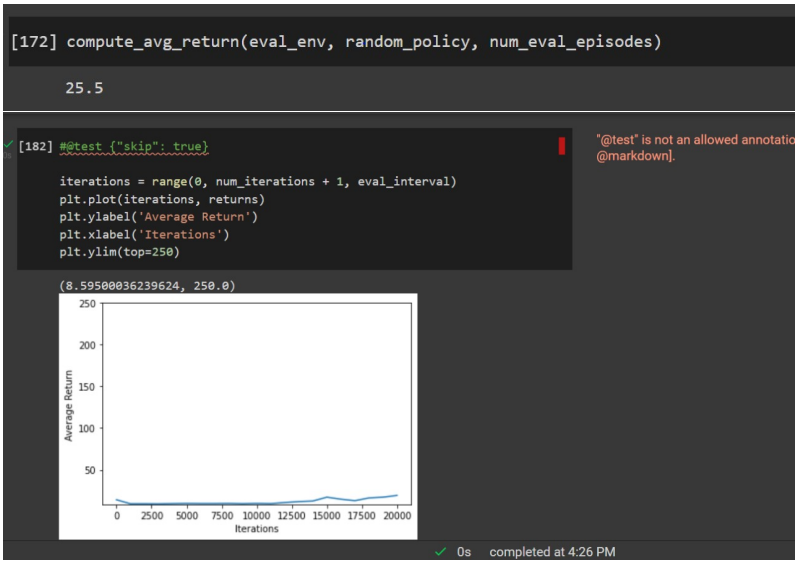

Figure 7: Activation Function: NMC

Figure 8: Activation Function: Sine

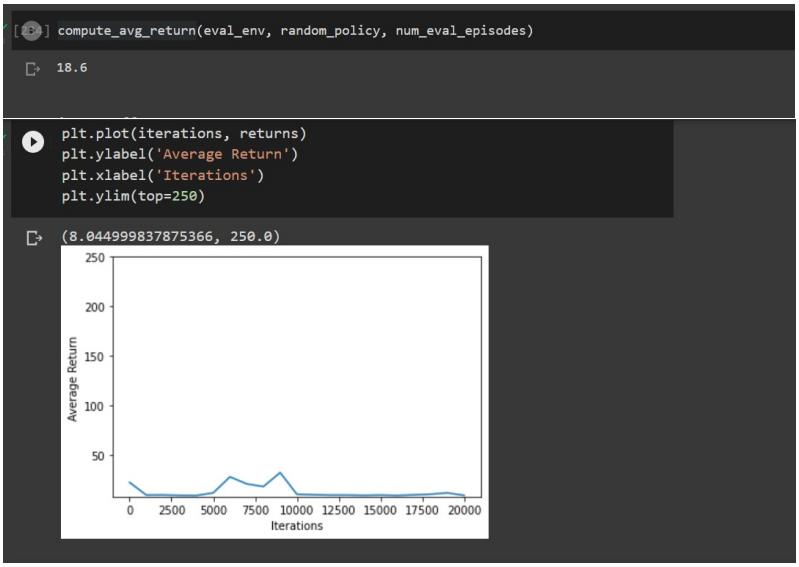

Figure 9: Activation Function: Cos_2

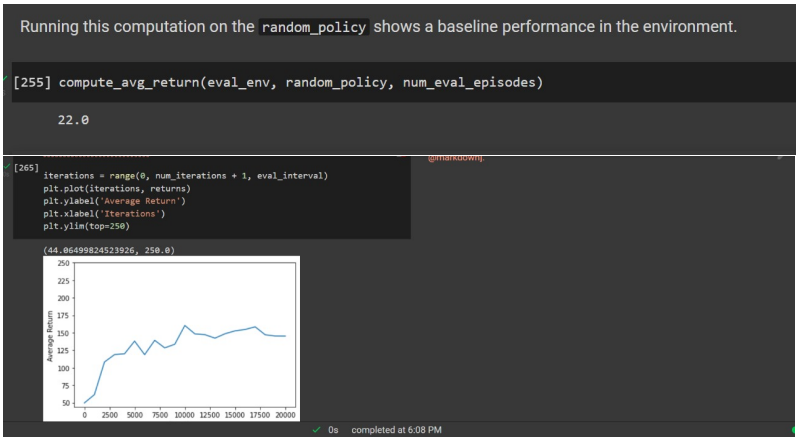

Figure 10: Activation Function: SSU

5 Conclusion

References

[1] Lennart Ante. Non-fungible token (nft) markets on the ethereum blockchain: Temporal development, cointegration and interrelations. *SSRN Electronic Journal*, 2021.

[2] Shubham Bharadwaj. Using convolutional neural networks to detect compression algorithms. *arXiv preprint arXiv:2111.09034*, 2021.

[3] Sarah Bouraga. On the popularity of non-fungible tokens: Preliminary results. *2021 3rd Conference on Blockchain Research amp; Applications for Innovative Networks and Services (BRAINS)*, 2021.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Logan Kugler. Non-fungible tokens and the future of art. *Communications of the ACM*, 64(9):19–20, 2021.

[7] Mathew Mithra Noel, Advait Trivedi, Praneet Dutta, et al. Growing cosine unit: A novel oscillatory activation function that can speedup training and reduce parameters in convolutional neural networks. *arXiv preprint arXiv:2108.12943*, 2021.

[8] Matthew Mithra Noel, Shubham Bharadwaj, Venkataraman Muthiah-Nakarajan, Praneet Dutta, and Geraldine Bessie Amali. Biologically inspired oscillating activation functions can bridge the performance gap between biological and artificial neurons. *arXiv preprint arXiv:2111.04020*, 2021.

[9] Andrew Park, Jan Kietzmann, Leyland Pitt, Amir Dabirian, and Amir Dabirian. The evolution of nonfungible tokens: Complexity and novelty of nft use-cases. *IT Professional*, 24(1):9–14, 2022.

[10] Wajiha Rehman, Hijab e Zainab, Jaweria Imran, and Narmeen Zakaria Bawany. Nfts: Applications and challenges. *2021 22nd International Arab Conference on Information Technology (ACIT)*, 2021.

[11] Sakib Shahriar and Kadhim Hayawi. Nftgan: Non-fungible token art generation using generative adversarial networks, 2021.