"Simulating Agent Behavior Using Reinforcement Learning"

Ofer Drori[1], Nir Sassy[1], Ben Engelhard[2]
[1] Department of Biomedical Engineering, Technion - IIT, Haifa, Israel
[2] Department of Medicine, Engelhard Lab, Technion - IIT, Haifa, Israel

Introduction

Reinforcement learning (RL) is a subfield of machine learning (ML) that focuses on developing algorithms that enable agents to learn optimal behavior by interacting with their environment. Both RL and **biological learning** utilize reward signals as behavior guidelines. RL agents and biological learners receive feedback from the environment in the form of rewards or punishments, enabling the learner to understand which actions lead to positive outcomes [1]. In ML, this feedback is commonly provided by assigning a reward signal for the agent's actions. In biological learning, the reward is provided by the release of neurotransmitters like dopamine in response to positive outcomes. According to the reward predicting error hypothesis, dopamine neurons encode the difference between expected reward and actual reward, also known as the "**Temporal Difference**", a concept utilized in RL loss functions [2]. This has led to the belief that RL could be a fundamental mechanism in bio-learning methods.

Objective

- Extract learning signals from RL agents trained using **DQN** and **PPO** algorithms to serve as a baseline for future comparisons in the lab.
- Establish an easy-to-use platform for conducting RL experiments.

Methods

This project mainly focused on the ML aspect of the research. It involved:

- Python Programming
- Reinforcement Learning algorithms – Deep Q Network (DQN) and Proximal Policy Optimization (PPO)
- PFRL's open-source library as a baseline [3]

RL algorithms are (mainly) divided into two categories: Value-based and Policy-based algorithms. We chose one 'representative' from each category which are DQN for the value-based and PPO for the policy-based

In **value-based** learning, we are giving a value for each state-action pair in the environment, denoted as Q(s, a), which represents the expected future rewards for taking an action 'a' in state 's'. In DQN, the Q-function is represented as a neural network that receives a state and returns the value for each available action as seen in Figure 3. The loss function for the algorithm is called the bellman's error which is a very similar concept to the temporal difference.

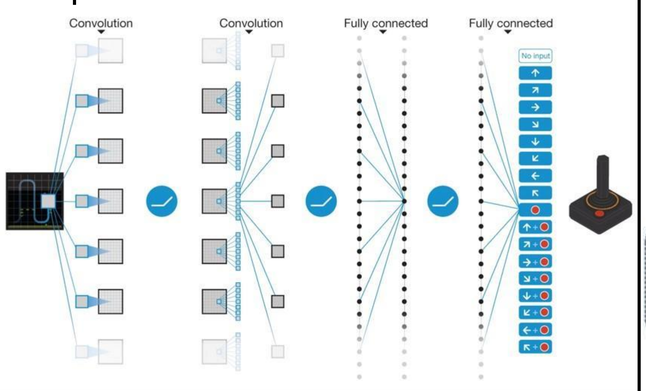

Fig 1 – The Q-Network architecture [4]

In **policy-based** learning, we are optimizing the policy that maps states to actions directly. PPO is an 'Actor-Critic' algorithm, where the critic learns the environment dynamics and constantly evaluates the actor's performance, while the actor explores the environment and takes actions based on the current policy [5]. The difference between the critic's prediction of the actor returns and the actual returns given is called the advantage (a TD error in a way) and is used to optimize the policy-gradient loss function.

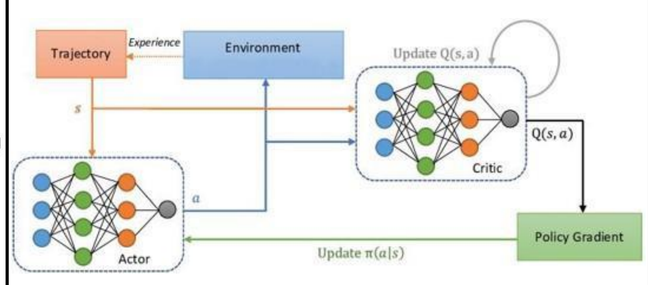

Fig 2 – The PPO algorithm flowchart

Results

Established a framework for conducting RL experiments using vast RL algorithms on several tasks.
We were able to train RL agents and extract their artificial learning signals. Those signals will be used for comparisons with learning signals extracted from biological learners.

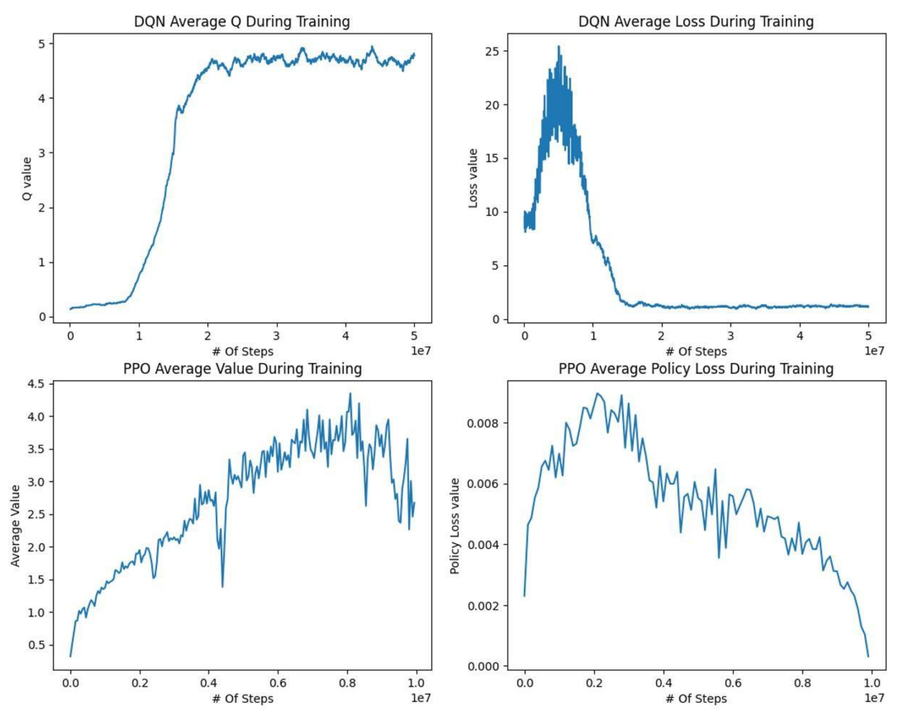

Fig 3 – Learning signals we retrieve from agents that played 'Atari – BreakOut' using PPO (down) and DQN (up).

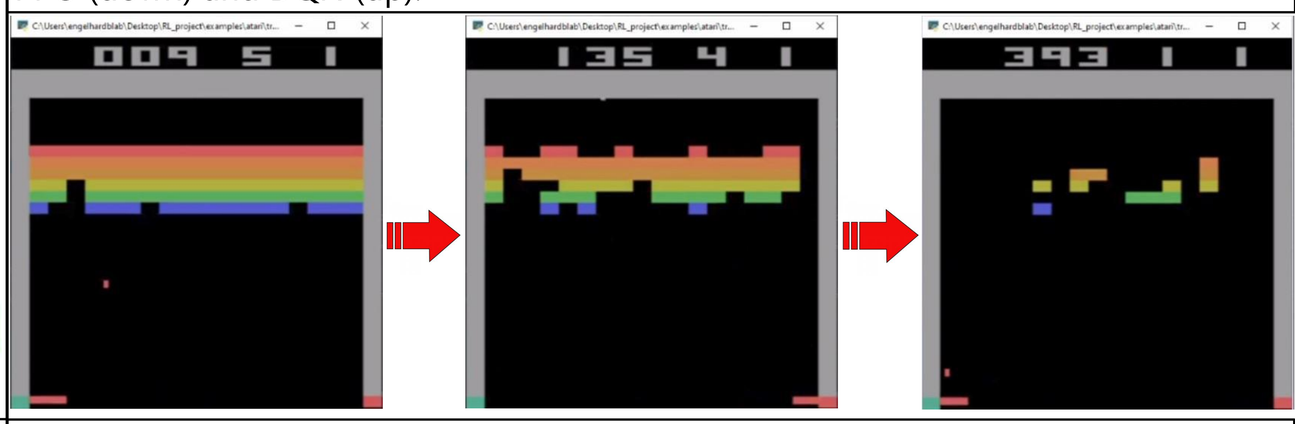

Fig 4 – RL agent we trained to play 'Atari – BreakOut' game using PPO algorithm

Discussion

The learning signals we managed to extract set an important milestone in the research that seeks to understand biological learning better. The signals presented here are time-dependent, meaning they are influenced by the training parameters (e.g., exploration decay rate). Therefore, our next step involves choosing the most informative learning signals based on the biological aspect of the research, training RL agents on the same task as the biological learners, and comparing the artificial learning signals with the dopamine secretion signals.

Summary

Deep RL algorithms were influenced by the principles of biological learning, and by that succeeded to train computer models to take optimal actions in unfamiliar environments. We utilized value-based and policy-based algorithms to train agents on various tasks and extracted artificial learning signals which will make a baseline for future comparisons with biological learning signals, such as dopamine secretion. Furthermore, our platform enables the conducting of further RL experiments if needed.

Acknowledgments

- Dr. Ben Engelhard
- Rotem Shapira
- Guy Sassy (A.I Researcher – ECE, Technion)
- Uriel Nusbaum (Algorithms Developer – Elbit)
- Behrooz Omidvar-Tehrani (Contributor - PFRL library)

References

[1] R. S. Lee, B. Engelhard, I. B. Witten, and N. D. Daw, "A vector reward prediction error model explains dopaminergic heterogeneity," bioRxiv, p. 2022.02.28.482379, Mar. 2022
[2] K. Doya, "Reinforcement learning: Computational theory and biological mechanisms," HFSP J, vol. 1, no. 1, p. 30, 2007.
[3] "PFRL: a PyTorch-based deep reinforcement learning library." https://github.com/pfnet/pfrl
[4] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature 2015 518:7540, vol. 518, no. 7540, pp. 529–533, Feb. 2015
[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. K. Openai, "Proximal Policy Optimization Algorithms," Jul. 2017