Biological Neurons vs Deep Reinforcement Learning: Sample efficiency in a simulated game-world

Forough Habibollahi *
Department of Biomedical Engineering
University of Melbourne
Melbourne, Australia

Moein Khajehnejad *
Department of Data Science and AI
Monash University
Melbourne, Australia

Amitesh Gaurav
Cortical Labs Pty Ltd
Melbourne, Australia

Brett J. Kagan
Cortical Labs Pty Ltd
Melbourne, Australia

Abstract

How do synthetic biological systems and artificial neural networks compete in their performance in a game environment? Reinforcement learning has undergone significant advances, however remains behind biological neural intelligence in terms of sample efficiency. Yet most biological systems are significantly more complicated than most algorithms. Here we compare the inherent intelligence of in vitro biological neuronal networks to state-of-the-art deep reinforcement learning algorithms in the arcade game 'pong'. We employed DishBrain, a system that embodies *in vitro* neural networks with in silico computation using a high-density multielectrode array. We compared the learning curve and the performance of these biological systems against time-matched learning from DQN, A2C, and PPO algorithms. Agents were implemented in a reward-based environment of the 'Pong' game. Key learning characteristics of the deep reinforcement learning agents were tested with those of the biological neuronal cultures in the same game environment. We find that even these very simple biological cultures typically outperform deep reinforcement learning systems in terms of various game performance characteristics, such as the average rally length implying a higher sample efficiency. Furthermore, the human cell cultures proved to have the overall highest relative improvement in the average number of hits in a rally when comparing the initial 5 minutes and the last 15 minutes of each designed gameplay session.

1 Introduction

The concept of reinforcement learning dates back to the early days of cybernetics and has been studied in statistics, psychology, neuroscience, and computer science. In the past decade, its use has become increasingly popular in the fields of machine learning and artificial intelligence. Its promise is highly convincing - a way of programming agents by rewarding and punishing them without having to specify how the task is to be accomplished. However, to deliver on this promise, formidable computational obstacles must be overcome. Reinforcement learning (RL) implies learning the best policy to maximize an expected cumulative long-term reward throughout many steps in order to achieve complex objectives (goals) [1]. A deep reinforcement learning (deep RL) approach integrates artificial neural networks with a reinforcement learning framework that helps the system to achieve its goals [2]. That is, it maps states and actions to the rewards they bring, combining

36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, USA.

*Indicates equal contribution.

function approximation and target optimization. Reinforcement algorithms that incorporate deep neural networks have been developed to beat human experts playing numerous Atari video games [3], poker [4], multiplayer contests [5], and complex board games, including go and chess [6, 7, 8]. Nevertheless, reinforcement learning still faces real challenges including but not limited to complexities in the selection of reward structure, sample inefficiency [9, 10], reproduciblity issues [11], as well as requiring high levels of computing power [12]. All of these suggest that deep RL algorithms may differ fundamentally from the underlying mechanisms of human learning while also being too inefficient to be accepted as plausible models of human learning [10].
It was recently demonstrated that by using electrophysiological stimulation and recording in a real-time closed-loop system with a monolayer of living biological neurons, these cells could be trained to significantly improve performance in the simulated 'pong' gameworld [13]. The question arises as to whether this observed performance is notable in comparison to that of reinforcement learning at the same task. To compare the performance and efficiency of such a biological neuronal network (BNN) to that of deep RL, we use data gathered from the *DishBrain* system [13] against time-matched learning from DQN, A2C & PPO algorithms. *DishBrain* is a novel system shown to display biological intelligence by harnessing the inherent adaptive computation of neurons within a simulated gameplay environment in real time through closed-loop stimulation and recordings. In this system, *in vitro* neuronal networks are integrated with *in silico* computing via high-density multi-electrode arrays (HD-MEAs). We investigate whether these elementary learning systems achieve performance levels which can compete with state-of-the-art deep RL algorithms while varying the input information density required for training the RL algorithms to also determine the impact of information sparsity and ensure suitable comparisons to the biological system. This is the first comparison between a synthetic biological intelligence system and state-of-the-art RL algorithms.

2 Methods

2.1 DishBrain System

To investigate the learning efficiency of the BNNs in the task-present state, recordings from cultures integrated onto an MEA were used. The *DishBrain* environment is a low latency, real-time system which interacts with the MEA (Maxwell Biosystems, Switzerland) software to allow closed-loop stimulation and recording. *DishBrain* was utilised to embody neural cultures in a virtual game-world, to simulate the arcade game 'Pong'. Sensory stimulation was applied into a predefined bounded two-dimensional sensory area consisting of 8 sensory electrodes to communicate ball's position on the x and y-axis using a combination of rate coding (4Hz - 40Hz) electrical pulses and place coding, respectively. The movement of the paddle was controlled by the level of electrophysiological activity measured in a predefined "motor area", which was recorded in real time. The cells also received information about the closed-loop response to their control of the paddle.
It was possible to deliver five types of input. Either the sensory stimulation as explained above, or one of four feedback protocols: Unpredictable, Predictable, Silent, or No-feedback. The reported results in this work are obtained using the unpredictable feedback protocol. Cultures received unpredictable stimulation when they missed connecting the paddle with the 'ball', i.e. when a 'miss' occurred. Using a feedback stimulus at a voltage of 150 mV and a frequency of 5 Hz, unpredictable external stimulus could be added to the system. Random stimulation took place at random sites over the 8 predefined sensory electrodes at random timescales for a period of four seconds, followed by a configurable rest period of four seconds where stimulation paused, then the next rally began. Each recording session of the cultures was 20 minutes. This equaled an average number of 70 training episodes.

Figure 1 illustrates the the input information, feedback loop setup, and electrode configurations in the DishBrain system.

2.2 Deep Reinforcement Learning Algorithms

We use three state-of-the-art deep reinforcement learning algorithms: Deep Q Network (DQN) [3], Advantage Actor-Critic (A2C) [14] and Proximal Policy Optimization (PPO) [15], established to have good performance in Atari games. Benefiting from deep learning advantages in automated feature extraction, specifically exploiting Convolutional Neural Networks (CNN) in their structures, these methods are robust tools in reinforcement tasks, particularly in games where the system's input

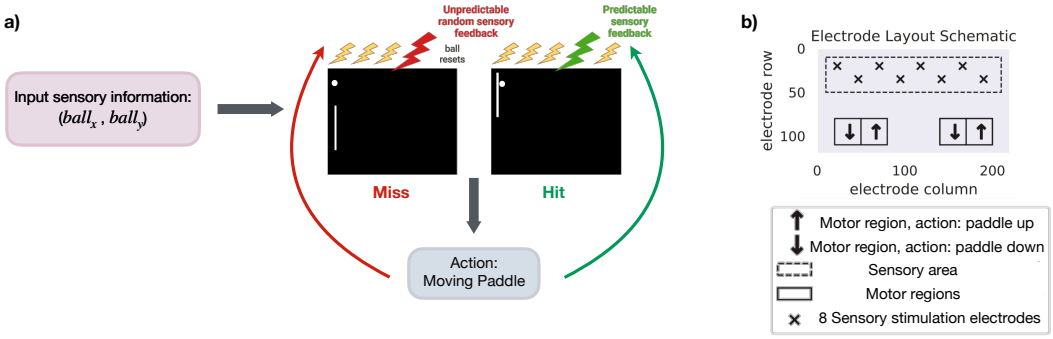

Figure 1: a) DishBrain feedback loop setup. **b)** Electrode configuration and predefined sensory and motor regions. Figures adapted and modified from [13]

is an image. We only report the results of the deep RL algorithms in a design where the current state is a tensor of the difference of pixel values from the two most recent frames (i.e. another 40×40 grayscale pixel image). This current state is then input into the CNN to obtain the selected action. However, to account for potential adversaries resulting from the high dimensionality [16] of the `IMAGE` input, we also designed two additional types of low-dimensional input information called `PADDLE & BALL POSITION`: $[ball_x, ball_y, paddle_{top_y}, paddle_{bottom_y}]$; and `BALL POSITION`: $[ball_x, ball_y]$, where $ball_y \in \{1, 2, \cdots, 8\}$ as we divide the y-axis to 8 equal segments mimicking the 8 sensory electrodes which place code the $ball_y$ in the biological cultures. We compared all three different designs with the performance of biological cultures and observed no significant difference in the outcome of these comparisons. In the training phase of all RL algorithms, we ran them for 40 random seeds and a total of 70 episodes for each seed (similar to BNNs). These seeds imply 40 different neural networks trained separately, resembling 40 different recorded cultures. We report the average value of each metric among all seeds.

Figure 2 illustrates the comparison between the input information in the DishBrain system and the deep RL algorithms.

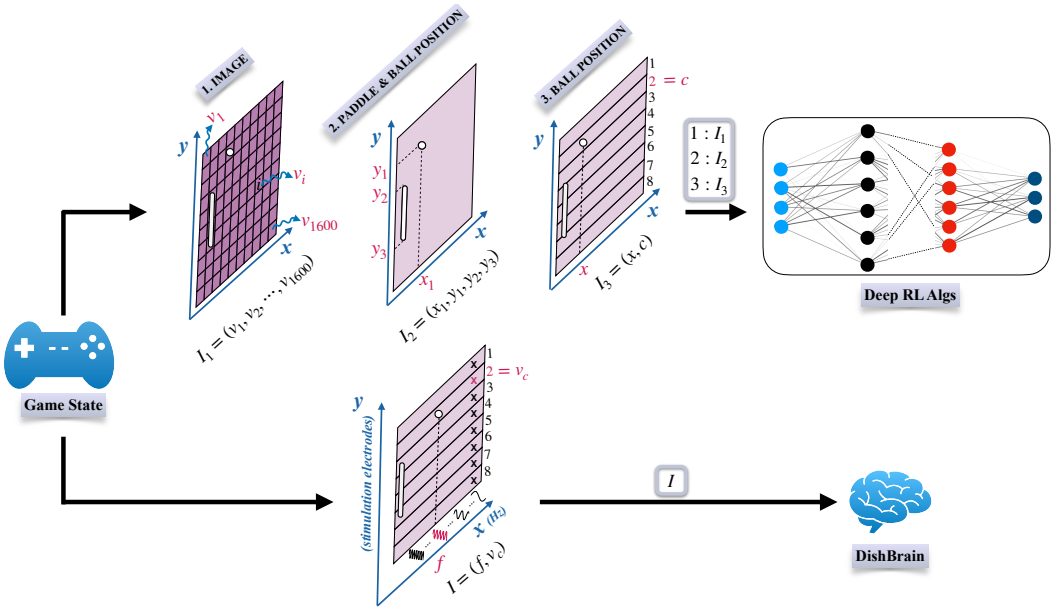

Figure 2: Schematic comparing the information feeding routes in the DishBrain system (bottom) and the three implementations of the deep RL algorithms (top). In each design, the input information to the computing module (deep RL algorithms or DishBrain) is denoted by a vector I.

3 Results

We studied both human cortical cells (HCCs; 174 sessions) and mice cortical cells (MCCs; 110 sessions) and compared them to the introduced RL baseline methods. The reported results in this section are obtained using the `IMGAE` input design for the RL methods. To determine how the learning arises both in the cultures and the baseline methods, key gameplay characteristics were examined. The hit counts in the gameplay in each episode before the ball was missed for the first time, the number of times the paddle failed to intercept the ball on the initial serve (aces), and the number of long rallies (> 3 consecutive hits) were calculated for this data. For comparison purposes, we first mapped every 70-episode run of each RL algorithm to a real-time equivalent of 20 minutes by first normalizing to the actual total length of each run in minutes and then multiplying by 20 minutes.

The DQN algorithm is outperformed by all groups in the highest level of average hits per rally

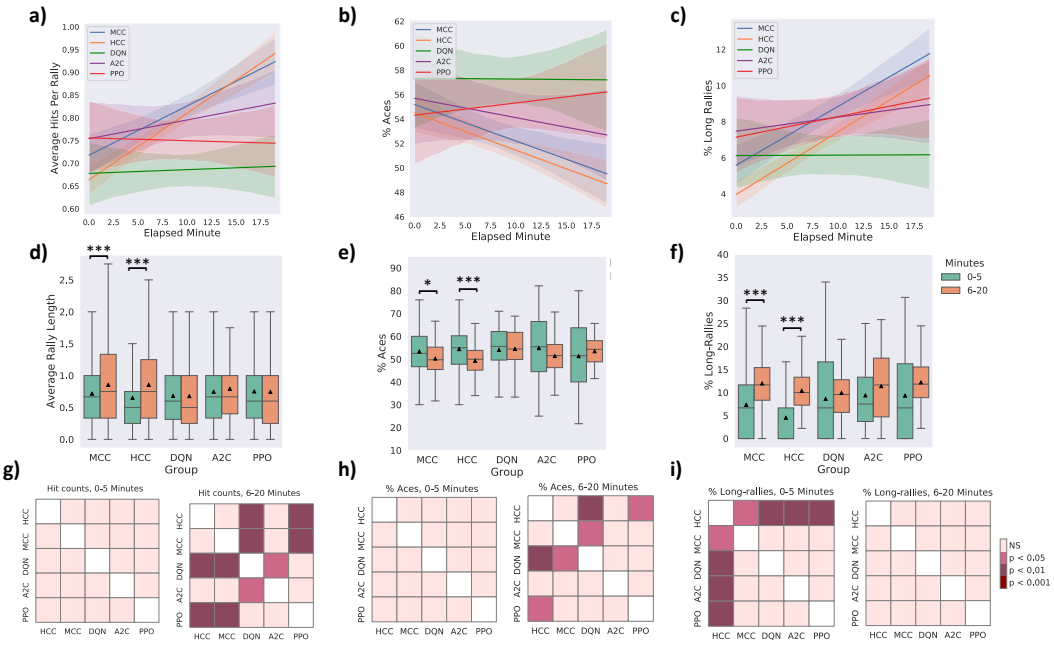

Figure 3: **a)** Average number of hits per rally, **b)** % aces, and **c)** % long rallies over 20 minutes real-time equivalent of training RL algorithms and biological cultures. A regressor line on the mean values with a 95% confidence interval highlights the learning trends. **d)** Average rally length in the first 5 minutes and last 15 minutes of the sessions. **e)** Average % of aces within groups and over time. **f)** Average % of long-rallies (>3) performed in each interval. **g, h, and i)** Pairwise Tukey's post hoc test among groups in each time interval and for **g)** average rally length, **h)** % aces, and **i)** % long rallies. Box plots show interquartile range, with bars demonstrating 1.5X interquartile range, the line marks the median and ▲ marks the mean. Error bands = 1 SE

achieved, while the biological cultures (i.e. HCC and MCC) outperform all the RL baselines (see Subfigure 3.a). This indicates that the cultures represent faster growing learning rates. Subfigure 3.b compares the % of missed balls on the initial serve, i.e. aces. HCC and MCC achieve the lowest % of aces in Subfigure 3.b. The % of long rallies has an increasing trend in all groups with the highest levels achieved by MCC and HCC as illustrated in Subfigure 3.c.

Next, for all the groups, we compared the key activity metrics in the first 5 minutes versus the last 15 minutes in each session. Our aim was to identify any significant improvement in the learning process within each group. Subfigure 3.d compares the average rally length between the two defined time intervals. The results imply that the intra-group improvement in the length of rallies is significant only in the biological groups (One-way ANOVA test). Subfigure 3.e represents the change in the average % of aces over time. A significant decrease in the number of aces implies an improved game performance. Only MCC and HCC groups had a significant decrease in the average % of aces (One-way ANOVA test). Subfigure 3.f shows that % of long rallies in the first 5 minutes versus the last 15 minutes only significantly increased for the biological cultures (One-way ANOVA test).

Pairwise inter-group comparison was carried out for both time intervals and all metrics using Tukey's post hoc test represented in Subfigures 3.g, h, and i for hit counts, % of aces, and % of long rallies.

It should be noted that while certain metrics of the performance of the deep RL methods comes closets to the biological cultures, the density of input information is starkly different between RL methods and the biological cultures. While RL agents receive pixel data with a density of 40×40 pixels, biological cultures only receive input from 8 stimulation points with a given integer rate code of 4Hz–40Hz, highlighting important efficiency differences in informational input between these learning systems. The possibility of the higher input information dimensionality having adverse effects on the overall sample efficiency of these RL algorithms was further nullified by evaluating the two alternative input structures as discussed above.

To account for potential effects of paddle movement speed on the success rate of paddle control, we derived the average paddle movement (in pixels) for all groups. Subfigure 4.a represents these results with DQN having a significantly higher average paddle movement compared to biological cultures (Pairwise Tukey's post hoc). Interestingly, the higher paddle movement speed of the RL algorithms is not reflected as better game performance according to our results.

Subfigure 4.b compares the relative improvement in the performance of different groups over time. This measure identifies the relative increase in the average accurate hit counts in the second 15 minutes of the game compared to the first 5 minutes. The HCC group shows the highest improvement in time and performing Tukey's post hoc tests showed that the difference in this measure is significant between HCC and PPO, as well as HCC and DQN. The MCC group also outperforms DQN.

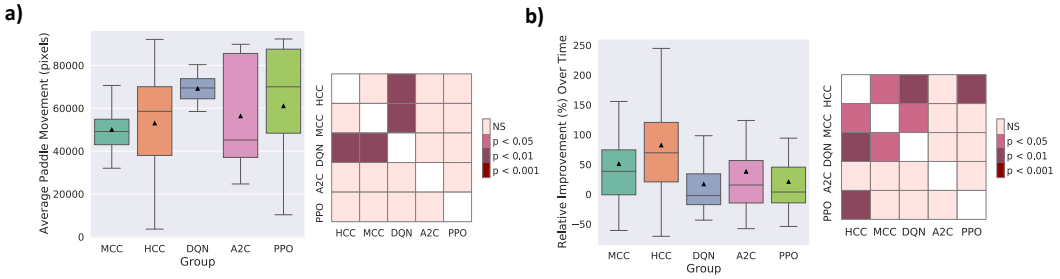

Figure 4: **a)** The average paddle movement in pixels and pairwise Tukey's post hoc test representing the significance of the differences. **b)** Relative improvement (%) in the average hit counts between the first 5 minutes and the last 15 minutes of all sessions in each separate group and pairwise Tukey's post hoc test.

4 Discussion

In this work, we compared the performance of BNNs with that of state-of-the-art deep RL algorithms in the game environment of *pong*. The results show that the game performance of the deep RL algorithms in terms of relative learning improvement in time and the ultimate number of average hits per rally is outperformed by biological cultures. Furthermore, their performance in the average rally length and percentage of aces only matches those of neuronal cultures at best. The RL algorithms showed the lowest sample efficiency having the lowest improvement in learning given the 70 episode training duration provided for all the groups.

This is the first comparison between a synthetic biological intelligence system and state-of-the-art RL algorithms. This early work establishes that even the most rudimentary SBI systems with limited informational input are a viable learning system that can compete and even defeat the established RL algorithms which receive significant more information input. Coupled with the promise of significant gains in power efficiencies, flexibility of tasks, and as data representation to the SBI system is improved, these biological intelligence systems present a compelling pathway for realizing real-time learning unachievable by current silicon-based approaches.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Matteo Hessel et al. "Rainbow: Combining improvements in deep reinforcement learning". In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[3] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[4] Matej Moravčík et al. "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker". In: *Science* 356.6337 (2017), pp. 508–513.

[5] M Jaderberg et al. "Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. arXiv". In: *arXiv preprint arXiv:1807.01281* (2018).

[6] David Silver et al. "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815* (2017).

[7] David Silver et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[8] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[9] Pedro A Tsividis et al. "Human learning in Atari". In: *2017 AAAI spring symposium series*. 2017.

[10] Gary Marcus. "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631* (2018).

[11] Elizabeth Gibney et al. "This AI researcher is trying to ward off a reproducibility crisis". In: *Nature* 577.7788 (2020), pp. 14–14.

[12] Seyed Sajad Mousavi, Michael Schukat, and Enda Howley. "Deep reinforcement learning: an overview". In: *Proceedings of SAI Intelligent Systems Conference*. Springer. 2016, pp. 426–440.

[13] Brett J Kagan et al. "In vitro neurons learn and exhibit sentience when embodied in a simulated game-world". In: *Neuron* (2022).

[14] Kai Arulkumaran et al. "Deep reinforcement learning: A brief survey". In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 26–38.

[15] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[16] Richard Bellman and Robert Kalaba. "Dynamic programming and statistical communication theory". In: *Proceedings of the National Academy of Sciences* 43.8 (1957), pp. 749–751.