Exploring Advanced Techniques in Reinforcement Learning for Atari Breakout

Adam Grabowski, Siqin Li, Vasilii Gorbunov, Antony Garcia

Abstract

In this study, we address performance stagnation in a Deep Q-Network (DQN) agent with a convolutional neural network (CNN) architecture, tasked with playing Atari Breakout. Initially, the agent demonstrated learning progress for the first 10 million sampling cycles, achieving up to 400 points. Subsequent performance plateaued, attributed to the game's advanced stages presenting fewer bricks and a propensity for the agent to get trapped in local optima or exhibit suboptimal play. Our approach entailed two primary strategies: refining the training loop and experimenting with advanced neural network models. We enhanced the training loop by prioritizing high-scoring events and simplifying the state inputs—stripping away non-critical features such as score display, color, and borders. This led to improved agent performance in later game stages, indicating a better understanding of critical game states. We also explored the integration of more sophisticated models like Vision Transformers, U-Nets, and Spatial-Shift MLPs (S2-MLP). These models did not outperform the enhanced CNN-DQN, but the Convolutional Vision Transformer showed promise, suggesting that with extended training, it could surpass the baseline model.

1. Introduction

The computational problem at the core of our research is the plateau in performance of a Deep Q-Network (DQN) agent trained to play Atari Breakout. This issue is significant because it highlights a limitation in the learning capability of DQN agents, which are foundational to many applications of reinforcement learning (RL). Previous research has addressed such stagnation by adjusting the reward function, enhancing exploration strategies, or increasing complexity. However, these solutions often lead to incremental improvements without addressing the underlying issue of the agent's inability to learn from advanced stages of the game.

Existing approaches tend to focus on model tuning or reward shaping, which do not fundamentally alter the agent’s engagement with critical game states. As such, these methods may not be adequate for scenarios where the game's dynamics change significantly as progress is made, such as the final stages of Atari Breakout, where fewer bricks remain, and strategic play becomes essential. The necessity for a new approach is underscored by the need to overcome the inherent limitations of CNN-based DQN models. These models struggle with complex, sparse environments that require advanced strategy over brute-force pattern recognition.

Our research presents a novel approach, focusing on the qualitative aspects of the agent's interaction with the game, could provide a more substantial advance. By redefining the agent's learning process to prioritize meaningful, high-scoring experiences and simplifying input functions, we aim to encourage the agent to develop a more nuanced understanding of game mechanics. Also, we explore the potential of leveraging advanced neural network architectures, which could offer more sophisticated processing of game states. This approach promises to not only address the challenge of performance stagnation, but also contribute to the broader field of RL.

1.1 Research Contributions

We implemented a prioritized replay memory approach within a single CNN-based DQN model to achieve high scores in Atari Breakout that have not been documented prior to our research. This represents a significant advancement. Expanding on this innovation, we explored the application of image classification architectures—U-Nets, Spatial-Shift MLPs (S2-MLP), and Vision Transformers—in reinforcement learning settings. These architectures

have been predominantly applied in supervised learning contexts and, in the case of S2-MLP and U-Nets, have not been previously applied to RL tasks. The Vision Transformer, while it has seen limited use in RL, has not been leveraged in this manner. Our research is the first to test these models against the premise that if a model can reduce its performance loss to under 0.1% relative to a trained DQN model, it can be considered effective for RL tasks.

2. Related Work

The RL landscape for Atari Breakout has seen a variety of DQN adaptations aimed at improving agent performance. Notably, [1] introduced the Double DQN (DDQN) model, which mitigates the overestimation of action values by decoupling selection and evaluation of the action-value function. This yielded a score of 418.5 points, demonstrating improved stability over the DQN model [2].

An Actor-Shared-Learner architecture, known as ASL DDQN, introduced in [3], combined policy gradient and value function methods, achieving a score of 621 points. The Quantile DQN (QR-DQN-1), proposed in [4] employed distributional RL to predict a full value distribution instead of an expected value, leading to a performance of 742 points. This approach provided a richer learning signal at the cost of increased computational demand.

While these approaches improved upon the DQN's performance through architectural enhancements, they do not focus on optimizing the data quality used for training. So, the models may still rely on potentially noisy or uninformative experiences during training. This could hinder the agent's ability to learn optimal strategies, especially in the complex end-game scenarios of Atari Breakout.

In the context of Transformers, their application has been limited to the reinforcement learning domain for Atari games. Papers reporting the use of Decision Transformers [6] and State-to-Go Transformers [7] have shown modest success, with scores of 267 and 288 points, respectively. These Transformer-based approaches aim to leverage the model's ability to handle sequential data. However, they have not been fully explored in the context of Atari Breakout nor have they been optimized for the kind of data efficiency and focus that our research proposes. U-Net can be employed for processing video input in RL. Here, video summarization through RL with a 3D spatio-temporal U-Net was implemented [8].

Considering these observations, our research is motivated by the potential to leverage both data optimization and advanced architectures, such as Transformers, to address the shortcomings of prior work. By refining the training data quality and exploring the application of typically non-RL models like U-Nets, S2-MLPs [9], and Vision Transformers [10], we aim to achieve new high scores.

3. Proposed Method

3.1 Deep Q-Learning

Deep Q-network is the baseline model used here. Reinforcement learning is a subfield of machine learning by training agent (appendix a). The network consists of three convolutional layers used for feature extraction, followed by two fully connected layers. Our approach to enhancing a pretrained DQN model, adept at scoring 420 points in Atari Breakout (appendix b), revolved around two strategic modifications: optimizing the training data and streamlining the model's input. We employed a novel probability function in the replay memory system, which favored high-scoring episodes (+400 points).

Simultaneously, we simplified the model's input by removing extraneous visual elements such as the scoreboard, frame borders, and color (appendix c). This reduction to black and white frames focuses the model's attention on the game's critical components - the ball, paddle, and bricks. This streamlined input not only reduces computational complexity but also sharpens the model's capacity for strategic gameplay, promising enhanced learning efficiency and improved performance.

3.2 Vision Transformer

Initial experimentation with Vision Transformers (ViTs) in the context of Atari Breakout, a game that utilizes four consecutive 84x84 pixel frames to represent the state, met challenges. The conventional ViT, as proposed in [9], was integrated to replace the CNN in our DQN model (appendix d). However, this replacement did not yield the desired results. The

ViT struggled to minimize the loss in comparison to the pre-trained CNN-based DQN, indicating a difficulty in capturing the necessary spatial and temporal features for the game. This was a critical insight, as the dynamics of Atari Breakout heavily rely on understanding game elements over time.

To improve performance, various approaches were tested to enhance the representation of the game's state. These included utilizing color channels, arranging the four frames in a 2x2 grid to create a larger 168x168 input, and stacking images with varying transparencies to track the ball's trajectory. Despite these efforts, the model failed to show much improvement in learning the game's dynamics. This led us to explore an alternative approach using a Convolutional Vision Transformer (Convolutional ViT) developed by Facebook Research. This variant was more adept at recognizing the temporal aspects of the game, particularly through its use of color channels. We propose that the Convolutional ViT could effectively learn in an online training setup with the Atari Breakout game.

3.3 Spatial-Shift MLP

The S2-MLP architecture is a variant of the MLP (Multilayer Perceptron) designed specifically for spatial information processing (appendix e). The patch size was set at 6×6 so the input image (84×84) is divided into $(84/6)^2=196$ non-overlapping patches. Each patch undergoes a fully connected layer to obtain a patch embedding vector. Each S2-MLP block (N=36) consists of four fully connected layers (M=384), two GELU (Gaussian Error Linear Unit) layers, two layer normalizations, two skip connections, and a spatial-shift module. The spatial-shift module involves splitting channels into groups and shifting these groups in different directions, creating a local reception field. The operation is parameter-free, making it computationally efficient (appendix f). S2-MLP processes non-overlapping patches, utilizing the spatial-shift module to allow for communication between adjacent patches.

3.4 U-net for moving object detection

With this approach we address an input sequence of 2D images as a 3D array with time as the third dimension. We claim that 3D convolution will train the model to recognize static and moving objects via different profiles in 3D. We also line out the importance of positional recognition that is lost in classic max-pooling processes. This architecture includes skip connections that help to keep the positional data for action selection (appendix f). The encoder includes one 3D convolution with 16 filters size (channels=1,time=4, x=5, y=5) followed by max pooling. The convolution feature map and polling feature map are then concatenated and passed to a one-layer softmax classifier. We pretrained the model based on performance of a previous one, then trained it with DQN approach. We chose to trade model complexity for higher performance.

4. Experiment

To evaluate our techniques, we utilized a standardized DQN training regimen for the Vision Transformer (ViT), U-Net, and S2-MLP models. Training involved a lengthy process spanning 20 million episodes, segmented into an initial exploration phase with 500,000 random samples, followed by an epsilon decay phase over 2 million samples, and culminating in an exploitation phase for the remaining 17.5 million episodes with a consistent 5% probability for random actions. For the DQN, we employed a pretrained version. Initially, the model underwent a data collection phase, accumulating 500,000 samples from episodes scoring above 400 points. Once this specialized replay memory was filled, we applied a refined training strategy. This strategy employed a weighted probability system for incorporating data into the model, ensuring an emphasis on higher-scoring experiences.

5. Results

5.1 CNN-based DQN model

The implementation of the weighted experience replay memory significantly enhanced the performance of our pretrained model. This strategic adjustment resulted in the model consistently achieving an impressive average score of 490 points over five episodes during evaluation. Moreover, in certain individual test runs, the model remarkably reached scores as high as 816 points. This achievement not only represents a notable improvement over the original model's capabilities

but also surpasses the outcomes of several previously established DQN-based models, demonstrating the efficacy of our refined approach in optimizing reinforcement learning performance.

5.2 Convolutional Vision Transformer

Experimentation with the Convolutional Vision Transformer yielded encouraging results. Initially configured with 8 attention heads, a 7-pixel patch size, an embed size of 512, and a depth of 4 layers, the model demonstrated a significant learning capacity of over 20 million episodes. This setup enabled the Convolutional ViT to achieve a score exceeding 260 points across five episodes. Further tuning of its parameters led to even better outcomes. Adjusting the configuration to a 7-pixel patch size, 36 attention heads, a reduced embed size of 72, and doubling the depth to 8 layers, showed marked improvement. This refined setup allowed the model to score over 300 points on average after 10 million sampled experiences. While these results did not surpass the performance of the CNN-based DQN, they represented great advancement over previous transformer-based models applied to Atari Breakout (appendix g and h). This result underscores the potential of Convolutional ViT as a viable alternative in RL scenarios, particularly in tasks that require sophisticated spatial and temporal feature extraction.

5.3 U-network

The model demonstrated an inherent ability to learn and adapt (95% accuracy in pretraining), yet it encountered a notable limitation in its performance. After an extensive training period spanning 20 million episodes, the model consistently plateaued, unable to exceed an average score of 30 points across five games. While there were occasional instances where the model achieved scores above 75 points, these occurrences were infrequent and not indicative of sustained improvement or learning consistency (appendix i). While the model possesses basic learning capabilities, it struggles to advance beyond a certain proficiency level in the game, highlighting a potential area for further investigation.

5.4 Spatial-Shift MLP

The performance of our S2-MLP model was limited, as it struggled to surpass the basic threshold of random movement, typically scoring between 2 to 4 points (appendix j). This outcome indicates a fundamental challenge in the model's learning capability, particularly in adapting to the strategic requirements of this task. Also, the complexity of training this model further compounded the issues, suggesting its unsuitability for the specific demands of this reinforcement learning scenario.

6. Discussion

Our work, while pioneering in certain aspects, is not without limitations. A primary constraint lies in the scalability and generalizability of our models across different reinforcement learning environments. The techniques we employed, especially the weighted experience replay memory and specific architectural optimizations, were tailored to the Atari Breakout game. This specificity raises questions about how these methods would perform in other contexts.

The utilization of U-Net for moving object detection presents a promising approach to RL with video input. To address our computational limitations and achieve a desirable speed, the model underwent significant simplification. Despite this, the model attained a reasonable level of accuracy. The success of the weighted experience replay memory in enhancing the performance of the CNN-based DQN model suggests that focusing on the quality of training data can be as critical as architectural innovations. Our exploration of advanced models in RL contexts opens new avenues for using them beyond conventional applications.

7. Conclusion and Future Work

Our paper's key findings highlight the effectiveness of an enhanced CNN-based DQN model using weighted experience replay memory, achieving remarkable scores in Atari Breakout. For future research, integrating this refined data approach with the Convolutional Vision Transformer (ViT) could be promising, as ViT may benefit significantly from higher quality data. Additionally, exploring ways to improve feature extraction capabilities of the U-Net model presents another intriguing direction, potentially unlocking new levels of performance in complex reinforcement learning environments.

8. References

[1] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling Network Architectures for Deep Reinforcement Learning," in Proc. 33rd Int. Conf. Mach. Learn. - Volume 48, ICML'16, pp. 1995-2003, 2016.

[2] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," in Proc. AAAI Conf. Artif. Intell., vol. 30, no. 1, 2016.

[3] Y. Zhou, Z. Liu, and H. Sun, "Train a Real-world Local Path Planner in One Hour via Partially Decoupled Reinforcement Learning and Vectorized Diversity," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 12345-12354, 2021.

[4] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional Reinforcement Learning with Quantile Regression," in Proc. 32nd AAAI Conf. Artif. Intell., pp. 2892-2901, 2018.

[5] Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep Exploration via Bootstrapped DQN," in Proc. Advances in Neural Inf. Process. Syst., vol. 29, 2016.

[6] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Goodfellow, "Decision Transformer: Reinforcement Learning via Sequence Modeling," arXiv preprint arXiv:2106.01345, 2021.

[7] J. Yu, X. Tan, B. Gong, C. Wang, J. Yang, D. Tao, and Q. Wu, "StARformer: Transformer with State-Action-Reward Representations for Visual Reinforcement Learning," arXiv preprint arXiv:2112.00162, 2021.

[8] B. Zhao, X. Li, and X. Lu, "Video Summarization through Reinforcement Learning with a 3D Spatio-Temporal U-Net," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), pp. 1-6, 2020.

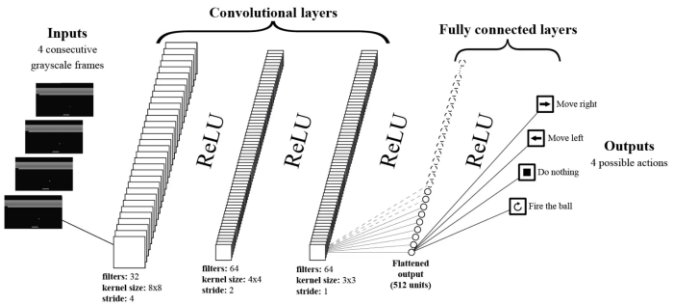
[9] F. Fan, F. Meng, J. Yang, M. Li, P. Zhou, A. T. S. Wee, and X. Liu, "S2-MLP: Spatial-Shift MLP Architecture for Vision," arXiv preprint arXiv:2106.07477, 2021.

[10] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

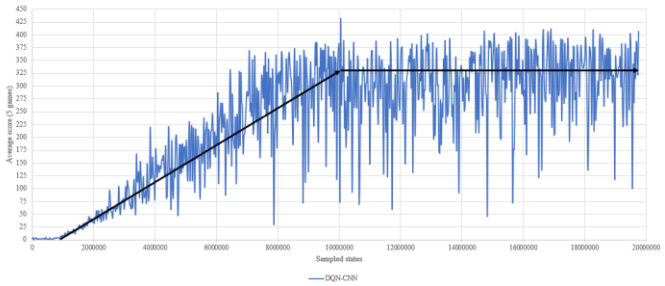
[11] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases," arXiv preprint arXiv:2103.10697, 2021.

Appendices

a. DQN base model Convolutional Neural Network Architecture

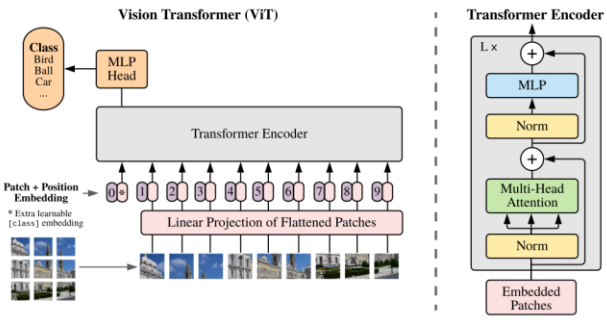


b. Base model performance over sampled states

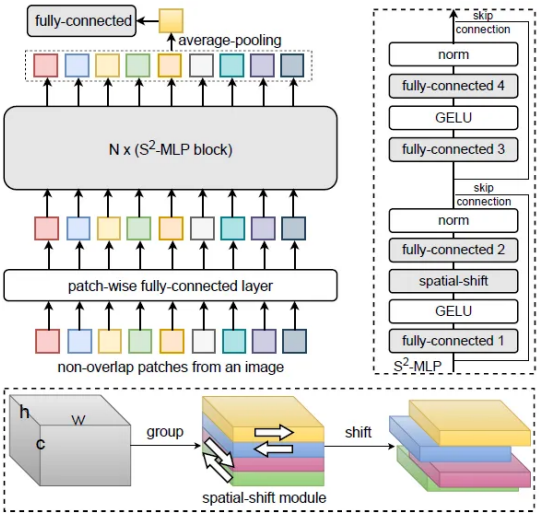


c. Input simplification techniques used to improve CNN-based DQN model

d. Vision Transformer (ViT) architecture



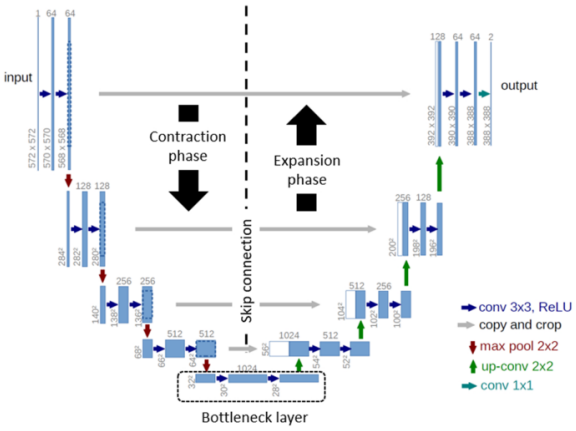
e. Spatial-Shift MLP (S2-MLP) architecture



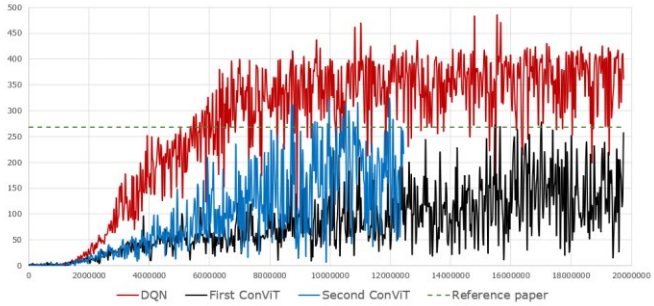
f. Spatial-shift operation

$$\mathcal{T}_1[1:w,:,:] \leftarrow \mathcal{T}_1[0:w-1,:,:],$$
$$\mathcal{T}_2[0:w-1,:,:] \leftarrow \mathcal{T}_2[1:w,:,:],$$
$$\mathcal{T}_3[:,1:h,:] \leftarrow \mathcal{T}_3[:,0:h-1,:],$$
$$\mathcal{T}_4[:,0:h-1,:] \leftarrow \mathcal{T}_4[:,1:h,:].$$

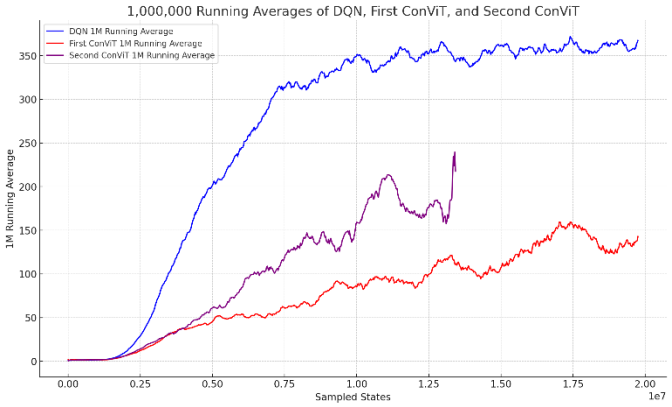
g. U-net architecture



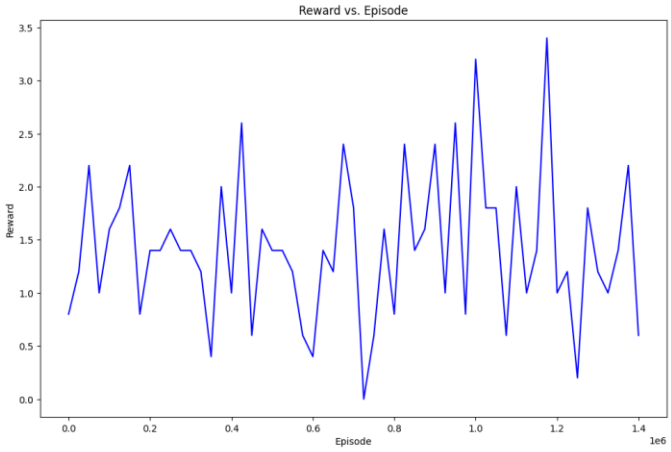
h. Results obtained with multiple ConViT tests



i. Averaged results obtained with the multiple ConViT tests



j. Results obtained with the S2-MLP model



k. Results obtained with the U-Net architecture (pretraining and DQN)