

Optimizing Deep Q-Learning Experience Replay with SHAP Explanations: Exploring Minimum Experience Replay Buffer Sizes in Reinforcement Learning

Robert S. Sullivan^{1,*}, Luca Longo^{1,2}

¹Artificial Intelligence and Cognitive Load Research Lab

¹School of Computer Science, Technological University Dublin

Abstract

Explainable Reinforcement Learning (xRL) faces challenges in debugging and interpreting Deep Reinforcement Learning (DRL) models. A lack of understanding for internal components like Experience Replay, which samples and stores data from the environment, risks burdening resources. This paper presents an xRL-based Deep Q-Learning (DQL) system using SHAP (SHapley Additive exPlanations) to explain input feature contributions. Data is sampled from Experience Replay, creating SHAP Heatmaps to understand how it influences the neural network Q-value approximator's actions. The xRL-based system aids in determining the smallest Experience Replay size for 23 simulations of varying complexities. It contributes an xRL optimization method, alongside traditional approaches, for tuning the Experience Replay size hyperparameter. This visual and creative approach achieves over 40% reduction in Experience Replay size for 18 of the 23 tested simulations, smaller than the commonly used sizes of 1 million transitions or 90% of total environment transitions.

Keywords

Deep Reinforcement Learning, Experience Replay, SHapley Additive exPlanations, eXplainable Artificial Intelligence

1. Introduction

Deep Reinforcement Learning (DRL) can optimise complex control and decision-making processes. However, it lacks explainability, limiting its widespread use in regulated environments like manufacturing, finance and medicine, where rising cost, safety and ethical concerns exist. Experience Replay is an internal DRL sampling technique, inspired by neurons during sleep [1], to break data correlation and stabilise deep off policy learning. Although Explainable Reinforcement Learning (XRL) is emerging, Deep Q-Learning (DQL) is challenging to debug and interpret with inefficiencies that burden resources, cause unnecessary energy consumption and carbon emissions. SHapley Additive exPlanations (SHAP values) are a popular tool to explain model predictions. This paper aims to create an XRL-based system that produces SHAP

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.


✉ robssully@gmail.com (R. S. Sullivan); luca.longo@tudublin.ie (L. Longo)

🌐 lucalongo.eu/about (L. Longo)

🆔 0009-0007-6240-6080 (R. S. Sullivan); 0000-0002-2718-5426 (L. Longo)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

heat-maps to explain how input samples from Experience Replay affect the actions taken by a DQL Agent. These SHAP heat-maps are further used as an additional tool to investigate the impact of reducing Experience Replay on an Agent’s performance in simulations of varying complexity.

2. Related work

[2, 3] introduced DQL, using the Bellman Equation [4] to optimise the control process in complex environments. The agent uses equation 1 through trial and error to learn the quality of taking an action in a given Markov Decision making Process (MDP) state to find an optimal policy that maximises its total reward.

$$Q_t(s, a) = Q_{t-1}(s, a) + \alpha(R(s, a) - \gamma \max_{a'} Q(s', a') - Q_{t-1}(s, a)) \quad (1)$$

Simulated environments, mimic real-world problems and generate valuable training data in a secure manner [5]. Evaluation of performance is comparing the Agent to a handcrafted, human expert, or random policy. Approximating Q-values with a neural network in large and complex states destabilises learning, so Mnih used Experience Replay [6] to sample data and store from the environment for the approximator to later reuse. However, drawbacks included correlated samples, limited capacity causing an agent to forget information, outdated samples from non-stationary environments and overfitting from samples memorised. Prioritized Experience Replay (PER) [7] and Attention based Experience Replay [8] attempted to solve these.

Understanding Experience replay is crucial for efficiency. Deepmind’s Agent57 [9] which beat human champions in Atari contained 80 billion frames of experience to achieve optimal performance. Consequently many consider Experience Replay flawed with most wanting it replaced. Asynchronous Actor-Critic (A3C) by [10] is a popular alternative. It trains multiple agents in parallel, to explore the environment, and update a shared network, requiring more resources but converging faster. Experience replay, although slower, is more memory efficient only requiring stored transitions and not multiple copies of the network. [11] highlighted that the size of Experience Replay M is a neglected hyperparameter and if large hurts performance, but [12] stated to keep it high using 90% of total environment transition steps as a rule of thumb. [13] stated most default to Mnih’s 1M transitions for the capacity size. Experiments in Atari showed increasing Experience Replay from 1 million to 10 million transitions while also decreasing the age of the oldest Policy did improve performance. However, any increase in size of Experience Replay further burdens resources.

The minimum experience replay size allowed is not known but explainability can help find it. [14, 15, 16]. Custom explainers exist [17, 18] to understand simulation events but not Experience Replay. Within XRL [19, 20, 21], SHAP (SHapley Additive exPlanations)[22] is a popular choice to explain black-box models [23, 24, 25]. It assign feature importance values for a particular prediction. RL-SHAP diagram explains environment features effect on action selection. Similarly Experience Replay is partitioned based on Rule Density into clusters and labelled to select environment features[26]. This paper proposes the use of SHAP for Experience Replay aiding replay capacity size reduction.

3. Design

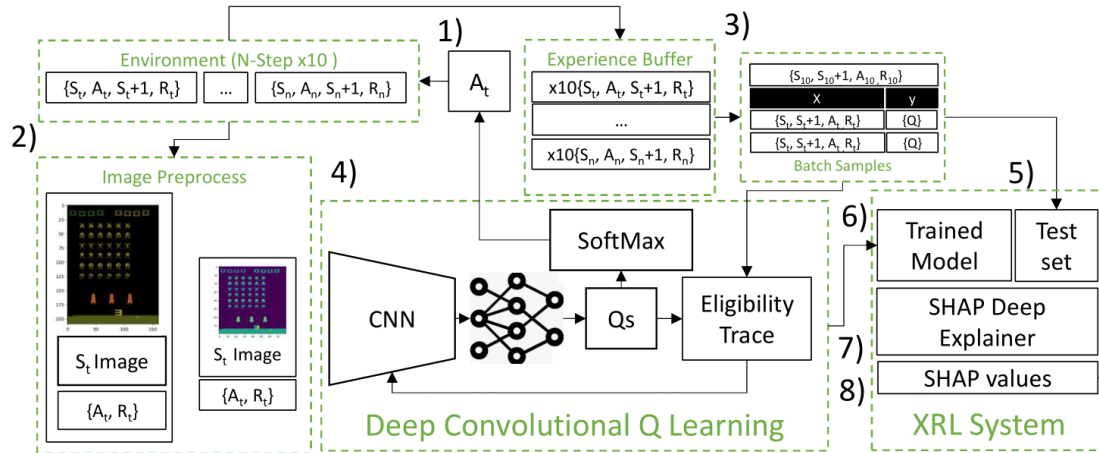


Figure 1: DCQL Architecture: 1) Agent takes action. 2) Environment returns reward + next state (10 processed images). 3) Experiences are sampled. 4) Neural network outputs q-values for Softmax. 5) Unseen test set created. 6) model extracted. 7) Deep Explainer creates interpretable SHAP values.

A DQL Agent, either neural network (layers: observations input, 30 neuron hidden, q value output) or convolutional (Figure 1, layers: 80x80px grayscale input, 32 features 5x5px, 32 features 3x3px, 64 features 2x2px, flattening layer, stride: 2, 30 neuron hidden, q value output), is placed in 23 simulations, using Adam's optimiser and SoftMax Policy. 128 samples of previous states, next states, actions and rewards are stored in Experience Replay. Previous and next states contain 10 images each if the convolutional Agent is used. Regardless of Agent chosen, 10% is set aside for SHAP Deep Explainer. These samples are not seen by the trained agent model. Hyperparameters ($\gamma = 0.9$, $\alpha = 0.001$, $T = 100\%$) held constant, the Experience Replay capacity was reduced: 1M to 500k, 100k, 50k, 10k, 5k, 1k, and 500 transitions respectively. The alternative hypothesis tests for a difference ($p < 0.05$) in reward scores when experience replay capacity is reduced. Agent is initialised with an Experience Replay set to 1 million transitions, then three steps occur. Firstly the environment simulates for 200 episodes. Secondly a reward is stored and graphed to learn how the Agent performed. Finally, either a SHAP summary plot for state vector data or a SHAP heat map is generated for state image data to explain from Experience Replay why the Agent took an action in a given state. These are repeated until the size reaches 500 transitions. When all simulations are complete a density plot of Experience Replay sizes is created. Shapiro-Wilk test is used to confirm if ANOVA and Tukey or Kruskal-Wallis test with Dunn's post hoc test can be used.

4. Results and discussion

Table 1
Minimum Experience Replay Size Allowed Table

Sim Type	Sim Name	State Input	Discrete Output	Model Type	Preprocessed Input	Default Cap	Min Cap	% Reduction
Atari	asterix	(210, 160, 3)	9	CNN DQL	(80,80,1)	1000000	500	99.95%
Atari	james bond	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	500	99.95%
Atari	asteroids	(210, 160, 3)	14	CNN DQL	(80,80,1)	1000000	1000	99.90%
Atari	breakout	(210, 160, 3)	4	CNN DQL	(80,80,1)	1000000	1000	99.90%
Atari	space invaders	(210, 160, 3)	6	CNN DQL	(80,80,1)	1000000	1000	99.9%
Atari	wizard of wor	(210, 160, 3)	10	CNN DQL	(80,80,1)	1000000	1000	99.9%
Atari	air raid	(250, 160, 3)	6	CNN DQL	(80,80,1)	1000000	5000	99.50%
Atari	pong	(210, 160, 3)	6	CNN DQL	(80,80,1)	1000000	5000	99.5%
Atari	ms pack-man	(210, 160, 3)	9	CNN DQL	(80,80,1)	1000000	10000	99%
Atari	private eye	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	50000	95%
Atari	bowling	(210, 160, 3)	6	CNN DQL	(80,80,1)	1000000	100000	90%
Atari	QBert	(210, 160, 3)	6	CNN DQL	(80,80,1)	1000000	100000	90%
Atari	demon attack	(210, 160, 3)	6	CNN DQL	(80,80,1)	1000000	500000	50%
Atari	gravitar	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	500000	50%
Atari	yars's revenge	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	500000	50%
Atari	zaxxon	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	500000	50%
custom	rat cocaine addiction	(2,)	3	NN DQL	N/A	1000000	500000	50%
box2d	lunar lander	(8,)	4	NN DQL	N/A	1000	500	50%
Atari	freeway	(210, 160, 3)	3	CNN DQL	(80,80,1)	1000000	1000000	0%
Atari	sequest	(210, 160, 3)	18	CNN DQL	(80,80,1)	1000000	1000000	0%
classic	cartpole	(4,)	2	NN DQL	N/A	1000	1000	0%
Atari	montezuma's revenge	(210, 160, 3)	18	CNN DQL	(80,80,1)	N/A	N/A	N/A
Atari	venture	(210, 160, 3)	18	CNN DQL	(80,80,1)	N/A	N/A	N/A

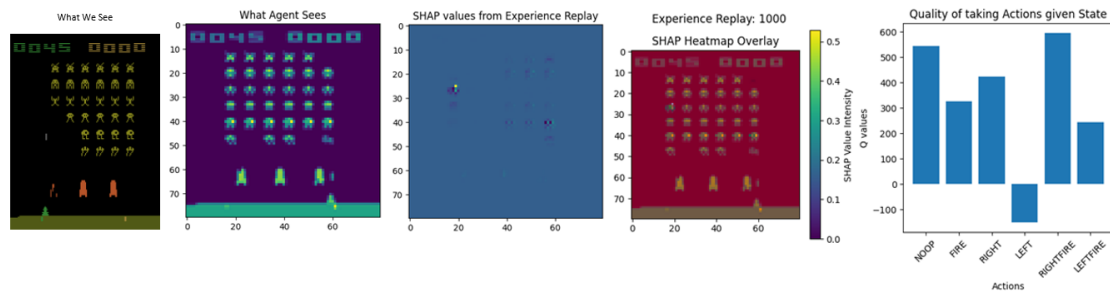


Figure 2: Space Invaders 1,000 experiences Heatmap, shows bright & dark Q-value importance. We see the agent has learned enemy location & movement. Q-values: Right-left highest, Go-Left lowest.

DQL or DCQL Agents was placed in 23 simulations. Table 1 shows the smallest Experience Replay size allowed. Shap heatmaps is used to explain why. A Kruskal-Wallis test and Dunn's post hoc test is used due to reward data failing Shapiro-Wilk test of normality. The null hypothesis is rejected. There is a difference ($p < 0.05$) in reward scores when Experience Replay is reduced. It is found that in 18 of 23 simulations the Agent is tested in, Experience Replay can be reduced over 40% smaller than the default 1 million transitions or the 90% rule-of-thumb for total transitions. Some simulations proved too challenging to get a suitable result (Montezuma's Revenge and Venture) or had to be kept high (Freeway, Seaquest and CartPole) in order to receive the highest reward. In future work Rule Density [26] will be considered to maintain experience quality as capacity is reduced. In conclusion, the proposed XRL-based system using SHAP values for Experience Replay can provide a more transparent, interpretable explanation of actions taken by a DQL agent, which can aid in optimisation for a better use of resources.

References

- [1] T. L. Hayes, G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski, C. Kanan, Replay in deep learning: Current approaches and missing biological elements, *Neural Computation* 33 (2021) 2908–2950. doi:10.1162/neco_a_01433.
- [2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, *Nature* 518 (2015) 529–533. URL: <https://doi.org/10.1038/nature14236>. doi:10.1038/nature14236.
- [3] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, second ed., The MIT Press, 2018. URL: <http://incompleteideas.net/book/the-book-2nd.html>.
- [4] R. Bellman, *Dynamic Programming*, Dover Publications, 1957.
- [5] M. G. Bellemare, Y. Naddaf, J. Veness, M. Bowling, The arcade learning environment: An evaluation platform for general agents, *J. Artif. Int. Res.* 47 (2013) 253–279.
- [6] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, *Mach. Learn.* 8 (1992) 293–321. doi:10.1007/BF00992699.
- [7] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized experience replay, in: Y. Bengio, Y. LeCun (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.
- [8] M. Ramicic, A. Bonarini, Attention-based experience replay in deep q-learning, *Association for Computing Machinery*, 2017, pp. 476–481. URL: <https://doi.org/10.1145/3055635.3056621>. doi:10.1145/3055635.3056621.
- [9] S. Kapturowski, V. Campos, R. Jiang, N. Rakićević, H. van Hasselt, C. Blundell, A. P. Badia, Human-level atari 200x faster, 2022.
- [10] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, volume 48, PMLR, 2016, pp. 1928–1937. URL: <https://proceedings.mlr.press/v48/mniha16.html>.
- [11] S. Zhang, R. Sutton, A deeper look at experience replay (2017).
- [12] T. D. Bruin, J. Kober, K. Tuyls, R. Babuška, Experience selection in deep reinforcement learning for control, *J. Mach. Learn. Res.* 19 (2018) 347–402.
- [13] W. Fedus, P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, W. Dabney, Revisiting fundamentals of experience replay, *JMLR.org*, 2020.
- [14] L. Longo, R. Goebel, F. Lécué, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. R. Weippl (Eds.), *Machine Learning and Knowledge Extraction - 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020*, Dublin, Ireland, August 25-28, 2020, Proceedings, volume 12279 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 1–16. URL: https://doi.org/10.1007/978-3-030-57321-8_1. doi:10.1007/978-3-030-57321-8_1.
- [15] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, *Frontiers in Artificial Intelligence* 4 (2021) 160. URL: <https://www.frontiersin.org/article/10.3389/frai.2021.717899>. doi:10.3389/frai.2021.717899.

- [16] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, *Machine Learning and Knowledge Extraction* 3 (2021) 615–661. URL: <https://www.mdpi.com/2504-4990/3/3/32>. doi:10.3390/make3030032.
- [17] M. Keramati, A. Durand, P. Girardeau, B. Gutkin, S. H. Ahmed, Cocaine addiction as a homeostatic reinforcement learning disorder, *Psychol. Rev.* 124 (2017) 130–153.
- [18] L. Miralles-Pechuán, F. Jiménez, H. Ponce, L. Martínez-Villaseñor, A methodology based on deep q-learning/genetic algorithms for optimizing covid-19 pandemic government actions, *Association for Computing Machinery*, 2020, pp. 1135–1144. URL: <https://doi.org/10.1145/3340531.3412179>. doi:10.1145/3340531.3412179.
- [19] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, *Knowledge-Based Systems* 214 (2021) 106685. URL: <https://www.sciencedirect.com/science/article/pii/S0950705120308145>. doi:<https://doi.org/10.1016/j.knsys.2020.106685>.
- [20] G. Ras, N. Xie, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, *J. Artif. Int. Res.* 73 (2022). URL: <https://doi.org/10.1613/jair.1.13200>. doi:10.1613/jair.1.13200.
- [21] G. A. Vouros, Explainable deep reinforcement learning: State of the art and challenges, *ACM Comput. Surv.* 55 (2022). URL: <https://doi.org/10.1145/3527448>. doi:10.1145/3527448.
- [22] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Curran Associates Inc.*, 2017, pp. 4768–4777.
- [23] S. Kumar, M. Vishal, V. Ravi, Explainable reinforcement learning on financial stock trading using shap, 2022.
- [24] A. N. Thirupathi, T. Alhanai, M. M. Ghassemi, A machine learning approach to detect early signs of startup success, *Association for Computing Machinery*, 2022. URL: <https://doi.org/10.1145/3490354.3494374>. doi:10.1145/3490354.3494374.
- [25] K. Zhang, J. Zhang, P.-D. Xu, T. Gao, D. W. Gao, Explainable ai in deep reinforcement learning models for power system emergency control, *IEEE Transactions on Computational Social Systems* 9 (2022) 419–427. doi:10.1109/TCSS.2021.3096824.
- [26] F. Sovrano, A. Raymond, A. Prorok, Explanation-aware experience replay in rule-dense environments, *CoRR abs/2109.14711* (2021). URL: <https://arxiv.org/abs/2109.14711>.