Balancing Abstraction and Spatial Relationships for Robust Reinforcement Learning

Jannis Blüml*
hessian.AI
Technische Universität Darmstadt
64289 Darmstadt, Germany
jannis.blueml@tu-darmstadt.de

Cedric Derstroff*
hessian.AI
Technische Universität Darmstadt
cedric.derstroff@tu-darmstadt.de

Elisabeth Dillies
Sorbonne Université
elisabeth.dillies@gmail.com

Quentin Delfosse
Technische Universität Darmstadt
quentin.delfosse@tu-darmstadt.de

Kristian Kersting
Hessian Center for Artificial Intelligence (hessian.AI)
Centre for Cognitive Science
German Research Center for Artificial Intelligence (DFKI)
Technische Universität Darmstadt
kersting@cs.tu-darmstadt.de

Abstract

Reinforcement learning (RL) agents trained on raw pixel inputs struggle with irrelevant visual features, overfitting, and poor generalization to visual perturbations. Inspired by human cognitive strategies, we explore how abstraction can enhance robustness, interpretability, and adaptability in RL. We address these issues by simplifying inputs while preserving task-critical features, enabling agents to focus on meaningful environmental elements.

We propose an abstraction method, which we call object-centric hard attention, to evaluate trade-offs between simplifying inputs and retaining critical spatial relationships. Experiments in the Atari Learning Environment demonstrate its efficacy under simplified and visually modified conditions. Results show that appropriate abstraction improves robustness to perturbations like background changes and occlusions and enhances generalization without modifying existing RL architectures or training algorithms. However, excessive compression of input information, as seen in object-centric representations, can hinder spatial reasoning and performance in tasks with global spatial dependencies. Our findings emphasize balancing abstraction and information retention to maximize RL performance and robustness.

This work emphasizes the potential of adaptive abstraction strategies enabling agents to generalize more effectively across diverse scenarios.

Keywords: Reinforcement Learning, Object-Centric Attention, Feature Representation

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research, the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) within their joint support of the National Research Center for Applied Cybersecurity ATHENE, via the "SenPai: XReLeaS" project as well as their cluster project within the Hessian Center for AI (hessian.AI) "The Third Wave of Artificial Intelligence - 3AI".

*These authors contributed equally.

1 Introduction

Similar to fast and slow thinking (Kahneman, 2011), visual processing in humans appears to integrate two phases: an automatic process covering the entire visual field in parallel to quickly capture salient information (Treisman, 1985) and a sequential screening of relevant areas requiring focused attention to extract more complex representations (Treisman and Gelade, 1980) consciously. The former, known as the pre-attentive phase, notably enables humans, when playing video games, to rapidly capture the essential features of the visual scene while filtering out irrelevant information, such as the background. It is suggested that abstract representations like this—derived from key entities and their relationships—are central to human reasoning and planning (Baars, 1993, 2002; Bengio, 2017; Goyal and Bengio, 2022).

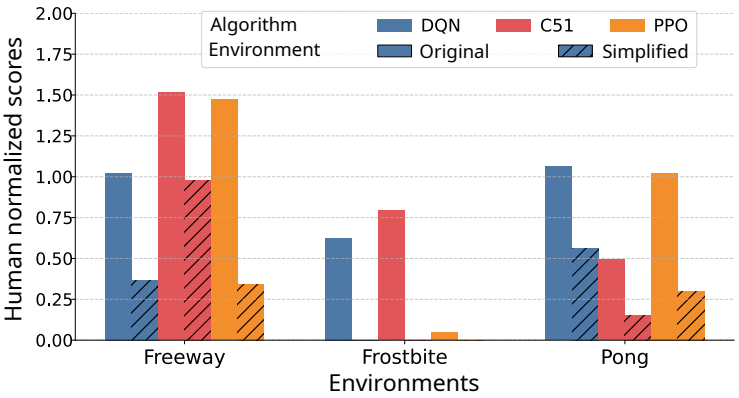

Figure 1: **Deep agents cannot generalize to simpler scenarios.** Testing deep agents in simpler versions of the environments reveals a significant performance drop, highlighting a critical limitation in their robustness and adaptability, as shown by Delfosse et al. (2024a). DQN and C51 agents are taken from Gogianu et al. (2022), and PPO from Delfosse et al. (2024b). The human normalized scores are calculated using the human results by Delfosse et al. (2024a).

In reinforcement learning (RL), however, directly learning from raw pixels of visual states has become the most common approach, contrasting with human-like information processing systems and results in significant challenges, including noise and confounders (Agnew and Domingos, 2021; Yoon et al., 2023). Therefore, inspired by cognitive science, recent studies have proposed integrating abstraction processes into RL to enhance generalization (Bertoin et al., 2022; Zhao et al., 2021). Farebrother et al. (2020) highlights this, showing that agents often overfit to specific environments, failing to adapt to minor variations. This issue is exacerbated by misalignment problems, where agents exploit unintended shortcuts during training, leading to failures in novel scenarios.

A well-documented example in Pong illustrates how RL agents frequently base their actions on the position of the opponent's paddle—a spurious shortcut—instead of focusing on the ball (Delfosse et al., 2024a,c). Such dependencies on spurious correlations compromise generalization, as shown in Fig. 1, where deep RL agents exhibit significant performance drops when tested outside their training settings. These observations underscore the need for approaches prioritizing meaningful and transferable RL representations by balancing abstraction and spatial relationships.

Object-centric approaches offer a promising solution by disentangling scenes into object-level attributes and relationships. Advances like Slot Attention (Locatello et al., 2020), notably applied to Atari RL environments (Delfosse et al., 2023) demonstrate the potential of unsupervised object extraction methods, that can extract object representations to improve efficiency and generalization (Delfosse et al., 2024b; Patil et al., 2024). However, these methods can overly compress information, potentially hindering an agent's ability to learn spatial relationships as effectively as traditional convolutional neural networks (CNNs), as illustrated in Table 1.

We address this challenge by proposing to use objects to filter irrelevant information from visual inputs while preserving critical elements, such as the relative positions of objects in the environment. We call this **object-centric hard attention**. Building on ideas from MinAtar (Young and Tian, 2019) and Davidson and Lake (2020), our approach abstracts the original scene into a semanticized representation, emphasizing essential elements and their interactions. Unlike MinAtar, which simplifies entire tasks and also positional information, we retain the original task complexity while providing semantically meaningful input to the agent.

Our method employs object-centric hard attention to mask irrelevant information while preserving task-relevant features and spatial relationships, creating interpretable representations that align with principles of conscious planning. By removing the background and retaining bounding boxes of relevant objects, we reduce the impact of confounders, simplify learning, and enhance generalization. This approach enables RL agents to focus on critical elements and relationships within their environment, paving the way for robust and efficient learning. Our results highlight the potential of integrating abstraction and object-centric representations into RL frameworks, demonstrating efficiency, robustness, and improved interpretability.

In summary, this work evaluates a novel abstraction method that balances simplicity and spatial awareness to improve robust and generalizable learning across diverse tasks without changing network architectures or training algorithms.

2 Methodology

RL agents that rely solely on pixel-based representations often ground their decisions on irrelevant visual features, which can lead to brittle behavior, as spurious correlations between background features and rewards may emerge during training. Consequently, minor environmental changes—such as lighting, texture, or occlusions—can significantly degrade the agent's performance. Such vulnerabilities highlight the limitations of pixel-based RL agents and the importance of developing representations that prioritize semantically meaningful and task-relevant information. We propose removing unnecessary information to address these challenges to enhance classical pixel-based RL methods, such as DQN or PPO. Our approach preserves relevant object features while filtering out background information, reducing input complexity while maintaining the spatial relationships within the image. This ensures that RL agents can focus on task-critical elements and their interactions, improving robustness to visual perturbations.

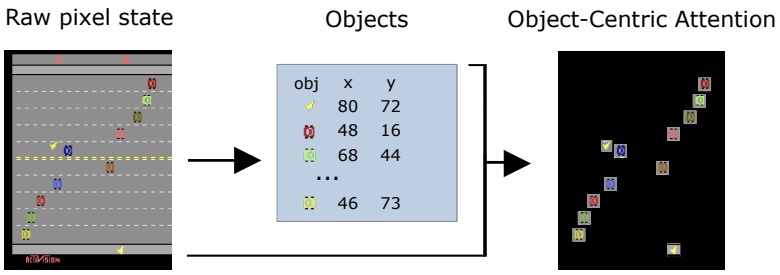

Figure 2: **Object-centric hard attention.** Task-relevant features within the input frame are isolated by identifying and masking all irrelevant background details. This ensures that RL agents focus on spatially and semantically meaningful components.

We utilize the Atari Learning Environment (ALE) (Bellemare et al., 2013; Machado et al., 2018) as the testbed for evaluating our proposed framework. The ALE provides diverse Atari games that challenge RL algorithms with high-dimensional state spaces, dynamic interactions, and long-term planning requirements. To assess the robustness of our method, we test it on a subset of games with varying gameplay dynamics and task complexities, including variations introduced by Delfosse et al. (2024a).

Our method begins by extracting all game-relevant objects' positions, attributes, and, specifically, their position and size. These representations can be derived using advanced visual extraction techniques or directly accessed from the environment's internal state, as demonstrated by Delfosse et al. (2024b). The next step is to use the objects to filter out irrelevant parts of the visual input using object masks, as illustrated in Fig. 2. This is done by setting all pixels outside the bounding boxes of identified objects to 0, isolating game-relevant elements while discarding confounding background features. This masking step ensures that the agent's attention is directed to critical areas of the environment, minimizing distractions from irrelevant details.

3 Results

To evaluate the effectiveness of our approach, we conducted experiments on a diverse set of Atari games using PPO agents with both pixel-based and object-centric input representations. The pixel-based representation follows the method by Mnih et al. (2013), while the object-centric representation is based on Delfosse et al. (2024b), extracting the positions and attributes of game-relevant objects. Our approach applies object-centric hard attention to filter irrelevant visual features and constructs abstractions for decision-making. Agents were trained for 40 million frames using PPO with hyperparameters from Huang et al. (2022b), following the training guidelines of Huang et al. (2022a). Performance was evaluated using average episodic rewards across three random seeds, with results aggregated over 10 games per seed using 3 seeds. Robustness was assessed by testing agents on environments with visual perturbations, such as changes in object colors or layouts.

As shown in Table 1, balancing abstraction and spatial relationship preservation is critical for RL. Object-centric methods, while effectively filtering irrelevant details, often compress input representations too strongly, resulting in the loss of critical spatial information. This limits their adaptability to tasks requiring fine-grained spatial reasoning, leading to lower performance than pixel-based PPO in certain scenarios. However, their abstraction provides notable robustness to visual modifications, such as color changes or occlusions.

To investigate the trade-off between abstraction and task performance, we conducted an ablation study comparing two masking strategies: *Object Masks* and *Binary Masks*. *Object Masks*, as illustrated in Fig. 2, retain object-specific details such as position, color, and shape by setting all non-object pixels to black, preserving spatial relationships and context. In contrast, *Binary Masks* simplify the input further by replacing objects with uniform blocks the size of their bounding boxes, maximizing abstraction while maintaining spatial relationships. Results in Table 1 demonstrate that *Binary Masks* often perform comparably to, or better than, *Object Masks* in tasks involving visual modifications. It can also be seen that, not surprisingly, Object Masks are more robust against changes in color since they are missing in the representation. Further, it can be seen that masking is not the solution to all our problems, as the performance still drops if the change is correlated to game logic or movement, similar to PPO.

Game	Tested in	PPO			
		Standard DQN	Object-Centric	Object Masks	Binary Masks
Boxing	Standard	91.13 ± 5.16	86.27 ± 9.60	93.07 ± 4.22	$\mathbf{93.40 \pm 3.95}$
	Red vs Blue	3.13 ± 8.59	89.60 ± 7.60	45.37 ± 26.00	$\mathbf{94.27 \pm 3.83}$
	Switched Positions	$\mathbf{56.70 \pm 40.51}$	36.40 ± 35.31	-6.10 ± 48.23	41.30 ± 57.83
Breakout	Standard	152.73 ± 80.47	34.30 ± 13.20	$\mathbf{217.87 \pm 105.71}$	213.70 ± 107.05
	Red Blocks	228.23 ± 113.44	37.50 ± 10.66	237.77 ± 111.91	$\mathbf{247.50 \pm 95.66}$
	Red Player	8.23 ± 3.55	44.63 ± 11.20	17.53 ± 13.21	$\mathbf{229.20 \pm 100.27}$
Freeway	Standard	31.93 ± 1.18	31.83 ± 0.58	$\mathbf{33.07 \pm 0.89}$	33.00 ± 0.82
	Black Cars	22.90 ± 3.04	31.43 ± 0.88	22.47 ± 5.11	$\mathbf{33.07 \pm 0.51}$
Frostbite	Standard	$\mathbf{294.33 \pm 11.16}$	252.67 ± 16.52	283.00 ± 18.47	271.67 ± 11.28
	Stopped Ice	61.00 ± 68.67	$\mathbf{154.67 \pm 61.20}$	82.67 ± 68.21	99.67 ± 97.66
Pong	Standard	15.40 ± 2.85	18.17 ± 2.32	17.17 ± 2.27	$\mathbf{18.57 \pm 1.63}$
	Lazy Enemy	-10.17 ± 5.09	-17.63 ± 2.86	$\mathbf{8.60 \pm 6.50}$	7.17 ± 10.89
MsPacman	Standard	3073.33 ± 906.36	2578.67 ± 1137.46	$\mathbf{4293.33 \pm 1526.33}$	4264.33 ± 1036.15
	2nd Maze	365.67 ± 220.15	463.33 ± 432.40	404.67 ± 435.00	$\mathbf{545.67 \pm 387.89}$

Table 1: **Reduced input representations outperform baseline approaches, particularly in modified and challenging environments.** This table compares the average episodic rewards achieved by PPO using 4 different input representations: the standard DQN-like representation, OCAtari's object-centric vector representation, and our proposed masking approaches: Object Masks and Binary Masks. Results are averaged over 3 random seeds; the error margins indicate standard deviations. All agents are trained in the original game environment and tested in the same, simpler game variants. **Bold** values indicate the highest performance for each game and modification, highlighting the effectiveness of our methods in enhancing performance and generalization. Game variants are taken from Delfosse et al. (2024a).

Going even further, we can look at the object-centric input representation (Delfosse et al., 2024b) (cf. Table 1). This representation provides even stronger abstraction by reducing object attributes such as position and size into a single vector, which complicates the inference of spatial relationships. Consequently, performance suffers in tasks requiring spatial precision, underscoring the limitations of excessive compression while highlighting the robustness of such methods against color modifications.

These results emphasize the need to tailor abstraction strategies to task requirements. Detailed representations can enhance performance in tasks demanding spatial precision, while more abstract representations improve robustness and generalization under environmental changes. Adaptive abstraction strategies that strike this balance hold significant promise for RL.

4 Conclusion

This work introduces an object-centric input abstraction method for reinforcement learning, leveraging *object-centric hard attention* to filter irrelevant information while preserving task-relevant features and spatial relationships. Our approach improves the robustness of RL agents to visual perturbations without sacrificing performance in standard settings, as demonstrated through experiments in the Atari Learning Environment. By focusing the agent's attention on meaningful environmental elements, we address critical challenges of pixel-based RL, such as sensitivity to irrelevant features and poor generalization.

While promising, this work is an intermediate step toward fully leveraging abstraction in reinforcement learning. The experiments focus on PPO and controlled benchmarks, leaving the exploration of other algorithms and more complex environments open. Also, the static abstraction levels used in this study may not optimally suit all task scenarios. Therefore, research should also investigate dynamic and more diverse abstraction approaches, e.g., taking more than position and size into account, including features like object type or orientation. Further, the abstraction should adapt to varying task requirements to enhance efficiency and scalability.

By bridging the gap between perception and decision-making, our findings emphasize the potential of abstraction-based methods to create RL agents that are more robust, interpretable, and adaptable while concentrating on the effect of spatial relationships in the input representation.

References

William Agnew and Pedro Domingos. Relevance-guided modeling of object dynamics for reinforcement learning, 2021.

Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1993.

Bernard J Baars. The conscious access hypothesis: origins and recent evidence. *Trends in cognitive sciences*, 2002.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2013.

Yoshua Bengio. The consciousness prior, 2017.

David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning. *Advances in Neural Information Processing Systems*, 2022.

Guy Davidson and Brenden M. Lake. Investigating simple object representations in model-free deep reinforcement learning, 2020.

Quentin Delfosse, Wolfgang Stammer, Thomas Rothenbacher, Dwarak Vittal, and Kristian Kersting. Boosting object representation learning via motion and object continuity. 2023.

Quentin Delfosse, Jannis Blüml, Bjarne Gregori, and Kristian Kersting. Hackatari: Atari learning environments for robust and continual reinforcement learning. *Interpretable Policies Workshop @ The Reinforcement Learning Conference*, 2024a.

Quentin Delfosse, Jannis Blüml, Bjarne Gregori, Sebastian Sztwiertnia, and Kristian Kersting. OCAtari: Object-centric Atari 2600 reinforcement learning environments. *Reinforcement Learning Journal*, 2024b.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c.

Jesse Farebrother, Marlos C. Machado, and Michael Bowling. Generalization and regularization in dqn, 2020.

Florin Gogianu, Tudor Berariu, Lucian Buşoniu, and Elena Burceanu. Atari agents, 2022. URL `https://github.com/floringogianu/atari-agents`.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 2022.

Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. The 37 implementation details of proximal policy optimization. In *ICLR Blog Track*, 2022a. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, pages 1–18, 2022b.

Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew J. Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5573–5577, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning, 2013.

Vihang Patil, Markus Hofmarcher, Elisabeth Rumetshofer, and Sepp Hochreiter. Contrastive abstraction for reinforcement learning, 2024.

Anne Treisman. Preattentive processing in vision. *Computer vision, graphics, and image processing*, 1985.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 1980.

Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning, 2023.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments, 2019.

Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, and Yoshua Bengio. A consciousness-inspired planning agent for model-based reinforcement learning. *Advances in neural information processing systems*, 2021.