Learning Representations and Robust Exploration for Improved Generalization in Reinforcement Learning

Doctoral Consortium

Nasik Muhammad Nafi
Kansas State University
Manhattan, KS, USA
nnafi@ksu.edu

ABSTRACT

Deep Reinforcement Learning agents typically aim to learn a task through interacting in a particular environment. However, training on such singleton RL tasks, where the agent interacts with the same environment in every episode, implicitly leads to overfitting. Thus, the agent fails to generalize to minor changes in the environment, especially in image-based observation. Generalization is one of the main contemporary research challenges and recently proposed environments that enable diversified episode generation opens up the possibility to investigate generalization. My initial work towards this objective includes representation learning through the partial decoupling of policy and value networks and hyperbolic discounting in a single-agent setting. Efficient exploration is another crucial aspect of achieving generalization when learning from limited data. My dissertation would focus on proposing and evaluating methods that enable better representation learning and exploration for unseen scenarios. Another key objective is to extend my work to multi-agent generalization which is comparatively less studied.

KEYWORDS

reinforcement learning, generalization, representation learning, discounting, exploration, multi-agent systems

ACM Reference Format:
Nasik Muhammad Nafi. 2023. Learning Representations and Robust Exploration for Improved Generalization in Reinforcement Learning: Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

The recent advances in deep learning, in particular, deep neural networks, have enabled intelligent agents that are capable of mastering complex tasks. As these learning agents are built based on the idea of minimizing empirical prediction error, they tend to memorize the underlying data distribution which is often termed as overfitting. In Reinforcement Learning (RL), where an agent learns from sampled state-action pairs, the agent overfits to the training trajectories [16] [2]. Thus the agents' ability to generalize in unseen contexts is compromised, however, generalization and faster adaptation capabilities are paramount factors to ensure reliable performance in the real-world application where the agent regularly gets exposed to unseen scenarios or environments not encountered during training.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

The issue of overfitting and subsequent lack of generalization becomes more critical while learning from visual data. Learning control tasks directly from visual observation is challenging, especially in a real-world setting, where the observations are particularly more unstructured and diverse [21]. Given infinitely many samples of realistic images the agent may learn an optimal behavior. However, an unlimited supply of real images beyond simulation is often expensive and impossible. Thus, learning representation, a smaller vector encoding that can capture adequate information, remains an efficient solution to learning from limited data.

While in RL algorithms the main objective is to maximize the cumulative reward over the episodes, optimizing the policy, represented through a deep neural networks-based function approximator, just for that objective doesn't necessarily guarantee better representation learning. Implicitly learning the representation can be sufficient to learn the sampled training trajectory and predict actions, however, a small perturbation to the visual state brings a drastic change in the policy and causes failure. Recently, learning task-relevant representation, which is invariant to task-agnostic factors, received much attention as this helps to achieve generalization. To learn a representation that is task-relevant and agnostic of the irrelevant dynamic elements present in the observation, explicit measures are often necessary such as minimizing the bisimulation distances of the states, decoupling the policy and value representation, and learning contrastive behavioral similarity [16] [21].

My dissertation research focuses on learning representations to avoid spurious correlations between the learned policy and task-irrelevant information and evaluate them on environments under the Contextual Markov Decision Process Framework, which provides a structured way to quantify an agent's ability to generalize. My current research includes investigating non-trivial sources of task-relevant information and uncertainty handling such as attention mechanisms, complex discounting schemes, and value target augmentation. Another aspect of my research is to enable representation learning with minimum overhead as opposed to the most proposed approaches that compromise computing time, introduce complex design choices, or bring in a lot of new hyperparameters.

2 PROBLEM FORMULATION

We consider a Contextual Markov Decision Process (CMDP) given by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, C, \mathcal{T}, r, \mu_C, \mu_S)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, C is the context space, $\mathcal{T}(s'|s,a)$ is the transition function, r is the reward function, μ_C is the context distribution, and μ_S is the context-dependent initial state distribution. At the beginning of each episode, a context is sampled according to $c \sim \mu_C$. Then an initial state is sampled according to $s_0 \sim \mu(\cdot|c)$ and the successive

states within that episode are sampled based on $s_{t+1} \sim \mathcal{T}(\cdot|s_t, a_t, c)$. Let d_π^c denotes the state distribution resulting from acting with the policy π in context c. During training, the agent is exposed to a limited number of contexts which is far smaller than the whole set of contexts on which the agent is being tested. The goal is to learn a policy π that maximizes the expected return over all possible contexts such that $\mathcal{G} = \mathbb{E}_{c\sim\mu_C, s\sim d_\pi^c, a\sim\pi(s)}[r(s,a)]$ while learning from those limited contexts. The context parameters generally take the form of an initial seed, identifier, or parameter vector that determines the variation of an episode. Those parameters are not observable by the agent. The context remains the same within an episode but varies between episodes. This formulation can easily be extended to multi-agent scenarios where the agents share the same action space, however, the state observation varies agent-wise.

3 RELATED WORK

Various techniques from the general deep learning domain have been employed to improve the generalization ability of deep RL agents. Such approaches include regularization techniques such as dropout, batch normalization, and data augmentation [3, 4, 9, 10, 17, 19, 20, 22]. To encourage task-specific policy representation, Raileanu and Fergus [16] proposed decoupling the policy and value networks to disentangle the policy representation from value representation. Cobbe et al. [5] used decoupled architecture but with phase-wise training to avoid the interference between policy and value. Furthermore, bisimulation metrics and policy similarity embeddings have been proposed to measure state similarity, leading to task-relevant representations [1, 21]. Recently, language models have been used for history compression to enable memory to store abstractions of the observations [15]. Previous works that address the issue of exploration for generalization include [7, 8, 14]

4 PRELIMINARY WORKS

4.1 Partial Decoupling of Actor and Critic

This research aims to address policy-value representation asymmetry, a major cause for the lack of generalization [16], in an efficient way. I introduce an Attention-based Partially Decoupled Actor-Critic (APDAC) that shares some early convolutional layers of the network while separating the later (downstream) ones into policy and value sub-networks [13]. This partial separation of networks acknowledges the asymmetry in representation between policy and value, and thus encourages distinct feature learning for each, while still allowing shared low-level (edges, dots) representation learning through early shared layers. Sharing some of the layers enables the joint optimization of policy and value, reducing the high cost of two separate optimizations for policy and value networks as needed in IDAAC [16]. The inclusion of the channel-wise attention mechanism ensures learning minimal and compact representations that eliminate spurious correlations between generic features (e.g., background color) and the value/policy function. Using a gradient-based heatmap generation technique, this work produces visualizations that reveal crucial insights into the learned policies and value representations. Being an architectural contribution this hybrid network architecture can be used with any actor-critic algorithm to achieve generalization. An extension of this work analyzes how the extent of policy-value decoupling impacts generalization [11].

4.2 Hyperbolic Discounting

This aspect of my research investigates the influence of learning from hyperbollically discounted advantage estimates on the generalization ability of an agent [12]. We identify that variations of contexts come with an additional phenomenon that the completion time may vary drastically based on the contexts. This completion time can be interpreted as survival time and linked to the hazard rate of the environment. Exponential discounting of future rewards implies the assumption that the agent in the environment encounters a fixed, known risk or hazard [18]. The hazard rate is defined as the per-time-step risk the agent incurs as it acts in the environment. My work hypothesizes that this assumption of a fixed hazard rate for the agent in an environment does not hold in a procedurally generated environment that is based on CMDP (levels are analogous to contexts). This is because the variation in the environment's dynamics and attribute distribution across levels introduce a greater degree of uncertainty and stochasticity compared to typical reinforcement learning scenarios. Fedus et al. [6] shows that when an agent holds uncertainty over the environment's hazard rate, a non-exponential (such as hyperbolic) reward discounting scheme is more appropriate. Thus, this work proposes to leverage hyperbolically discounted advantage estimates in policy optimization to enable learning representation that considers the unknown hazard due to a variety of contexts and achieves generalization.

5 RESEARCH PLAN

5.1 Robust Exploration

Efficient exploration is a prerequisite for better representation learning. If the agent fails to explore the state space adequately, especially in limited data settings, it will lose important information pertinent to state representation. [8] shows that the exploration task poses more difficulty in the CMDP setting as the chance of visiting the same state reduces drastically. This is mainly due to the fact that added variation or even noise can change the state in terms of visual appearance while keeping the states semantically similar. Thus, learning a representation that preserves semantic meaning in the encoding will facilitate generalization. Future work will include developing a soft count for the states based on semantic similarity.

5.2 Multi-Agent Systems

Going forward and based on the successful outcome in a single-agent setting, my research plan is to extend the representation learning and exploration framework to multi-agent settings. I want to explore how representation learning can be improved in a procedurally generated multi-agent system. The broad idea is to learn a joint representation using different value estimates predicted by the agents. I plan to incorporate attention mechanisms that can prioritize spatial objects relevant to the task and the corresponding agent. Also, I will develop efficient exploration strategies for multi-agent systems based on shared knowledge of the agents about the state space. As a test bed, I will use Neural MMO, a multi-agent environment that includes procedural generation of tile-based terrain, a food and water foraging system, and a strategic combat system. The environment also provides an easier way to visualize learned value functions, visitation maps, and inter-agent dependencies.

REFERENCES

[1] Rishabh Agarwal, Marlos C Machado, Pablo Samuel Castro, and Marc G Bellemare. 2021. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. *arXiv preprint arXiv:2101.05265* (2021).

[2] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2048–2056.

[3] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 2019. Quantifying Generalization in Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1282–1289. https://proceedings.mlr.press/v97/cobbe19a.html

[4] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 2019. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1282–1289.

[5] Karl W Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. 2021. Phasic policy gradient. In *International Conference on Machine Learning*. PMLR, 2020–2027.

[6] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865* (2019).

[7] Mikael Henaff, Minqi Jiang, and Roberta Raileanu. [n.d.]. Integrating Episodic and Global Bonuses for Efficient Exploration. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

[8] Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. 2022. Exploration via Elliptical Episodic Bonuses. *arXiv preprint arXiv:2210.05805* (2022).

[9] Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. 2021. Regularization Matters: A Nonparametric Perspective on Overparametrized Neural Network. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 829–837.

[10] Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. 2019. Generalization in reinforcement learning with selective noise injection and information bottleneck. *arXiv preprint arXiv:1910.12911* (2019).

[11] Nasik Muhammad Nafi, Raja Farrukh Ali, and William Hsu. [n.d.]. Analyzing the Sensitivity to Policy-Value Decoupling in Deep Reinforcement Learning Generalization. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

[12] Nasik Muhammad Nafi, Raja Farrukh Ali, and William Hsu. 2022. Hyperbolically Discounted Advantage Estimation for Generalization in Reinforcement Learning. In *Decision Awareness in Reinforcement Learning Workshop, ICML*.

[13] Nasik Muhammad Nafi, Creighton Glasscock, and William Hsu. 2021. Attention-based Partial Decoupling of Policy and Value for Generalization in Reinforcement Learning. In *Deep RL Workshop NeurIPS 2021*.

[14] Ian Osband, Benjamin Van Roy, and Zheng Wen. 2016. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*. PMLR, 2377–2386.

[15] Fabian Paischer, Thomas Adler, Vihang Patil, Angela Bitto-Nemling, Markus Holzleitner, Sebastian Lehner, Hamid Eghbal-Zadeh, and Sepp Hochreiter. 2022. History Compression via Language Models in Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 17156–17185.

[16] Roberta Raileanu and Rob Fergus. 2021. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 8787–8798.

[17] Roberta Raileanu, Max Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2020. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862* (2020).

[18] Peter D Sozou. 1998. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265, 1409 (1998), 2015–2020.

[19] Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. 2020. Improving generalization in reinforcement learning with mixture regularization. *Advances in Neural Information Processing Systems* 33 (2020), 7968–7978.

[20] Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*.

[21] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2020. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742* (2020).

[22] Hanping Zhang and Yuhong Guo. 2021. Generalization of reinforcement learning with policy-aware adversarial data augmentation. *arXiv preprint arXiv:2106.15587* (2021).