Analyzing the Sensitivity to Policy-Value Decoupling in Deep Reinforcement Learning Generalization

Extended Abstract

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

Raja Farrukh Ali
Kansas State University
Manhattan, KS, USA
rfali@ksu.edu

William Hsu
Kansas State University
Manhattan, KS, USA
bhsu@ksu.edu

ABSTRACT

Shared policy-value representations in traditional actor-critic architectures have been shown to limit the generalization capabilities of a reinforcement learning (RL) agent. Fully decoupled/separated networks for policy and value avoid overfitting by addressing this representation asymmetry; however, this introduces additional computational overhead. Partial separation has been shown to reduce this overhead while still achieving the same level of generalization. This raises questions regarding the exact need for two separate networks and whether increasing the degree of separation in a partially separated network improves generalization. To investigate these questions, this paper compares four different degrees of network separation (fully shared, early separation, late separation, and full separation) on the RL generalization benchmark Procgen. Our results indicate that for environments without a distinct or explicit source of value estimation, partial late separation captures the necessary policy-value representation asymmetry and achieves better generalization performance than other architectural options in unseen scenarios, while early separation fails to perform adequately. This also gives us a model selection mechanism for those cases where full separation performs best.

KEYWORDS

reinforcement learning; generalization; policy-value representation asymmetry; decoupling policy and value; representation learning

ACM Reference Format:
Nasik Muhammad Nafi, Raja Farrukh Ali, and William Hsu. 2023. Analyzing the Sensitivity to Policy-Value Decoupling in Deep Reinforcement Learning Generalization: Extended Abstract. 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

Deep reinforcement learning (RL) agents generally suffer from poor generalization performance when applying learned policies to unseen scenarios [3][2]. Raileanu and Fergus [7] demonstrate that asymmetry between the policy and value representation contributes to poor generalization. Shared actor-critic architectures fail to account for distinct features of policy and value components, leading policies to overfit to training instances and causing the agent to fail to generalize. To address this asymmetry, separate networks for policy and value functions have been used that enable distinct feature learning [4][7]. However, such *decoupling* requires greater computation time and introduces design complexity to ensure the flow of value gradients to the policy network. *Partial separation* of policy and value networks has been proposed to achieve comparable performance to the fully separated approach at a lower computational cost [6]. Furthermore, this architecture does not require an additional value (or advantage) head in the policy network as used in fully decoupled approaches to maintain value gradients.

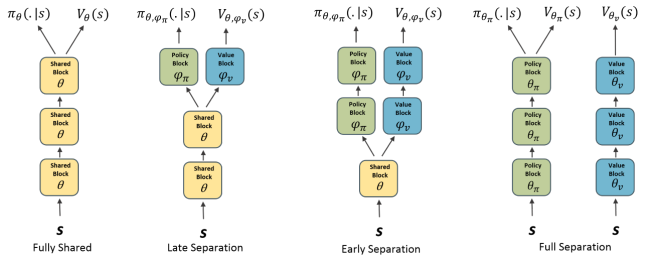

Figure 1: Architectures with different extents of decoupling for policy and value networks

In this work, we analyze the sensitivity of an agent's generalization performance to different extents of decoupling in policy and value networks. We also analyze the learned representations for individual environments to infer whether and how much the degree of separation helps. Our findings suggest that in most cases, partial late separation is sufficient and potentially the best choice; however, full separation is necessary when a clear distinction exists between the sources of value and policy features. Additionally, contrary to the expectation for a monotonic trend, we observe early separation decreases performance while increasing computation costs.

2 METHODOLOGY

We analyze the four distinct architectures shown in Figure 1, denoting varying degrees of separation: fully shared (no separation), early separation, late separation, and full separation. Our categorization of the degrees of separation is exhaustive in the sense that it partitions the possible separations into distinct categories, each with a unique separation point. We do not instantiate every possible separation point for a deep model with k layers, but instead focus on discrete, mutually exclusive categories. We now describe how we use the large IMPALA-CNN model [5] to instantiate the architectures and their corresponding loss functions. We also make our code publicly available. [1]

[1] https://github.com/nasiknafi/sensitivity-to-decoupling

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.

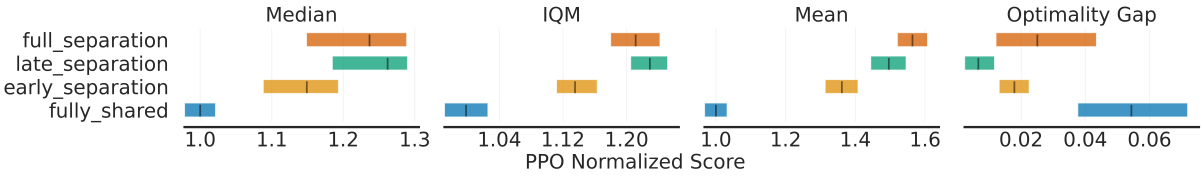

Figure 2: PPO normalized aggregate metrics for all four variants across all the 16 environments in the Procgen benchmark.

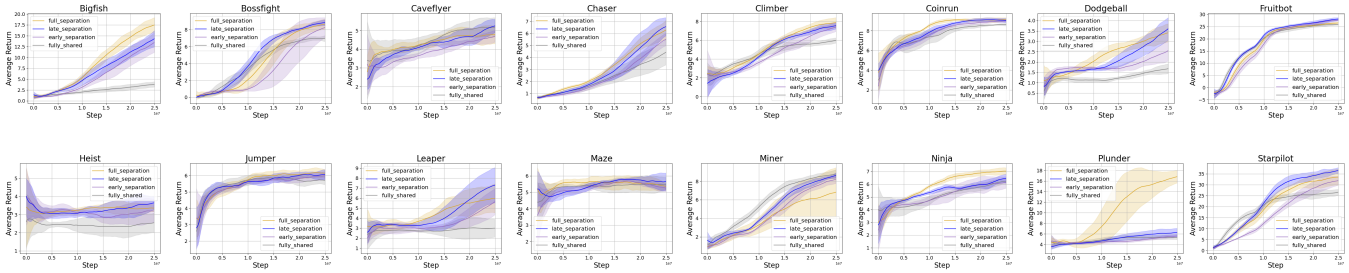

Figure 3: Test performance of fully shared (gray), early separation (purple), late separation (blue), and full separation (yellow) for all 16 environments from Procgen. Mean and standard deviation are calculated over 5 trials, each with a different seed.

Fully Shared: This variant shares all convolution layers of the IMPALA-CNN network while separating only the final fully connected layers used as the policy and value heads. As in Proximal Policy Optimization (PPO) [8], we optimize the joint objective:

$$J_{NS}(\theta) = J_\pi(\theta) - \alpha_v L_V(\theta) + \alpha_s S_\pi(\theta) \tag{1}$$

where $J_\pi(\theta)$ is the policy gradient objective, $L_V(\theta)$ is the value loss, $S_\pi(\theta, \phi_\pi)$ is the entropy bonus for exploration, α_v is the value loss coefficient, and α_s is the coefficient for entropy bonus.

Early Separation: We implement *early separation* in the IMPALA-CNN architecture by sharing the first layer block and separating the subsequent two for policy and value. This approach results in only 5 convolutional layers being shared. Let θ be the shared parameters, ϕ_π are the parameters of the separated policy subnetwork, and ϕ_v are the parameters of the separated value subnetwork. We then jointly optimize the objective:

$$J_{ES}(\theta, \phi_\pi, \phi_v) = J_\pi(\theta, \phi_\pi) - \alpha_v L_V(\theta, \phi_v) + \alpha_s S_\pi(\theta, \phi_\pi) \tag{2}$$

where $J_\pi(\theta, \phi_\pi)$ is the policy gradient objective, $L_V(\theta, \phi_v)$ is the value loss and $S_\pi(\theta, \phi_\pi)$ is the entropy bonus.

Late Separation: We share two initial layer blocks of the IMPALA-CNN architecture instead of just one as in early separation, for a total of 10 convolutional layers. This leads to distinct third blocks for the policy and value subnetworks. The parameterization and objectives are thus the same as for early separation, but the number of shared parameters θ is higher.

Full Separation: Here the policy network is isolated from the value network. The policy and value networks are specified separately and trained for independent objectives. The policy network parameterized by θ_π is optimized for:

$$J_{FS}(\theta_\pi) = J_\pi(\theta_\pi) - \alpha_A L_{A_\pi}(\theta_\pi) + \alpha_s S_\pi(\theta_\pi) \tag{3}$$

Here $L_{A_\pi}(\theta_\pi)$ is the advantage loss for the additional advantage head used to support the policy network [7]. The value network (θ_v) optimizes the value loss where $\hat{V}_t^{targ}$ is the target value function:

$$L_v(\theta_v) = \hat{\mathbb{E}}_t[(V_{\theta_v}(S_t) - \hat{V}_t^{targ})^2] \tag{4}$$

3 RESULTS AND DISCUSSIONS

In our experiments, we adopt the same approach to training and testing as outlined by [2] for *easy* settings. Figure 2 shows the PPO normalized scores for the four variants across all 16 Procgen environments, while Figure 3 compares their returns achieved in individual environments. We report aggregate metrics: Median, Interquartile Mean (IQM), Mean, and Optimality Gap (OG) over the average return for each trained variant, showing that the *late separation* architecture outperforms others with a higher IQM and lower optimality gap, at lower computational overhead. IQM and OG are considered to be statistically significant metrics, robust to outlier trials [1]. Figure 3 shows that late separation also achieves competitive average return and outperforms full separation in most scenarios, with the early separation model performing competitively in some cases, though inconsistently. Detailed investigation regarding exceptions to the above-documented trends indicates that where there is an additional information source for value estimation (e.g., a life bar in the *Plunder* environment), policy networks may prioritize this channel, limiting the policy learned through any degree of network sharing. Full separation succeeds in such cases by disentangling the policy network from the value network.

Our contribution thus provides practical guidelines to RL practitioners regarding the degree of separation to use for better generalization, suggesting partial late separation as a default. In addition, our findings show that two fully separate networks are crucial in the presence of an explicit source of value representation.

REFERENCES

[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems* 34 (2021), 29304–29320.

[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 *International Conference on Machine Learning*. PMLR, 1282–1289.

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

[5] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*. PMLR, 1407–1416.

[6] 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*.

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

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).