Towards Understanding How to Reduce Generalization Gap in Visual Reinforcement Learning

Extended Abstract

Jiafei Lyu[1]
SIGS, Tsinghua University
Shenzhen, China
lvjf20@mails.tsinghua.edu.cn

Le Wan
IEG, Tencent
Shenzhen, China
vinowan@tencent.com

Xiu Li[†]
SIGS, Tsinghua University
Shenzhen, China
li.xiu@sz.tsinghua.edu.cn

Zongqing Lu[†]
School of Computer Science, Peking University
Beijing, China
zongqing.lu@pku.edu.cn

ABSTRACT

It is vital to learn a *generalizable* policy in visual reinforcement learning (RL). Many algorithms are proposed to handle this problem while none of them theoretically show what affects the generalization gap and why their methods work. In this paper, we bridge this issue by theoretically answering the key factors that contribute to the generalization gap when the testing environment has distractors. Our theories indicate that minimizing the representation distance between training and testing environments is the most critical. Our theoretical results are supported by the empirical evidence in the DMControl Generalization Benchmark.

KEYWORDS

Visual reinforcement learning; Generalization gap; RL theory

ACM Reference Format:

Jiafei Lyu[1], Le Wan, Xiu Li[†], and Zongqing Lu[†]. 2024. Towards Understanding How to Reduce Generalization Gap in Visual Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

It is critical for visual RL algorithms to be able to *generalize* to unseen scenarios. Unfortunately, it is challenging as the difference between the (possibly) clean training environment and the unseen environments is not predictable. Existing methods remedy such mismatch by leveraging data augmentation [6, 9, 10, 12, 16, 19, 20, 24], domain randomization [2, 4, 21–23, 27, 34], self-supervision [1, 8, 10, 28, 30, 35], pre-trained image encoders [5, 33], normalization [17], etc. Despite their success, *none of them explain why their methods work in practice from a theoretical perspective.* In this paper, we aim at bridging this gap. We focus on the following generalization setting: the algorithm is trained in a clean environment with visual input, while deployed in an unseen environment where the color or the background of the agent changes. Since the policy keeps evolving during training, we resort to *reparameterization trick* to decouple the randomness in the environment from the policy, the transition dynamics, and the initial state distribution. Under some mild assumptions, we establish concrete theoretical bounds on the generalization gap when deploying the policies in testing environments with distractors. Our results suggest that the most crucial factor that influences the test performance is the representation deviation before and after adding the distractor. We examine the theoretical conclusions by conducting experiments of different algorithms in DMControl Generalization Benchmark (DMC-GB) [10]. The empirical evidence is consistent with the theoretical insights.

1: Work done while working as an intern at Tencent. †: Corresponding authors.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 REPARAMETERIZABLE VISUAL RL

We consider episodic MDPs with a finite horizon. Denote the trajectory τ with length $T+1$ as $\tau=\{s_0,s_1,\ldots,s_T\}$. Denote the joint distribution of the trajectories in an episode as $\mathcal{D}_{\pi,p,p_0}$, which is jointly determined by the transition probability p, initial state distribution p_0, and the learned policy π. We assume p and p_0 are fixed and the policy is deterministic. Then, $\mathcal{D}_{\pi,p,p_0}$ becomes $\mathcal{D}_\pi$. Our goal is $\max_{\pi\in\Pi}\mathbb{E}_{\tau\sim\mathcal{D}_\pi}[J(\tau;\theta)]=\mathbb{E}_{\tau\sim\mathcal{D}_\pi}[\sum_{t=0}^T\gamma^t r(s_t,\pi(s_t))]$. We define the generalization gap as: $\|\mathbb{E}_{\tau\sim\mathcal{D}'_{\hat{\pi}}}[J(\tau)]-\frac{1}{n}\sum_{i=1}^n J(\tau_i)\|_2^2$, where $\mathcal{D}'_{\hat{\pi}}$ is the state sequence distribution in the testing environment, $\hat{\pi}=\arg\max_{\pi\in\Pi,\tau_i\in\mathcal{D}_\pi}\frac{1}{n}\sum_{i=1}^n J(\tau_i)$, where n is the number of training episodes, and Π is the policy class. It is difficult to quantify this gap since the underlying sample distribution in the training environment $\mathcal{D}_\pi$ changes as the policy evolves.

In visual RL, we denote the observation as s and the encoder as $\phi(\cdot)$. The reward gives $r(s,\pi(\phi(s)))$ and the policy is $\pi(\phi(s)):\Phi\mapsto\mathcal{A}$, where Φ is the representation space. During testing, we assume there exists the distractor $f(\cdot)$ that transforms the vanilla image s into a new image. We name $f(\cdot)$ as the *transpose function*, which can take an arbitrary form. We assume that both the transition dynamics and the state initialization process can be reparameterized, then by using the reparameterization trick [3, 7, 11, 13–15, 18, 25, 26, 29, 31], we can rewrite the objective function as follows:

$$\mathbb{E}_{\tau\sim\mathcal{D}_\pi}[J(\phi(\tau))]=\mathbb{E}_{\xi\sim q(\xi)}\left[J(\phi(\tau(\xi;\pi_\theta)))\right], \qquad (1)$$

where $q(\xi)$ is the distribution of the random variable ξ. This objective no longer depends on $\mathcal{D}_\pi$ and π. That is, we isolate the randomness of the policy π from the expected return. We denote $\mathcal{T}(s, \pi(s)) = p(s, \pi(s), s')$ as the state transition probability, and $\mathcal{I} : \Xi \mapsto \mathcal{S}$ is the initialization function, where Ξ is the space of the random variable ξs. We present the pseudo code of reparameterizable visual RL in Algorithm 1, where we reparameterize the *transition dynamics* of the system.

Algorithm 1 Reparameterizable Visual RL

```
Sample ξ_0, ξ_1, ..., ξ_T
Get s_0 = I(ξ_0) and initialize R = 0
Set encoder φ(·), policy π(·)
for t = 0 to T do
    R = R + γ^t r(s_t, π(φ(s_t)))
    s_{t+1} = T(s_t, π(φ(s_t)), ξ_t)
end for
```

Note that the random variables $\xi_0, \xi_1, \ldots, \xi_T$ can be drawn from some distributions before the episode starts, hence isolating the randomness of the policy. The above formulation also applies when the policy evolves during training, since the trajectory can be decided deterministically by executing $\mathcal{T}(s_t, \pi(\phi(s_t)), \xi_t)$ repeatedly.

3 THEORETICAL ANALYSIS ON THE GENERALIZATION ERROR

Under Lipschitz assumptions on the transition dynamics, the policy, the encoder, and the reward function, and assume $\|\phi(f(s)) - \phi(s)\| \le \varrho, \forall s$, the discrepancies of transition dynamics and the initialization function between training and testing environments gives at most ζ and ϵ, we present the generalization gap bound below (**check the full version in Arxiv for details**).

Theorem 1. *Under some mild assumptions, we have with probability at least $1 - \delta$, the generalization error gives,*

$$\left\| \mathbb{E}_\xi \left[J\left(\phi\left(f\left(\tau\left(\xi; \pi, \mathcal{T}', \mathcal{I}'\right)\right)\right)\right) \right] - \frac{1}{n}\sum_{i=1}^{n} J(\phi(\tau(\xi_i; \pi, \mathcal{T}, \mathcal{I}))) \right\|$$

$$\le \lambda\zeta \sum_{t=0}^{T} \gamma^t \frac{\nu^t - 1}{\nu - 1} + \lambda\epsilon \sum_{t=0}^{T} \gamma^t \nu^t + \frac{L_{r_2} L_{\pi_1} \varrho}{1-\gamma}(1 - \gamma^{T+1})$$

$$+ O\left(L_J K \sqrt{\frac{m}{n}}\right) + O\left(r_{\max}\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

where $\lambda, \nu, L_{r_2}, L_{\pi_1}, L_J, K$ are constants, m is the dimension of the policy parameter, $\mathcal{T}', \mathcal{I}'$ are the transition dynamics and initialization function in the testing environment.

Remark: We summarize a key insight based on the above bound, *the generalization gap can only be small if the representation distance between the training and testing environments is small*, since ϱ is the only factor that one can control in the bound. This is somewhat consistent with a human's intuition: the representations before and after involving distractors are similar and hence the policy can retrieve good behaviors it learned in the training environment.

4 EXPERIMENTAL SUPPORT

We examine whether our theory applies to existing algorithms and explains why they work in practice. We choose DrQ [32], SVEA [9], and PIE-G [33]. PIE-G and SVEA exhibit better generalization performance than DrQ [33]. We expect that the representation deviation $\|\phi(f(s)) - \phi(s)\|$ (as well as the policy deviation $\|\pi(\phi(f(s))) - \pi(\phi(s))\|$) of PIE-G and SVEA are smaller than DrQ. We verify this by conducting experiments on two environments from DMC-GB, `walker-walk` and `finger-spin`. We run these algorithms under their default hyperparameters on the clean training environment first and then replace the background with playing videos (i.e., *video-easy* setting). Our experimental setting is, the trajectory remains the same, and only backgrounds are changed. This generally meets our formulation. We evaluate the representation deviation using the learned encoder and the policy deviation with the policy network of each algorithm on the clean training trajectory and the testing trajectories with distractors for 100 episodes and 5 different random seeds. We summarize the results in Figure 1, where the empirical results are unanimously in line with our expectations. Hence, we believe our theory explains in part why these algorithms work in practice.

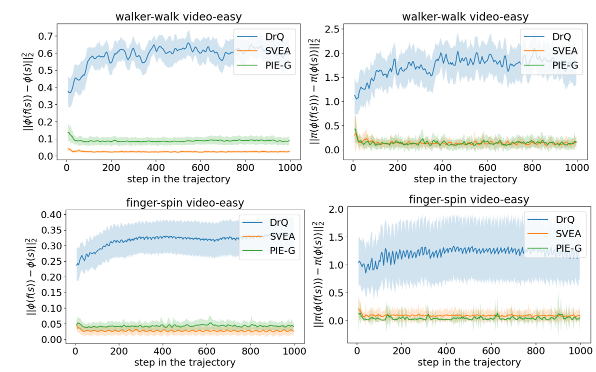

Figure 1: Comparison of representation deviation and policy deviation of SVEA, PIE-G, and DrQ on video-easy setting of walker-walk and finger-spin tasks from DMC-GB. The results are averaged over 5 varied random seeds.

5 CONCLUSIONS

Despite there are many practical algorithms for enhancing the generalization capability of visual RL policies, a clear and instructive theoretical analysis on the generalization gap, and how to minimize the generalization gap are absent. Our work aim to provide a theoretical bound on the generalization gap in visual RL when there exist distractors in the testing environment, and explain why previous methods work. However, directly analyzing the generalization gap is difficult since the policy keeps evolving. We isolate the randomness from the policy by resorting to the reparameterization trick. Our bound indicates that the key to reducing the generalization gap is to minimize the representation deviation between the training and testing environments. We further provide empirical evidence, which we find is consistent with the theoretical results.

ACKNOWLEDGMENTS

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

REFERENCES

[1] Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G Bellemare. 2021. Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=qda7-sVg84

[2] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan D. Ratliff, and Dieter Fox. 2018. Closing the Sim-to-Real Loop: Adapting Simulation Randomization with Real World Experience. In *2019 International Conference on Robotics and Automation (ICRA)*.

[3] Kamil Ciosek and Shimon Whiteson. 2020. Expected Policy Gradients for Reinforcement Learning. *Journal of Machine Learning Research* 21, 52 (2020), 1–51.

[4] Tianhong Dai, Kai Arulkumaran, Samyakh Tukra, Feryal M. P. Behbahani, and Anil Anthony Bharath. 2019. Analysing Deep Reinforcement Learning Agents Trained with Domain Randomisation. *Neurocomputing* 493 (2019), 143–165.

[5] Andrea Dittadi, Frederik Träuble, Manuel Wüthrich, Felix Widmaier, Peter Gehler, Ole Winther, Francesco Locatello, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. 2021. The Role of Pretrained Representations for the OOD Generalization of Reinforcement Learning Agents. *arXiv preprint arXiv:2107.05686* (2021).

[6] Linxi (Jim) Fan, Guanzhi Wang, De-An Huang, Zhiding Yu, Li Fei-Fei, Yuke Zhu, and Anima Anandkumar. 2021. SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies. In *International Conference on Machine Learning*.

[7] Michael Figurnov, Shakir Mohamed, and Andriy Mnih. 2018. Implicit Reparameterization Gradients. In *Neural Information Processing Systems*.

[8] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. 2021. Self-Supervised Policy Adaptation during Deployment. In *International Conference on Learning Representations*. https://openreview.net/forum?id=o_V-MjyyGV_

[9] Nicklas Hansen, Hao Su, and Xiaolong Wang. 2021. Stabilizing Deep Q-Learning with ConvNets and Vision Transformers under Data Augmentation. In *Neural Information Processing Systems*.

[10] Nicklas Hansen and Xiaolong Wang. 2020. Generalization in Reinforcement Learning by Soft Data Augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*.

[11] Nicolas Manfred Otto Heess, Greg Wayne, David Silver, Timothy P. Lillicrap, Tom Erez, and Yuval Tassa. 2015. Learning Continuous Control Policies by Stochastic Value Gradients. In *Neural Information Processing Systems*.

[12] Yangru Huang, Peixi Peng, Yifan Zhao, Guangyao Chen, and Yonghong Tian. 2022. Spectrum Random Masking for Generalization in Image-based Reinforcement Learning. In *Neural Information Processing Systems*.

[13] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkE3y85ee

[14] Diederik P. Kingma, Tim Salimans, and Max Welling. 2015. Variational Dropout and the Local Reparameterization Trick. In *Neural Information Processing Systems*.

[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).

[16] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. 2019. Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning. In *International Conference on Learning Representations*.

[17] Lu Li, Jiafei Lyu, Guozheng Ma, Zilin Wang, Zhen Yang, Xiu Li, and Zhiheng Li. 2023. Normalization Enhances Generalization in Visual Reinforcement Learning. *ArXiv* abs/2306.00656 (2023).

[18] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *ArXiv* abs/1611.00712 (2016).

[19] Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. 2021. THDA: Treasure Hunt Data Augmentation for Semantic Navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[20] P. Mitrano and Dmitry Berenson. 2022. Data Augmentation for Manipulation. *ArXiv* abs/2205.02886 (2022).

[21] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and P. Abbeel. 2017. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*.

[22] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and P. Abbeel. 2017. Asymmetric Actor Critic for Image-Based Robot Learning. *ArXiv* abs/1710.06542 (2017).

[23] Riccardo Polvara, Massimiliano Patacchiola, Marc Hanheide, and Gerhard Neumann. 2020. Sim-to-Real Quadrotor Landing via Sequential Deep Q-Networks and Domain Randomization. *Robotics* 9 (2020), 8.

[24] Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2021. Automatic Data Augmentation for Generalization in Reinforcement Learning. In *Neural Information Processing Systems*.

[25] Tim Salimans and Diederik P. Kingma. 2016. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Neural Information Processing Systems*.

[26] Frank Sehnke, Christian Osendorfer, Thomas Rückstiess, Alex Graves, Jan Peters, and Jürgen Schmidhuber. 2008. Policy Gradients with Parameter-Based Exploration for Control. In *International Conference on Artificial Neural Networks*.

[27] Reda Bahi Slaoui, William R. Clements, Jakob N. Foerster, and S'ebastien Toth. 2019. Robust Domain Randomization for Reinforcement Learning. *ArXiv* abs/1910.10537 (2019).

[28] Yu Sun, X. Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2019. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *International Conference on Machine Learning*.

[29] Huan Wang, Stephan Zheng, Caiming Xiong, and Richard Socher. 2019. On the Generalization Gap in Reparameterizable Reinforcement Learning. In *International Conference on Machine Learning*.

[30] Xudong Wang, Long Lian, and Stella X. Yu. 2021. Unsupervised Visual Attention and Invariance for Reinforcement Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Ming Xu, Matias Quiroz, Robert Kohn, and Scott Anthony Sisson. 2018. Variance Reduction Properties of the Reparameterization Trick. In *International Conference on Artificial Intelligence and Statistics*.

[32] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. 2021. Reinforcement Learning with Prototypical Representations. In *International Conference on Machine Learning*.

[33] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. 2022. Pre-Trained Image Encoder for Generalizable Visual Reinforcement Learning. In *Neural Information Processing Systems*.

[34] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto L. Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. 2019. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[35] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *International Conference on Learning Representations*. https://openreview.net/forum?id=-2FCwDKRREu