PADDLE: Logic Program Guided Policy Reuse in Deep Reinforcement Learning

Extended Abstract

Hao Zhang
College of Intelligence and Computing, Tianjin University
Tianjin, China
3018216216@tju.edu.cn

Tianpei Yang*
University of Alberta and Alberta Machine Intelligence Institute
Edmonton, Canada
tpyang@tju.edu.cn

Yan Zheng
College of Intelligence and Computing, Tianjin University
Tianjin, China
yanzheng@tju.edu.cn

Jianye Hao*
College of Intelligence and Computing, Tianjin University
Tianjin, China
jianye.hao@tju.edu.cn

Matthew E. Taylor
University of Alberta and Alberta Machine Intelligence Institute
Edmonton, Canada
matthew.e.taylor@ualberta.ca

ABSTRACT

Learning new skills through previous experience is regular in human life, which is the core idea of Transfer Reinforcement Learning (TRL). TRL requires the agent to learn *when* and *which* source policy is the best to reuse as the target task's policy and *how* to reuse the source policy. Most TRL methods learn, transfer, and reuse black-box policies, which is hard to explain: 1) when to reuse, 2) which source policy is effective, and reduces transfer efficiency. In this paper, we propose a novel TRL method called **P**rogr**A**m gui**D**e**D** po**L**icy r**E**use (PADDLE). PADDLE can measure the logic similarities between tasks and transfer knowledge which reflects the logic behind the target task. To achieve this, we propose a hybrid decision model that synthesizes high-level logic programs and learns low-level DRL policy to learn source tasks. Second, we propose a transferability metric that can measure the logic similarity between the target task and source tasks. Last, we combine it with the low-level policy similarity to select the appropriate source policy as the guiding policy for the target task. Experimental results show that PADDLE can effectively select the appropriate source tasks to guide learning on the target task, outperforming black-box TRL methods.

KEYWORDS

Deep Reinforcement Learning; Transfer Learning, Neuro-Symbolic Learning; Policy Reuse

ACM Reference Format:
Hao Zhang, Tianpei Yang*, Yan Zheng, Jianye Hao*, and Matthew E. Taylor. 2024. PADDLE: Logic Program Guided Policy Reuse in Deep 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

Although Deep Reinforcement Learning (DRL) has achieved success in various domains, it suffers from the sample inefficiency problem, making learning from scratch difficult [10, 14]. Transfer Learning (TL) has shown great potential to accelerate DRL by leveraging prior knowledge from past learned tasks [4, 6, 15, 16]. However, most existing transfer learning methods mentioned above focus on learning, extracting, and reusing black-box knowledge, which makes it difficult to reveal internal connections between source tasks and target tasks at appropriate granularity. Thus, learning when and which knowledge is effective requires significant learning costs, limits transfer effectiveness, and even fails in situations where only slight logical changes occur between different tasks.

To this end, we propose the **P**rogr**A**m gui**D**e**D** po**L**icy r**E**use (PADDLE) algorithm to address the above challenges. PADDLE incorporates a hybrid decision model to learn the policy, a two-level model combining the advantages of DRL and program synthesis, where the high-level uses program synthesis to generate logic programs [1, 3, 5, 8], and the low-level adopts arbitrary DRL methods to learn primitive policies. Then PADDLE estimates the logic similarity between each source task and the target task and combines it with the low-level policy similarity to determine which source policy should be reused at different stages (i.e., in subtasks or for subgoals). In this way, PADDLE abstracts and aligns the target task's state space and the source tasks' state spaces at a more granular representation, effectively selecting appropriate source task policies. Furthermore, the proposed similarity measurement is easily computed and relies less on the value function than advantage-based methods [13, 16], avoiding the negative influence of value estimation errors. Our contributions are summarized as: 1) A hybrid decision model is proposed that demonstrates excellent performance and logical reasoning ability. 2) A logic program-based transfer method is proposed that enables efficient knowledge transfer when learning the target task. 3) Experimental results show that PADDLE outperforms state-of-the-art transfer baselines in both discrete and continuous domains, exhibiting the advantages of knowledge with interpretable logic in TRL.

* 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).

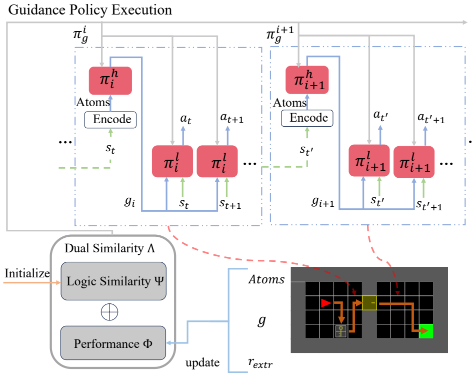

Figure 1: The workflow of PADDLE: the dual similarity and guidance policy selection.

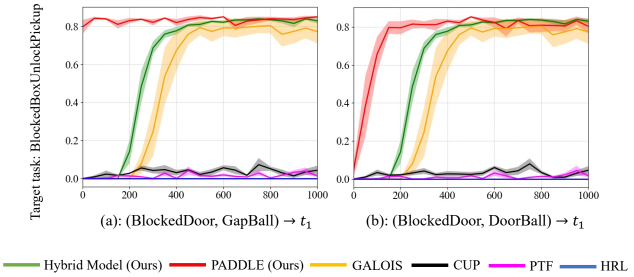

Figure 2: Results of transfer experiments in MiniGrid; The x-axis is the number of training episodes, and the y-axis is the normalized discounted reward converted according to the step size of the completed task.

2 LOGIC PROGRAM GUIDED POLICY REUSE

The details of the PADDLE algorithm are illustrated in Figure 1. Initially, a set of source policies is learned using the hybrid decision model, and a target policy is randomly initialized. The goal is to estimate the similarity between each source policy and the target policy to quickly find the appropriate source policies to guide target policy learning. PADDLE comprises two key components:

(1) A Hybrid Decision Model (HDM) uses a special hierarchical structure where the high-level policy uses the program synthesis method [8, 11, 17] to synthesize logic programs, while the low-level policy uses a DRL algorithm to learn the primitive policy. Figure 1 demonstrates how the policies are switched, and how interactions between the environment and different levels of HDM. The HDM leverages the ability of program synthesis to reveal the causal logic for a given task and different from the given sequence plan in [7] which has limited generalizability and needs more prior knowledge [5], while also releasing expert knowledge on low-level exploration tasks by lower-cost DRL methods.

(2) A transfer algorithm based on the dual similarity measurement Λ (a semantic coincidence degree Ψ and a performance function Φ), which directly optimizes the target policy by alternatively using knowledge from both the environment and from appropriate source policies. Specifically, Ψ records the most similar scene and value of each source task for each target task scene, and Φ records the average cumulative return for the execution of the selected source policy on the corresponding scene of the target task.

3 EXPERIMENTS

For the evaluation of PADDLE, we constructed some complex target tasks and source tasks that compose two sets of source tasks in MiniGrid [2], and each set consists of two source tasks. For BlockedBoxUnlockPickup (The agent's task is to move the yellow ball blocking the door, then open the box to retrieve the key to open the door, and finally put down the key to pick up the green ball behind the door.), the source tasks in the first set include all the skills required, to test whether PADDLE can quickly recombine past skills. The source tasks in the second set only include some of the skills required for BlockedBoxUnlockPickup. Still, the missing skills are in different positions in the complete skill chain, which tests whether PADDLE can recombine past skills and quickly learn new ones.

In addition, We compare various baseline algorithms, including TL methods CUP [16] and PTF [15], hierarchical algorithms [9], the original underlying algorithm, and the proposed HDM. PADDLE is applicable to all program synthesis methods and RL algorithms. In this paper, we use PPO [12] as the low-level algorithm, and GALOIS [1] as the high-level algorithm. We averaged all results on six random seeds, and the results are shown in Figure 2. Compared with baseline algorithms, PADDLE utilizes white-box knowledge of causal logic to quickly enable agents to understand which source policy may be effective at each stage, greatly reducing learning costs and improving transfer efficiency.

4 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel transfer framework, PADDLE, with a hybrid decision model as the backbone. Unlike most previous transfer methods that learned, extracted, and reused black box policies, we explore how to effectively measure the logic similarity and transfer knowledge reflecting the logic of tasks behind, which further improves transfer efficiency. PADDLE is simple to implement and easy to combine with existing DRL algorithms. Experimental results show that PADDLE outperforms previous state-of-the-art transfer methods. As for future work, it is worthwhile extending PADDLE to multiagent problems to capture the transferable knowledge among multiple agents, even heterogeneous agents. Another direction is to investigate how to learn the optimal logic program from human feedback. Specifically, logic programs with clear semantics and interpretability are intuitive to humans and can be fine-tuned by introducing human feedback in learning.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China(Grant No. 92370132) and the National Key R&D Program of China (Grant No. 2022ZD0116402). Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; Compute Canada; Huawei; Mitacs; and NSERC.

REFERENCES

[1] Yushi Cao, Zhiming Li, Tianpei Yang, Hao Zhang, Yan Zheng, Yi Li, Jianye Hao, and Yang Liu. 2022. GALOIS: boosting deep reinforcement learning via generalizable logic synthesis. *Advances in Neural Information Processing Systems* 35 (2022), 19930–19943.

[2] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. *CoRR* abs/2306.13831 (2023).

[3] Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research* 61 (2018), 1–64.

[4] Fernando Fernández and Manuela M. Veloso. 2006. Probabilistic policy reuse in a reinforcement learning agent. In *5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan, May 8-12, 2006*. ACM, New York, NY, USA, 720–727.

[5] Claire Glanois, Zhaohui Jiang, Xuening Feng, Paul Weng, Matthieu Zimmer, Dong Li, Wulong Liu, and Jianye Hao. 2022. Neuro-Symbolic Hierarchical Rule Induction. In *International Conference on Machine Learning, ICML 2022*, Vol. 162. PMLR, Baltimore, Maryland, 7583–7615.

[6] Steven R Guberman and Patricia M Greenfield. 1991. Learning and transfer in everyday cognition. *Cognitive Development* 6, 3 (1991), 233–260.

[7] León Illanes, Xi Yan, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2020. Symbolic Plans as High-Level Instructions for Reinforcement Learning. In *Proceedings of the 30th International Conference on Automated Planning and Scheduling (ICAPS)*. AAAI Press, 540–550.

[8] Zhengyao Jiang and Shan Luo. 2019. Neural Logic Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, Vol. 97. PMLR, Long Beach, California, USA, 3110–3119.

[9] Tejas D. Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. Curran Associates Inc., Red Hook, NY, USA, 3675–3683.

[10] Johan Samir Obando-Ceron and Pablo Samuel Castro. 2021. Revisiting Rainbow: Promoting More Insightful and Inclusive Deep Reinforcement Learning Research. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Vol. 139. PMLR, Virtual Event, 1373–1383.

[11] Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end Differentiable Proving. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. Curran Associates Inc., Red Hook, NY, USA, 3788–3800.

[12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017).

[13] Yunzhe Tao, Sahika Genc, Jonathan Chung, Tao Sun, and Sunil Mallya. 2021. REPAINT: Knowledge Transfer in Deep Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, Vol. 139. PMLR, Virtual Event, 10141–10152.

[14] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[15] Tianpei Yang, Jianye Hao, Zhaopeng Meng, Zongzhang Zhang, Yujing Hu, Yingfeng Chen, Changjie Fan, Weixun Wang, Wulong Liu, Zhaodong Wang, and Jiajie Peng. 2020. Efficient Deep Reinforcement Learning via Adaptive Policy Transfer. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. ijcai.org, Yokohama, Yokohama, Japan, 3094–3100.

[16] Jin Zhang, Siyuan Li, and Chongjie Zhang. 2022. Cup: Critic-guided policy reuse. *Advances in Neural Information Processing Systems* 35 (2022), 27537–27548.

[17] Matthieu Zimmer, Xuening Feng, Claire Glanois, Zhaohui Jiang, Jianyi Zhang, Paul Weng, Jianye Hao, Dong Li, and Wulong Liu. 2021. Differentiable Logic Machines. *CoRR* abs/2102.11529 (2021).