Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Doctoral Consortium Track

Deep Reinforcement Learning with Hierarchical Structures

Siyuan Li
Institute for Interdisciplinary Information Sciences, Tsinghua University
sy-li17@mails.tsinghua.edu.cn

Abstract

Hierarchical reinforcement learning (HRL), which enables control at multiple time scales, is a promising paradigm to solve challenging and long-horizon tasks. In this paper, we briefly introduce our work in bottom-up and top-down HRL and outline the directions for future work.

1 Introduction

Deep reinforcement learning (RL) has recently made significant progress in various domains such as games and continuous control for robotics [Gu *et al.*, 2017; Badia *et al.*, 2020]. Nevertheless, solving long-horizon tasks with sparse rewards remains a major challenge for these methods. Hierarchical reinforcement learning (HRL), which aims at decomposing complex tasks to easier subtasks with a hierarchical structure, has shown great potential in extending the successes of existing RL methods to more difficult and temporally extended tasks [Nachum *et al.*, 2018a].

In a two-level hierarchical policy, the high level communicates commands to the low level over a relatively longer time scale, and the low level takes a primitive action every timestep conditioned on the received command and the current state. When the high-level actions correspond to diverse low-level behaviors, following a low-level behavior for multiple timesteps could lead to better exploration ability. Furthermore, the hierarchical structure makes compositional generalization possible among related tasks. For example, both playing football and play basketball needs the running skill, so the low-level policy of running could be transferred between these two tasks, thus achieving more efficient learning.

2 Current Work

The hierarchical reinforcement learning approaches could be roughly divided to two categories, bottom-up methods and top-down methods. The bottom-up methods propose to firstly train low-level policies with unsupervised or self-supervised objectives [Eysenbach *et al.*, 2018; Sharma *et al.*, 2019], and then learn a high-level policy to compose them to solve a difficult downstream task. The top-down methods decompose long-horizon tasks with subgoals [Nachum *et al.*, 2018a; Li *et al.*, 2021a] or termination functions [Bacon *et al.*, 2017], and learn multi-level policies simultaneously. The following of this section briefly introduces our work in bottom-up and top-down HRL.

2.1 Bottom-Up HRL

Our work in bottom-up HRL focuses on the problem of low-level policy selection [Li and Zhang, 2018; Li *et al.*, 2019a] and adaptation [Li *et al.*, 2019b]. The previous work either selects low-level policies via a soft-max method [Fernández and Veloso, 2013], or with human knowledge [Taylor *et al.*, 2007]. To efficiently select the most appropriate low-level policy from a policy set, we propose to formulate the policy selection problem as a Multi-Armed Bandit problem, and utilize the Upper Confidence Bound algorithm to achieve the optimal low-level policy selection [Li and Zhang, 2018]. The selected low-level policy is combined with a random policy to guide the exploration in the environment. However, this method assumes that there is one low-level policy much similar to the target policy to be learned. To release this assumption, we propose a context-aware policy selection method [Li *et al.*, 2019a], which formulates the low-level policies as options [Sutton *et al.*, 1999]. With a call-and-return execution model, our method not only learns when to select which policy, but also learns when to terminate the selected policy.

Most bottom-up HRL methods only learn the high-level policies in the downstream tasks with fixed low-level policies [Florensa *et al.*, 2017; Eysenbach *et al.*, 2018]. However, using fixed low-level policies without further adaptation may be insufficient for solving complex tasks. To adapt the low-level policies in tasks with sparse reward signals, we proposed a novel auxiliary reward function for low-level policy learning [Li *et al.*, 2019b]. Since we would like the low-level agent to explore the promising states with larger high-level values, the auxiliary rewards are defined with the high-level advantage function. Experimental results on the benchmark MuJoCo tasks [Todorov *et al.*, 2012] demonstrate that the learning performance is greatly improved with the help of the proposed auxiliary rewards.

2.2 Top-Down HRL

In top-down goal-conditioned HRL, the high-level policy sets subgoals to the low level, and the low-level policies are trained to reach those subgoals. A crucial problem is how to learn an effective subgoal representation, since explorative

low-level behaviors could be induced by setting subgoals in this representation space. Previous methods learn the subgoal space either by bounding the sub-optimality of the hierarchical policy [Nachum *et al.*, 2018b], or in an end-to-end manner with policy learning [Dilokthanakul *et al.*, 2019]. Notice that the high-level agent makes decision at a low temporal resolution and the subgoal space is the high-level action space, we proposed to learn the subgoal representation with the slowness objective [Li *et al.*, 2021b], which is optimized with the contrastive loss [Chopra *et al.*, 2005]. Furthermore, we provide a theoretical grounding for the proposed slowness objective that selecting slow features as the subgoal representation is the optimal for exploration when the dimension of the subgoal space is fixed.

3 Future Directions

In the future, I would like to further improve the subgoal representation learning in large scale problems with the Long-Short Term Memory networks and the reconstruction loss. In addition, accelerating the high-level policy learning with model-based or episodic control methods could be an interesting future direction as well.

References

[Bacon *et al.*, 2017] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Badia *et al.*, 2020] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pages 507–517. PMLR, 2020.

[Chopra *et al.*, 2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[Dilokthanakul *et al.*, 2019] Nat Dilokthanakul, Christos Kaplanis, Nick Pawlowski, and Murray Shanahan. Feature control as intrinsic motivation for hierarchical reinforcement learning. *IEEE transactions on neural networks and learning systems*, 30(11):3409–3418, 2019.

[Eysenbach *et al.*, 2018] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[Fernández and Veloso, 2013] Fernando Fernández and Manuela Veloso. Learning domain structure through probabilistic policy reuse in reinforcement learning. *Progress in Artificial Intelligence*, 2(1):13–27, 2013.

[Florensa *et al.*, 2017] Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012*, 2017.

[Gu *et al.*, 2017] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

[Li and Zhang, 2018] Siyuan Li and Chongjie Zhang. An optimal online method of selecting source policies for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Li *et al.*, 2019a] Siyuan Li, Fangda Gu, Guangxiang Zhu, and Chongjie Zhang. Context-aware policy reuse. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 989–997, 2019.

[Li *et al.*, 2019b] Siyuan Li, Rui Wang, Minxue Tang, and Chongjie Zhang. Hierarchical reinforcement learning with advantage-based auxiliary rewards. In *Advances in Neural Information Processing Systems*, pages 1409–1419, 2019.

[Li *et al.*, 2021a] Siyuan Li, Jin Zhang, Jianhao Wang, and Chongjie Zhang. Efficient hierarchical exploration with stable subgoal representation learning. *arXiv preprint arXiv:2105.14750*, 2021.

[Li *et al.*, 2021b] Siyuan Li, Lulu Zheng, Jianhao Wang, and Chongjie Zhang. Learning subgoal representations with slow dynamics. In *International Conference on Learning Representations*, 2021.

[Nachum *et al.*, 2018a] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:1805.08296*, 2018.

[Nachum *et al.*, 2018b] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. *arXiv preprint arXiv:1810.01257*, 2018.

[Sharma *et al.*, 2019] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

[Sutton *et al.*, 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[Taylor *et al.*, 2007] Matthew E Taylor, Peter Stone, and Yaxin Liu. Transfer learning via inter-task mappings for temporal difference learning. *Journal of Machine Learning Research*, 8(9), 2007.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.