Hierarchical Reinforcement Learning with Human-AI Collaborative Sub-Goals Optimization

Extended Abstract

Haozhe Ma
National University of Singapore
Singapore
haozhe.ma@comp.nus.edu.sg

Thanh Vinh Vo
National University of Singapore
Singapore
votv@comp.nus.edu.sg

Tze-Yun Leong
National University of Singapore
Singapore
leongty@comp.nus.edu.sg

ABSTRACT

Hierarchical reinforcement learning requires identifying relevant sub-goals to guide low-level decision-making, but this process can be time-consuming and challenging. Moreover, manually specifying sub-goals may introduce bias or mislead agents. To address these issues, we propose a collaborative human-AI algorithm that automatically optimizes candidate sub-goals and refines prior knowledge. Our algorithm can be integrated into various hierarchical frameworks and effectively prevent negative inferences that may arise from conflicting sub-goals. Our approach is robust in the face of different levels of human knowledge and able to accelerate convergence to optimal sub-goals and hierarchical policies.

KEYWORDS

Human-AI Collaboration; Hierarchical Reinforcement Learning; Sub-Goals Discovery; Variational Inference

ACM Reference Format:
Haozhe Ma, Thanh Vinh Vo, and Tze-Yun Leong. 2023. Hierarchical Reinforcement Learning with Human-AI Collaborative Sub-Goals Optimization: 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

Hierarchical reinforcement learning (HRL) is a promising approach for solving complex problems involving long-duration tasks with delayed and sparse rewards. By modeling problems at different levels of abstraction, HRL can improve learning efficiency and reduce computational burden. It can also facilitate transfer learning by enabling the reuse of high-level policies. One common approach to designing hierarchical structures is to divide the overall target into multiple sub-tasks by setting corresponding sub-goals. Many popular efforts focus on the two-level hierarchical structure [1, 3, 8, 9]: the high level optimizes the policy to select a sub-goal representing a short-term task; the low level learns the policies to achieve the targeted sub-goals. However, defining appropriate sub-goals often requires extensive domain knowledge. Moreover, the sub-goal space introduces bias and in severe cases, some confusing sub-goals may lead to sub-optimal policies.

To automatically detect and correct misleading human knowledge or confusing sub-goals in different solution contexts, we propose a Human-AI collaborative sub-Goal Optimization (HAI-GO) algorithm. Unlike the approaches that rely entirely on automatic discovery [4–7, 10, 12], our algorithm leverages human-AI cooperation, where humans encode general and domain-specific knowledge in defining the sub-goals, while machines optimize sub-goal selection in deriving optimal policies. Given a candidate sub-goal space, HAI-GO maintains a critic function to evaluate the utility of selecting each sub-goal. The algorithm can be flexibly embedded into a wide range of HRL frameworks without modifying their original structures, enabling the agent to determine an optimal sub-goal space and converge to the corresponding optimal hierarchical policies.

We evaluate our HAI-GO algorithm in complex maze environments and finds that it effectively identifies optimal sub-goals based on relevant performance measures. Our results also show that HAI-GO is robust in detecting and filtering out potentially confusing sub-goals across different degrees of human knowledge integration. Our algorithm outperforms state-of-the-art HRL baselines even when pre-defined knowledge includes misleading sub-goals.

2 METHODOLOGY

We consider an environment $\mathcal{E}$ that our intelligent agent interacts with. Suppose $G = \{g_1, g_2, ..., g_N\}$ is a candidate sub-goal space defined based on prior knowledge. We assume that these sub-goals are responsibly defined and cover a subset of positive decomposition of the overall task. We define a critic function represented as a set of independent Bernoulli distributions for each candidate sub-goal as $\mathbf{q} = \{q_1(w_1; \lambda_1), q_2(w_2; \lambda_2), ..., q_N(w_N; \lambda_N)\}$. For each $q_i(w_i; \lambda_i)$, the random variable $w_i \in \{0, 1\}$ indicates to select sub-goal g_i by $w_i = 1$ or not to select it by $w_i = 0$. We initialize an HRL agent with the high-level module and low-level module. Our HAI-GO algorithm simultaneously learns both the critic function and the hierarchical policies. The optimal sub-goal space G^* can be finally obtained based on the learned critic function after training.

2.1 Hierarchical Structure with Sub-Goal Policy

HAI-GO is designed as an additional component in HRL agents that learns a high-level policy to select one sub-goal as a short-term target. One simple prototype consists of two levels: at each time step, the high level selects a sub-goal based on its policy π^h representing a short-term task that the agent is expected to complete in next stage; the low level selects an elementary action based on its policy π^l in the following N steps, where $N > 1$ is a hyper-parameter representing the expected steps for the low level to complete a particular sub-goal. The high level revises a new sub-goal after N steps or after the low level completes the current one.

The high level interaction model is defined by a Markov Decision Process (MDP) $\langle S, G, T^h, R^h, \gamma^h \rangle$. The main target is to learn

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.

the policy $\pi^{h^*} : S \rightarrow G$ to maximize the discounted high-level return $R_t^h = \sum_{\tau=t}^{\infty} \gamma^{h^{\tau-t}} r_t^h$. We implement Q-learning-based algorithms to approximate the $Q^h(s, g; \theta)$ by minimizing the temporal difference error (TD-error), i.e., the distance between temporal difference target (TD-target) $y_t = r_t^h + \gamma^h \max_{g'} Q^h(s_{t+N}, g'; \theta)$ and the predicted Q-value $Q^h(s_t, g_t; \theta)$. The low level learns a policy to select the elementary action given a state s_t as well as the sub-goal g_i instructed by the high level, $a_t = \pi^l(s_t|g_i)$, which can be trained by any applicable flat RL algorithms. We will show how HAI-GO can be embedded into this general framework.

2.2 HRL with Sub-Goal Optimization

Our proposed HAI-GO algorithm integrates human expertise with automatic calculation, enabling the agent to start from a human-specified sub-goal space and gradually refine the candidate knowledge. The agent learns the critic function $\mathbf{q} = \{q_i(w_i; \lambda_i)\}$ to generate filtered sub-goal spaces $\hat{G}$ during training. $\hat{G}$ only contains the sub-goals whose entry $w_i = 1$ dominates the entry $w_i = 0$, where the high level will select one sub-goal from. We define $Q_{\hat{G}}(s, g; \theta)$ as the Q-function conditional on the filtered sub-goal space $\hat{G}$. Similarly, we denote the conditional TD-target and the loss function as $y_{\hat{G}}$ and $L_{\hat{G}}$ respectively. Based on the conditional assumption, we have:

$$L_{\hat{G}} = 0.5\big(y_{\hat{G}} - Q_{\hat{G}}(s, g; \theta)\big)^2. \tag{1}$$

The main objective of HAI-GO to optimize the critic function is to update $q_i(w_i; \lambda_i)$ to be one best approximation to the real posterior $p_i(w_i|y_{\hat{G}})$. The posterior gives the distribution of indicator w_i conditional on the corresponding TD-target. We adopt a variational inference approach [2, 13] to optimize the parameters λ_i. We minimize the KL-divergence of $q_i(w_i; \lambda_i)$ and $p_i(w_i|y_{\hat{G}})$ for $i = 1, 2, ..., N$, which is

$$D_{KL}\big(q_i(w_i; \lambda_i)||p_i(w_i)\big) - \mathbb{E}_{w_i \sim q_i(w_i;\lambda_i)}[\log p(y_{\hat{G}}|w_i)],$$

where $p_i(w_i) \sim \text{Bernoulli}(\delta_i)$ is a prior, and δ_i is a hyper-parameter. Eq. (1) indicates that $y_{\hat{G}} = Q_{\hat{G}}(s, g; \theta) + \epsilon_{\hat{G}}$, where $\epsilon_{\hat{G}} \sim \mathcal{N}(0, \sigma^2)$. Hence, we have $\log p(y_{\hat{G}}|w_i) = -L_{\hat{G}} + constant$. Thus, the loss function for each candidate is:

$$L(\lambda_i) = \mathbb{E}_{w_i \sim q_i(w_i;\lambda_i)}[L_{\hat{G}}] + D_{KL}\big(q_i(w_i; \lambda_i)||p_i(w_i)\big).$$

As we assume that all Bernoulli distributions are independent, we minimize the total loss: $L(\lambda) = \sum_{i=1}^{N} L(\lambda_i)$. To timely influence the agent training, we adopt an ϵ-greedy strategy to control our HAI-GO component to gradually affect the high-level policy learning by providing the filtered $\hat{G}$. With an increasing probability of ϵ, the high level selects one sub-goal only from $\hat{G}_t$, with the probability of $1 - \epsilon$, from the initial candidate space. The HAI-GO embedded HRL will converge to both the optimal critic function and the optimal policies. After learning, we derive the optimal sub-goal space $G*$ based on the learned distributions.

3 EXPERIMENTS

3.1 Sub-Goal Discovery

In this section, we show the sub-goal discovery of our HAI-GO compared with two baselines: the L-Cut [11], a graph-theory-based approach and the HADS [4], a pre-trained process before HRL learning. We compute the normalized difference between the two entries $\phi_{g_i} = q(w_i = 1; \lambda_i) - q(w_i = 0; \lambda_i)$ to indicate the intensity of selecting each candidate. The optimized distribution is shown in Figure 1. In addition to indicating the two paths as the most important sub-goals, our results present an interesting feature, that is, the closer to the final state the more important the candidate is, which is more reasonable from the human perspective.

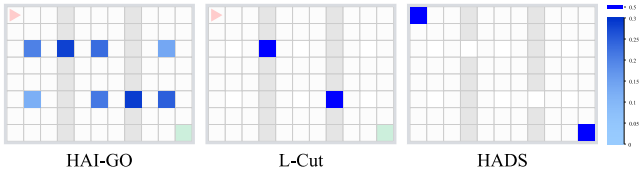
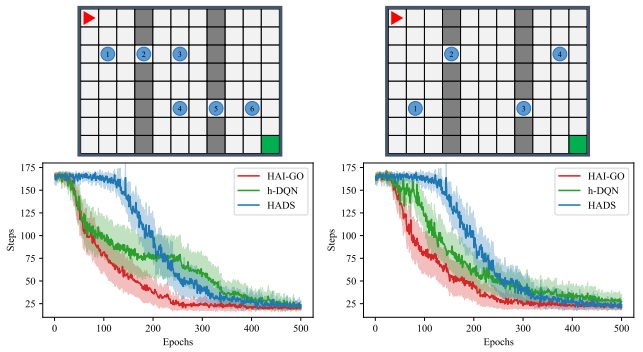


Figure 1: Comparison of the discovered sub-goals.



(a) general configuration. **(b) confusing configuration.**

Figure 2: Comparison of human knowledge refinement.

3.2 Human Knowledge Refinement

In this section, we compare our HAI-GO with two HRL baselines: h-DQN [3] and HADS [4], to evaluate the learning performance and the ability to refine the encoded human knowledge. We designed two configurations representing different degrees of prior knowledge: one with **general** candidates which include grids located in the rooms; the other one with **confusing** sub-goals that may mislead the agent to useless exploration. The two configurations with their corresponding convergence are shown in Figure 2. As compared to the baselines, our approach learns the importance of each sub-goal and applies it to the agent training. The unimportant and confusing sub-goals can be filtered out and only the optimal ones are retained, thus resulting in the fastest convergence.

4 CONCLUSION

We proposed HAI-GO, a human-AI collaborative sub-goals optimization algorithm that integrates human expertise into intelligent agent learning. HAI-GO maintains a critic function to eliminate biases and refine encoded human knowledge, resulting in faster convergence and stable learning performance. The optimal sub-goal space derived provides a better understanding of complex environments, and the algorithm is highly expandable. Future work should focus on real-time human-AI collaboration, defining better performance measures, and accurately defining sub-goal spaces.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016) and a Research Scholarship from the Ministry of Education in Singapore.

REFERENCES

[1] Bram Bakker, Jürgen Schmidhuber, et al. 2004. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In *Proceedings of the 8-th Conference on Intelligent Autonomous Systems*. Citeseer, 438–445.

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.

[3] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems* 29 (2016).

[4] Chenghao Liu, Fei Zhu, Quan Liu, and Yuchen Fu. 2021. Hierarchical reinforcement learning with automatic sub-goal identification. *IEEE/CAA journal of automatica sinica* 8, 10 (2021), 1686–1696.

[5] Sridhar Mahadevan and Mauro Maggioni. 2007. Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. *Journal of machine learning research* 8, 10 (2007).

[6] Amy McGovern and Andrew G Barto. 2001. Automatic discovery of subgoals in reinforcement learning using diverse density. (2001).

[7] Ishai Menache, Shie Mannor, and Nahum Shimkin. 2002. Q-cut—dynamic discovery of sub-goals in reinforcement learning. In *European conference on machine learning*. Springer, 295–306.

[8] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. 2018. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems* 31 (2018).

[9] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–35.

[10] Özgür Şimşek and Andrew Barto. 2008. Skill characterization based on betweenness. *Advances in neural information processing systems* 21 (2008).

[11] Özgür Şimşek, Alicia P Wolfe, and Andrew G Barto. 2005. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on Machine learning*. 816–823.

[12] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. 2018. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play. In *International Conference on Learning Representations*.

[13] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. 2018. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 2008–2026.