Interactive RL via Online Human Demonstrations

Extended Abstract

Chao Yu
School of Data & Computer Science, Sun Yat-Sen University, Guangzhou, China
yuchao3@mail.sysu.edu.cn

Tianpei Yang
Computer Science Department, New York University, New York, U.S.A
1464439923@qq.com

Wenxuan Zhu
School of Computer Science & Technology, Dalian University of Technology, Dalian, China
zhuwenxuan@mail.dlut.edu.cn

Yinzhao Dong
School of Computer Science & Technology, Dalian University of Technology, Dalian, China
1447866357@qq.com

Guangliang Li
College of Information Science & Engineering, Ocean University of China, Qingdao, China
guangliangli@oue.edu.cn

ABSTRACT

In this paper, we propose a general approach that uses online human demonstrations to directly shape an agent's behaviors. This approach can alleviate the uncertainties caused by human critiques, while at the same time, removing the offline pre-training in most existing learning from demonstration approaches. Using this approach, we also investigate the interplay among different shaping methods for more robust and efficient interactive learning between humans and agents.

KEYWORDS

Reinforcement Learning; Interactive Learning; Human-Agent Interaction; Shaping.

ACM Reference Format:
Chao Yu, Tianpei Yang, Wenxuan Zhu, Yinzhao Dong, and Guangliang Li. 2020. Interactive RL via Online Human Demonstrations. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Interactive Reinforcement Learning (InterRL) provides a human-in-the-loop computing paradigm that enables the integration of human knowledge (in terms of advice, preference or demonstrations) into agent learning process such that the overall learning cost can be reduced [11]. Thus far, plenty of work has investigated how humans can help RL agents to learn more efficiently through different assumptions of interaction modes, combination methods or transferred knowledge [3, 5–7, 10, 12]. However, these approaches require constantly monitoring and labeling the agent's behaviors, which is effort-consuming due to human's cognitive burdens, as well as the complexity of states and actions. Compared to observing and critiquing an agent's behaviors, a more direct way is to let humans provide more explicit examples of desired behaviors. *Learning from demonstration* (LfD) [14] enables an agent to learn a complex task by using human demonstrations in solving the same task. However, most existing LfD methods often rely on sophisticated supervised learning techniques to fit the demonstrator's behavior, but typically do not use environmental reward signal to improve the agent's learning policy. While some studies combine human demonstrations with RL rewards [4] , it requires a prior offline process to collect human demonstrations in order to derive a potential function for reward shaping. This assumption is not always possible in agent-human interaction settings, when an agent must interact with humans in a timely online manner.

The other common issue of current research is the generalization problem. The human knowledge is integrated into agent learning process by shaping a specific component of RL, i.e., the value function, action, reward or policy [1]. While this specific design can leverage explict learning algorithm or representation to derive powerful InterRL methods, it inevitably faces generalization and interpretation problems. The methods may perform well for some types of algorithms or domains, but poorly on others, thus, general observations or conclusions claimed may not hold consistently. Thus, due to lack of deep and consistent understanding of their benefits and shortcomings, existing InterRL methods usually cannot be readily applied in various domain settings without substantial effort and further hand-engineering.

To this end, this paper proposes a new InterRL approach that resorts to direct human demonstrations to speed up learning efficiency by biasing the agent's exploration process, but at the same time, uses the agent's environmental reward signals to guarantee final learning performance. A human trainer provides immediate demonstrations using the same environmental inputs as the agent. Then, these demonstrations are directly integrated into the agent' learning process in order to shape its behaviors. Using this approach, this paper then proposes an adaptive shaping algorithm that is capable of combining the benefits of several shaping methods for more robust and efficient interactive learning between the agent and human. By implementing the algorithm in two

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand. © 2020 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

classic RL domains and analyzing the interplay dynamics among different InterRL methods, some interesting conclusions can be achieved, which provide valuable insights into understanding the role and impact of various shaping methods and human factors (the likelihood, correctness and weight of human feedbacks) in human-agent interactive learning.

2 DESIGN

In the existing InterRL methods, human trainers perceive the agents' states, actions or trajectories and provide scalar numeric feedbacks or preferences that indicate how good of the current behavior. The agent, in response, tries to learn a policy directly based on the human feedback signals, or estimate a hidden function that approximates the human feedback such that its behavior can be adjusted to align with this estimate. Unlike these *observing-and-critiquing* InterRL methods, we let human trainers step in the agent's learning process while providing demonstrations at the same time. At each time step, with certain possibility, a human trainer makes decisions based on the same environmental inputs and action choices as the agent. A reward vector is employed to indicate the human's reward signals regarding all the action choices in a state. When the human has chosen an action at this time step, the corresponding value in the reward vector is filled with a positive value of r_h, and the values for other actions are filled with $-r_h$. Therefore, the value of r_h indicates the influence magnitude of human's demonstrations on the agent's learning process. The reward vector is then applied to update the accumulated human reward function H, which is then intergraded into the agent's learning process in order to shape its behaviors using various InterRL methods.

Since different kinds of InterRL methods embody various characteristics and advantages in different domains, it is natural to explore the combinatoric space of these learning methods in order to derive more robust InterRL methods. Using the above design methodology, we then investigate how different kinds of InterRL methods can interact with each other, and how this interplay can impact the final learning performance. Assuming a set of InterRL methods, in each learning episode, we choose an InterRL method based on its weight. Then, the chosen InterRL method runs for one episode and returns the similarity value for each method in the portfolio and the total accumulated reward for the current running method. Since the similarity maintains the policy similarity between each method and the current running method, the weights of all methods can be updated proportionally to the similarity between these two methods.

3 EXPERIMENT

We evaluate our method in the two benchmark RL domains Pac-Man and Cart-Pole. Following [9], four explicit InterRL methods are considered: the *Action Biasing* (AB), the *Control Sharing* (CS), the *Reward Shaping* (RS) and the *Q Augmentation* (QA). In order to quantitatively study the advantages and disadvantages of different methods in various parameter settings, we first obtained the optimal policy using standard Q-learning to simulate completely correct human guidance, and introduced two extra parameters: L (the *likelihood* of feedback) which represents the probability that the human provides guidance at each time step, and C (the *consistency* of feedback) which represents the probability that the human provides the optimal instructions correctly [2, 5, 8, 13]. Results under different settings of L and C show that none of the four individual InterRL methods can achieve a parameter-independent performance. Especially, human factors such as correctness and influence of human guidance play a crucial role in biasing the performance of each InterRL method. Generally, when the correctness of human guidance is not high enough, the individual InterRL methods are more likely to fail and may lead to divergence of learning process. Moreover, for the value-based and reward-based methods to be more efficient, human reward should be set to a relatively low value to reduce its influence. The proposed adaptive shaping algorithm, due to the interplay between different methods, can take the benefits of each method to achieve a more robust and efficient learning performance.

We then report how real human trainers will perform using the proposed InterRL approach and the adaptive shaping algorithm. Results show that AB and CS are generally more efficient than RS and QA in real human scenarios. Although the individual InterRL methods perform variously in the same parameter setting, the dynamic interplay of these methods can greatly promote learning performance. The interplay of InterRL methods presents an interesting phenomenon. In Pac-Man, although AB alone can already guarantee a good performance, it still faces the problem that an occasional error action given by the human can potentially bias the agent's learning process. By slightly choosing other InterRL methods during the early stage, particularly those indirectly affecting the actions and policies (RS and QA), the learning performance can be greatly promoted. Situations are a bit different in Cart-Pole where the four individual methods perform similarly and only slightly better than Q-learning. Since the benefit of human learning is not as apparent as that in simpler discrete domains, no specific InterRL methods can dominate the dynamic interplay process.

Results also show that the simulated oracle produces lower quality performance than the real human. The result is a bit surprising since it is believed that the optimal strategy generated by the oracle should be more useful than the flawed data by the human. This phenomenon suggests the important role of human factors in an agent's RL process, that is, although human reinforcement is generally flawed, the informationally rich knowledge of human learning can bring about significant benefits over the flawless yet poor agent learning. Compared to agents, humans optimize for outcomes (e.g., exploration, novelty, or near-danger experiences), which are not always directly related to task performance [7].

ACKNOWLEDGMENTS

This work is supported by the Dalian Science and Technology Innovation Fund under Grant 2018J12GX046.

REFERENCES

[1] David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. 2017. Agent-Agnostic Human-in-the-Loop Reinforcement Learning. *arXiv preprint arXiv:1701.04079* (2017).

[2] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara Grosz. 2016. Interactive Teaching Strategies for Agent Training. In *IJCAI2016*. 804–811.

[3] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. 2019. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257* (2019).

[4] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, and Matthew E. Taylor. 2015. Reinforcement learning from demonstration through shaping. In *IJCAI2015*. 3352–3358.

[5] Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea Lockerd Thomaz. 2015. Policy Shaping with Human Teachers.. In *IJCAI2015*. 3366–3372.

[6] Carlos Celemin and Javier Ruiz-del Solar. 2018. An interactive framework for learning continuous actions policies based on corrective feedback. *Journal of Intelligent & Robotic Systems* (2018), 1–21.

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *NIPS2017*. 4299–4307.

[8] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *NIPS2013*. 2625–2633.

[9] W Bradley Knox and Peter Stone. 2010. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *AAMAS2010*. 5–12.

[10] W Bradley Knox and Peter Stone. 2012. Reinforcement learning from simultaneous human and MDP reward. In *AAMAS2012*. 475–482.

[11] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. 2019. Human-Centered Reinforcement Learning: A Survey. *IEEE Transactions on Human-Machine Systems* 49, 4 (2019), 337–349.

[12] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *ICML2017*. 2285–2294.

[13] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. 2017. Where to Add Actions in Human-in-the-Loop Reinforcement Learning.. In *AAAI2017*. 2322–2328.

[14] Matthew E Taylor and AI Borealis. 2018. Improving Reinforcement Learning with Human Input. In *IJCAI2018*. 5724–5728.