Internal knowledge representation on trainer agents by using interactive reinforcement learning

Anonymous Author(s)

Affiliation
Email

Abstract

Interactive reinforcement learning has become an important apprenticeship approach to speed up convergence in classic reinforcement learning problems. On some occasions, the trainer may be another artificial agent which in turn was trained using reinforcement learning methods to afterward become an advisor for other learner-agents. In this work, we analyze internal representation and characteristics of artificial trainer-agents to determine which agent may outperform others to become a better advisor.

1 Motivation and Research Problem

Reinforcement learning (RL) [1] is a behavior-based approach which allows an agent, either an infant or a robot, to learn a task by interacting with its environment and observing how the environment responds to the agent's actions. RL has been shown in robotics [2, 3] and in infant studies [4, 5] to be successful in terms of acquiring new skills, by mapping situations to actions [6].

Interactive RL allows to speed up the apprenticeship process by using a parent-like advisor to support the learning by giving useful advice in selected episodes. This allows to reduce the search space and thus to learn the task faster in comparison to an agent exploring fully autonomously [7]. In this regard, the parent-like advisor guides the learning robot, enhancing its performance in the same manner as external caregivers may support infants in the accomplishment of a given task, with the provided support frequently decreasing over time.

By using artificial agents as trainers, some properties have been studied so far such as different effects of giving advice in different episodes and with different strategies during the learning process [8,9] and effects of different probabilities and consistency of feedback [10,11]. In this work, we study possible implications of utilizing artificial trainers with different characteristics and different internal representations of the knowledge based on their previous experience.

The scenario consists of two robots in front of a table to learn a cleaning task (Fig. 1a). We defined *objects*, *locations*, and *actions*. The scene includes two objects: a *sponge* and a *goblet*. The table is divided into three zones, the *left* and *right* table sides and an additional position called *home* where we place the sponge during the execution of the task. We allow a robot to perform seven actions: *get*, *drop*, *go home*, *go left*, *go right*, *clean*, and *abort*.

As long as the agent successfully finishes the task, a reward equal to 1 is given to it, whereas a reward of -1 is given if a failed-state was reached. In this context, a failed-state is a state from where the robot cannot continue the expected task execution, for instance attempting to pick-up an object when another is already held. Furthermore, in intermediate states, it is given a small negative reward of -0.01 to encourage the agent to take shorter paths towards a final state.

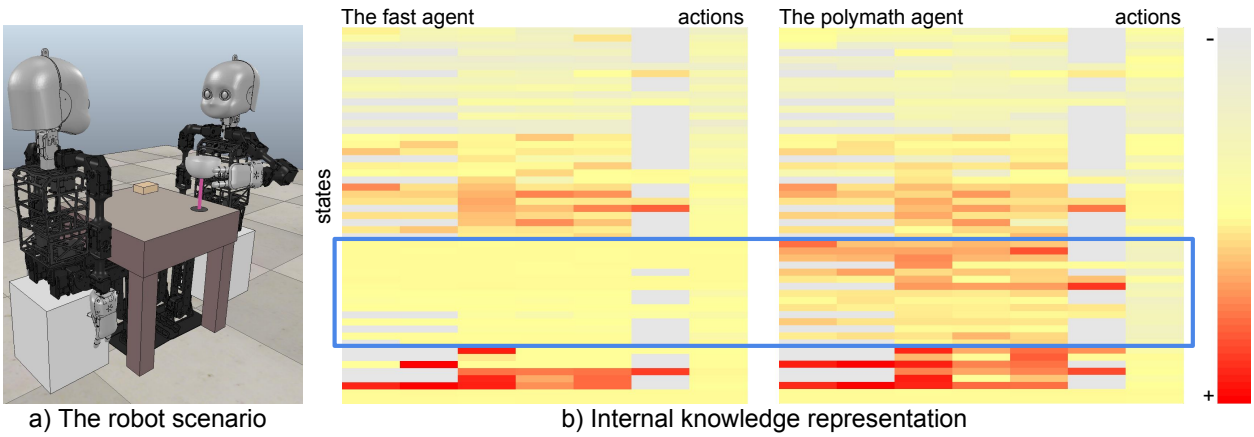

Figure 1: a) Simulated cleaning scenario. b) Internal knowledge representation for two possible parent-like advisors, namely the fast and the polymath agent. The fast agent shown in left, despite collecting more reward, does not have enough knowledge to advise a learner in every situation represented by the light blue box. Whereas the polymath agent, shown in right, has overall much more distributed knowledge which allows it to better advise a learner-agent.

2 Contribution and Discussion

In the presented scenario, there are agents with diverse behaviors which differ mostly in the path they choose until reaching a final state. There exist agents that regularly take the shortest path and others that take the longest one, we refer to them as the fast and the slow agents respectively. In both cases, agents successfully accomplish the task, although they accumulate different amounts of average reward. Obviously, the fast agents are the ones with better performance in terms of collected reward. Additionally, there is a third kind of agents with a more homogeneously distributed experience, meaning that they do not have a favorite sequence to follow and have equally explored both paths. We refer to such agents as polymath agents.

Accumulating plenty of reward does not necessarily lead an agent to becoming a good trainer, in fact it only means that the agent is able to select the shortest path most of the time from the initial state, but the experience collected in other states not involved in that route is absent or barely present and therefore such an agent cannot give good advice in those states where it does not know how to act optimally.

We recorded the internal representation of the knowledge through the Q-values to confirm the lack of learning in a subset of states. Fig. 1b shows a heat map of the internal Q-values of two agents, the fast and the polymath agent. Warmer regions represent larger reward and colder regions lower values. In fact, the coldest regions are associated with failed-states from where the agent should start a new episode, obtaining a negative reward of $r = -1$. It can be observed that the fast agent may be an inferior advisor since there exists a whole region uniformly yellow as highlighted by the blue box in Fig. 1b, which shows no knowledge about what action to prefer. It is important to note that the region on top is in both cases colder than the rest because it is the most distant from the final states where a positive reward $r = 1$ is given, but in spite of that, the polymath agent is still able to select a suitable action according to the learned policy.

In this regard, the learned policy by the fast agent is partially incomplete. On the contrary, the policy learned by the polymath agent is much more complete when observing the same region. Therefore, the polymath agent is a better candidate to become an advisor for other RL agents.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: Bradford Book, 1998.

[2] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, pp. 1–37, 2013.

[3] P. Kormushev, S. Calinon, and D. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, pp. 122–148, 2013.

[4] D. Hämmerer and B. Eppinger, "Dopaminergic and prefrontal contributions to reward-based learning and outcome monitoring during child development and aging," *Developmental Psychology*, vol. 48, pp. 862–874, 2012.

[5] G. O. Deak, A. M. Krasno, J. Triesch, J. Lewis, and L. Sepeta, "Watch the hands: Infants can learn to follow gaze by seeing adults manipulate objects," *Developmental Science*, vol. 17, pp. 270–281, 2014.

[6] A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA, USA: MIT Press, 2015.

[7] H. B. Suay and S. Chernova, "Effect of human guidance and state space size on interactive reinforcement learning," in *Proceedings of IEEE International Symposium on Robot and Human Interactive Communication RO-MAN*, pp. 1–6, 2011.

[8] L. Torrey and M. Taylor, "Teaching on a budget: Agents advising agents in reinforcement learning," in *Proceedings of International Conference on Autonomous Agents and Multi-agent Systems AAMAS*, pp. 1053–1060, 2013.

[9] M. E. Taylor, N. Carboni, A. Fachantidis, I. Vlahavas, and L. Torrey, "Reinforcement learning agents providing advice in complex video games," *Connection Science*, vol. 26, pp. 45–63, 2014.

[10] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 2625–2633, 2013.

[11] F. Cruz, S. Magg, C. Weber, and S. Wermter, "Training agents with interactive reinforcement learning and contextual affordances," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, pp. 271–284, 2016.