CMP9764M Comaring a Q-learning driven agent to an IRL based agent

UNIVERSITY OF LINCOLN

line 1: 2nd Stephen Rerri-Bekibele
line 2: *School of Computer Science. University of Lincoln (of Affiliation)*
line 3: Lincoln, United Kingdom
line 4: 16663359@students.lincoln.ac.uk

Table of Contents

***Abstract*—since the discovery of Artificial Intelligence there has been a range of applications of said AI for research purposes and commercial entertainment such as in video games. In advanced Robotics AI has been researched on in the subject of Human Robot Interactions which seeks to improve the relationship between humans and robots and make it more common place. HRI is a combination of social science, cognitive science, AI and Robotics. HRI involves robots learning from humans through interactions, specifically in terms of skills, expertise, and knowledge or via their general characteristics in terms of psychology, physical properties, etc. This research paper looks at two methods for Robots learning from humans via Interactive Reinforcement Learning. Standard Q-Learning and an Interactive Reinforcement Learning Variant and then compares and contrasts them.**

Keywords—Robotics, AI, Cognitive science, social science, human robot interactions, research, applications, Q-Learning (*key words****)***

INTRODUCTION

Research into Interactive Reinforcement Learning has shown that “teaching by demonstration and teaching by assigning rewards” [1] are two popular methods of

knowledge transfer in humans as well as in a reinforcement learning algorithm such as Q-learning. In the context of robot learning, the preference between assessing a learning agent's performance and assigning a reward or punishment vs assessing if a demonstration is good or not has not been studied extensively.

LITERATURE REVIEW

A research paper by titled "Human Feedback as Action Assignment in Interactive Reinforcement Learning" (Raza S.A et al) suggests replacing the traditional method of reward assignment with action assignment (which is similar to providing a demonstration) in interactive reinforcement learning, with computing a reward only if the suggested action is followed by a self-acting agent or not. This "action assignment" method was compared against a traditional reward assignment method via a human study with a two-dimensional maze game.

The results of this was that the action assignment method significantly improved the human's ability to teach the right behaviour.

However, the study showed that the reward assignment required more mental effort, as repeatedly assigning rewards and seeing the agent disobey commands annoyed the human teachers and many humans desired to control the agent's behaviour directly.

Another paper, "DQN-Tamer: Human-in-the-loop Reinforcement Learning with intractable Feedback" designed a system to train an agent with real-time feedback from a human observer who immediately gives rewards for some actions [2]. The method used was a Reinforcement Learning algorithm called DQN-Tamer, a combination of the DQN and Deep Tamer algorithms which now used both human feedback and distant rewards. The results showed that the DQN tamer was able to outperform the baseline in a simulated Maze environment. Interestingly enough the feedback in this case was the user's facial expressions while the agent explored the maze. The main problems with this approach of human feedback in a real application were: BINARY, DELAY, STOCHASTICITY, UNSUSTAINABILITY and NATURAL REACTION. Nonetheless the experiment showed that the proposed DQN-Tamer model was robust against inconvenient feedback and outperformed existing algorithms like DQN and Deep Tamer.

In both of the cases of "DQN-Tamer: Human-in-the-loop Reinforcement Learning with intractable Feedback" and "Human Feedback as Action Assignment in Interactive Reinforcement Learning" A key factor that affected the efficiency of the Interactive Reinforcement Learning Agent was the capabilities of the human teacher. In "Human feedback as action" the teachers felt that they did not have enough control over the agent's actions as they could only assign rewards and it was up to the agent to act on that reward or not, perhaps the algorithm was lacking in the instantiation of the importance of the reward to the agent and perhaps not.

Nonetheless the implementation of action assignment did increase the human's ability to give the right behaviour, the right demonstration as mentioned at the start of this research paper, yet this did not improve the agent's ability to listen, which is a different issue that is based on how the agent reacts to rewards.

In the second paper DQN-Tamer, the results showed that the experiments with a human teacher were successful however they also showed that they could be improved due to some hardware and software limitations such as the CNN classification for detecting the teachers actions as feedback and the information delay form the feedback to the agent but also the reaction of the teacher was also a problem because human beings have many actions and this would have made it hard for a teacher to always default to a set number of expression that the agent could recognise as good or bad otherwise the agent would need to have access to a broader understanding of human expressions, which are classified as good and bad and a scale value proportional to a reward value would need to have been devised.

A third paper "Exploration from Demonstration for Interactive Reinforcement Learning" Proposed a model-free policy-based approach akin to Q-learning covered in this Assignment. The difference being that the approach was called "Exploration from Demonstration" [3] EfD used an on-policy approach with human demonstrations to guide search space exploration, statistical measures of the RL algorithm were provided to a human teacher as feedback, alerting them of the agent's uncertainty. This feedback was then used to solicit targeted demonstrations useful from the agent's perspective.

This loop of feeding agent uncertainty at each state to a human teacher that then provides demonstrations for the agent at that state allow the agent to learn an exploration policy that actively guides the agent towards important aspects of a problem.

When the method was employed in a maze, the results showed that the EfD approach provides convergence speed-ups over traditional exploration and interactive learning methods such as Q-Learning. the benefits of using a policy-based approach for exploration in RL in the context of EfD showed that when faced with large domains with sparse rewards and long horizons, a

policy-based approach is less vulnerable to the large sample requirements of value-based methods as the information acquired from a single demonstration allows the agent to extends its range of exploration over multiple time steps.

Additionally such a method does not concern itself solely with reward information like the Q-Learning approach. The statistical measures used in EfD (leverage and discrepancy) focus on different aspects of the MDP which allow the algorithm to function well across a wider class of problems. This is in contrast to value-based methods which rely on large samples of reward information to estimate the uncertainty in the value function often made complicated in sparse reward and long horizon domains. Also EfD does not require optimal demonstrations to learn but instead demonstrations that serve to connect two regions of the agent's choice. As these demonstrations are used for exploration, they can be potentially noisy (which may in some cases help the agent).

CONCLUSION

In conclusion, the results of the aforementioned papers all show that the Implementation of Interactive Reinforcement Learning in context of grid solving can be accomplished in a variety of ways with each method focusing on the shortcoming of another.

EVALUATION

The Comparison of Q-Learning vs the IRL agent showed that the Q-learning is much faster than an IRL agent because of the need for human confirmation at each step as not only could the human also make mistakes at time since there is no policy being used aside from the human's choice of what they think is the right action, the human can also carelessly classify bad actions as good and good as bad. Moreover the IRL agent needs to suggest a good alternative to an action if the human says an action is bad otherwise the IRL agent ends up stuck in its current position until a good move is suggested, this means the agent takes a lot more time to move if it keeps suggesting wrong actions, there needs to be a way to let the agent know an action is bad due. This can be done by looking at the last actions suggested by the agent and if they are bad actions then the agent suggests a new action at random that is not one of the previous actions. This reduces how long it takes the agent to move but also doesn't take the freedom of choice from the agent because if it given exactly 3 chances to pick a right action then in a true deterministic world the agent would have picked a different potentially right action as there are only 4 actions.

When the Q-learning and IRL agent were tested for 40 iterations, the mean number of states the IRL agent took to get to the goal was 8.45, the maximum number was 46 and the minimum was 5. This showed that in most cases the IRL agent performed better. The maximum time taken to get to the goal only occurred on the first run of the experiment. This demonstrates that the agent leant the best actions to suggest as time went on.

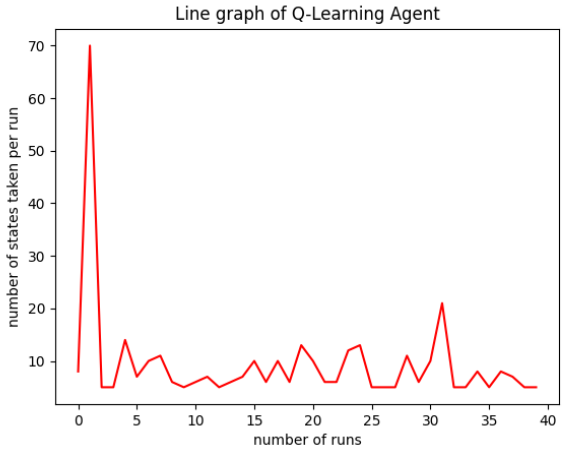

Figure 1. Plot of Q-Learning

The above figure shows that on the first run the agent took the most amount of steps to get to the goal but this reduced in the succeeding steps.

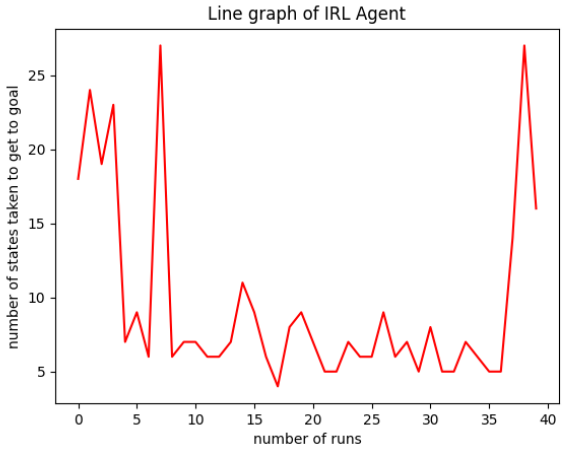

Figure 2. Plot of IRL Agent

The above graph showed that the performance of the Interactive Agent Fluctuated and there was no clear increase in performance as the experiment carried on therefore although the agent made some improvements halfway through the run, this progress was lost on the last runs.

Figure 3. Plot of Q-learning Agent

Above is a Histogram of the Performance of the Q-Learning Agent. The results show that out of the 40 runs the mean number of actions taken to get to the goal was 3 around 3.

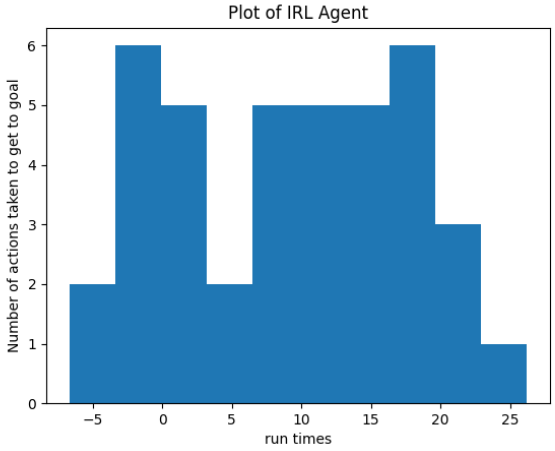

Figure 4. Plot of IRL Agent

The above plot for the IRL agent is a histogram demonstrating the distribution of actions taken in each run for 40 runs.

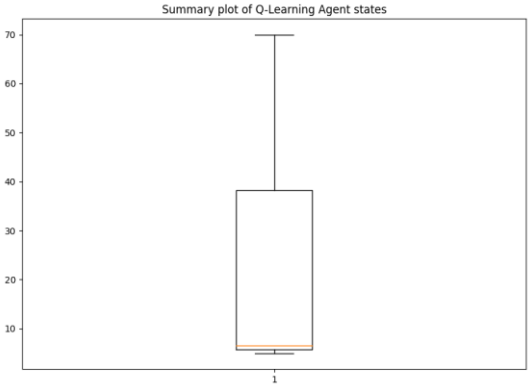

Figure 5. Boxplot of Q-Learning Agent states

The above is a boxplot of the Maximum, minimum and median value of the number of states that the Q-learning Agent took over the 40 runs. The Median value is quite close to the minimum value which means that the maximum value was an outlier and resulting in the large contrast.

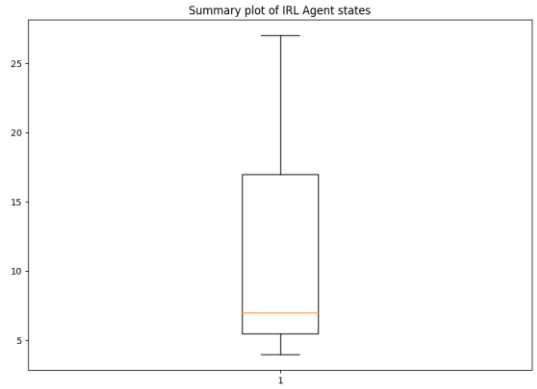

Figure 6. Boxplot of IRL Agent states

The above graph showed that the Median value was around the same as the median value of the Q-Learning Agent however the maximum number of states the IRL agent took across the 40 runs was significantly smaller than the maximum number of steps the Q-Learning Agent took at under 30 states. For the Q-Learning Agent however, this value was around 70. The upper quartile value of the IRL Agent was also lower between the ranges of 15 and 20 while the upper quartile of the Q-learning Agent was around 40 actions.

ACKNOWLEDGMENT

I would like to acknowledge The Lecturers for their explanation of HRI principles and how they could be applied.

REFERENCES

[1] Raza, S.A. and Williams, M.A., 2020. Human feedback as action assignment in interactive reinforcement learning. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, *14*(4), pp.1-24.

[2] Arakawa, R., Kobayashi, S., Unno, Y., Tsuboi, Y. and Maeda, S.I., 2018. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*.

[3] Subramanian, K., Isbell Jr, C.L. and Thomaz, A.L., 2016, May. Exploration from demonstration for interactive reinforcement learning. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* (pp. 447-456).

[4]