Guiding Reinforcement Learning Exploration Using Natural Language

Extended Abstract

Brent Harrison
University of Kentucky
Lexington, KY
harrison@cs.uky.edu

Upol Ehsan
Georgia Institute of Technology
Atlanta, GA
ehsanu@gatech.edu

Mark O. Riedl
Georgia Institute of Technology
Atlanta, GA
riedl@cc.gatech.edu

ABSTRACT

In this work we present a technique for using natural language to help reinforcement learning generalize to unseen environments using neural machine translation techniques. These techniques are then integrated into policy shaping to make it more effective at learning in unseen environments. We evaluate this technique using the popular arcade game, Frogger, and show that our modified policy shaping algorithm improves over a Q-learning agent as well as a baseline version of policy shaping.

KEYWORDS

Interactive Machine Learning; Reinforcement Learning; Policy Shaping

ACM Reference Format:
Brent Harrison, Upol Ehsan, and Mark O. Riedl. 2018. Guiding Reinforcement Learning Exploration Using Natural Language. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

Interactive machine learning (IML) [3] algorithms allow human teachers to help algorithms learn faster by providing targeted feedback [2, 4, 5] or behavior demonstrations [1].One of the primary limitations of IML algorithms is that human feedback is tightly coupled with state information, meaning that it can be difficult to generalize feedback to unseen states without retraining. In this work, we seek to use natural language to enable IML algorithms to better generalize to unseen environments.

Humans are extremely proficient at generalizing over many states, and often language aids in this endeavor. In this work, we aim to use neural machine translation techniques—specifically encoder-decoder networks—to learn generalized associations between natural language behavior descriptions and state/action information. We then use this model, which can be thought of as a model of generalized action advice, to augment a state of the art interactive machine learning algorithm, policy shaping [4] to make it more effective in unseen environments.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

2 POLICY SHAPING

In this paper, we build upon the policy shaping framework [2, 4], which is a technique that incorporates human critique into reinforcement learning. Unlike other techniques such as reward shaping, policy shaping considers critique to be a signal that evaluates whether the *action* taken in a state was desirable rather than whether the resulting *state* was desirable. Policy shaping utilizes human feedback by maintaining a *critique policy* to calculate the probability, $Pr_c(a)$, that an action $a \in A$ should be taken in a given state according to the human feedback signal. During learning, the probability that an agent takes an action is calculated by combining both $Pr_c(a)$ and $Pr_q(a)$:

$$Pr(a) = \frac{Pr_q(a)Pr_c(a)}{\sum_{a' \in A} Pr_q(a')Pr_c(a')} \quad (1)$$

The critique policy used in policy shaping is generated by examining how consistent the feedback for certain actions are. If an action receives primarily positive or negative critique, then the critique policy will reflect this with a greater or lower probability, respectively, to explore that action during learning.

3 USING LANGUAGE TO GENERALIZE HUMAN CRITIQUE

To make IML more generalizable, we show how an encoder-decoder network [6] can be used to learn a *language-based critique policy*. Our technique works by first having humans generate a set of annotated states and actions by interacting with a single learning environment offline and providing explanations of their actions. These annotations are then used to train an encoder-decoder network to create the language-based critique model that can be queried while the agent explores new environments to receive general action advice. Each of these steps will be discussed in greater detail below.

3.1 Creating the Language-Based Critique Policy

Typically, training an agent using critique requires a large amount of consistent online feedback to create the *critique policy*, which provides little opportunity to generalize. To address this, our technique uses natural language as a means to generalize feedback across many, possibly unknown, states by using an encoder-decoder model to construct a general *language-based critique policy*. This model learns to reconstruct symbolic state-action information based on natural language descriptions describing the action taken and the

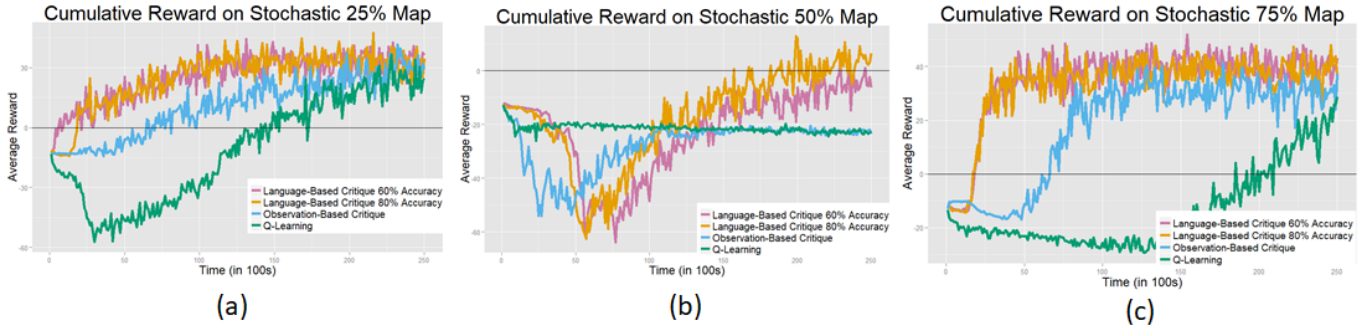

Figure 1: Learning rates for agents on stochastic versions of the 25% map (a), 50% map (b), and 75% map(c).

reason that it was taken. This enables the agent to receive action advice for any potential state it finds itself in using the language-based critique policy.

3.2 Utilizing the Language-Based Critique Policy

To help agents generalize human feedback to unseen environments, we will use the language-based critique policy learned earlier to take the place of the standard critique policy normally used by policy shaping.

To determine what piece of previously observed advice the agent should follow, we calculate probability of reconstructing the agent's current state given each piece of advice in our training set. We use the utterance that best describes the agent's current state to create the action distribution as follows:

Specifically, this is calculated as:

$$Pr_{lc}(a) = \frac{e^{Pr_l(s,a,i)/\tau}}{\sum_{a'} e^{Pr_l(s,a',i)/\tau}} \quad (2)$$

where $Pr_l(s, a, i)$ is the log probability of performing action a in state s according to the language-based critique policy using sequence i as input. We also make use the τ parameter in Equation 2 to control how much weight we place on the knowledge extracted from the language-based critique policy.

Having done this, the RL agent now explores its environment as it normally would using policy shaping; however, the probability of the agent performing an action in a given state is defined as:

$$Pr(a) = \frac{Pr_q(a)Pr_{lc}(a)}{\sum_{a' \in A} Pr_q(a')Pr_{lc}(a')} \quad (3)$$

where $Pr_c(a)$ is replaced with the probability obtained from the language-based critique policy.

4 EVALUATION

To evaluate our technique, we examine how it can be used to speed up learning in the popular arcade game, Frogger. We compare our system against the following baseline agents: a standard Q-learning algorithm with no additional information and a policy shaping algorithm that only has access to behavior observations with no additional language data.

To provide some measure of control over the data used for training, we used a semi-synthetic grammar to generate the humanlike explanations needed for training. There is also either a 60% chance or an 80% chance that the teacher will provide incorrect feedback to better simulate unreliable human teachers.

For this evaluation we examine how each agent performs on three different frogger maps with a 25%, 50%, and 75% chance of an obstacle occupying a given space. In this environment, there is also a 20% chance that the agent takes a random action instead of its intended action.

In all test cases, the learned policy was evaluated every 100 episodes and then the total cumulative reward earned during each episode was averaged over 100 total runs.

4.1 Results

As can be seen from Figure 1, both language-based critique agents converge much faster than the Q-learning agent and the observation-based critique agent under all conditions. For this set of experiments, both language-based critique agents outperform the other two agents on each map used for testing. This supports our claim that the addition of language-based critique helps agents generalize advice to unseen environments. It is also interesting to note that on the 50% map, the language-based critique agents are the only ones to converge after the 25,000 training episodes.

We performed another experiment using a deterministic version of Frogger to test how this would affect the learning rate of agents trained using our technique. While the general outcomes were similar, we found that the performance difference between the language-based critique agents and all other agents was more pronounced than in the stochastic environments.

5 CONCLUSIONS

Language is a powerful tool that humans use to generalize knowledge across a large number of states. In this work, we explore how language can be used to augment machine intelligence and give intelligent agents an expanded ability to generalize knowledge to unknown environments. Specifically, we show how neural machine translation techniques can be used to give action advice to reinforcement learning agents that generalizes across many different states, even if they have not been seen before. As our experiments have shown, this generalized model of advice enables reinforcement learning agents to quickly learn in unseen environments.

REFERENCES

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems*

57, 5 (2009), 469–483.
[2] Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. 2015. Policy Shaping with Human Teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
[3] Sonia Chernova and Andrea L Thomaz. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8, 3 (2014), 1–121.
[4] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems*. 2625–2633.
[5] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2014. Learning something from nothing: Leveraging implicit human feedback strategies. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 607–612.
[6] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1412–1421.