Joint Inference of Reward Machines and Policies for Reinforcement Learning

Zhe Xu¹, Ivan Gavran², Yousef Ahmad¹, Rupak Majumdar², Daniel Neider², Ufuk Topcu¹, Bo Wu¹

¹University of Texas at Austin, Austin, TX  
²Max Planck Institute for Software Systems, Kaiserslautern, Germany

Abstract
We investigate an RL problem where the high-level knowledge is in the form of reward machines, i.e., a type of Mealy machine that encodes non-Markovian reward functions.

We present an iterative algorithm that performs joint inference of reward machines and policies (JIRP) for RL (more specifically, q-learning).

We implement the proposed JIRP approach and compare it with three baseline methods. We evaluated them in three scenarios: an autonomous vehicle scenario, an office world scenario, and a Minecraft world scenario, running a set of tasks for every scenario.

The experiments show that the proposed JIRP approach significantly outperforms the three baseline algorithms in all three scenarios.

Methodology

The Algorithm (JIRP)
initialize $H, Q; X$; repeat
for all $(i, \rho, \beta) \in QRM(H, Q)$ do
    if $H(i) \neq \rho$ then
        add $(i, \rho)$ to $X$;
        $H = \text{infer}(X)$;
        reinitialize $Q$;
end for
end repeat

Optimizations
1. Batching of counterexamples
2. Transfer of q-functions

Results

Challenges in RL
- Complex task
- Sparse reward
- Large search space over history of states
- Exploration is inefficient

Related Work
- Carte et al., NeurIPS 2019
- Gaon & Brafman, AAAI 2020
- Furelos-Blanco et al., AAAI 2020

Acknowledgements
This work is supported in part by grants DFG 389792660-TRR 248, DFG 434592664, ERC 610150, DARPA D19A1P00004, ONR N000141712623, and NASA 80NSSC19K0209.

Conclusion
We proposed an iterative approach that alternates between reward machine inference and RL for the inferred reward machine.

Future work:
First, we are interested in exploring a scenario in which the reward machine is not known, but a number of hints (properties) about the reward are given.

Second, we will explore methods that can infer the reward machines incrementally.

Finally, the method to transfer the q-functions between equivalent states of reward machines can be also used for transfer learning between different tasks where the reward functions are encoded by reward machines.