Learning Symbolic Representations for Reinforcement Learning of Non-Markovian Behavior

Phillip Christiansen  Andrew C. Li  Rodrigo Toro Icarte  Sheila A. McIlraith
Department of Computer Science, University of Toronto, Toronto, Canada
Vector Institute for Artificial Intelligence, Toronto, Canada
1Schwartz Reisman Institute for Technology and Society, Toronto, Canada

Overview

Many real-world RL tasks require temporally extended behavior. Such reward-worthy behavior requires the execution of a pattern of actions that yields reward only upon completion of the pattern.

Standard deep RL solution:
- Use a recurrent neural network (RNN) — hidden state summarizes state-action history.
- Disadvantage: takes large number of samples to train.

Standard KR-based solution:
- Use a pre-determined abstraction of the state space to define reward-relevant features, realized via a labelling function.
- Disadvantage: requires a priori domain knowledge.

Our solution:
- Automatically learn reward-relevant features from the state-action history!
- Use the learned features to accelerate learning.

In our experiments with non-Markovian goals, we outperform state-of-the-art RL based on RNNs.

Example

A DFA learned by AutRL for the example.

Algorithm 1: AutRL

```
dfa ← empty_automata;
π ← uniform_random_policy;
traces ← Ø;
while true do
  if sample_traces ← sample(π, N) then
    append traces with sample_traces;
    if sample_traces inconsistent with dfa then
      dfa ← autLearn(traces);
  end
  π ← markov_learn(sample_traces × dfa);
end
```

Main idea:
- Train a DFA to predict whether a state-action sequence receives reward 0 or 1.
- Augment the agent with the DFA state and use any standard RL algorithm to learn policies which are non-Markovian in the original problem.
- If a learned DFA can predict this reward perfectly, then the problem becomes completely Markovian.
- Alternate between DFA learning and Markovian learning, until a consistency condition is met for the learned DFA.

Details:
- Recent advances in automata learning (Shve et al., 2020) allow us to efficiently learn small DFAs that are robust to noise.
- In practice, the DFA need not perfectly classify the reward. For example, it is enough to be able to partition the DFA states into ones where have_keys is true, and where have_keys is false.
- We prove AutRL optimally converges under mild assumptions on the Markovian learning policy, exploratory policy, and on the structure of the goal-based reward.

Analysis of Results:
- AutRL is, at most, over an order of magnitude more sample-efficient than Recurrent-PPO, to 95% confidence.
- AutRL exhibited much more consistent and stable learning than the Recurrent-PPO.
- In practice, checking for high performance of Markov learning given a DFA, instead of perfect reward classification, yields better stability and sample efficiency.

Selected Related Work:
- Prior works also attempt to learn reward structure, but are sensitive to noise (e.g. Toro Icarte et al., 2019; Xu et al., 2020).
- (Goon & Braffman, 2020) shares our high-level idea but we leverage advances in automata learning and show this can outperform standard state-of-the-art RL.

Discussion

Limitations and Future Work:
- This method does not perform well if the goal histories 𝑓 do not form a regular language in 𝑆 × 𝑋 (e.g. counting-related tasks).
- We focus on non-Markovian reward, future work can consider partial observability.
- DFA learning method requires the state-action space to be small and discrete: future work should focus on continuous states and scalability.
- We only support non-Markovian goals rather than general reward functions.

Bibliography:
- [Link to reference]
- [Link to reference]
- [Link to reference]
- [Link to reference]