Symbolic Plans as High-Level Instructions for Reinforcement Learning

León Illanes  Xi Yan  Rodrigo Toro Icarte  Sheila A. McIlraith
Department of Computer Science, University of Toronto  Vector Institute

Experiments

We ran experiments in two different environments. In each, we train agents to solve four tasks of order of complexity. The figures below show the rewards obtained by standard q-learning (qi), hierarchical RL using the options identified in the high-level models (hrl), and our approaches using sequential (seq) and partial-order (pop) plans. Our approaches find good solutions much faster than the alternatives, and effectively transfer learned information from one task to the next.

Motivation: Tell an RL agent to solve a specific task.

$$
\begin{array}{|c|c|c|c|}
\hline
B & * & * & C \\
\hline
* & * & * & * \\
\hline
A & * & * & D \\
\hline
\end{array}
$$

Symbol Meaning
\begin{itemize}
\item \(\Delta\) Robot
\item \(\text{\textbullet}\) Furniture
\item \(\square\) Coffee machine
\item \(\bigtriangleup\) Mail room
\item \(\bigtriangledown\) Office
\end{itemize}

A, B, C, D Marked locations

Task examples
- T1. Deliver mail to the office
- T2. Deliver coffee and mail to the office
- T3. Visit A, B, C, and D (in any order)

Q: How do we specify a new task?
A: An expert designs a reward function.

We specify a goal for an abstract, high-level model.

Propositions: have-mail/coffee, delivered-mail/coffee, visited-A/B/C/D

Actions: get-mail/coffee, deliver-mail/coffee, visit-A/B/C/D

We have additional experiments that evaluate an execution monitoring approach that may help when models are incorrect.

We can use plans to guide the RL algorithms!

- T1. \(\langle\text{get-coffee},\text{deliver-coffee}\rangle\)
- T2. \(\langle\text{get-coffee},\text{get-mail},\text{deliver-coffee},\text{deliver-mail}\rangle\)
- T3. \(\langle\text{go-to-A},\text{go-to-B},\text{go-to-C},\text{go-to-D}\rangle\)

- This is hierarchical RL with a fixed high-level policy
- Every action maps to an option
- Many actions can map to a single option
  - e.g., deliver-mail and deliver-coffee

Blindly following the plans is suboptimal...

Relax the ordering constraints!

- Use partial-order plans instead of sequential ones
- The high-level policy is no longer fixed, but very restricted
- The agent also has to learn how to order the plans

\begin{itemize}
\item G1. \{ delivered-mail \}
\item G2. \{ delivered-mail, delivered-coffee \}
\item G3. \{ visited-A, visited-B, visited-C, visited-D \}
\end{itemize}