RePReL: Integrating Relational Planning and Reinforcement Learning for Effective Abstraction

Harsha Kokel, Arjun Manoharan, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli

UT Dallas · IIT Madras · Oregon State University

Aim Learn to act in relational domains, with varying task, number of objects and relations.

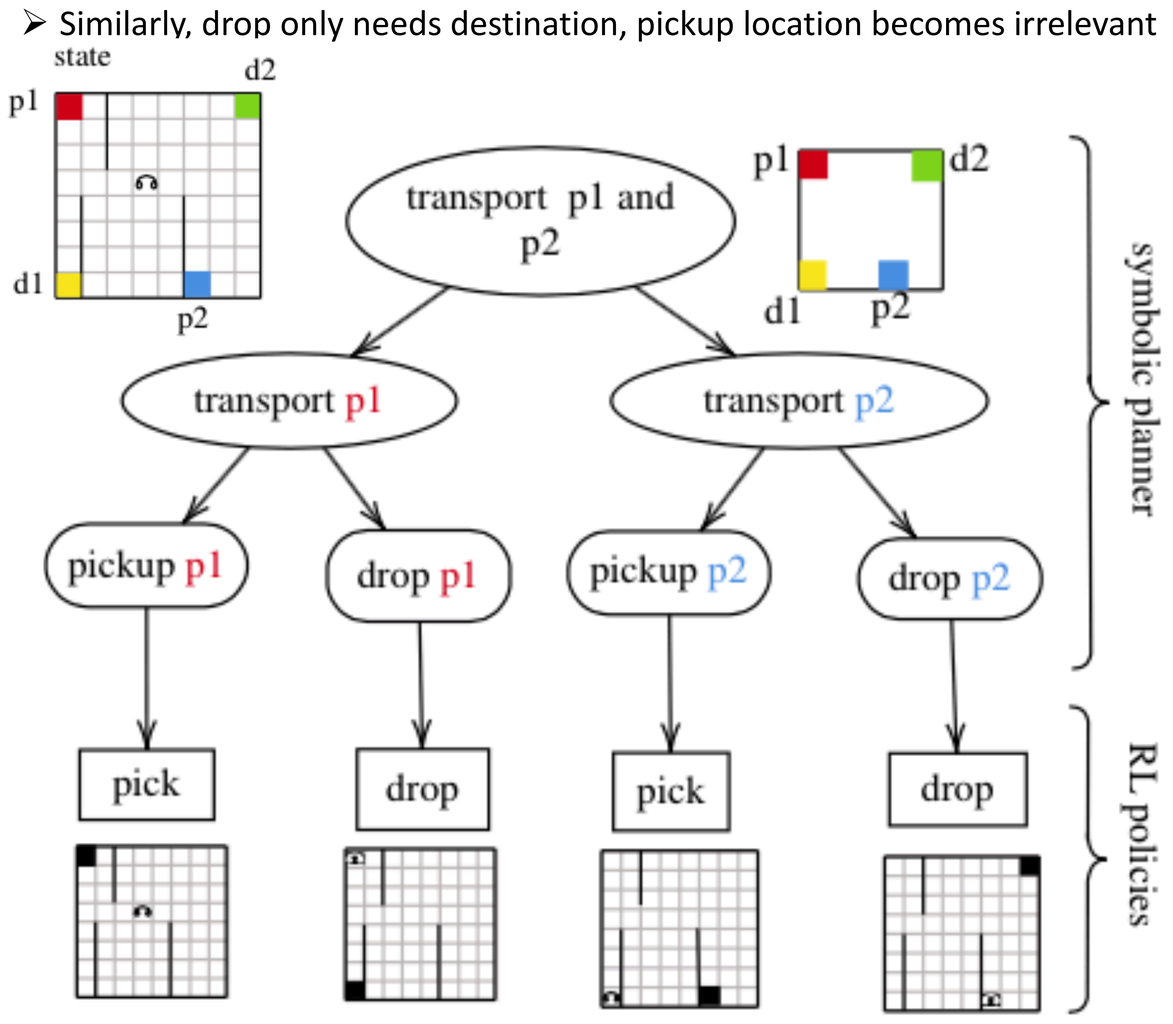
Motivation Humans are quite natural at creating an abstract representation for efficient planning. They use different abstractions for different tasks. Can we leverage their domain knowledge and provide such abstractions to the RL agent?

RePReL

- Hierarchical Framework with Planner and RL
- Plan sequence of operators at High level and learn to execute each operator at lower level
- Advantages
  - Planner provides compositionality
  - Different RL agent for execution allows separate state representations
- Abstraction is most useful when we have multiple relations and objects, so consider Relational MDP.
- Use D-FOCI statements to identify relevant objects and relations for a given task.

Toy Example

- Taxi domain with multiple passengers
- Acquire the sequence of passenger pickup and drop operators from planner
- Learn to execute the operators using RL
- Use different goal-conditioned RL for each operator
- Pickup operator only needs pickup location, drop location is irrelevant
- Similarly, drop only needs destination, pickup location becomes irrelevant



Model-agnostic Abstraction

Definition 4 (Li, Walsh, and Littman (2006)). *A model-agnostic abstraction $\phi(s)$ is such that for any action a and abstract state $\overline{s}$, $\phi(s_1) = \phi(s_2)$ if and only if*

$$\sum_{\{s_1' \mid \phi(s_1')=\overline{s}\}} R_o(s_1, a, s_1') = \sum_{\{s_2' \mid \phi(s_2')=\overline{s}\}} R_o(s_2, a, s_2')$$

$$\sum_{\{s_1' \mid \phi(s_1')=\overline{s}\}} P_o(s_1, a, s_1') = \sum_{\{s_2' \mid \phi(s_2')=\overline{s}\}} P_o(s_2, a, s_2')$$

- This is the bisimulation condition, where the abstraction function preserves the transition and the reward distribution of ground MDP in the abstract MDP

D-FOCI Statements

- First Order Conditional Influence (FOCI) statement is of the form
  "if *condition* then *X1* influence *X2*"
- FOCI statements encode the information that literal X2 is influenced only by the literals in X1 when the stated condition is satisfied
- We present Dynamic FOCI (D-FOCI) statements,

$$\text{operator} : \{\mathrm{p}(X_1), q(X_1)\} \xrightarrow{+1} \mathrm{q}(X_1)$$

- With +1, the D-FOCI encode the influence of literals in current time step on the literal in next time-step
- With operator, D-FOCI statement constraints the task being executed
- Given the current state and operator being executed we can obtain complete set of relevant state variables by grounding and rolling out the D-FOCI statements. This forms our abstract state
- Guarantees safe model-agnostic abstraction, if MDP satisfies the D-FOCI statement with fixed-depth unrolling

state : $\{\mathrm{at(p1, r)}, \text{taxi-at}(13), \mathrm{dest(p1, y)}, \neg\,\text{at-dest}(\mathrm{p1}), \neg\,\text{in-taxi}(\mathrm{p1}), \mathrm{at(p2, b)}, \mathrm{dest(p2, g)}, \neg\,\text{at-dest}(\mathrm{p2}), \neg\,\text{in-taxi}(\mathrm{p1})\}$

[at-p1, dest-p1, in-taxi-p1, at-dest-p2, at-p2, dest-p2, in-taxi-p2, at-dest-p2, taxi-at, move]

operator $\langle\ \text{pickup}(P), \{P/p1, L/r\}\ \rangle$

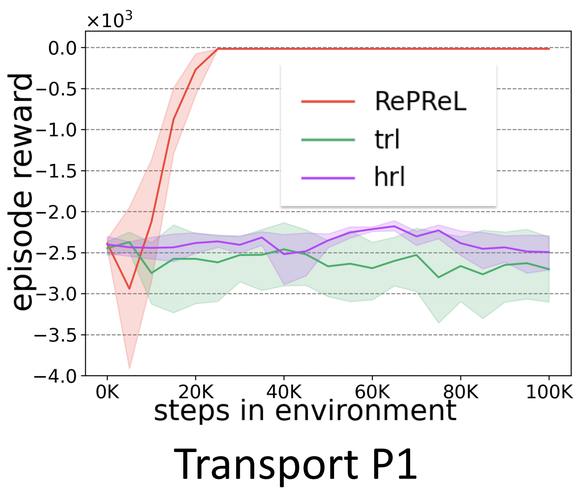
D-FOCI :
$\{\text{taxi-at(L1)}, \text{move(Dir)}\} \xrightarrow{+1} \text{taxi-at(L2)}$
$\{\text{taxi-at(L1)}, \text{move(Dir)}\} \longrightarrow R$
pickup(P):
$\{\text{taxi-at(L1)}, \text{at(P, L)}, \text{in-taxi(P)}\} \xrightarrow{+1} \text{in-taxi(P)}$
pickup(P): in-taxi(P) $\longrightarrow R_o$

abstract state: $\{\mathrm{at(p1, r)}, \text{taxi-at}(13), \neg\,\text{in-taxi}(\mathrm{p1}), \text{move(Dir)}\}$

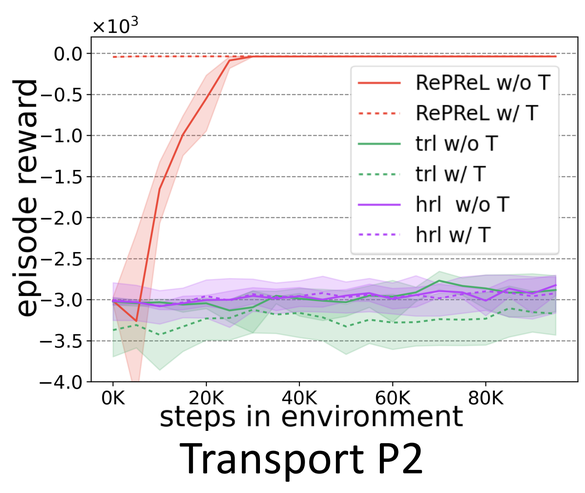
[at, taxi-at, in-taxi, move]

Experiments

Taxi domain

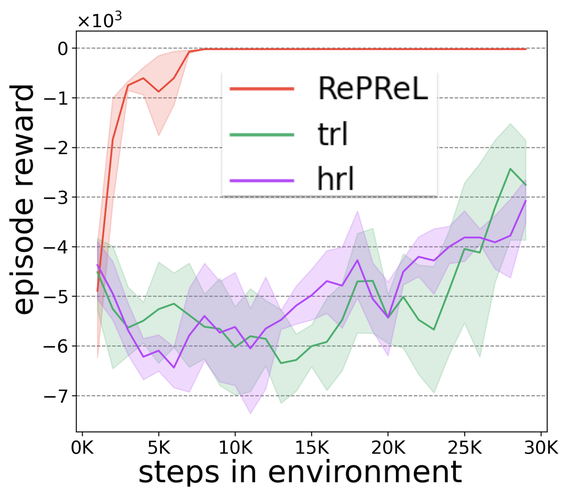


Transport P1

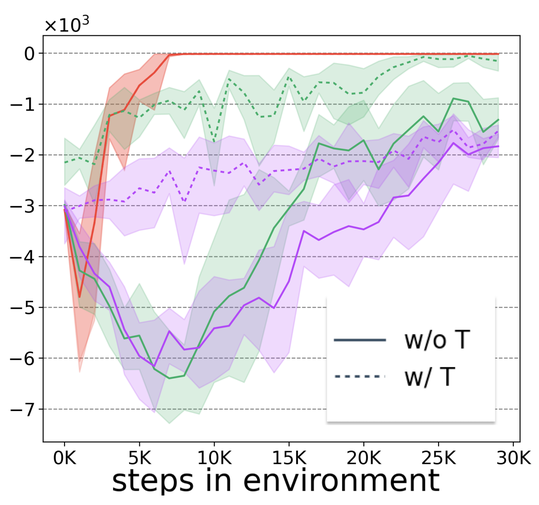


Transport P2

Office domain



Deliver mail



Deliver coffee

Human knowledge can help provide effective abstractions
that enable efficient learning and effective generalization across tasks and objects.

HK & SN gratefully acknowledge the support of ARO award W911NF2010224. SN acknowledges AFOSR award FA9550-18-1-0462. PT acknowledges support of DARPA contract N66001-17-2-4030 and NSF grant IIS-1619433. AM gratefully acknowledge the travel grant from RBCDSAI. Any opinions, findings, conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the ARO, AFOSR, NSF, DARPA or the US government.

More details on https://starling.utdallas.edu/papers/RePReL

STARLinG LAB