Survey on Hierarchical and Modular Reinforcement Learning

Shun Zhang

September 26, 2014

Markov Decision Process

MDP:

- State: S.
- Action: A.
- Transition: $P : S \times A \times S \to \mathcal{R}$.
- Reward: $R : S \times A \times S \to \mathcal{R}$.

Abstraction on MDP

- Aggregate states: feature extraction.

Abstraction on MDP

- Aggregate states: feature extraction.
- Aggregate actions: **option**.

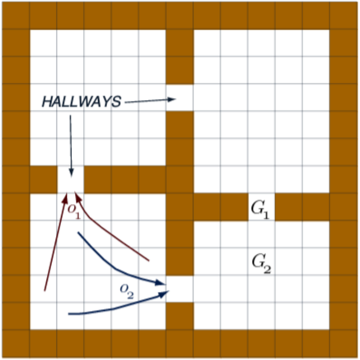
Abstraction on MDP

- Aggregate states: feature extraction.
- Aggregate actions: **option**.
- Factorize transition: factored MDP.

Abstraction on MDP

- Aggregate states: feature extraction.
- Aggregate actions: **option**.
- Factorize transition: factored MDP.
- Abstract MDP: **HAM, hierarchical RL, modular RL**.

MDP with Option



4 stochastic primitive actions

up

left *right*

down

Fail 33% of the time

8 multi-step options

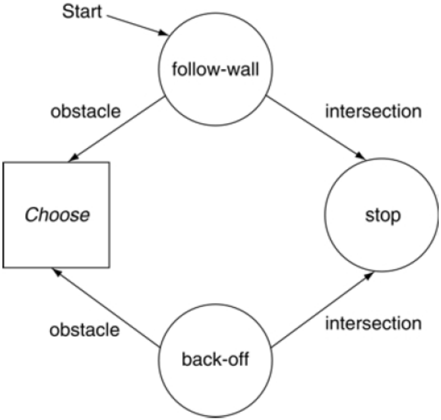
(to each room's 2 hallways)

MDP with Option

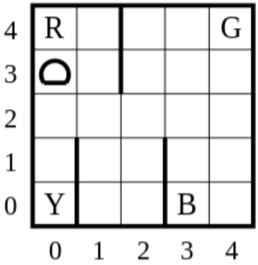
MDP:

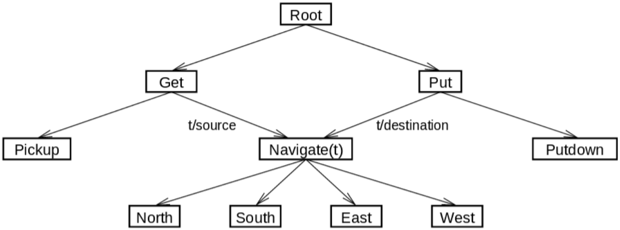
- State: S.
- Action: A, O.
- Transition: $P : S \times \{A, O\} \times S \to \mathcal{R}$.
- Reward: $R : S \times \{A, O\} \times S \to \mathcal{R}$.

Hierarchies of Abstract Machines (HAM)



Hierarchical RL





Hierarchical RL

MDP:

- State: $\mathcal{S}$.
- Action: $\mathcal{A}$.
- Transition: $\mathcal{T}$.
- Reward: $\mathcal{R}$.

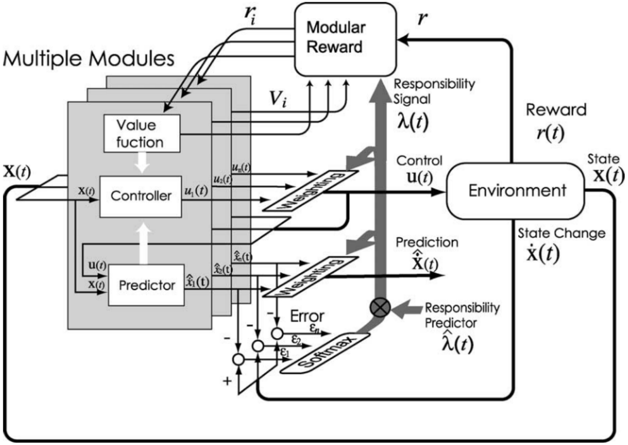

Multiple Modules
Modular Reward
r_i
r
V_i
Value fuction
Controller
Predictor
Responsibility Signal $\lambda(t)$
Reward $r(t)$
Control $\mathbf{u}(t)$
Environment
State $\mathbf{x}(t)$
State Change $\dot{\mathbf{x}}(t)$
Prediction $\hat{\mathbf{x}}(t)$
Weighting
Weighting
Error
Softmax
Responsibility Predictor $\hat{\lambda}(t)$
$\mathbf{x}(t)$

Fig. 1. MMRL.

Modular RL

MDP:

- State: $S_1 \times S_2 \cdots \times S_M$.
- Action: A.
- Transition: $P_1 \times P_2 \cdots \times P_M$.
- Reward: $R_1 \times R_2 \cdots \times R_M$.

Topics for Future Work

- Credit assignment.
- Learning task hierarchies.
- Dynamic abstraction.
- Integrating Deep Learning.