Learning Representations in Model-Free Hierarchical Reinforcement Learning

Jacob Rafati & David C. Noelle

Electrical Engineering and Computer Science, University of California, Merced

jrafatiheravi@ucmerced.edu — http://rafati.net/

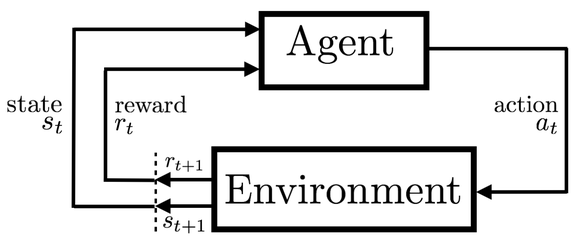
Abstract

Common approaches to Reinforcement Learning (RL) are seriously challenged by large-scale applications involving huge state spaces and sparse delayed reward feedback. Hierarchical Reinforcement Learning (HRL) methods attempt to address this scalability issue by learning action selection policies at multiple levels of *temporal abstraction*. Abstraction can be had by identifying a relatively small set of states that are likely to be useful as subgoals, in concert with the learning of corresponding skill policies to achieve those subgoals. Many approaches to *subgoal discovery* in HRL depend on the analysis of a model of the environment, but the need to learn such a model introduces its own problems of scale. Once subgoals are identified, skills may be learned through *intrinsic motivation*, introducing an internal reward signal marking subgoal attainment. In this paper, we present a novel model-free method for subgoal discovery using incremental unsupervised learning over a small memory of the most recent experiences (trajectories) of the agent. When combined with an intrinsic motivation learning mechanism, this method learns both subgoals and skills, based on experiences in the environment. Thus, we offer an original approach to HRL that does not require the acquisition of a model of the environment, suitable for large-scale applications.

Reinforcement Learning Problem

- $\mathcal{S}$: States Space, $\mathcal{A}$: Available Actions
- $e = (s, a, r, s')$ Agent's trajectory or transition experience.
- $\mathcal{D}$: Recent transition experience memory.

Objective: Find an optimal policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ to maximize the return, $G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'+1}$.



Model-Free: No knowledge of state transition probabilities, or the reward function, and no attempt to learn them.

When state space is huge, we use parameterized value function $Q(s, a; w) = \mathbb{E}[G_t | S_t = s, A_t = a]$.

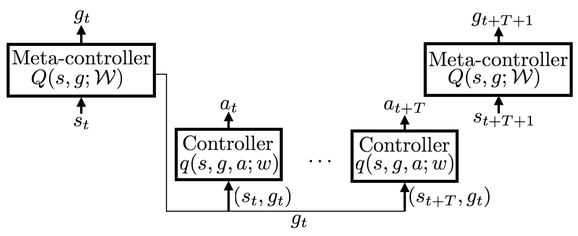
Q-Learning finds optimal values by minimizing:

$$L(w) \triangleq \mathbb{E}_{e \sim \mathcal{D}}\Big[\big(r + \gamma \max_{a'} Q(s', a'; w) - Q(s, a; w)\big)^2\Big].$$

Hierarchical Reinforcement Learning

The central goal is learning of representations at multiple levels of temporal abstraction.

Meta-controller/controller Framework



Intrinsic Motivation

Intrinsic motivation learning is the core idea behind the learning of value functions in the meta-controller and the controller. In some tasks with sparse delayed feedback, a standard RL agent cannot effectively explore the state space so as to have a sufficient number of rewarding experiences to learn how to maximize rewards. In contrast, the intrinsic critic in our HRL framework can send much more regular feedback to the controller, since it is based on attaining subgoals, rather than ultimate goals.

Unsupervised Subgoal Discovery

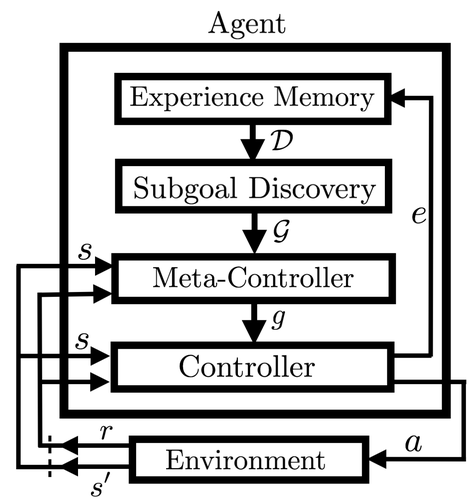
The performance of the meta-controller/controller frame- work depends critically on selecting good candidate sub- goals for the meta-controller to consider. What is a subgoal? In our framework, a subgoal is a state, or a set of related states, that satisfies at least one of these conditions:

1. It is close (in terms of actions) to a rewarding state.
2. It represents a set of states, at least some of which tend to be along a state transition path to a rewarding state.

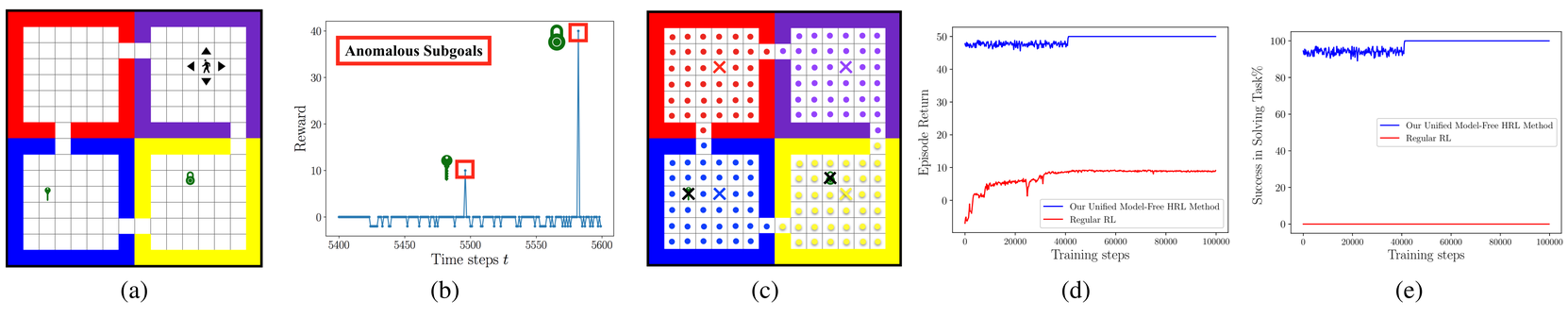
We hypothesize that good subgoals might be found by (1) attending to the states associated with anomalous transition experiences and (2) clustering experiences based on a similarity measure and collecting the set of associated states into a potential subgoal. Thus, our proposed method merges anomaly (outlier) detection with the K-means clustering of experiences.

Model-Free HRL Framework

These conceptual components can be unified into a single model-free HRL framework. At time t, the meta-controller observes the state, $s = s_t$, from the environment and chooses a subgoal, $g \in \mathcal{G}$ from a policy derived from $Q(s, \mathcal{G}; W)$. The controller receives an input tuple, (s, g), and is expected to learn to implement a subpolicy, $\pi(a|s, g)$, that solves the *subtask* of reaching from s to g. The controller selects an action, a, based on its policy, in our case directly derived from its value function, $q(s, g, a; w)$. After one step, the environment updates the state to s' and sends a reward r. The subgoal discovery mechanism exploits the underlying structure in the experience memory sets using unsupervised anomaly detection and experience clustering.

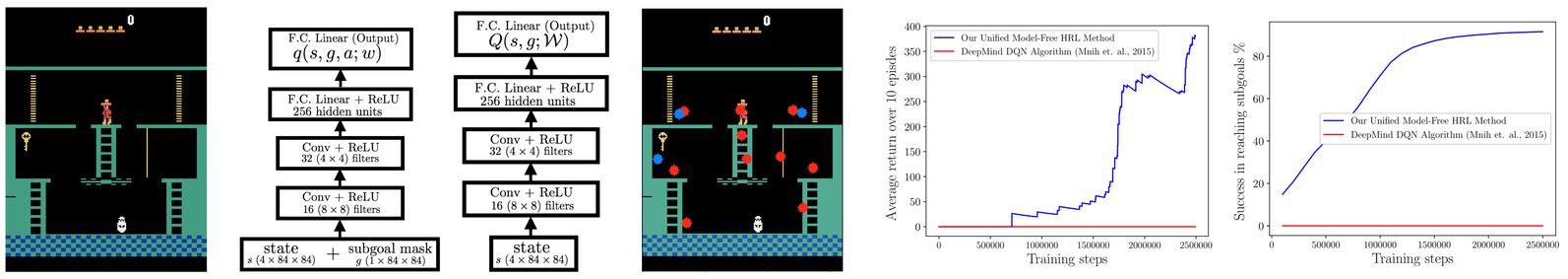


Numerical Experiments and Results



(a) (b) (c) (d) (e)

Figure 1: (a) The *rooms* task with a key and a lock. (b) Reward over an episode, with anomalous points corresponding to the key ($r = +10$) and the box ($r = +40$). (c) The results of the unsupervised subgoal discovery algorithm with *anomalies* marked with black Xs and *centroids* with colored ones. (d) The average episode return. (e) The success rate.



(a) (b) (c) (d) (e) (f)

Figure 2: (a) First room of ATARI 2600 *Montezuma's Revenge* game. (b),(c) Controller's and meta-controller's value function approximations. (d) The blue circles are the discovered anomalous subgoals and the red ones are the centroid subgoals. (e) The average episode return. (f) Intrinsic motivation success rate.

References

- Kulkarni et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. NeurIPS 2016.
- Rafati, J., and Noelle, D. C. 2018. Learning representations in model-free hierarchical reinforcement learning. (arXiv:1810.10096).
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; et al. 2015. Humanlevel control through deep reinforcement learning. Nature 518(7540):529533.