Near-Optimsl Representation learning for hierarchical reinforcement learning

Wei-Lun Kuo
Department of Computer Science
National Chiao Tung University
allenlike20.iie08g@nctu.edu.tw

1 Introduction

Please provide a clear overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper
  High-level controller solves tasks iteratively communicating goals which a lower-level policy is trained to reach. The mapping of observation space to goal space is crucial. They developed a notion of sub-optimality of a representation defined in terms of expected reward of the optimal hierarchical policy using this representation.
  They derive expressions which bound the sub-optimality and show how these expressions can be translated to representation learning objective which may be optimized in practice.

- The high-level technical insights into the problem of interest
  Hierarchical reinforcement learning (HRL) decomposes a reinforcement learning problem into a hierarchy of subproblems or subtasks such that higher-level parent-tasks invoke lower-level child tasks as if they were primitive actions.
- The main contributions of the paper (compared to the prior works)
  The contribution of the paper is that they use several approach to be the baseline. For example : XY,VAE,E2E,E2C,Whole obs,Resluts of their method and a number of variants on a suite of tasks in 10M steps of training,They find that outside of simple point environments, their method is the only one which can approach the performance of oracle x,y representations. These results show that our method can be successful, even when the representation is learned online concurrently while learning a hierarchical policy. They have presented a principled approach to representation learning in hierarchical RL. Their approach is motivated by the desire to achieve maximum possible return. This notion of sub-optimality is intractable to optimize directly.

- Your personal perspective on the proposed method
  I thought that this paper is totally different between another reinforcement learning. In this paper, Researcher designed the hierarchical RL, which means the target will be the higher level. But the action will begin in lower level.It can be more effective when there a lots of states to scale. Although fixing the representation to be the full state means that no information is lost, but this choice is difficult to scale to high dimension. So the hirarchical RL can solve this problem.

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

Preprint. Under review.

- Your notations (e.g. MDPs, value functions, function approximators,...etc)
  Following the previous work, they construct a two-level hierarchical policy on an MDP,the higher polcy modulates the behavior of a lower-level policy by choosing the desired goal state and rewarding the lower-level policy for reaching this state. Higher-policy π(g | s) where g $\in R^d$. the samples a high-level action $g_t \sim \pi_{hi}$(g|s_t),the lower-level policy $g_t \sim \pi_{lo}$(g|s_t,s_{t+k},k,) translates these high-level actions into lower level actions. ψ denotes this mapping from S x G to π
- The optimization problem of interest
  Two-level policies where a higher-level policy π_{hi}chooses goals g, which are translated into lower-level behaviors via ψ. The choice of ψ restricts the type and number of lower-level behaviors that the higher-level policy can induces. A notion of sub-optimality with respect to the form of ψ. They compared π^*_{hier} to an optimal higher-level policy π^{**}_{bi}(π|s) agnostic to ψ

  SubOpt(ψ) = sup $V^{\pi *}$(s)-$V^{\pi hier}$(s)
- The technical assumptions
  They assumes that they provide proxy expressions that bound the sub-optimality method by a specific choice of ψ. Whihc connects the sub-optimality of ψ to both goal-conditioned policy objectives and representation learning. They presents the results in the restricted case of c=1 and deterministic lower-level policies. If there exists : S x A $\rightarrow$ G such that, such that, sup D_{TV}(P(s'|s,a) || P(s'|s,ψ(s,(s,a))) $\leq \epsilon$

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight your new proofs or extensions (if any).
The common underlying formalism in hierarchical reinforcement learning is the semi-Markov decision process (SMDP). A SMDP generalizes a Markov decision process by allowing actions to be temporally extended. We will state the discrete time equations following Dietterich (2000), recognizing that in general SMDPs are formulated with real-time valued temporally extended actions .

Denoting the random variable N to be the number of time steps that a temporally extended action a takes to complete when it is executed starting in state s, the state transition probability function for the result state s and the expected reward function are given by (1) and (2) respectively, and we try to solve them.
(1) ${T^N}^a_{s\,s'}$ = Pr$\{s_{t+N} = s' | s_t = s, a_t = a\}$
(2)${R^N}^a_{s\,s'}$= E$\{\sum_{i=1}^{N} \gamma^{n-1} | s_t = s, a_t = a, s_{t+N} = s'\}$
As we just saw, the reinfocement learning problem suffers from serious scaling issues. Hierarchical reinforcement learning is a computational approach intend ed to address these issues by learning to operate on different levels of temporal abstraction.
The promise of HRL is to have :
1. Loog-term credit assignment : faster learning and better generalisation
2. Structured exploration : explore with sub-policies rather than primitive actions
3. Transfer learning : different levels of hierarchy can encompass different knowledge and allow for better transfer.
The HRL method demonstrates how to create a managerial leraning hierarchy in which lords learn to assign tasks to their serf learn to satisfy them. Sub-managers learn to maximize their reinforecment in the context of the command as pictured in the illustration below. The hierarchical learning takes advantage of two notions: Information hiding and Reward hiding.
A noteworthy effect of information and reward hiding is that the managers only need to know the state of the system at the granularity of their own choice of tasks.
MAXQ is a hierarchical learning algorithm in which the hierarchy of a task is obtained by decomposing the Q value of state-action pair into the sum of two components Q(p,s,a) = V(a,s)+ C(p,s,a) where V(a,s) is the total expected reward received when executing the action a in state s and C(p,s,a) is the total reward expected from the performance of the parent-task, noted by p.
And how to learn a good representation, consider a representation function f_θ: S $\rightarrow$ R^d,parameterized

by vector θ, In practice,these are seperate neural networks:f_θ =[θ_1,θ_2] Eventually, we want to learn a lower-level policy,which is standard in goal-conditioned hierarchical design ; Standard RL algorithms may be employes to maximize the lower-level reward implied by : -D(f($s_{tk}|g$) + $log\rho$(S_{t+k})) − $logP_\pi$($s_{t+k}|s_t$)

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
  The mapping of observation space to goal space is crucial. The sub-optimality bound can make the action more effective. When the lower policy translated into the higher policy. It can handle more useful information. Somebody will ask that why we don't use simple RL to take actions directly. However, the states will be too big to take actions from all. So we use the hierarchical reinforcement learning to get get the target.
- Any technical limitations
  The notion of sub-optimality is intractable to optimize directly,we are able to derive a mathematical relationship between it and a specific form of representation learning.
- Any latest results on the problem of interest

References