VIME:Variational Information Maximizing Exploration

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, Pieter Abbeel

Zahra Shekarchi F.

University of Toronto

Fall 2018

Introduction

- Exploitation: The agent maximizes rewards through behavior that is known to be successful.
- Exploration: The agent experiments with novel strategies that may improve returns in the long run.
- Trade-off?

Here

Scalable and effective exploration in the environment, to be applied in high-dimensional deep RL scenarios.

- continuous state and action spaces

- sparse rewards

Introduction (cont.)

To solve the trade-off:

- Baysian RL
- PAC-MDP
- Assume discrete state and action spaces

Heuristic exploration strategies:

- $\epsilon -$ *greedy*
- Boltzmann exploration
- Gaussian noise on the controls in policy gradient methods
- So, random walk behavior - highly inefficient

Establish Notation

A finite-horizon discounted MDP as $(S, A, P, r, \rho_0, \gamma, T)$:

- $S \subseteq \mathbb{R}^n$: state set
- $A \subseteq \mathbb{R}^m$: action set
- $P : SxAxS \rightarrow \mathbb{R}_{\geq 0}$: transition probability distribution
- $r : SxA \rightarrow \mathbb{R}$: reward function
- $\rho_0 : S \rightarrow \mathbb{R}_{\geq 0}$: initial state distribution
- $\gamma \in (0{,}1]$: discount factor
- T : horizon
- $\pi_\alpha : SxA \rightarrow \mathbb{R}_{\geq 0}$: a stochastic policy
- $\mu(\pi_\alpha) = \mathbb{E}_\tau[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$: expected discounted return

Curiosity

- $p(s_{t+1}|s_t, a_t, ; \theta)$: Θ agent model of the environment dynamics
  θ, a random variable Θ, prior $p(\theta)$, $\theta \in \Theta$
- $\xi_t = \{s_1, a_1, s_2, ..., a_t\}$: history of the agent up to step t
- should take actions maximize the reduction in uncertainty about the dynamics

  $\sum_t (H(\Theta|\xi_t, a_t) - H(\Theta|S_{t+1}, \xi_t, a_t))$
- this is mutual information between s_{t+1} and Θ

  $I(S_{t+1}; \Theta|\xi_t, a_t) =$
  $\mathbb{E}_{s_{t+1} \sim P(.|\xi_t, a_t)}[D_{KL}[p(\theta|\xi_t, a_t, s_{t+1})||p(\theta|\xi_t)]]$

Curiosity (cont.)

An approximate approach:

- taking action $a_t \sim \pi_\alpha(s_t)$
- sampling $s_{t+1} \sim P(.|s_t, a_t)$
- obtaining the new reward: add curiosity to the reward

  $r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}[p(\theta|\xi_t, a_t, s_{t+1})||p(\theta|\xi_t)]$

  where $\eta \in \mathbb{R}_+$ is a parameter to tune the exploration

Though, calculating the posterior dynamics distribution $(p(\theta|\xi_t, a_t, s_{t+1}))$ is intractable!

Variational Inference

Bayes' rule:

$p(\theta|\xi_t, a_t, s_{t+1}) = \frac{p(\theta|\xi_t)p(s_{t+1}|\xi_t, a_t; \theta)}{p(s_{t+1}|\xi_t, a_t)}$

where $p(\theta|\xi_t, a_t) = p(\theta|\xi_t)$ and

$p(s_{t+1}|\xi_t, a_t) = \int_\Theta p(s_{t+1}|\xi_t, a_t; \theta)p(\theta|\xi_t)\, d\theta$

- estimate $p(\theta|D)$ with $q(\theta; \phi)$
- through maximization of the variational lower bound: $L[q(\theta; \phi), D] = \mathbb{E}_{\theta \sim q(.;\phi)}[logp(D|\theta)] - D_{KL}[q(\theta; \phi)||p(\theta)]$

So, we have

$r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}[q(\theta; \phi_{t+1})||q(\theta|\phi_t)]$

Implementation

- learn $SxA \rightarrow S$ transition model via Bayesian Neural Network (BNN)
- The BNN weight distribution:

  $q(\theta;\phi) = \prod_{i=1}^{|\Theta|} \mathcal{N}(\theta_i|\mu_i;\sigma_i^2)$

Implementation (cont.)

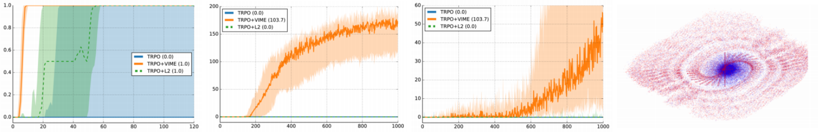
until convergence/goal:

- interact with the environment $\rightarrow < s_t, a_t, r, s_{t+1} >$
- compute curiosity reward
  $r'(s_t, a_t, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}[q(\theta; \phi_{t+1})||q(\theta|\phi_t)]$
- train agent $< s_t, a_t, r', s_{t+1} >$
- train BNN $< s_t, a_t, s_{t+1} >$

Experimental Setup

- continuous control tasks
- $S \in \mathbb{R}^3, \mathbb{R}^4, \mathbb{R}^6, \mathbb{R}^{20}, \mathbb{R}^{33}$
- $A \in \mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^6$
- TRPO [Schulman et al. 2015] used to learn policies
- sparse rewards

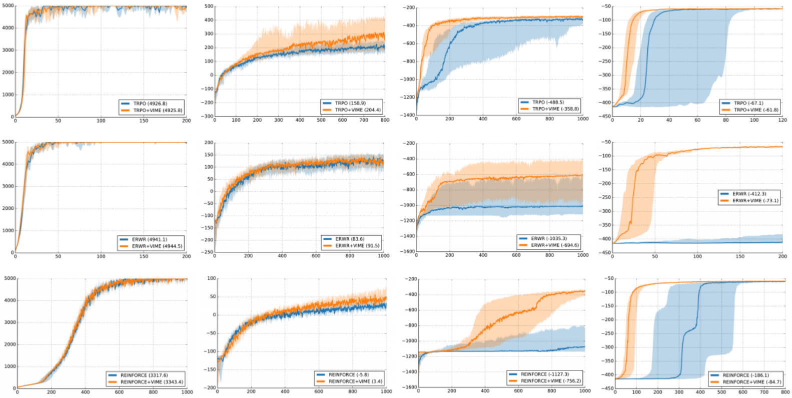
Results



(a) MountainCar (b) CartPoleSwingup (c) HalfCheetah (d) state space

Figure 1: (a,b,c) TRPO+VIME versus TRPO on tasks with sparse rewards; (d) comparison of TRPO+VIME (red) and TRPO (blue) on MountainCar: visited states until convergence

Results (cont.)

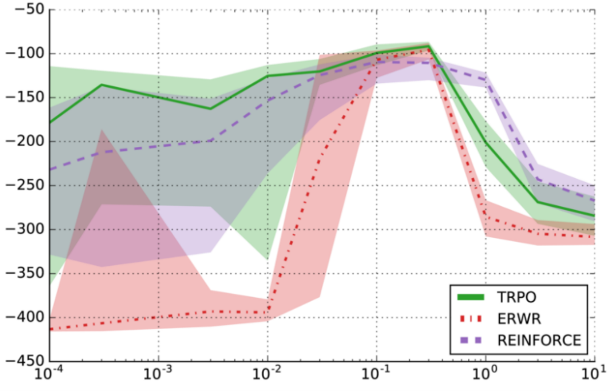


(a) CartPole (b) CartPoleSwingup (c) DoublePendulum (d) MountainCar

Figure 2: Performance of TRPO (top row), ERWR (middle row), and REINFORCE (bottom row) with (+VIME) and without exploration for different continuous control tasks.

Results (cont.)

The role of η:



Discussion

- VIME: a curiosity-driven exploration strategy for continuous control tasks
- used VI to approximate the posterior distribution of BNN
- BNN represent the environment dynamics
- using information gain as intrinsic rewards
- be curious about irrelevant things
  e.g. background in a game

Thank you