VIME: Variational Information Maximizing Exploration

Доклад Турышева Арсения

Exploration/exploitation trade-off

- **Exploration** – исследование среды, поиск новых способов решения
- **Exploitation** – использование уже известной информации о среде с целью максимизации награды

Эвристические методы:

- **ε-greedy**: с вероятностью ε равновероятно выбираем случайное действие
- **count-based**: штрафует за действия, которые выполнялись слишком часто

Curiosity-based exploration

Формализуем «любопытство» к исследованию среды

- Динамику среды зададим с помощью модели с параметрами θ: $p(s_{t+1}|s_t, a_t, \theta)$, $p(\theta)$.

- Пусть $\xi_t = \{s_0, a_0, \dots, s_t\}$ – история агента.
- При выборе очередного действия максимизируем сокращение энтропии:

$$H(\theta|\xi_t) - H(\theta|s_{t+1}, a_t, \xi_t) \to \max_{a_t}$$

старая информация

новая информация после очередного действия

Curiosity-based exploration

Распишем через взаимную информацию:

$$H(\theta|\xi_t) - H(\theta|s_{t+1}, a_t, \xi_t) = I(s_{t+1}; \theta) =$$

$$= \mathbb{E}_{s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[D_{KL}[p(\theta|s_{t+1}, a_t, \xi_t)||p(\theta|\xi_t)]]$$

новое распределение старое распределение

Прибавим наш exploration к награде:

Intrinsic reward

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

гиперпараметр – склонность к исследованию

Variational Bayes

Проблема: $p(\theta|s_{t+1}, a_t, \xi_t)$ – intractable.

Решаем задачу с помощью байесовских сетей. Приближаем $p(\theta|s_{t+1}, a_t, \xi_t)$ с помощью $q(\theta|\phi) = q(\theta|\mu, \sigma) = \prod_i \mathcal{N}(\theta_i|\mu_i, \sigma_i)$

Для нахождения параметров приближения $\boldsymbol{\phi}$ максимизируем **ELBO**:

$$L[q(\theta|\phi), \xi] = \underbrace{\mathbb{E}_{\theta\sim q(\theta|\phi)}[\log p(\xi|\theta)]}_{\text{оценка Монте-Карло}} - D_{KL}[q(\theta|\phi)||p(\theta)]$$

Награда:

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

VIME

Для каждой эпохи:

Для каждого момента времени каждой траектории:

Сохраняем тройку (s_{t+1}, a_t, s_t) в пул

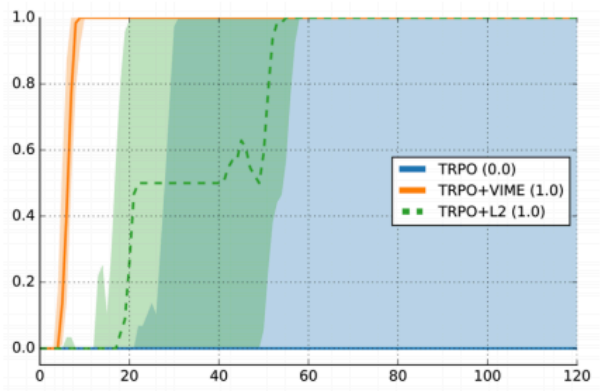
Вычисляем нормированные intrinsic reward

Обновляем агента любым стандартным RL-алгоритмом

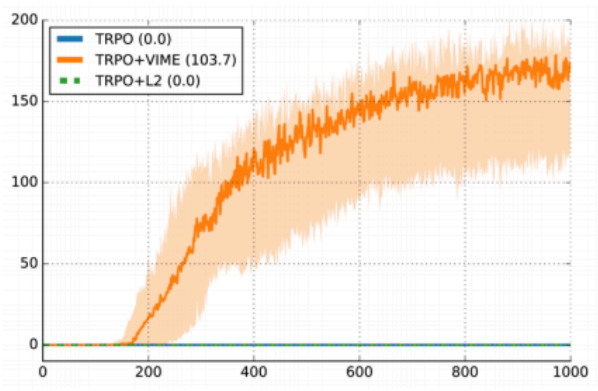
Обновляем веса ϕ, ξ сэмплируется из пула:

$$L[q(\theta|\phi), \xi] = \mathbb{E}_{\theta \sim q(\theta|\phi)}[\log p(\xi|\theta)] - D_{KL}[q(\theta|\phi)||p(\theta)]$$

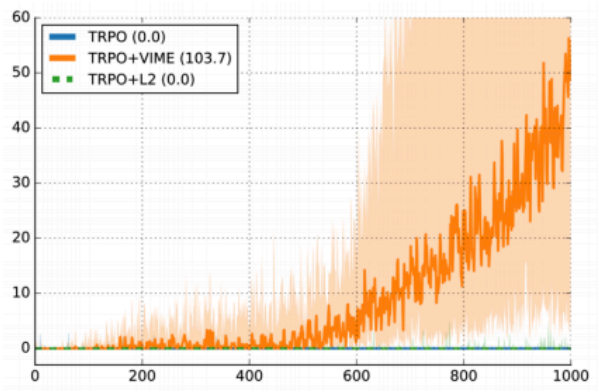
Результаты



(a) MountainCar



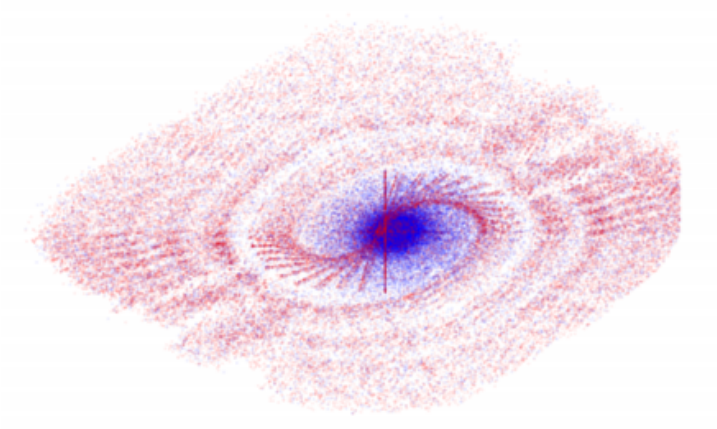
(b) CartPoleSwingup



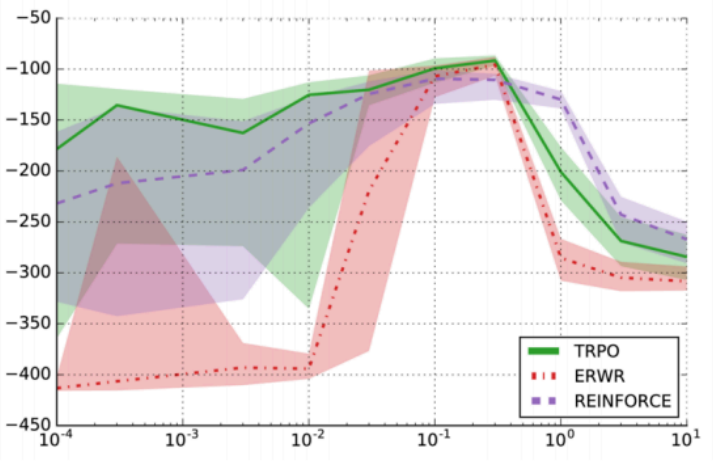
(c) HalfCheetah

Достигнутая награда в зависимости от числа итераций для трех сред разреженной наградой

Результаты



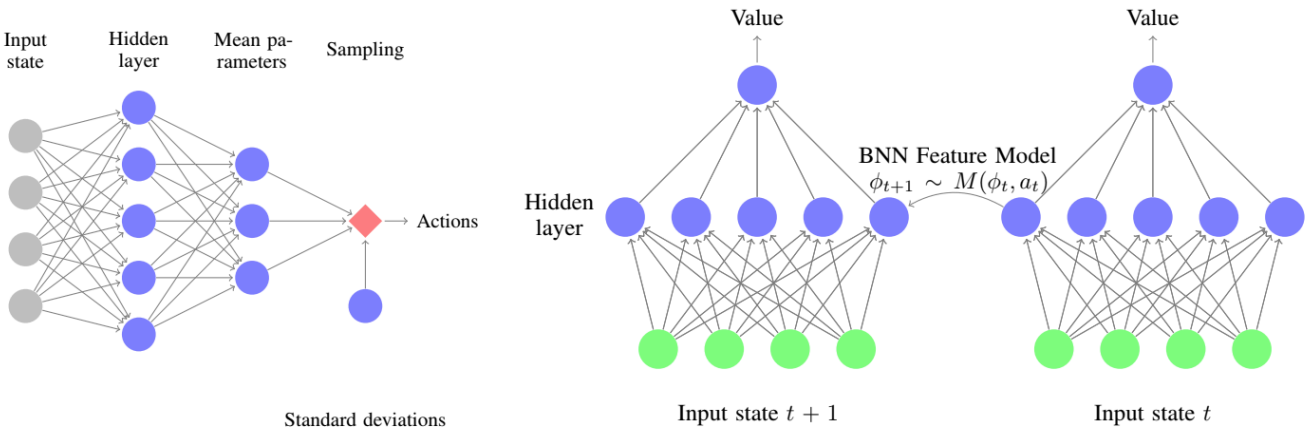
Множество посещенных состояний: VIME+TRPO (красный), TRPO (синий)



Зависимость результата от гиперпараметра η

Предложение об улучшении

- Не все данные о состоянии релевантны для агента.
- Будем использовать эмбеддинги с предпоследнего слоя агента



Ссылка. Trevor Barron, Heni Ben Amor. Information Maximizing Exploration with a LatentDynamics Model

Выводы

- VIME оптимизирует exploration, используя механизм intrinsic reward.
- Exploration задается через KL-дивергенцию между апостериорным распределением параметров агента при новых данных и апостериорным распределением при старых данных.
- Динамика среды задается байесовской нейронной сетью и оптимизируется с помощью максимизации ELBO.
- Метод является state-of-the-art методом в области.

Ссылка

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME:Variational information maximizing exploration. In Advances in Neural Information ProcessingSystems, pages 1109–1117, 2016.