

Объектно-центричное представление мира агента в обучении с подкреплением

Студент: Ярослав Ивченков¹

Научный консультант: Панов А. И., к. ф-м н.¹²

Научный руководитель: Матвеев И. А., д.т.н.¹



¹Московский Физико-Технический Институт

²Научно-Исследовательский Институт Искусственного
Интеллекта (AIRI)

Постановка задачи

- Среда: Марковский Процесс Принятия Решений $\langle S, A, R, p, \gamma \rangle$
 - S – пространство состояний
 - A – пространство действий
 - $p(s' | s, a)$ – функция перехода состояний среды
 - $R(s, s', a)$ – функция вознаграждения среды
- Задача: найти стратегию $\pi(a | s)$ максимизирующую ожидаемую отдачу $J(\pi, p, R) = \mathbb{E}_{p, \pi} \sum_t \gamma^t R_t$
- Модельное обучение с подкреплением: аппроксимировать функцию перехода состояний среды и функцию награды “моделью мира” $\hat{p}_\theta(s' | s, a)$, $\hat{R}_\theta(s, s', a)$ с которой агент может взаимодействовать с целью максимизации $J(\pi, \hat{p}_\theta, \hat{R}_\theta)$. Модель мира обучается на буфере опыта, полученного из взаимодействия агента со средой.
- Обобщение между задачами: пусть дано распределение задач (каждая из которых является МППР) $p_{\text{train}}(\tau)$. Обучая модельного агента на задачах $\tau \in p_{\text{train}}(\tau)$, необходимо максимизировать ожидаемую отдачу на задачах $\tau \in p_{\text{test}}(\tau)$.

Объектно-центричное обучение

- Базовое предположение: наблюдение состоит из N объектов, каждый из которых может быть смоделирован по отдельности
- Объекты могут взаимодействовать друг с другом и влиять друг на друга
- Объектная абстракция позволяет ввести структуру в наблюдения, представленные изображениями в таких задачах, как предсказание видео или визуальный контроль в обучении с подкреплением
- Может быть применено для повышения скорости обучения и обобщающих способностей в обучении с подкреплением

Объектно-центричное обучение

- Базовое предположение: наблюдение состоит из N объектов, каждый из которых может быть смоделирован по отдельности
- Объекты могут взаимодействовать друг с другом и влиять друг на друга
- Объектная абстракция позволяет ввести структуру в наблюдения, представленные изображениями в таких задачах, как предсказание видео или визуальный контроль в обучении с подкреплением
- Может быть применено для повышения скорости обучения и обобщающих способностей в обучении с подкреплением

Объектно-центричное обучение

- Базовое предположение: наблюдение состоит из N объектов, каждый из которых может быть смоделирован по отдельности
- Объекты могут взаимодействовать друг с другом и влиять друг на друга
- Объектная абстракция позволяет ввести структуру в наблюдения, представленные изображениями в таких задачах, как предсказание видео или визуальный контроль в обучении с подкреплением
- Может быть применено для повышения скорости обучения и обобщающих способностей в обучении с подкреплением

Объектно-центричное обучение

- Базовое предположение: наблюдение состоит из N объектов, каждый из которых может быть смоделирован по отдельности
- Объекты могут взаимодействовать друг с другом и влиять друг на друга
- Объектная абстракция позволяет ввести структуру в наблюдения, представленные изображениями в таких задачах, как предсказание видео или визуальный контроль в обучении с подкреплением
- Может быть применено для повышения скорости обучения и обобщающих способностей в обучении с подкреплением

Comparison

Differences:

- Presence of temporal context
- Presence of conditional factors
- Entity categories
- Interaction model
- Particular task peculiarities

Comparison

Differences:

- Presence of temporal context
- Presence of conditional factors
- Entity categories
- Interaction model
- Particular task peculiarities

Comparison

Differences:

- Presence of temporal context
- Presence of conditional factors
- Entity categories
- Interaction model
- Particular task peculiarities

Comparison

Differences:

- Presence of temporal context
- Presence of conditional factors
- Entity categories
- Interaction model
- Particular task peculiarities

Comparison

Differences:

- Presence of temporal context
- Presence of conditional factors
- Entity categories
- Interaction model
- Particular task peculiarities

Объектно Центричный Визуальный Контроль

Обычная задача визуального контроля включает в себя:

- Актуатор, напрямую контролируемый действиями
- Объекты, косвенно контролируемые актуатором
- Функцию награды, подразумевающую взаимодействие между актуатором и объектом(-ами).

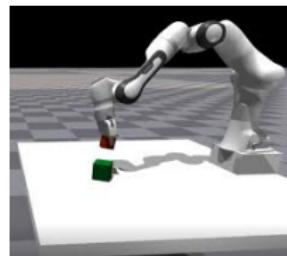
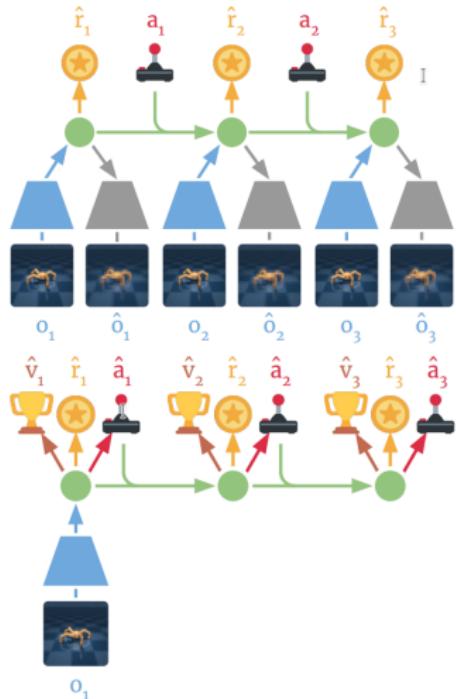


Figure: Различные среды визуального контроля

Базовая модель: Dreamer

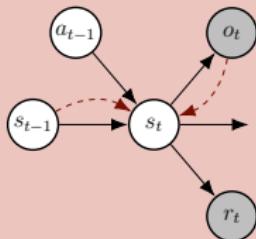
- Dreamer* представляет собой вариационный автоэнкодер (VAE), кодирующий наблюдения и награды
- Он состоит из
 - модели репрезентации (энкодера) $q(s_t | s_{t-1}, a_{t-1}, o_t)$
 - модель среды $p(s_t | s_{t-1}, a_{t-1})$
 - модель наблюдений (декодер изображений) $p(o_t | s_t)$
 - модели награды (декодер наград) $p(r_t | s_t)$
- Наблюдения $o_{1:T}$ считаются не марковскими, в связи с чем VAE выводит марковские состояния s_t
- Стратегия и критик тренируются на траекториях, полученных в "воображении" модели, полученных при помощи модели среды $p(s_t | s_{t-1}, a_{t-1})$



Модель мира СЕМА

Модель мира Dreamer[1]:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики среды
- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода
- $p(o_t | s_t)$ модель изображения
- $p(r_t | s_t)$ модель награды

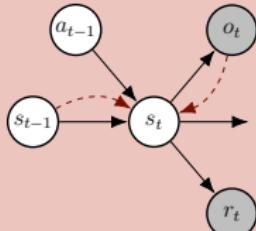


Динамика среды

Модель мира СЕМА

Модель мира Dreamer[1]:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики среды
- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода
- $p(o_t | s_t)$ модель изображения
- $p(r_t | s_t)$ модель награды

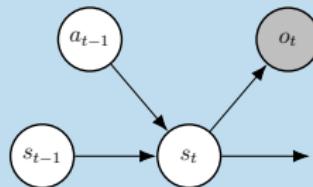


Динамика среды

Модель мира СЕМА:



Динамика объекта



Динамика робота

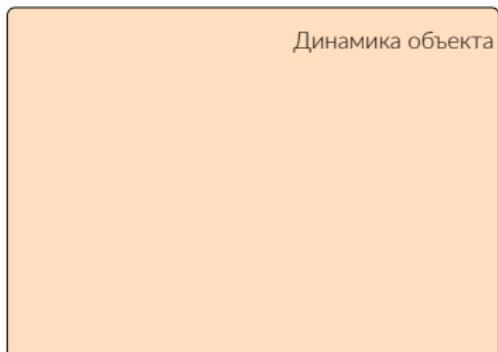
Модель мира СЕМА

Части модели мира:

Генеративная модель прямой причинности:

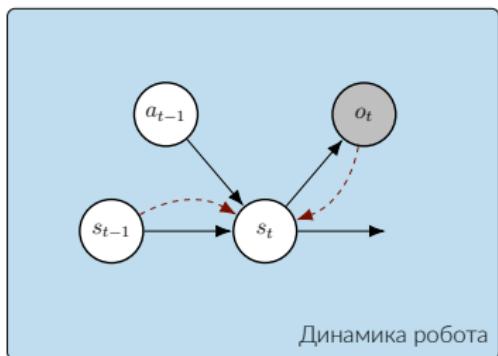
- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний

Модель мира СЕМА:



Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота



Выходы нейронной сети:

- $p(o_t | s_t)$ субъектный декодер

Модель мира СЕМА

Части модели мира:

Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности

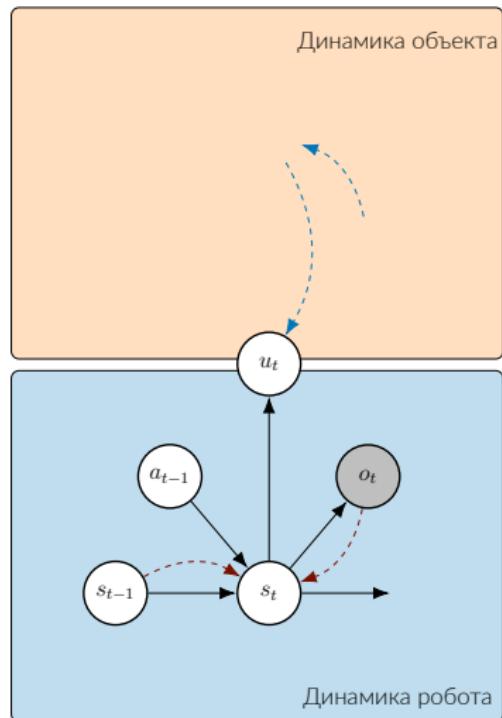
Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота

Выходы нейронной сети:

- $p(o_t | s_t)$ субъектный декодер

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

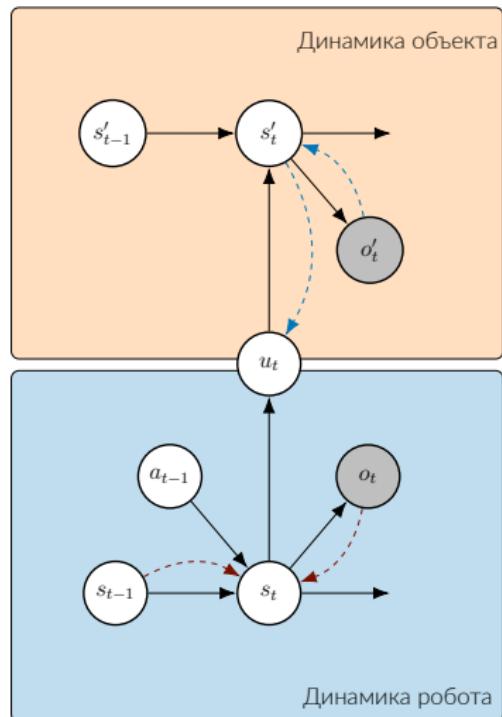
Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

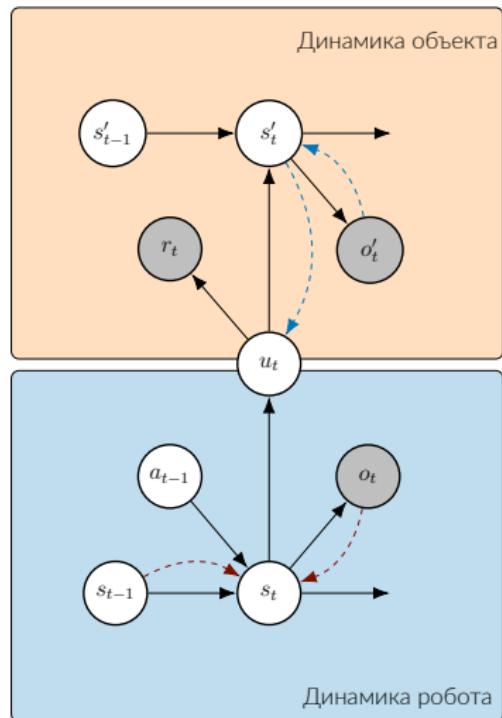
Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

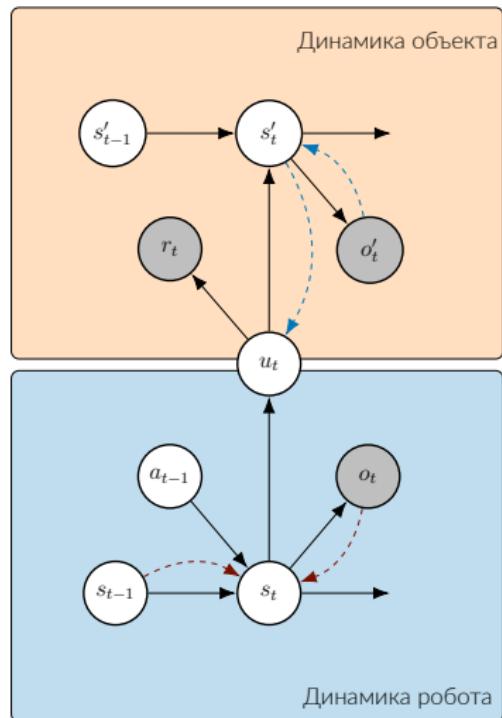
Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

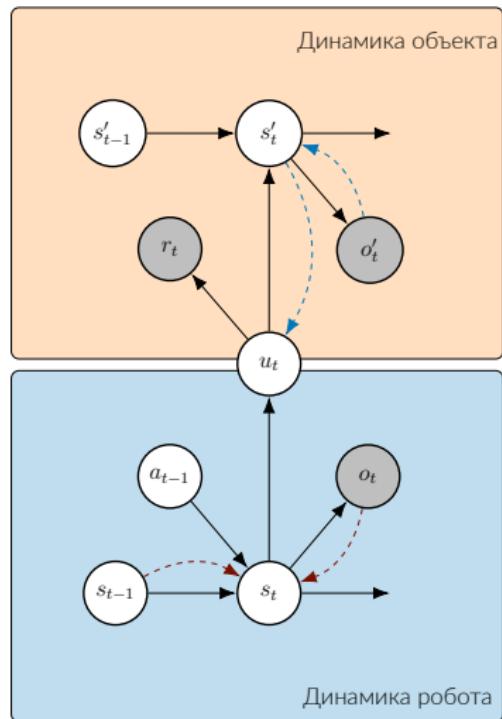
Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота
- $s'_t \sim q(s'_t | o'_t)$ модель вывода объектного состояния

Выходы нейронной сети:

- $p(o_t | s_t)$ субъектный декодер
- $p(o'_t | s'_t)$ объектный декодер
- $p(r_t | u_t)$ декодер награды

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

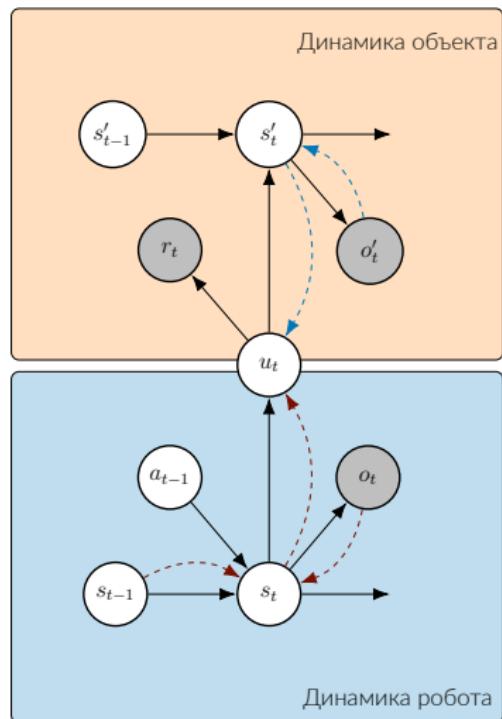
Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота
- $s'_t \sim q(s'_t | o'_t)$ модель вывода объектного состояния
- $u_t \sim q(u_t | s_t, s'_t)$ модель обратной причинности

Выходы нейронной сети:

- $p(o_t | s_t)$ субъектный декодер
- $p(o'_t | s'_t)$ объектный декодер
- $p(r_t | u_t)$ декодер награды

Модель мира СЕМА:



Модель мира СЕМА

Части модели мира:

Генеративная модель прямой причинности:

- $s_t \sim p(s_t | s_{t-1}, a_{t-1})$ модель динамики скрытых состояний
- $u_t \sim p(u_t | s_t, c)$ модель прямой причинности
- $s'_t \sim p(s'_t | s'_{t-1}, u_t)$ модель обновления

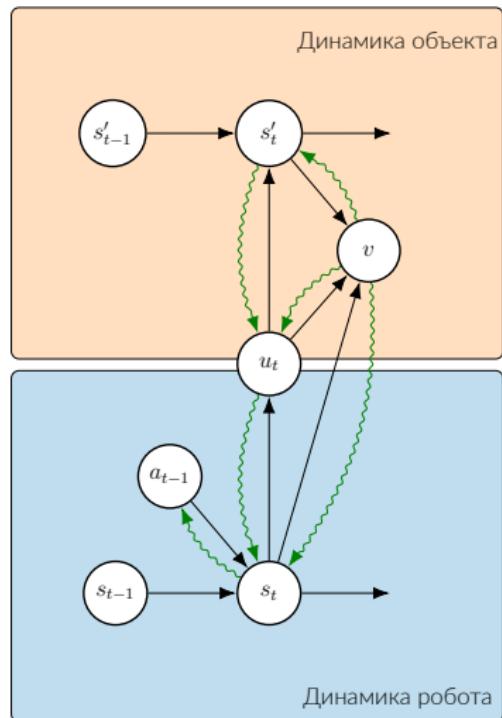
Обратная причинная модель (модель вывода):

- $s_t \sim q(s_t | s_{t-1}, a_{t-1}, o_t)$ модель вывода состояния робота
- $s'_t \sim q(s'_t | o'_t)$ модель вывода объектного состояния
- $u_t \sim q(u_t | s_t, s'_t)$ модель обратной причинности

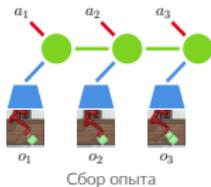
Выходы нейронной сети:

- $p(o_t | s_t)$ субъектный декодер
- $p(o'_t | s'_t)$ объектный декодер
- $p(r_t | u_t)$ декодер награды
- $\pi(a_t | s_t, c), v(s_t, u_t, s'_t)$ стратегия и функция полезности

Модель мира СЕМА:

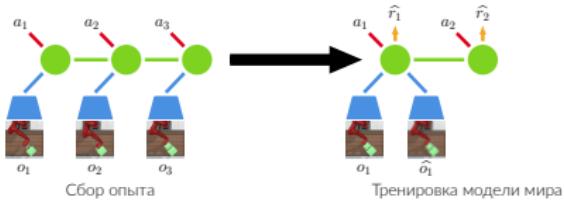


Тренировочный процесс

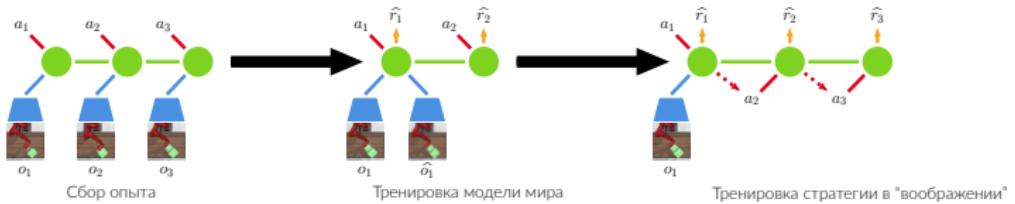


Сбор опыта

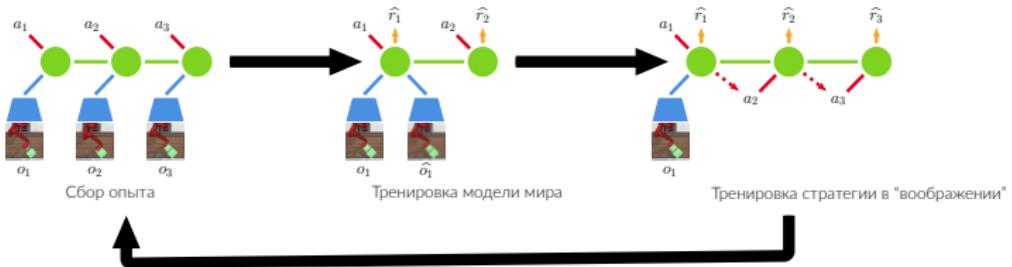
Тренировочный процесс



Тренировочный процесс



Тренировочный процесс



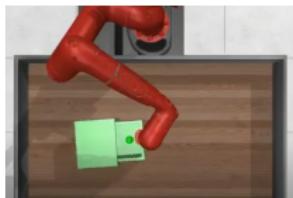
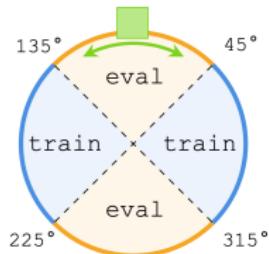
Эксперимент: Rotated Drawer World, описание среды

Среда для эксперимента:



Эксперимент: Rotated Drawer World, описание среды

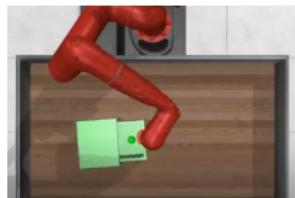
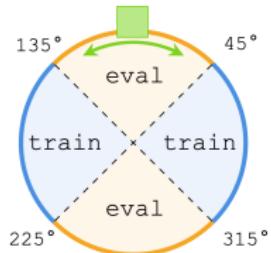
Среда для эксперимента:



Эксперимент: Rotated Drawer World, описание среды

Пример наблюдения:

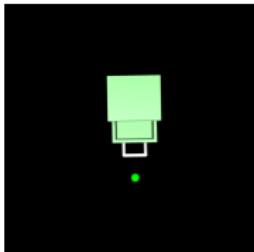
Среда для эксперимента:



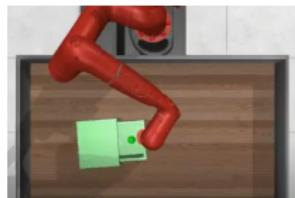
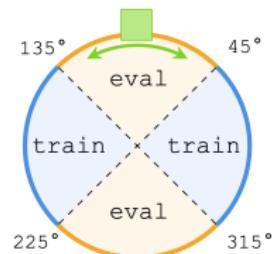
Эксперимент: Rotated Drawer World, описание среды

Пример наблюдения:

- Объектная часть



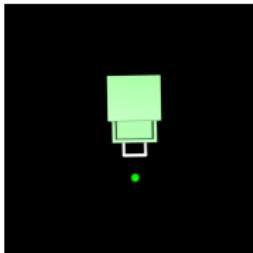
Среда для эксперимента:



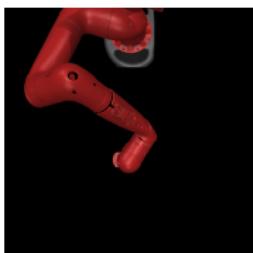
Эксперимент: Rotated Drawer World, описание среды

Пример наблюдения:

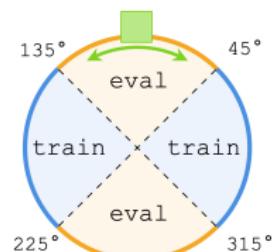
- Объектная часть



- Субъектная часть



Среда для эксперимента:



Эксперимент: Rotated Drawer World, результаты

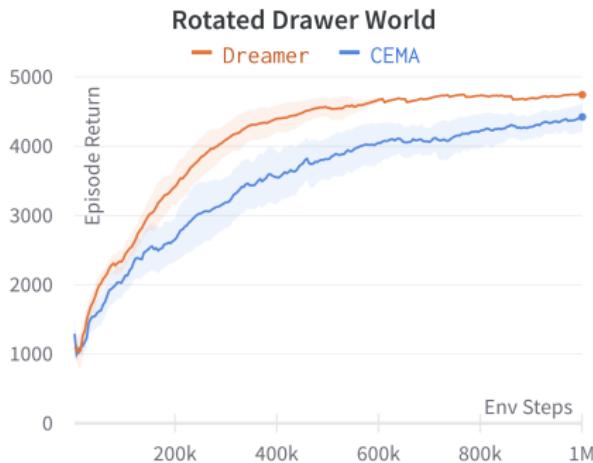


Figure: График наград на тренировочных задачах алгоритмов Dreamer и СЕМА.
Награды больше **3000** соответствуют решенной задаче.

Эксперимент: Rotated Drawer World, результаты

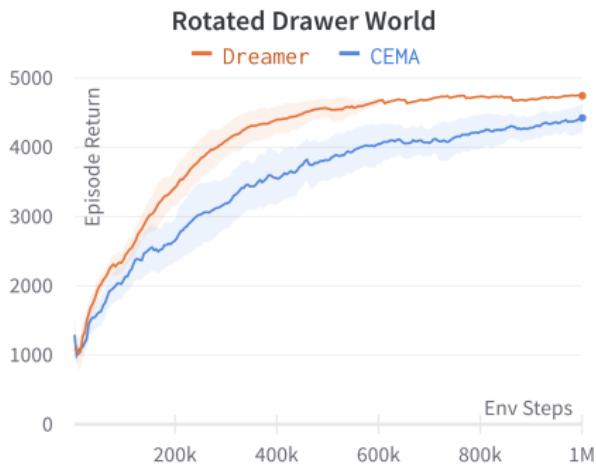


Figure: График наград на тренировочных задачах алгоритмов Dreamer и СЕМА.
Награды больше 3000 соответствуют решенной задаче.

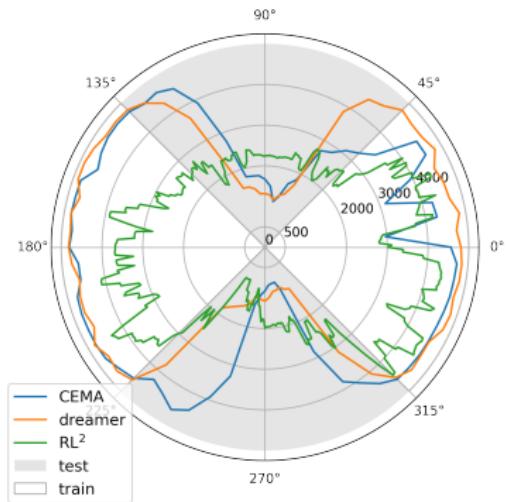


Figure: Результаты работы алгоритмов на каждой задаче. Белые регионы соответствуют тренировочным задачам, синие - тестовым.

Эксперимент: Взаимная Информация (MI)

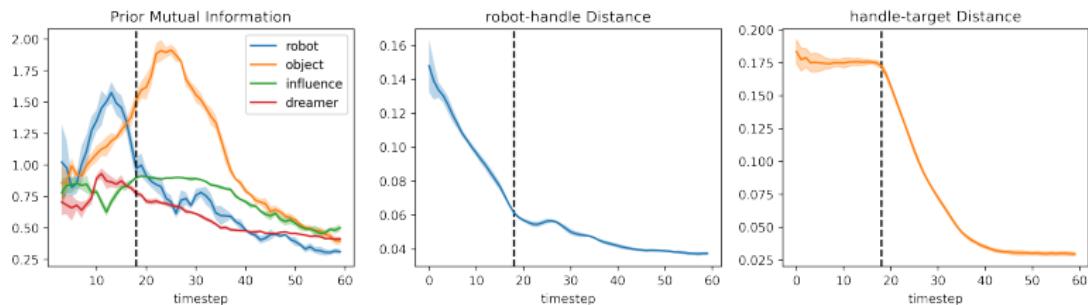


Figure: Взаимную информацию между входом и выходом различных частей модели мира агента.

Эксперимент: Взаимная Информация (MI)

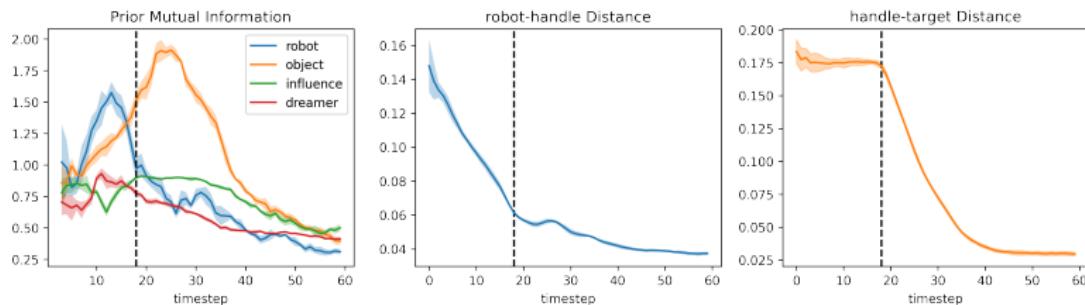


Figure: Взаимную информацию между входом и выходом различных частей модели мира агента.

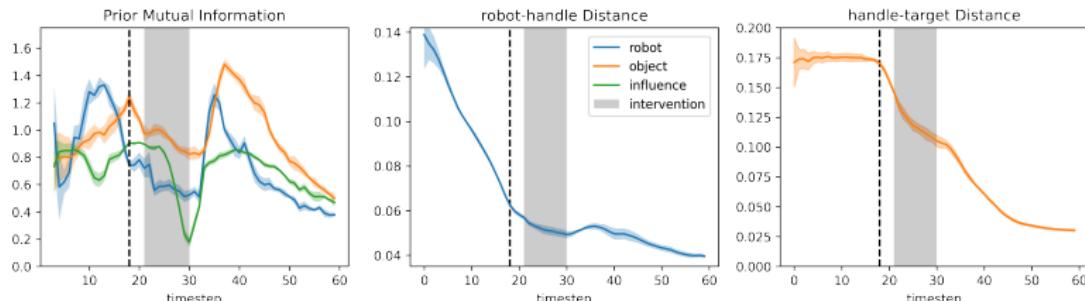
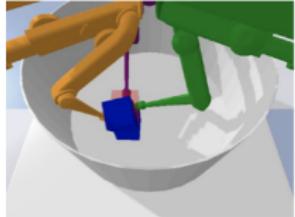
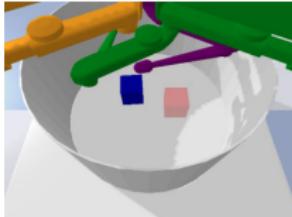
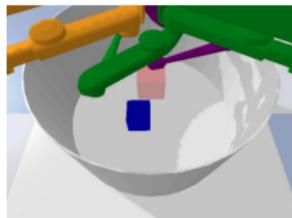


Figure: Взаимная информация в условиях искусственного вмешательства в эпизод*.

* В середине открытия ящика, движение всех объектов в среде заморожено путем искусственной вставки нескольких пустых действий.

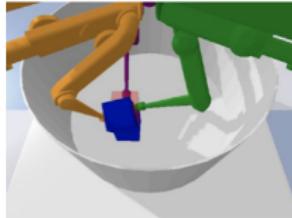
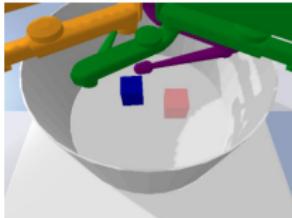
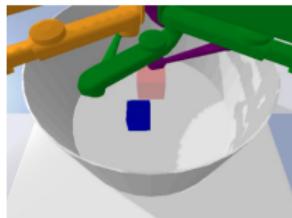
Эксперимент: Causal World, описание среды

Пример наблюдений:



Эксперимент: Causal World, описание среды

Пример наблюдений:

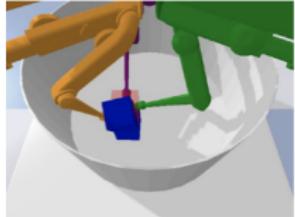
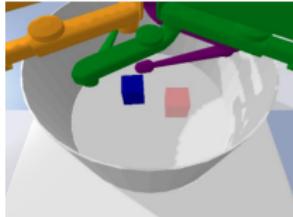
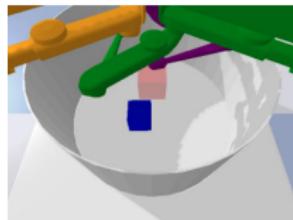


Эксперимент: Causal World, описание среды

Сырое изображение из
среды::



Пример наблюдений:



Эксперимент: Causal World, результаты

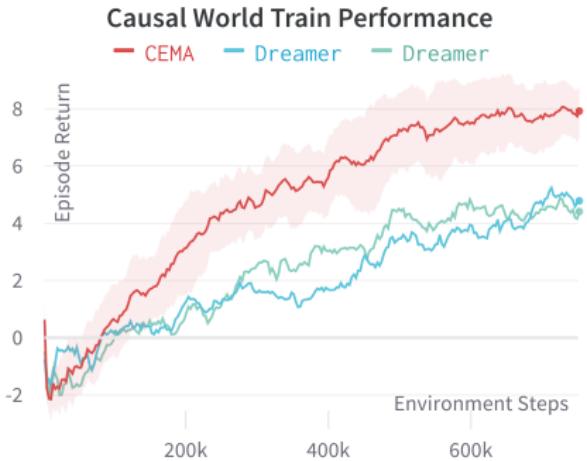


Figure: График наград на тренировочных задачах алгоритмов Dreamer и СЕМА.
Представленные награды являются неизмененными наградами среды.

Эксперимент: Causal World, результаты

Causal World Train Performance

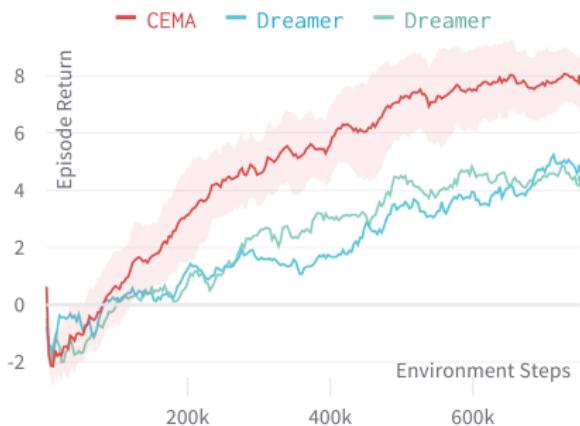


Figure: График наград на тренировочных задачах алгоритмов Dreamer и СЕМА. Представленные награды являются неизмененными наградами среды.

Causal World Evaluation Performance

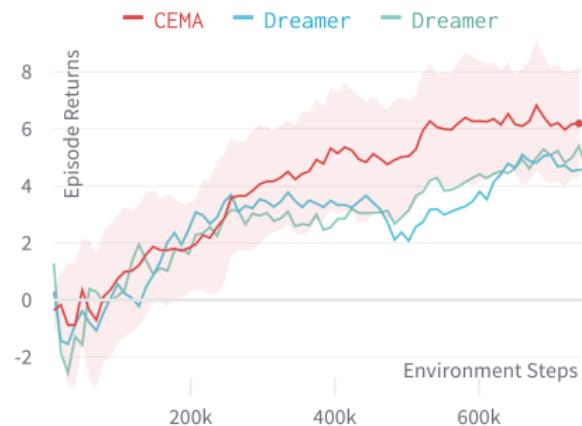


Figure: Сравнение результатов работы алгоритма на тестовых задачах.

Публикации

- **Принято:** Artem Zholus, Yaroslav Ivchenkov и Aleksandr Panov.
“Factorized World Models for Learning Causal Relationships”. В: ICLR2022
Workshop on the Elements of Reasoning: Objects, Structure and Causality.
2022.
- **Отправлено:** Artem Zholus, Yaroslav Ivchenkov, Aleksandr Panov “Casual
Factorized World Models in Model-based Reinforcement Learning”.
NeurIPS 2022 (under review)

Объектно-центричное представление мира агента в обучении с подкреплением

Код доступен по адресу:

github.com/artemZholus/generalization_with_world_models

Спасибо за внимание!