Обучение с подкреплением на основе модели и дискретных моделей мира

Вахитов Р.Р.
rodionvahitoff@yandex.ru

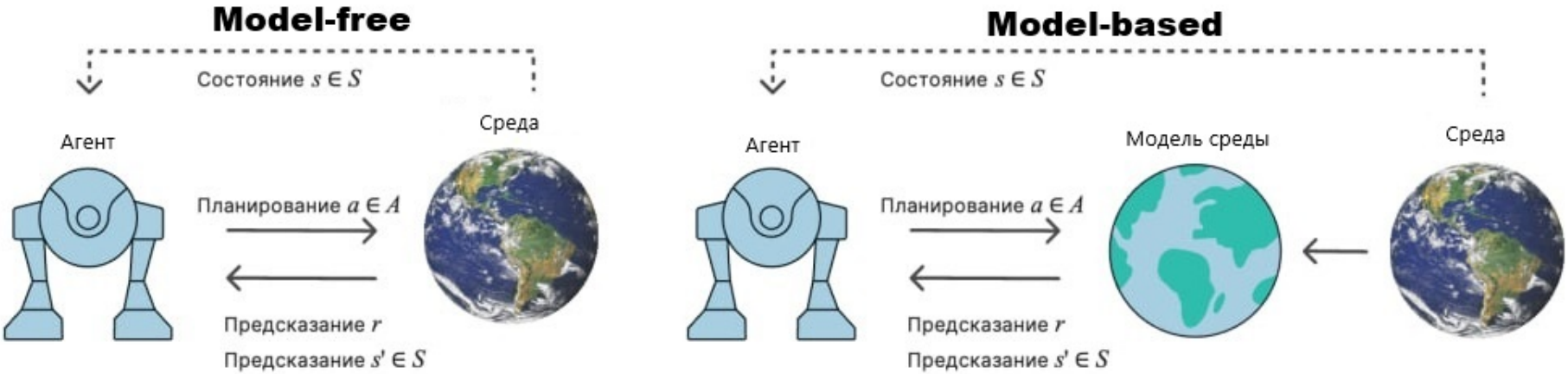
Центр когнитивного моделирования МФТИ

2024 – Обзор исследований





Введение



Цель данного обзора — исследовать современные методы обучения с подкреплением на основе модели, уделяя особое внимание представлению латентного пространства и анализу их роли в обеспечении обобщающей способности и эффективности алгоритмов.

Постановка задачи

(S, A, T, R, G, γ) - марковский процесс принятия решений, где:

- S: пространство состояний,
- A: множество действий,
- $T : S \times A \to S$: функция переходов,
- $R : S \times A \to \mathbb{R}$: функция вознаграждения,
- $G : S \to \{0, 1\}$: функция цели, определяющая состояние завершения,
- γ: коэффициент дисконтирования.

Агент имеет доступ к обновляемой модели среды $M = \langle \hat{T}, \hat{R} \rangle$ и может строить план для достижения цели $G(s_{n+1}) = 1$, моделируя переходы:

$$\text{Plan} = \langle s_i, r_i, a_i, \hat{s}_{i+1}, \hat{r}_{i+1}, a_{i+1}, \ldots, a_{i+n}, \hat{s}_{i+n+1} \rangle$$

Цель агента — максимизировать ожидаемое кумулятивное вознаграждение, определяемое стратегией π:

$$\mathbb{E}_\pi \sum_{t=0}^{\tau} \gamma^t R(s_t, a_t)$$

MuZero

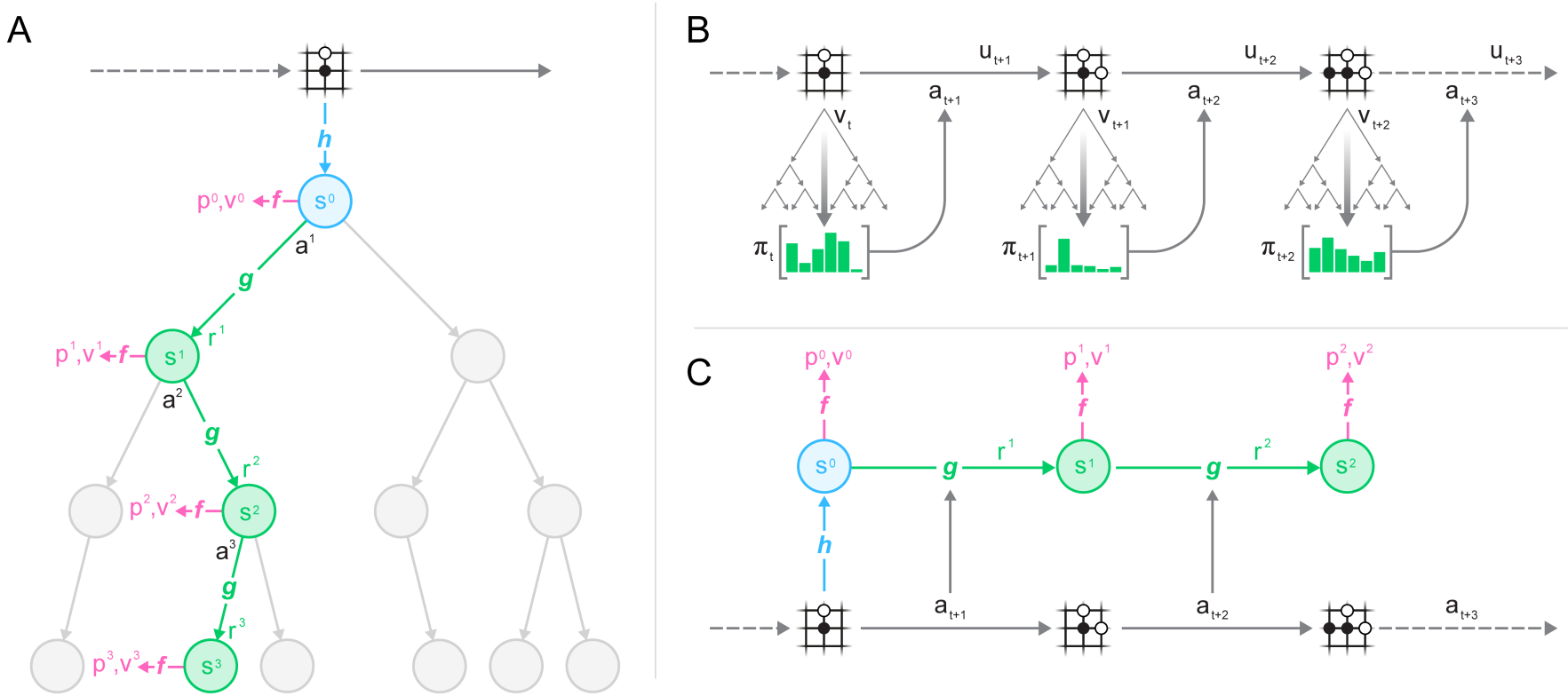
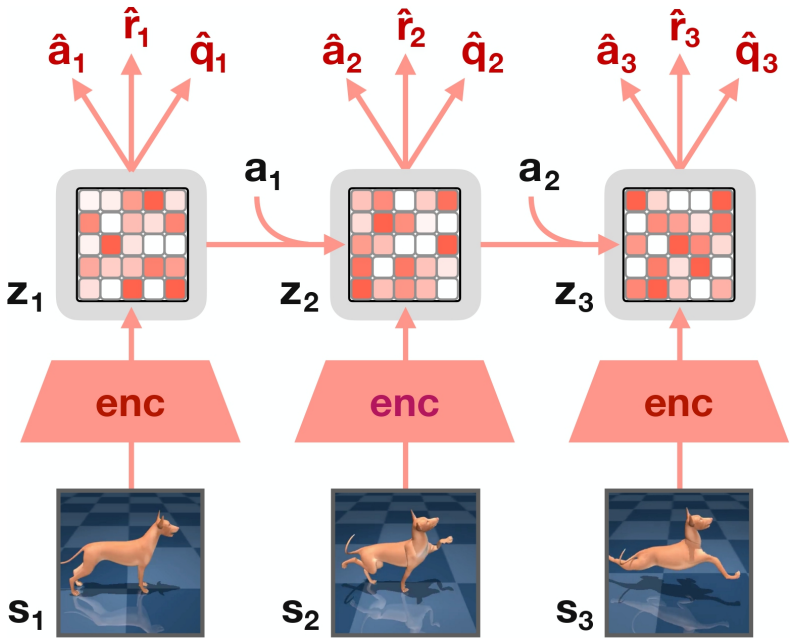


Рис.: **(A)** MCTS с моделями динамики, стратегии и предсказания. **(B)** MuZero взаимодействует с окружением через MCTS, выбирая действия по стратегии π_t для каждого состояния. **(C)** Параметры моделей обучаются совместно с использованием обратного распространения ошибки сквозь время для предсказания ключевых величин.

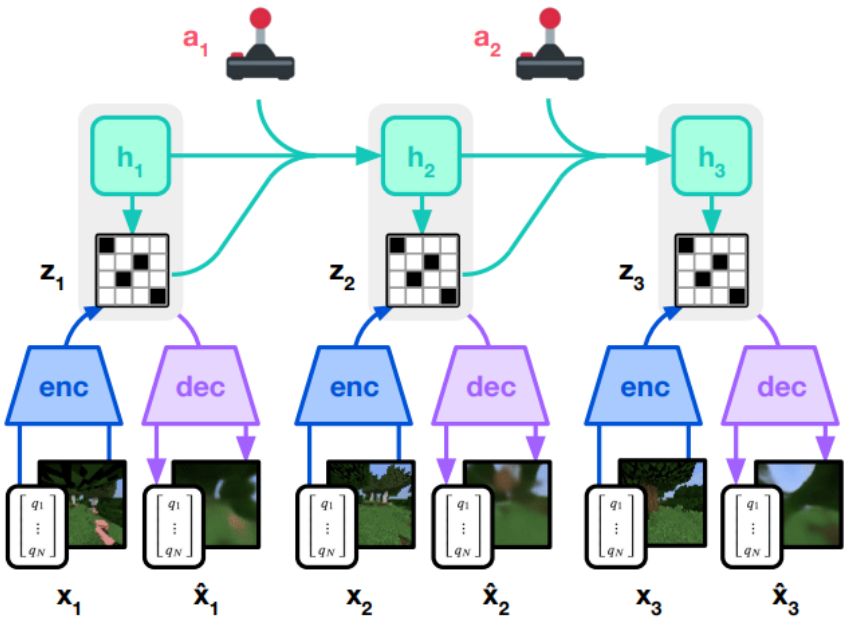
TD-MPC2



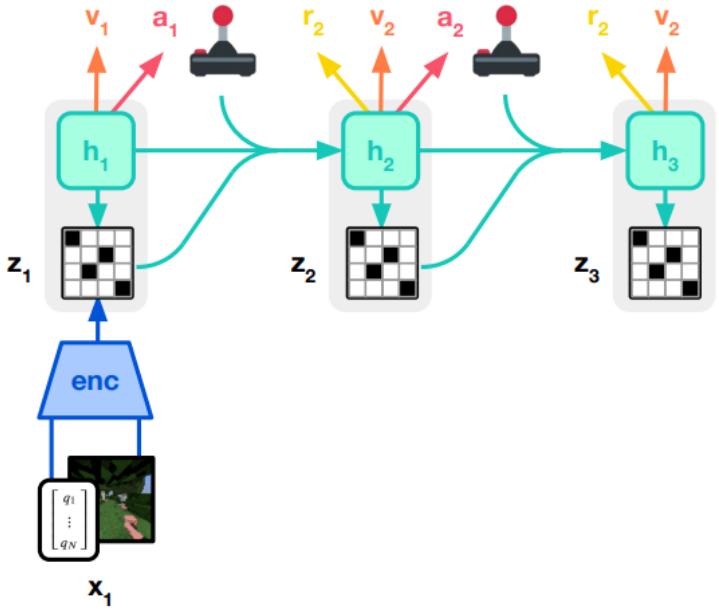
Encoder	$\mathbf{z} = h(\mathbf{s}, \mathbf{e})$	▷ Maps observations to their latent representations
Latent dynamics	$\mathbf{z}' = d(\mathbf{z}, \mathbf{a}, \mathbf{e})$	▷ Models (latent) forward dynamics
Reward	$\hat{r} = R(\mathbf{z}, \mathbf{a}, \mathbf{e})$	▷ Predicts reward r of a transition
Terminal value	$\hat{q} = Q(\mathbf{z}, \mathbf{a}, \mathbf{e})$	▷ Predicts discounted sum of rewards (return)
Policy prior	$\hat{\mathbf{a}} = p(\mathbf{z}, \mathbf{e})$	▷ Predicts action $\mathbf{a}^*$ that maximizes Q

Рис.: Архитектура TD-MPC2

DreamerV3



(a) World Model Learning



(b) Actor Critic Learning

$$
\text{RSSM}\begin{cases}
\text{Sequence model:} & h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1}) \\
\text{Encoder:} & z_t \sim q_\phi(z_t \mid h_t, x_t) \\
\text{Dynamics predictor:} & \hat{z}_t \sim p_\phi(\hat{z}_t \mid h_t) \\
\text{Reward predictor:} & \hat{r}_t \sim p_\phi(\hat{r}_t \mid h_t, z_t) \\
\text{Continue predictor:} & \hat{c}_t \sim p_\phi(\hat{c}_t \mid h_t, z_t) \\
\text{Decoder:} & \hat{x}_t \sim p_\phi(\hat{x}_t \mid h_t, z_t)
\end{cases}
$$

Рис.: Процесс обучения DreamerV3

EfficientZeroV2

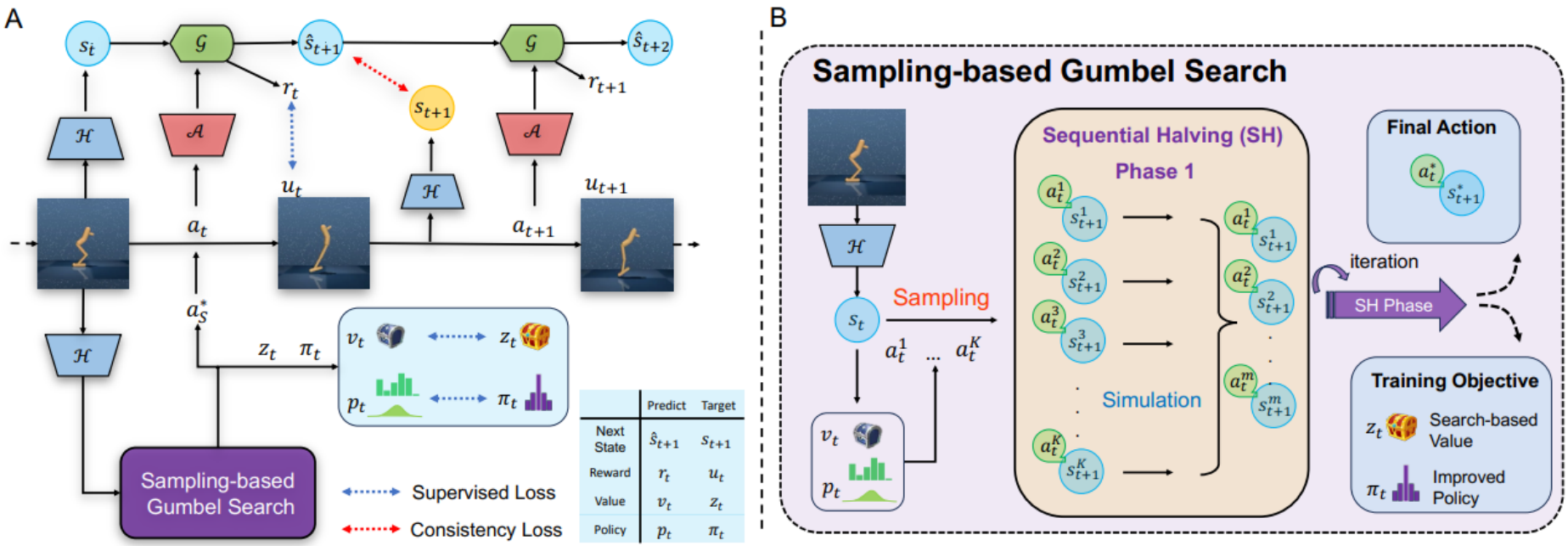
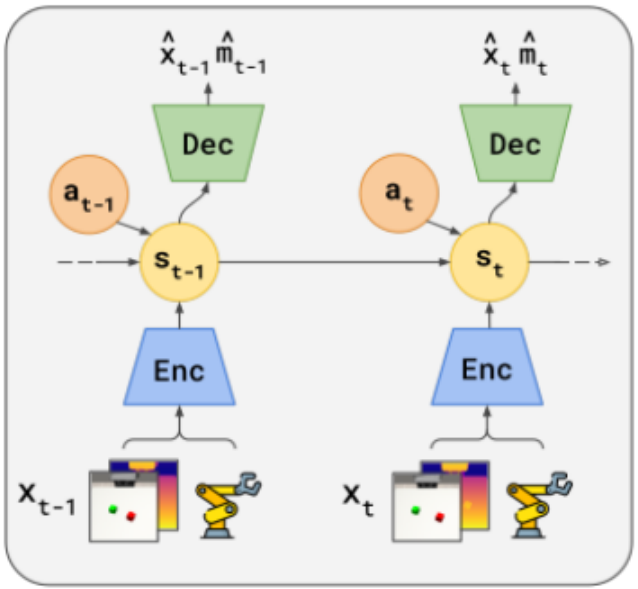
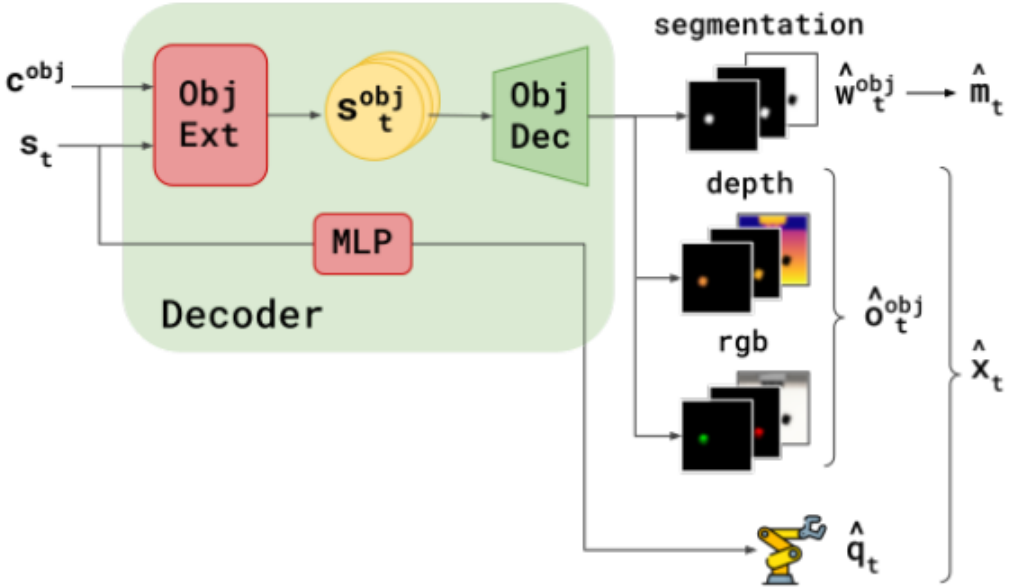


Рис.: Схема работы EZ-V2

FOCUS





$$\text{Encoder:} \quad e_t = f_\phi(x_t),$$
$$\text{Posterior:} \quad p_\phi(s_{t+1}|s_t, a_t, e_{t+1}),$$
$$\text{Prior:} \quad p_\phi(s_{t+1}|s_t, a_t),$$

$$\text{Proprio decoder:} \quad p_\theta(\hat{q}_t|s_t),$$
$$\text{Object latent extractor:} \quad p_\theta(s_t^{obj}|s_t, c^{obj}),$$
$$\text{Object decoder:} \quad p_\theta(\hat{o}_t^{obj}, w_t^{obj}|s_t^{obj})$$

Рис.: Схема работы FOCUS

Выводы

Center for Cognitive Modeling

- Выбор оптимального подхода к представлению латентного пространства зависит от конкретной задачи и ее особенностей
- Дальнейшие исследования в этой области направлены на разработку более устойчивых и универсальных методов создания векторных представлений и работающих с ними архитектур, которые смогут эффективно работать в различных условиях и обеспечивать высокую обобщающую способность алгоритмов обучения с подкреплением.

Спасибо за внимание!

cogmodel.mipt.ru

rodionvahitoff@yandex.ru