Continual Reinforcement Learning

JINYUAN SUN

Introduction

Agent Modeling

Implementation Methods

Imitations & Challenges

Introduction

Agent Modeling

Implementation Methods

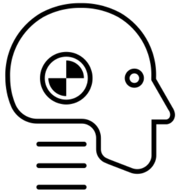
Imitations & Challenges

Goals of Never-Ending RL

- **long time goal:** to develop a machine to be able to learn and adapt to new scenarios over the duration of their lifetime.

- To long-term knowledge
- To adapt from previous skills

- To **Remember**
- To **Learn**



AI

Human

Constrains of Traditional RL

- computational expensive
- hard to train (convergence)
- hard to design a reward function
- hard to simulate the entire world
- unable to explore the world (touch, feel, smell, see)

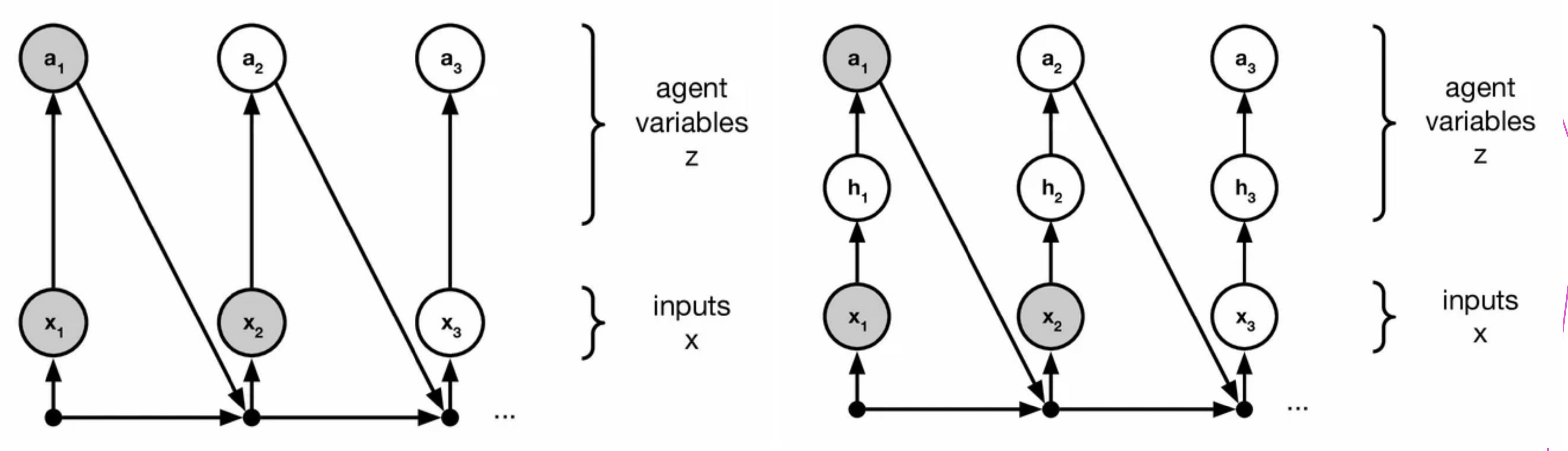
Introductions

Agent Modeling

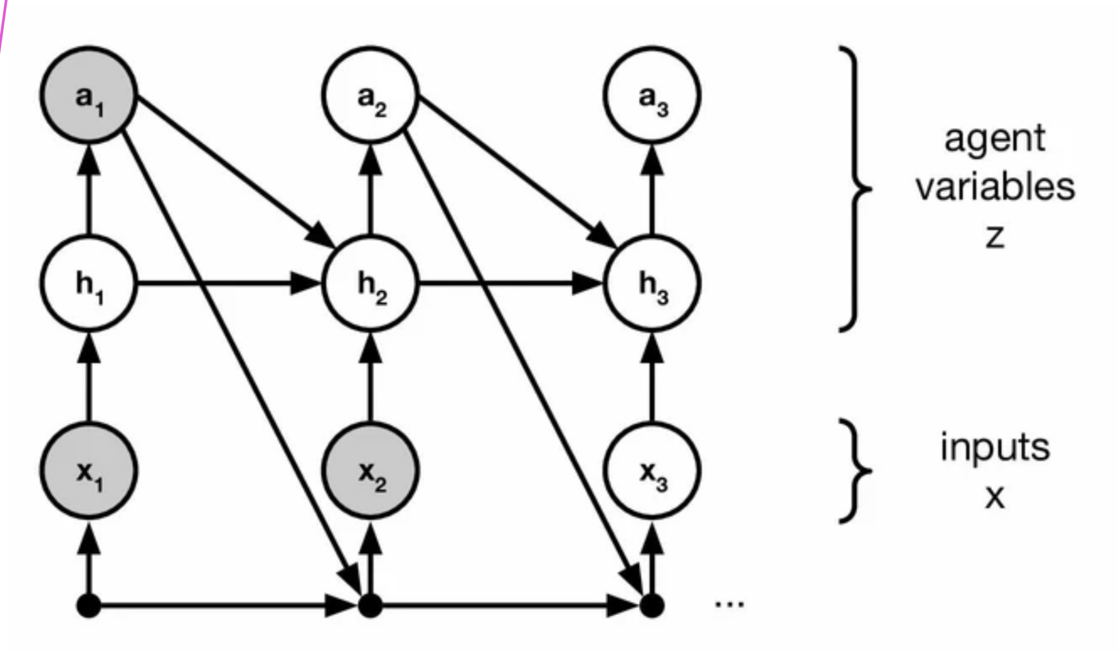
Implementation Methods

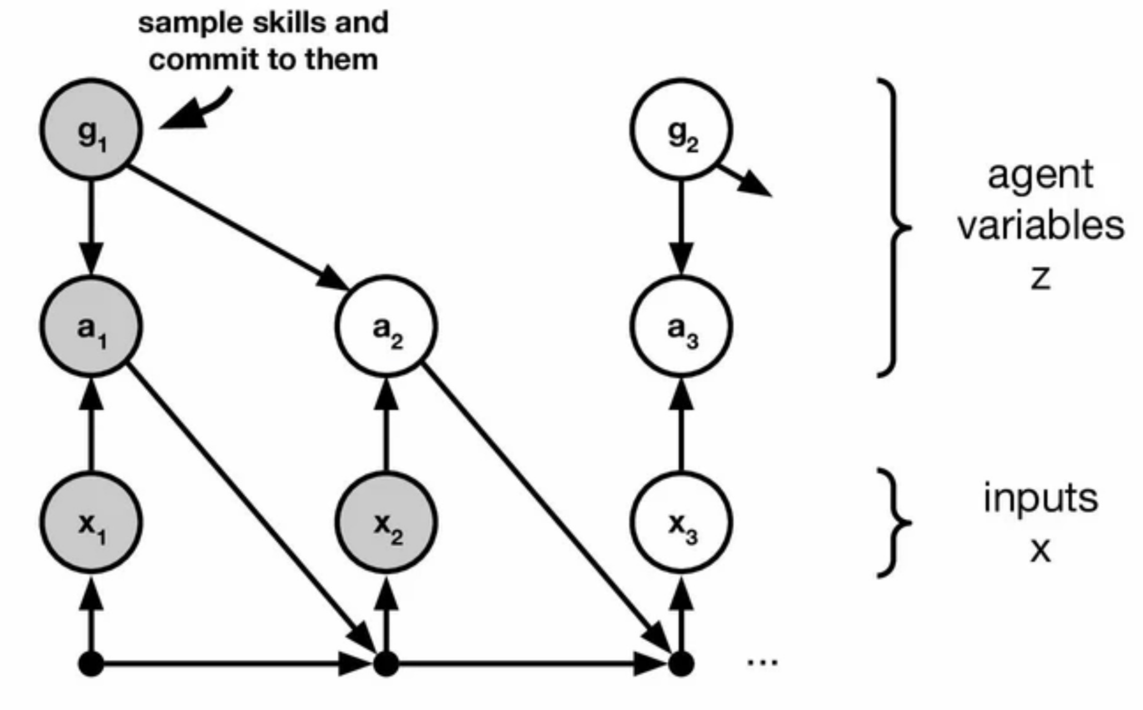
Imitations & Challenges

Stochastic Action Agent
Stochastic Representation Agent

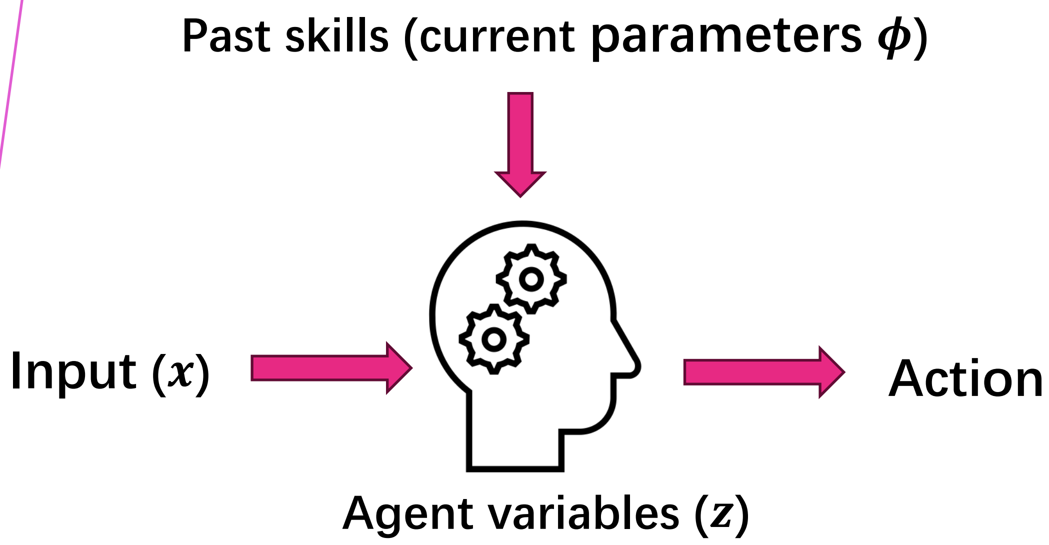


Stochastic Memory Agent
Stochastic Skill Agent



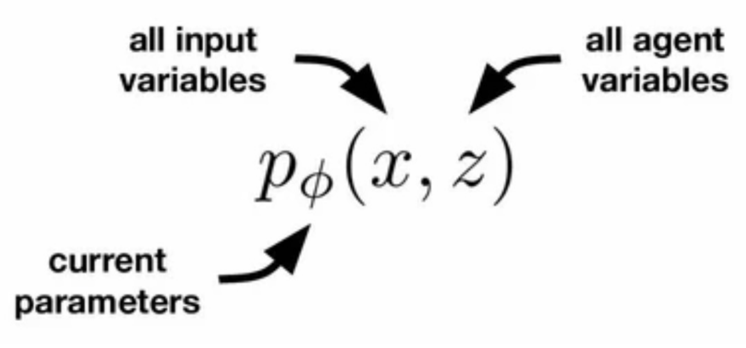


Agent Modeling



Modeling:

Implement agent as trying to follow a desired target distribution



Objectives:

$$\min_{\phi} \mathrm{KL}\left[\, \underbrace{p_\phi(x, z)}_{\text{actual}} \;\|\; \underbrace{\tau(x, z)}_{\text{target}} \,\right]$$

Correlation is all agents need!

FACTORIZED TARGETS

- Input and agent variables are **independent** under the target
- Actions are **task-dependent**

Task specific & Narrow

✕

EXPRESSIVE TARGETS

- Input and agent variables are **correlated** under the target
- Agent try to learn and explore the world **as a whole.**

Adaptive & General

✓

Maximize mutual information

Objectives of agents with expressive targets

$$\mathrm{I}\left[z;x\right] = \underbrace{\mathrm{I}\left[z;x_<\right]}_{\text{past infomax}} + \underbrace{\mathrm{I}\left[z;x_> \mid x_<\right]}_{\text{future infomax}}$$

- Self supervised representation learning
- Self supervised exploration
- Self supervised controllability

Policy Information Capacity(PIC)

$$\mathcal{I}(R;\Theta) = \mathcal{H}(R) - \mathbb{E}_{p(\theta)}\left[\mathcal{H}(R|\Theta=\theta)\right]$$

- *p(Θ)*: prior distribution of the policy parameter (including network architecture)
- Measure how controllable rewards are through parameter sampling
- PIC can be interpreted as *Reward Empowerment*.

Policy-optimal information capacity (POIC)

$$\mathcal{I}(\mathcal{O};\Theta) = \mathcal{H}(\mathcal{O}) - \mathbb{E}_{p(\theta)}\left[\mathcal{H}(\mathcal{O}|\Theta=\theta)\right]$$

- $p(\mathcal{O}=1|\tau) = \exp\left((r - r_{\max})/\eta\right)$
- Optimality variable, $O \in \{0, 1\}$, represents the optimality of trajectory [Levine 2018].
- POIC can be interpreted as *Optimality Empowerment*.

Furuta, H., Matsushima, T., Kozuno, T., Matsuo, Y., Levine, S., Nachum, O., & Gu, S. S. (2021, July). Policy information capacity: Information-theoretic measure for task complexity in deep reinforcement learning. In *International Conference on Machine Learning* (pp. 3541-3552). PMLR.

Introductions

Agent Modeling

Implementation Methods

Imitations & Challenges

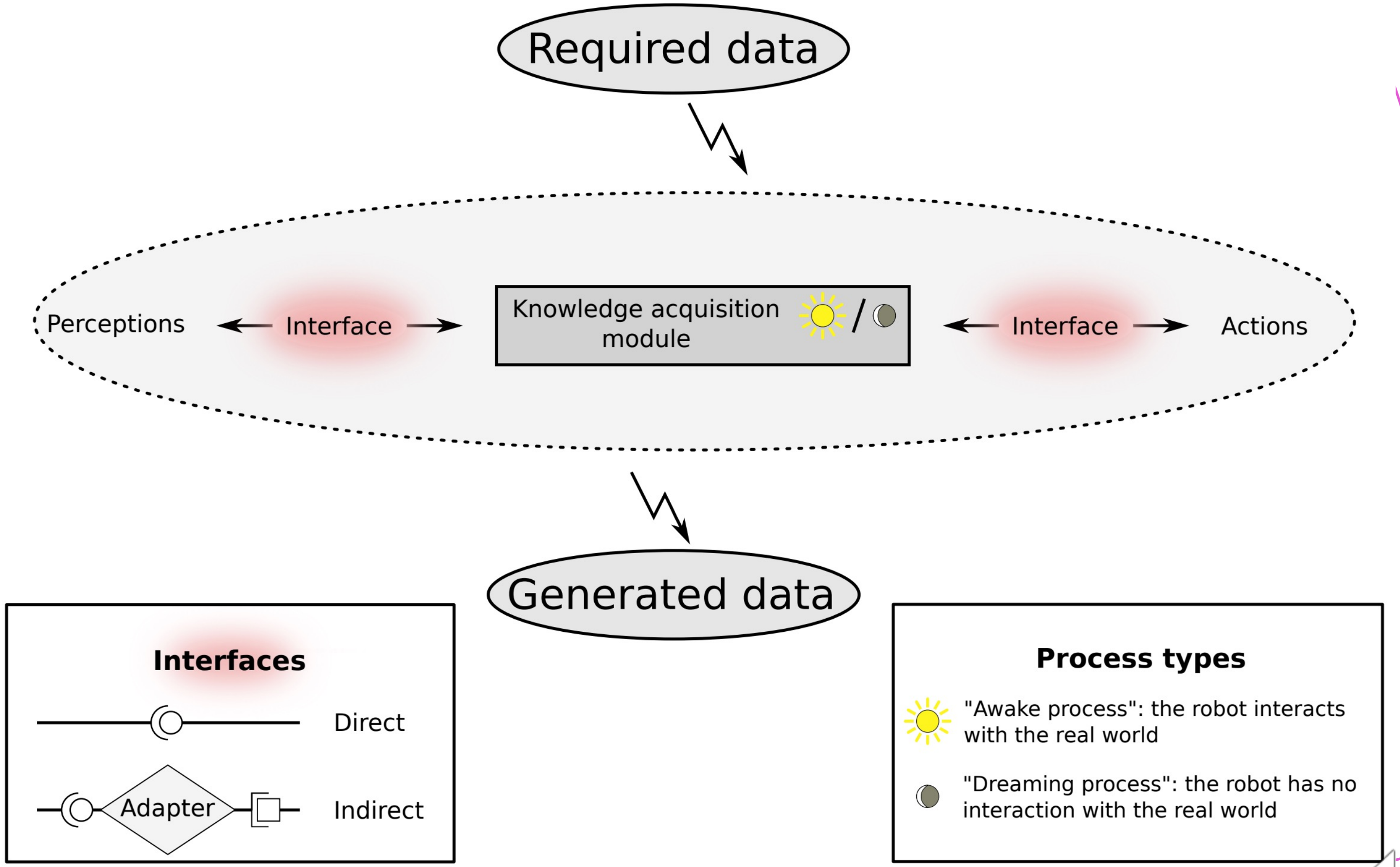
Implementation Methods

- Dreaming Based
- Diversity Based

Dreaming Process

Two Parts:

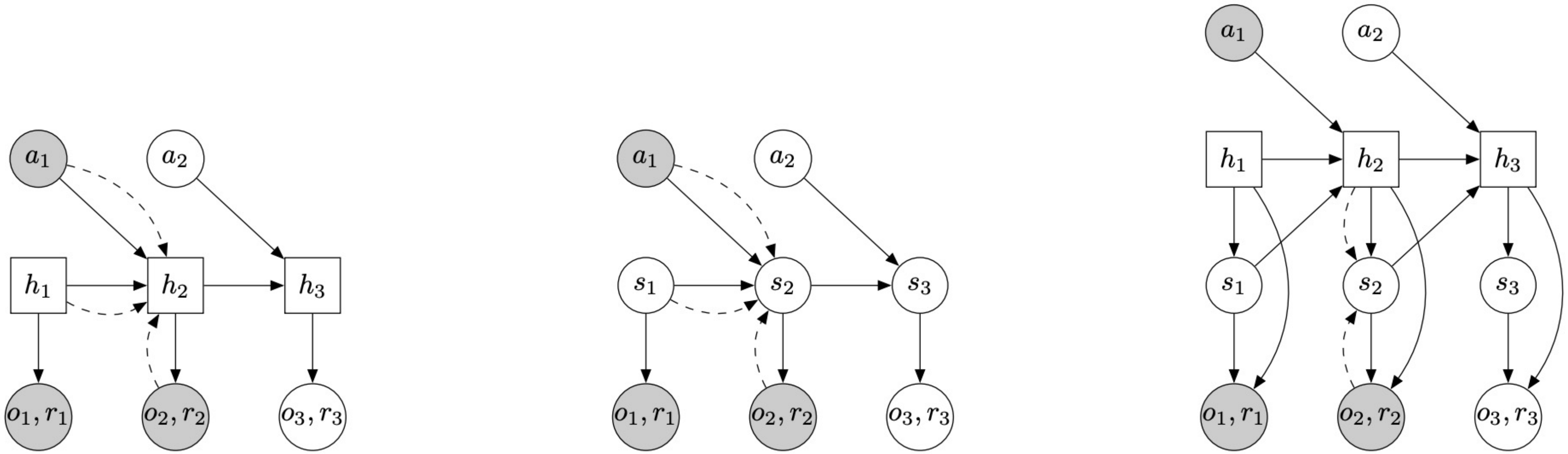
- Interact and learning from the real data
- Reflecting and predicting without the input data



Doncieux, S., Bredeche, N., Goff, L. L., Girard, B., Coninx, A., Sigaud, O., ... & Duro, R. (2020). DREAM architecture: a developmental approach to open-ended learning in robotics. *arXiv preprint arXiv:2005.06223*.

PlaNet: Learning World Model from Pixels

- Predict forward using compact latent states
- Using a combination of stochastic and deterministic process to robustly predict multiple futures



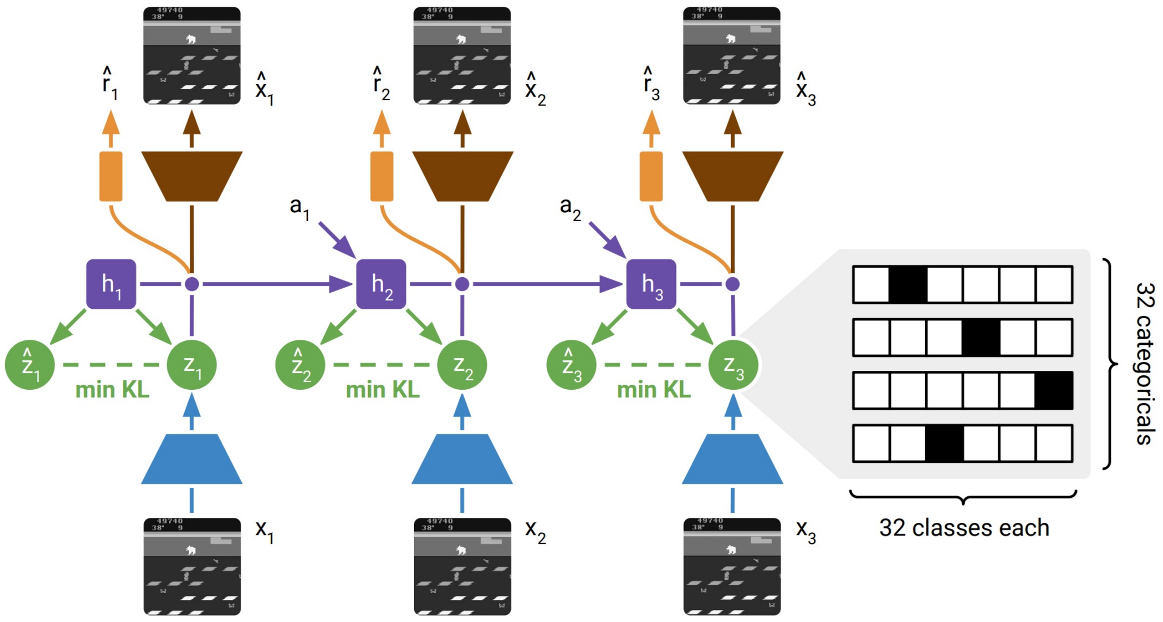
(a) Deterministic model (RNN) (b) Stochastic model (SSM) (c) Recurrent state-space model (RSSM)

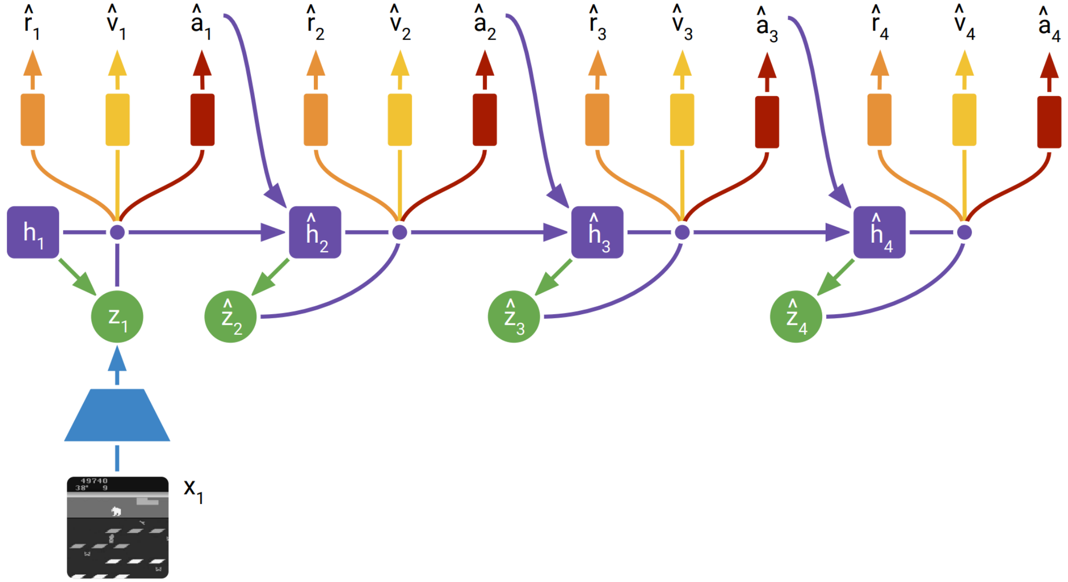
Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H. & Davidson, J.. (2019). Learning Latent Dynamics for Planning from Pixels. <i>Proceedings of the 36th International Conference on Machine Learning</i>, in <i>Proceedings of Machine Learning Research</i> 97:2555-2565 Available from https://proceedings.mlr.press/v97/hafner19a.html.

Dreamer V2

Achieves human-level performance on the Atari benchmark.

Improvements:
- Vectors of categorical
- KL balancing



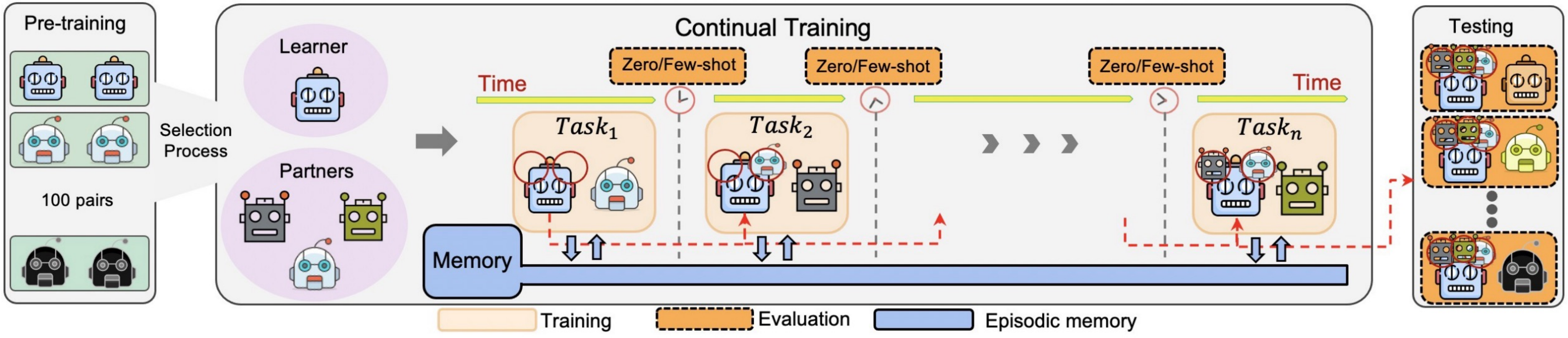


Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D., Lee, K. H., Fischer, I., & Abbeel, P. (2022). Deep hierarchical planning from pixels. *Advances in Neural Information Processing Systems*, *35*, 26091-26104.

Training with diverse agents

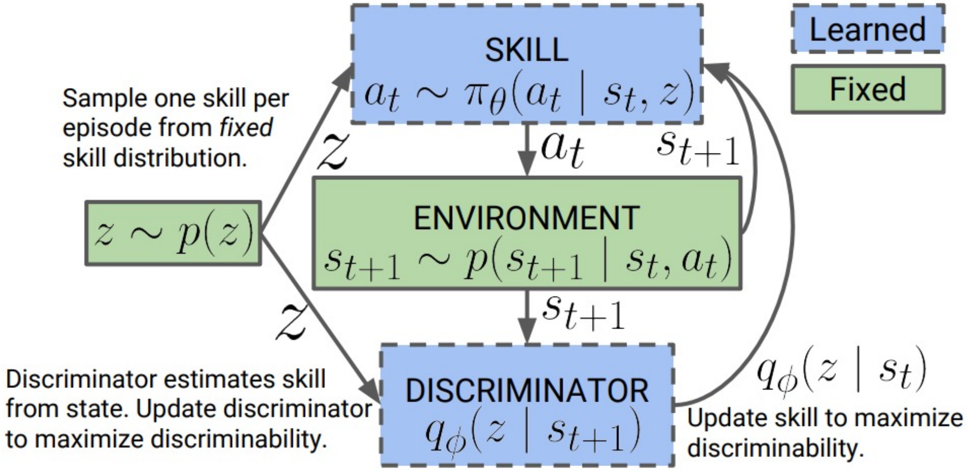
Continuous Coordination As a Realistic Scenario for Lifelong Learning



Nekoei, H., Badrinaaraayanan, A., Courville, A., & Chandar, S. (2021, July). Continuous coordination as a realistic scenario for lifelong learning. In *International Conference on Machine Learning* (pp. 8016-8024). PMLR.

Diversity is all you need

Learning skills without a reward function: DIAYN Algorithm



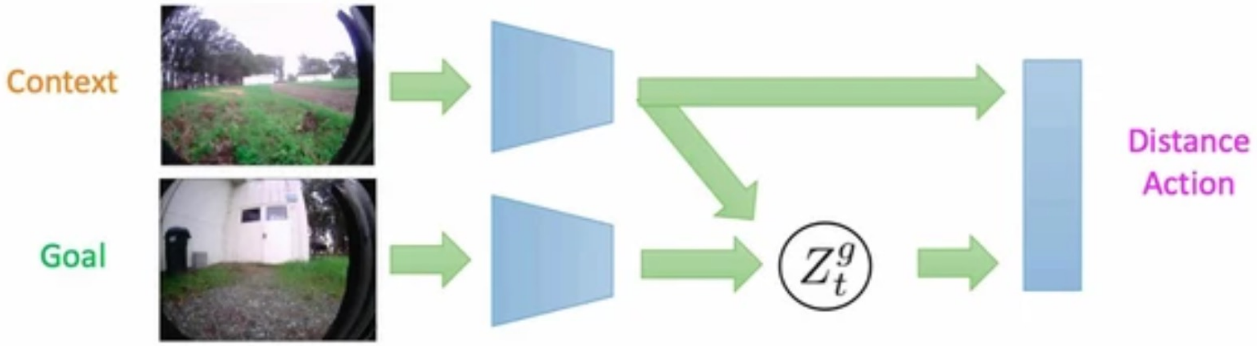
Algorithm 1: DIAYN

while *not converged* **do**
  Sample skill $z \sim p(z)$ and initial state $s_0 \sim p_0(s)$
  for $t \leftarrow 1$ **to** *steps_per_episode* **do**
    Sample action $a_t \sim \pi_\theta(a_t \mid s_t, z)$ from skill.
    Step environment: $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$.
    Compute $q_\phi(z \mid s_{t+1})$ with discriminator.
    Set skill reward $r_t = \log q_\phi(z \mid s_{t+1}) - \log p(z)$
    Update policy (θ) to maximize r_t with SAC.
    Update discriminator (ϕ) with SGD.

- the discriminator is updated to better predict the skill,
- the skill is updated to visit diverse states that make it more discriminable

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Rapid Exploration (RECON)



- a **learned latent variable model** of distances and actions, along with a non-parametric topological memory of images

Shah, D., Eysenbach, B., Rhinehart, N., & Levine, S. (2021). Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*.

Challenges & Limitations

- **Representations:** how to represent the world (formatting environments)
- **Memories:** how to preserve an already learned skill
- **Explorations:** how to automatically learn new skills (adaptive & general)

Continual Reinforcement Learning

JINYUAN SUN