Self-supervised Reinforcement Learning with Independently Controllable Subgoals

Andrii Zadaianchuk[1,2], Georg Martius[1], Fanny Yang[2]

[1]Max Planck Institute for Intelligent Systems [2]ETH Zürich

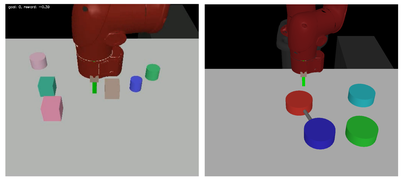


Abstract

- Self-supervised agents can learn manipulation skills in **multi-object environments**.
- Previous methods do not take the **dependencies between objects** into account.
- We propose to **estimate relations between objects** and use them to **independently control different objects**.
- Estimated relations between objects can be used to decompose a complex goal into **a compatible sequence of subgoals**.
- Our **SRICS agent** can efficiently and automatically learn manipulation tasks in multi-object environments with different relations between objects.

Object Manipulation in Multi-objects Environments

Multi-object Rearrange and Relational Rearrange environments

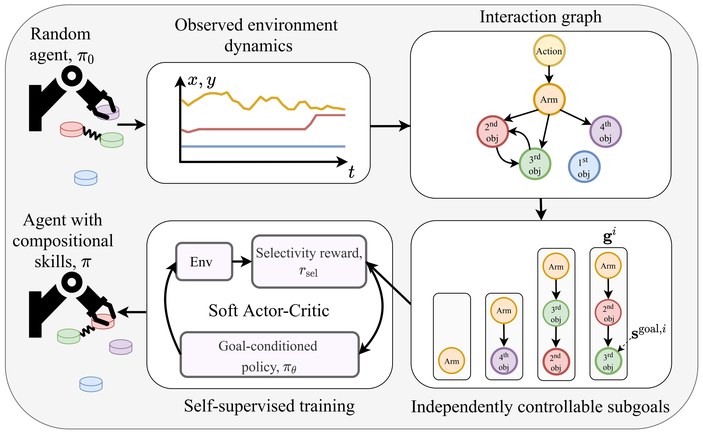


- Random initial state position and final goal position for each object
- Action space: robot end effector position
- In Relational Rearrange some objects can be connected (by spring connection) or static

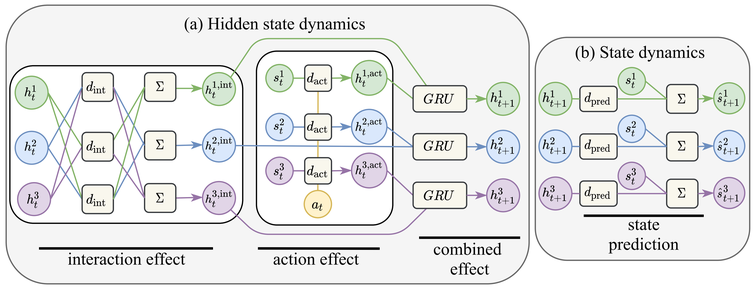
Object-centric Representations

- Scene is encoded as a set of vectors
- All the entities (including agent) are encoded with the same format
- Similar representation could be learned fully-unsupervised from high-dimensional observations

SRICS



Estimation of the Interaction Graph



- The interaction function d_{int} is further factorized into two components:

$$d_{\text{int}}(\mathbf{h}_t^i, \mathbf{h}_t^j) = w_t^{ij} \cdot d_{\text{int-eff}}(\mathbf{h}_t^i, \mathbf{h}_t^j); \quad q\left(w_t^{ij} \mid \mathbf{s}_t\right) = \text{softmax}\left(d_{\text{int-pres}}(\mathbf{h}_t^i, \mathbf{h}_t^j)\right).$$

- Training of the dynamical model using ELBO loss:

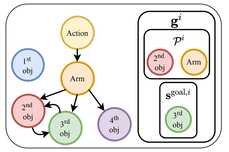
$$\mathcal{L} = \sum_{j=1}^{K} \sum_{t=1}^{T-1} \frac{\left\| \mathbf{s}_{t+1}^{j,\text{where}} - \hat{\mathbf{s}}_{t+1}^{j,\text{where}} \right\|^2}{2\sigma^2} + D_{\text{KL}}(q \,||\, p_{\text{prior}}).$$

- Local interaction weights w_t^{ij} are sparse due to sparsity prior p_{prior}.
- Global interaction graph G: aggregated over dataset and thresholded local interaction weights.

Independently Controllable Subgoals

$$\mathbf{g}^i = \left(\mathbf{s}^{\text{goal},i}, \mathcal{P}^i\right)$$

- $\mathcal{P}^i$ are the nodes that lie in a path from the action A and object i
- $\mathcal{P}^i$ correspond to the objects that could be used to control object i



Goal-directed Selectivity Reward

$$r_{\text{sel},i}\left(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{g}^i\right) = r_{\text{goal}} + \alpha\left(\text{sel}_i(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathcal{P}^i) - 1\right)$$

$$r_{\text{goal}} = -\left\|\mathbf{s}_t^i - \mathbf{s}^{\text{goal},i}\right\|$$

$$\text{sel}_i(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathcal{P}^i) = \begin{cases} \frac{\|\mathbf{s}_t^i - \mathbf{s}_{t-1}^i\|}{\sum_{j \notin \mathcal{P}^i} \|\mathbf{s}_t^j - \mathbf{s}_{t-1}^j\|}, & \text{if subgoal is not solved;} \\ 1 - \sum_{j \notin \mathcal{P}^i \cup \{i\}} \|\mathbf{s}_t^j - \mathbf{s}_{t-1}^j\|, & \text{otherwise.} \end{cases}$$

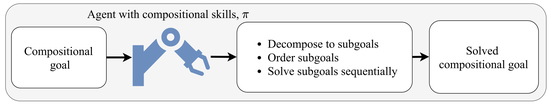
- Motivates agent to solve subgoal i without destroying subgoals in $\mathcal{P}^i$.

SRICS Training

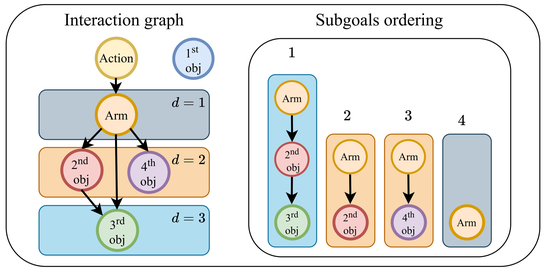
Given: GNN dynamical model D, goal-conditional attention policy π_θ, goal-conditional SAC trainer, number of training episodes K.

Train GNN dynamical model D on sequences from $\mathcal{D}$ using the ELBO loss and estimate the interaction graph G.
for $n = 1, ..., K$ episodes **do**
  Sample goal $\mathbf{s}^{\text{goal}}$ and construct subgoal $\mathbf{g}^i$ using G.
  Collect episode data with policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t, \mathbf{g}^i)$.
  Store transitions $\{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}, \mathbf{g}^i), \ldots\}$ into replay buffer $\mathcal{R}$.
  Sample transitions from replay buffer $(\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{g}^i) \sim \mathcal{R}$.
  Relabel $\mathbf{g}^i$ goal components to a combination of future states and goal sampling distribution.
  Compute selectivity reward signal $R = r_{\text{sel},i}(\mathbf{s}', \mathbf{s}, \mathbf{g}^i)$.
  Update policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t, \mathbf{g}^i)$ using R with SAC trainer.
end for

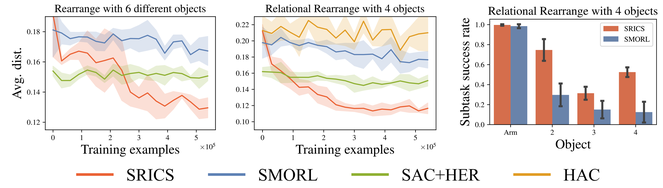
SRICS Evaluation



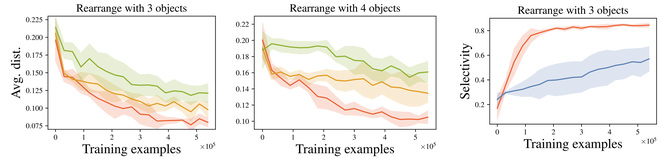
Subgoals ordering during evaluation



Comparative Analysis



Ablative Analysis



References

[1] Andrew Levy, George Dimitri Konidaris, Robert Platt Jr., and Kate Saenko. Learning multi-level hierarchies with hindsight. In *ICLR*, 2019.

[2] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Self-supervised visual reinforcement learning with object-centric representations. In *ICLR*, 2021.