ICRA 2022 Workshop
Reinforcement Learning for Contact-Rich Manipulation Workshop

Learning Dense Reward with Temporal Variant Self-Supervision

Yuning Wu, Jieliang Luo, Hui Li

May 27, Philadelphia USA

Carnegie Mellon University
AUTODESK

Background and Challenge

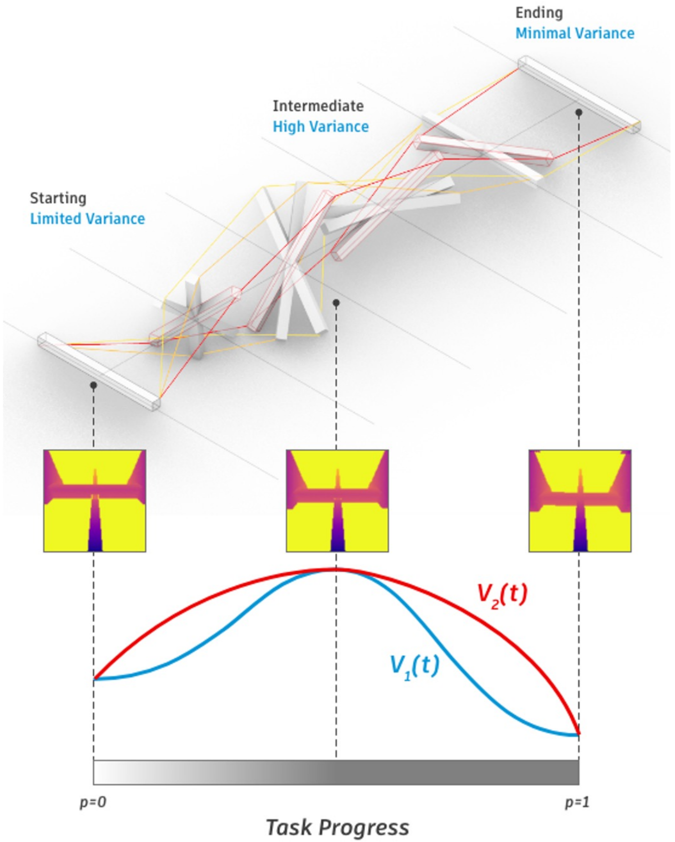
- Rewards play an essential role in reinforcement learning.
- In contrast to rule-based game environments with well-defined reward functions, real-world robotic applications, such as contact-rich manipulation, lack explicit reward.
- Previous effort has shown that it is possible to algorithmically **extract dense rewards directly from multimodal observations**.
- In this paper, we aim to extend this effort by **proposing a more efficient and robust way of sampling and learning**.

Core Idea

- Similar to method proposed in [1] by Wu et al, we aim to extract a ***task progress variable.***

$$p \in [0,\ 1]$$

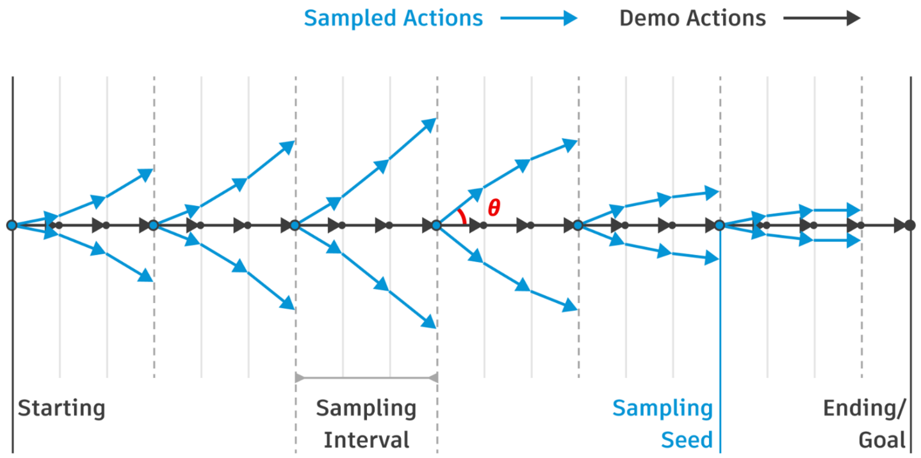
- It is extracted from multimodal sensory data (camera images, force/torque), by self-supervised learning.
- We use ***p*** as a dense reward to guide reinforcement learning in contact-rich manipulation tasks.



Our Approach

1. Temporal Variant Forward Sampling (TVFS)

- We aim to **sample a tree of multimodal observations** from **an expert demonstration** with a **physical simulator** for self-supervised learning.
- The sampling is controlled with temporal variance such that,
    - It captures common patterns of manipulation tasks.
    - Sampled actions do not diverge too much from the potential distribution of an expert demonstration.
    - Sampled actions are mostly progressing forward.



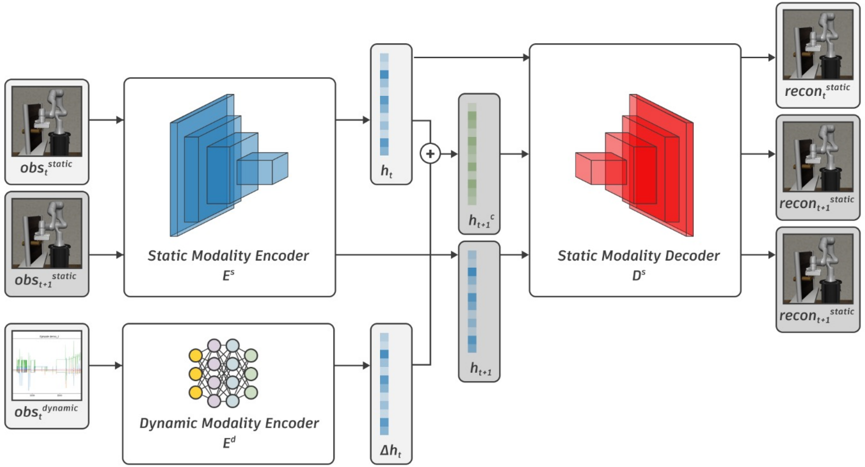
Our Approach

2. Self-Supervised Representation Learning

- Task progress is structured with distance measure ***d*** in a latent space ***H***

$$p = 1 - \frac{d(h_\phi(s), h_\phi(s_g))}{d(h_\phi(s_0), h_\phi(s_g))}$$

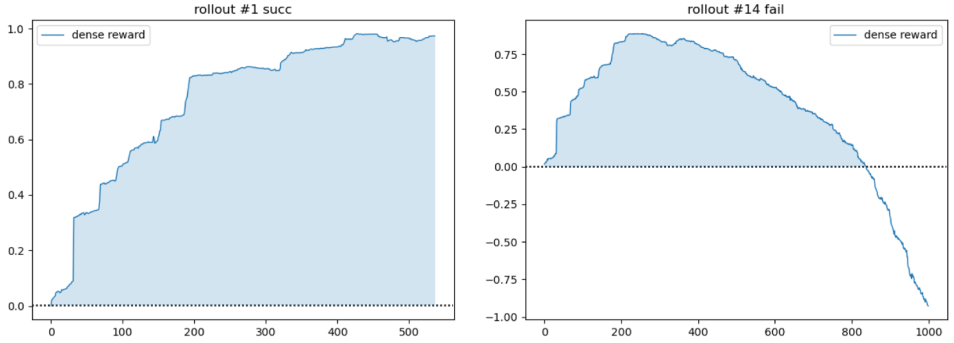
- Prior work propose to learn the representation through explicitly enforcing temporal order through a triplet loss function.
- We propose a novel architecture to learn representation by **utilizing dynamic relation among pairs of adjacent observations.**



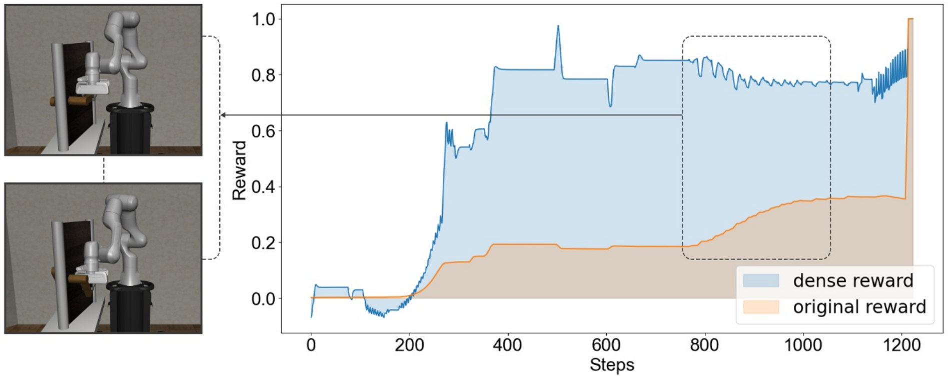
$$h_\phi(s_t) + \Delta h_\psi(s_t) = h_\phi(s_{t+1})$$

Experiments | Validation

- We visualize and validate our dense reward with a successful trajectory (*upper left*) and a failed trajectory (*upper right*).

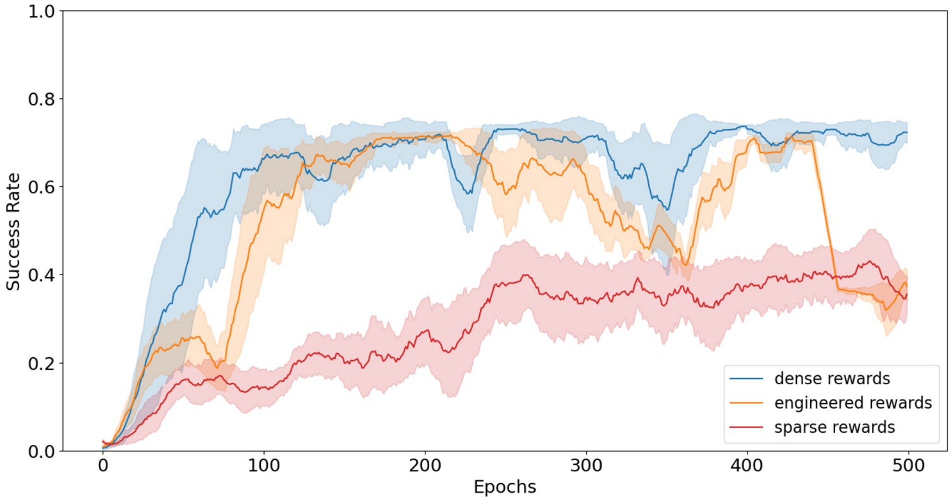


- We examine the case of an inexpert demonstration in door-opening (*bottom*). Our dense reward provide more feedback than distance reward in "plateau" trial stages.



Experiments | Benchmark

- We have chosen the door-opening task, and SAC [2] as the RL algorithm for benchmarking.
- We compared three types of rewards,
  - **our dense reward**,
  - **hand-crafted distance reward**
  - **sparse binary reward**.
- Preliminary results show that our dense reward leads to **faster convergence** and **more training stability**.



Conclusion and Future Work

- We propose an improved framework for learning dense reward for contact-rich manipulation tasks.
- For future work, we intend to conduct more ablation studies regarding the framework's adaptability and modalities.
- We are also curious about the framework’s performance in tasks with non-deterministic goal states.

Reference

[1] Wu, Zheng et al. “Learning Dense Rewards for Contact-Rich Manipulation Tasks.” *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021): 6214-6221.

[2] Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., ... & Levine, S. (2018). Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

ICRA 2022 Workshop
Reinforcement Learning for Contact-Rich Manipulation Workshop

Learning Dense Reward with Temporal Variant Self-Supervision

Yuning Wu, Jieliang Luo, Hui Li

May 27, Philadelphia USA

Carnegie Mellon University

AUTODESK