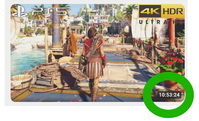
Towards Principled Representation Learning of Videos for Reinforcement Learning

Dipendra Misra*, Akanksha Saran*, Tengyang Xie, Alex Lamb, and John Langford

Microsoft Research
Sony Research

Foundation Models for Decision Making

- **Foundation models** trained on **abundant** text and image data have changed the ML landscape.
- How can we train AI agents to learn useful behavior for many tasks across domains such as robotics, games, software UI?



Challenges

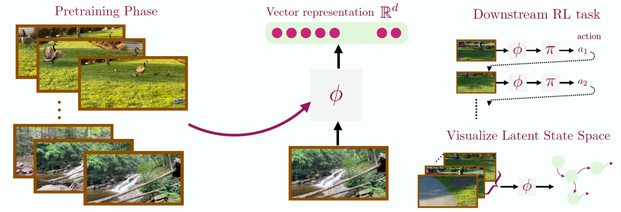
- Observations are **complex** requiring more samples
- Real-world samples are **expensive**.
- Typical offline data contains actions (and rewards) and is **not naturally**

Video data is naturally available in abundance for many domains such as gaming and software!

Problem Setup

Goal: Learn a representation model using readily available videos and use it to do RL in a downstream task.

Challenge: Videos might be about tasks **unrelated** to the downstream task. Cannot perform imitation learning.

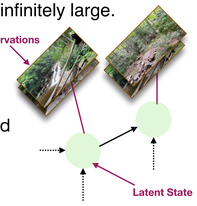


Block MDP

- Observation is generated from a finite latent state with iid noise.
- Observation space can be infinitely large.

Ex-Block MDP

- Extend Block MDP by adding temporally-correlated (exogenous) noise.



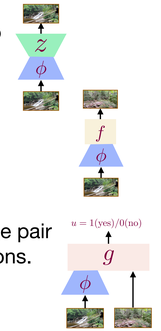
Research Questions

- Can one learn a representation that captures the latent state using just offline video data?
- Can such a representation be used to do efficient RL that outperforms doing RL from scratch?

Existing Representation Learning Algorithms

We study three existing representation learning methods for videos. **What do they learn**? **Do they reliably succeed in downstream RL**?

1. **Autoencoder:** Reconstruct an image x in the video from its representation $\phi(x)$.
$$\text{loss} = \|z(\phi(x)) - x\|_2^2$$
2. **Forward Modeling:** Reconstruct a future image x' from its representation $\phi(x)$.
$$\text{loss} = \ln f(x' \mid \phi(x))$$
3. **Temporal Contrastive Learning:** Predict if an image pair (x, x') is temporally causal using their representations.
$$\text{loss} = (g(\phi(x), x') - u)^2$$



Theoretical Results and Analysis

Video-based RL can provably succeed in Block MDP

Three assumptions for Block MDP:

1. **Coverage:** We have enough diverse data.
2. **Realizability:** Our models are expressive enough.
3. **Separability:** Information exists to separate states. E.g., for any two states, there is a state reachable from *only one* of them.

Theorem [Upper Bound]: Under assumptions 1-3, forward modeling and temporal contrastive

- Learn a representation ϕ that captures the underlying latent state.
- Can use ϕ to successfully do downstream RL
  Required RL episodes don't scale with the complexity of ϕ unlike RL-from-scratch.

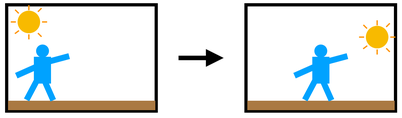
Downstream RL Sample complexity $n_{samp}(S, A, H, \varepsilon_o/2, \delta_o/4)$

No direct dependence on the complexity of function classes!

Video-based RL can fail with exogenous noise

Theorem [Lower Bound]: There exists an Ex-Block MDP such that no video-based representation learning can learn a compact representation that leads to efficient RL, no matter how many videos or RL episodes we use.

Intuition: From video data alone, we cannot distinguish between changes due to agent's action and changes due to exogenous noise.



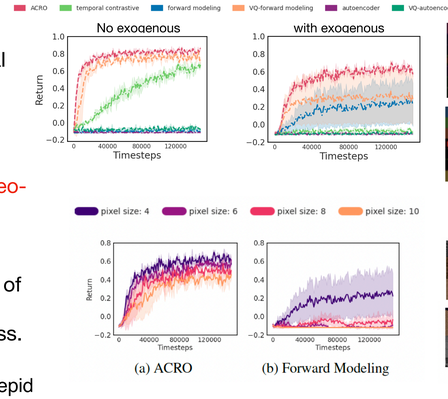
motion of blue figure: agent's action
movement of sun: exogenous noise

Experimental Results

Empirical results validate theory

- In the absence of exogenous noise both forward modeling and temporal contrastive representation learning work reliably **as predicted by our theory**.
- When there is exogenous noise, video-based methods suffer significantly **(as predicted by our theory)**, in particular, temporal contrastive learning. Whereas, the performance of ACRO — a trajectory-based representation method suffers far less.

Code: https://github.com/microsoft/Intrepid



Visualization of Learned Representations