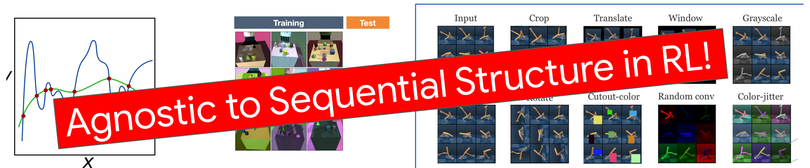
Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning

Rishabh Agarwal, Marlos C. Machado, Pablo S. Castro, Marc G. Bellemare

Motivation

- Agents should "do well" in environments *semantically* similar to training environments.
- Train agents that can generalize from a "few" environments rather than hundreds or thousands of environments.

Prior Work on Generalization



Regularization (ℓ2-reg., Dropout, Noise Injection)

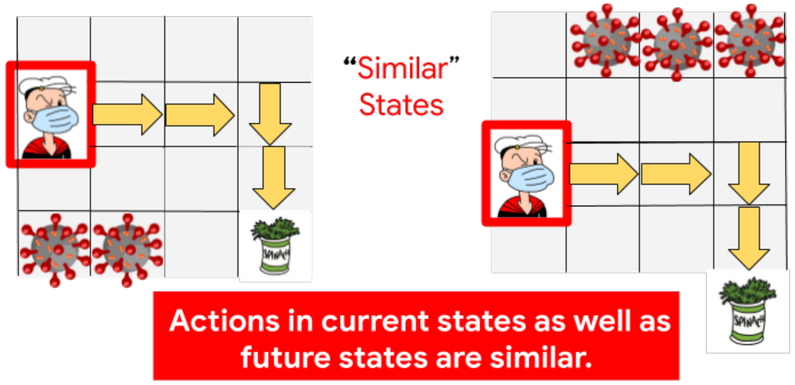
Domain Randomization

Data Augmentation (RandConv, RAD, DrQ ..)

Existing approaches are typically adapted from supervised learning and revolve around enhancing the learning process. These approaches rarely exploit properties of the sequential decision making problem such as similarity in actions across temporal observations

Behavioral Similarity for Generalization

- Exploit the sequential structure in RL to improve generalization.
- Learn state representations (PSEs) that encode "behavioral similarity" across states!



The agent needs to obtain spinach while maintaining social distancing. The states highlighted by red are behaviourally similar as the actions in these states as well as future states are similar.

Policy Similarity Metric (PSM)

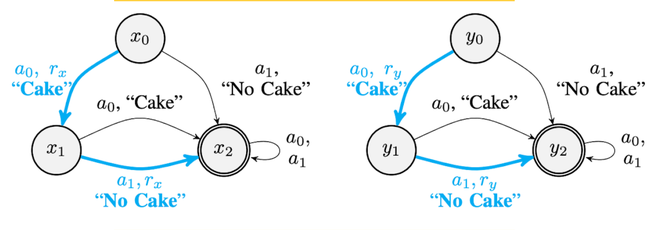
$$d^*(x, y) = \text{DIST}\big(\pi^*(x), \pi^*(y)\big) + \gamma \mathcal{W}_1(d^*)\big(P^{\pi^*}(\cdot \mid x), P^{\pi^*}(\cdot \mid y)\big)$$

Local Optimal Behavior Difference

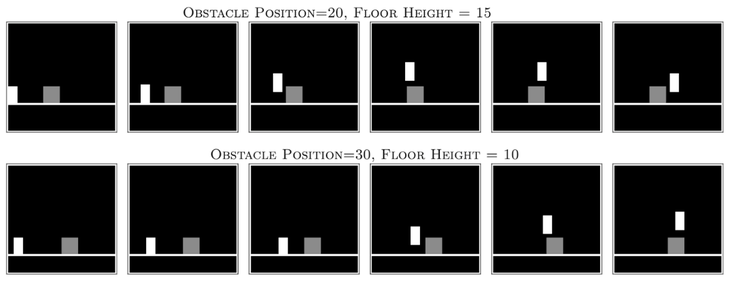
Long-term Optimal Behavior Difference

Policy Similarity vs Bisimulation (Reward Similarity)

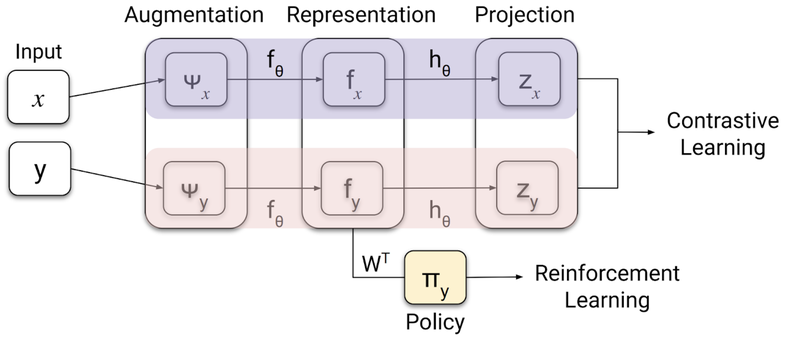
Similar Behavior, **Different** Rewards



Different Behavior, **Similar** Rewards



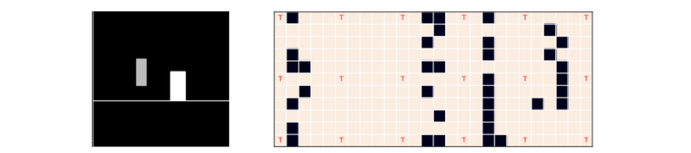
Policy Similarity Embeddings (PSEs)



agarwl.github.io/pse for details!

Jumping Task from Pixels

- PSEs significantly outperform state-of-the-art data augmentation (RandConv) and various bisimulation methods.

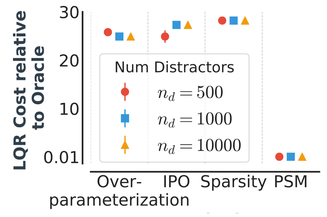


(a) Jumping task (b) "Wide" grid (c) "Narrow" grid (d) Random grid

Agents need to generalize to unseen obstacles (x-axis on grids below) and floor-heights (y-axis) from finite training tasks (T).

LQR With Spurious Correlations

- PSM ignores spurious information for generalization in a LQR task with non-image inputs.



Distracting DM Control

- Can agents can ignore visual distractors (random videos) irrelevant to the RL task?
- Shows scalability of PSEs without explicit access to optimal policies.