Unsupervised State Representation Learning in Atari

Ankesh Anand*, Evan Racah*, Sherjil Ozair*, Yoshua Bengio, Marc-Alexandre Côté, R Devon Hjelm
*Equal Contribution. {anandank, racaheva, ozairs}@mila.quebec

Overview

State representation learning, or the ability to capture **latent generative factors** of an environment, is crucial for building intelligent agents that can perform a wide variety of tasks. Learning such representations without supervision from rewards is a challenging open problem.

- We introduce **SpatioTemporal DeepInfoMax (ST-DIM)** which maximizes predictive mutual-information to learn high-level concepts in a scene:
  - without labels or rewards;
  - without modelling pixels directly.
- We also introduce the **Atari Annotated RAM Interface (AARI)**, which exposes the ground truth semantic information present in the RAM state. We use AARI to evaluate representations based on how well they capture the ground truth state variables.

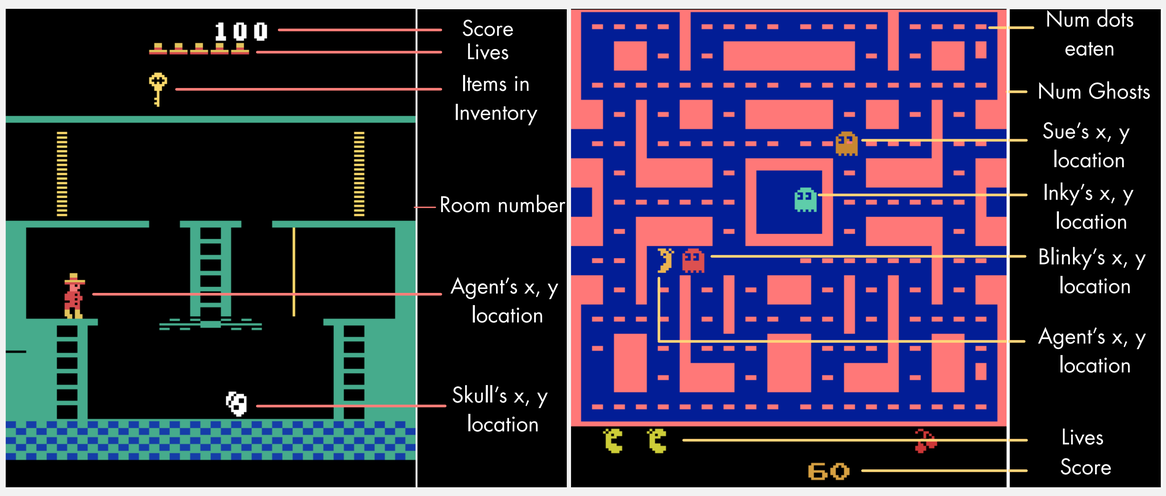
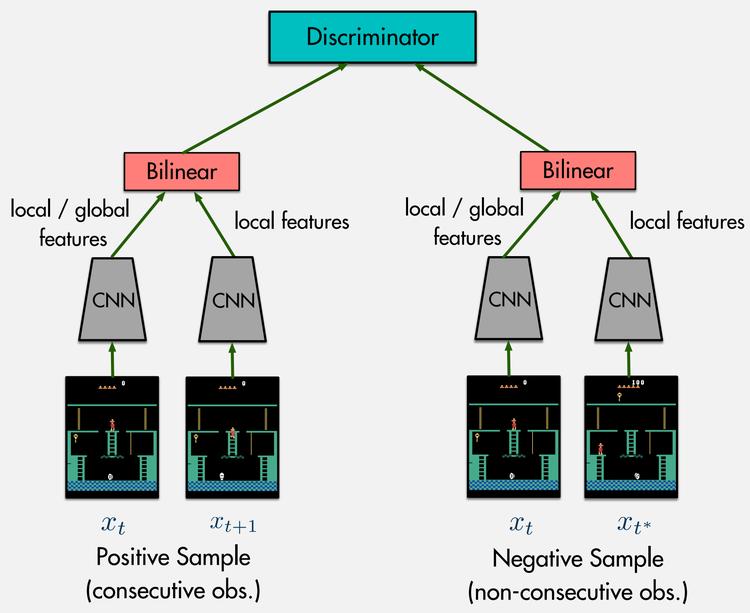


Figure 1: Ground truth semantic information present in the RAM state for Montezuma's Revenge (left) and MsPacman (right).

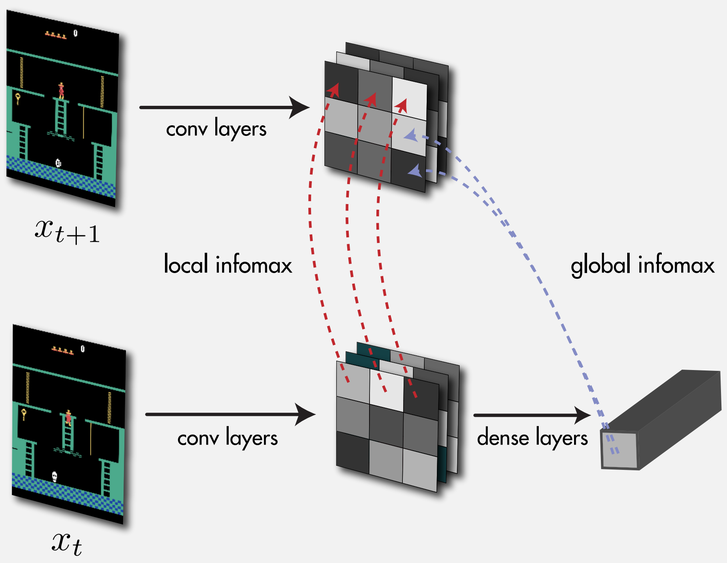
SpatioTemporal DeepInfoMax

Our method learns state representations by maximizing the mutual information between consecutive frame features across both spatial and temporal axes.

- To estimate the mutual information (MI), we setup a contrastive task that gives an InfoNCE bound on MI.



- Maximizing the MI across time can default to the encoder focusing only on an easily predictable feature (like the clock). Hence, we maximize the MI across each local feature.



$$\mathcal{I}_{NCE}(\{(x_t^k, x_{t+1}^k)\}_{k=1}^N) = \sum_{k=1}^{N} \log \frac{\exp f(x_t^k, x_{t+1}^k)}{\sum_{t^*=1}^{N} \exp f(x_t^k, x_{t^*})} \quad (1)$$

The Atari Annotated RAM Interface (AARI)

To systematically evaluate the ability of representation learning methods at capturing the true underlying factors of variation, we propose a benchmark based on Atari 2600 games using the Arcade Learning Environment (ALE).

- We identify important state variables from source code of 22 games.
- State variables include location of player, location of items of interest (keys, doors, etc.), non-player characters/enemies (see figure 1).
- Representations are evaluated using **linear probing**, i.e. the accuracy of linear classifiers trained to predict each latent generative factor from the learned representations.

  Gradients are never backpropagated through the encoder network.

Results

We consider two different modes for collecting the data:

- random agent (actions are picked uniformly at random);
- pretrained PPO agent (NB: results are in the paper).

Table 1: Probe F1 scores averaged across categories (data collected by random agents)

GAME	MAJ-CLF	RANDOM-CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
ASTEROIDS	0.28	0.34	0.36	0.34	0.42	**0.49**	N/A
BERZERK	0.18	0.43	0.45	**0.55**	**0.56**	0.53	0.68
BOWLING	0.33	0.48	0.50	0.81	0.90	**0.96**	0.95
BOXING	0.01	0.19	0.20	0.44	0.29	**0.58**	0.83
BREAKOUT	0.17	0.51	0.57	0.70	0.74	**0.88**	0.94
DEMONATTACK	0.16	0.26	0.25	0.32	0.57	**0.69**	0.83
FREEWAY	0.01	0.50	0.26	**0.81**	0.47	**0.81**	0.98
FROSTBITE	0.08	0.57	0.01	0.72	**0.76**	**0.75**	0.85
HERO	0.22	0.75	0.51	0.74	0.90	**0.93**	0.98
MONTEZUMAREVENGE	0.08	0.68	0.69	0.74	0.75	**0.78**	0.87
MSPACMAN	0.10	0.48	0.38	**0.74**	0.65	0.70	0.87
PITFALL	0.07	0.34	0.56	0.44	0.46	**0.60**	0.83
PONG	0.10	0.17	0.09	0.70	0.71	**0.81**	0.87
PRIVATEEYE	0.23	0.70	0.71	0.83	0.81	**0.91**	0.97
QBERT	0.29	0.49	0.49	0.52	0.65	**0.73**	0.76
RIVERRAID	0.04	0.34	0.26	**0.41**	**0.40**	0.36	0.57
SEAQUEST	0.29	0.57	0.56	0.62	**0.66**	**0.67**	0.85
SPACEINVADERS	0.14	0.41	0.52	**0.57**	0.54	**0.57**	0.75
TENNIS	0.09	0.41	0.29	0.57	**0.60**	**0.60**	0.81
VENTURE	0.09	0.36	0.38	0.46	0.51	**0.58**	0.68
VIDEOPINBALL	0.09	0.37	0.45	0.57	0.58	**0.61**	0.82
YARSREVENGE	0.01	0.22	0.08	0.19	0.39	**0.42**	0.74
MEAN	0.14	0.44	0.39	0.58	0.60	**0.68**	0.83

Table 2: Probe F1 scores averaged across all games (data collected by random agents)

CATEGORY	MAJ-CLF	CNN	VAE	PIXEL-PRED	CPC	ST-DIM	SUPERVISED
SMALL LOC.	0.14	0.19	0.17	0.31	0.42	**0.51**	0.69
AGENT LOC.	0.12	0.31	0.30	0.48	0.43	**0.58**	0.83
OTHER LOC.	0.14	0.50	0.36	0.61	0.66	**0.69**	0.81
SCORE/CLOCK/LIVES	0.13	0.58	0.53	0.76	0.83	**0.86**	0.93
MISC.	0.26	0.59	0.65	0.70	0.71	**0.74**	0.86

- ST-DIM largely outperforms other methods in terms of mean probe F1 score.
- Contrastive methods (ST-DIM and CPC) perform better than generative methods (VAE and PIXEL-PRED) across all categories.

Conclusion

- We present a new representation learning technique, **ST-DIM**, which **maximizes mutual information** of representations across **spatial** and **temporal** axes.
- We present a new **benchmark** for **state representation learning** by adding labels to **Atari** games to emphasize learning multiple generative factors.
- Our method excels at capturing latent factors for **small objects** and is **robust** to presence of easy-to-exploit features, which prove difficult for generative and other contrastive techniques, respectively.
- Our benchmark can be used to study qualitative/quantitative differences between representation learning techniques, and hope that it will encourage more research in the problem of state representation learning.