Prof. R. Wattenhofer

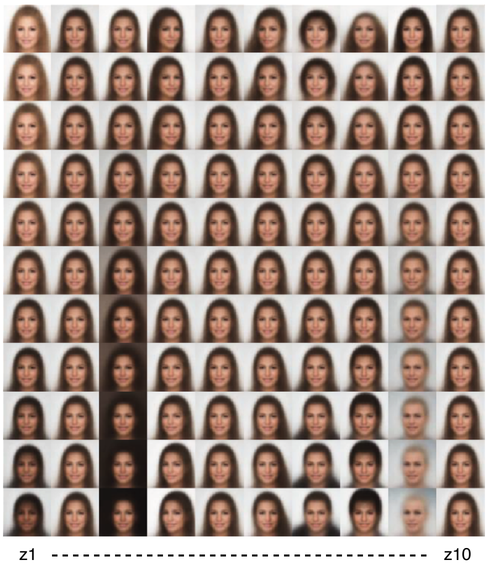
Disentangled Representations for RL

Disentanglement Learning is at the forefront of unsupervised learning, as disentangled representations of data are thought to improve generalization, interpretability, and performance in downstream tasks [2]. It works on the assumption of the existence of low-dimensional data generating factors for high-dimensional data and tries to recover these factors in an unsupervised fashion. Such a representation could for example serve as basis for the control of a robot arm [1]. We are specifically interested in the role disentangled representations play in downstream tasks and how strong their effect on the performance really is.

In this thesis, we will build on top of existing approaches for reinforcement and disentanglement learning and quantitatively evaluate the effect different representations have on the performance of a deep RL agent for simple games, e.g. Pong or the Atari games.

Requirements: Strong motivation, programming skills, and basic knowledge of machine and deep learning as well as reinforcement learning.

Interested? Please contact us for more details!



Contact

- Benjamin Estermann: besterma@ethz.ch, ETZ G60.1

References

[1] Irina Higgins et al. “Darla: Improving zero-shot transfer in reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1480–1490.

[2] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.