Towards explainable deep reinforcement learning through symbolic representations

Presenter: Hristofor Lukanov

This model is comprised of two parts. First the agent acts randomly and observes the environment in order to learn factors of variance. The model is encouraged to represent parts of the observation that seem to "behave" independently as separate representations, i.e. disentangled representations. There are two reasons for this: 1) it has been shown that such representations imply information minimality (Achille et al, 2017) and are therefore very efficient and 2) they are easily interpretable by humans and suggest learning of basic visual concepts (Whitney, 2016; Higgins et al, 2016). The model also learns which transformations in the observations are dependent on the actions of the agent and how actions result in transformations in the environment.

The second part is a reinforcement learning model that uses Actor-Critic in an unconventional way. The Actor learns how to manipulate representations that can be controlled by the agent such that they represent a "better" state, i.e. an augmented state that maximizes the expected reward. Based on the current state representation and the proposed change in some of its components - the agent decides on what action to take.