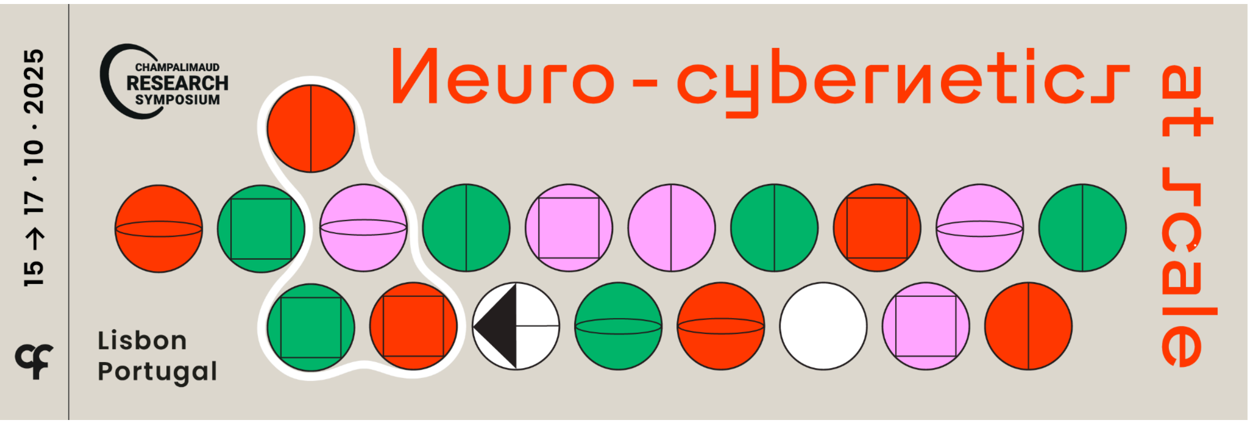

Exploring the Inner Workings and Internal Representations of Predictive Coding Networks in Comparison to Usual Feedforward Neural Networks

Aslan Satary Dizaji 1
1. NeuroAILab, Tehran, Iran

Predictive coding theory [1, 2] as a unified theory of brain formulating intelligence as a hierarchical process in which the brain builds an unsupervised world model and tries to predict the next states of the world by minimising the prediction errors. This process can be implemented in a variant of artificial neural networks, called energy-based networks or predictive coding networks, trained by a biologically plausible learning rule named prospective configuration [3, 4] .In each step of prospective configuration, first the neuronal activities of intermediate layers are adjusted to reflect the activities required to produce the targets and then synaptic weights of intermediate layers are adjusted to consolidate these neuronal activities [5]. While in recent years, there is a good progress on theoretical understanding of predictive coding networks trained by prospective configuration algorithm, the inner machinery and the internal representations [6] of these neural networks compared to usual feedforward neural networks are still unknown. This project aims to fill this gap using interpretability techniques for vision tasks. Basically, a few recently published methods are used to measure the internal representations of predictive coding networks to compare them with usual feedforward neural networks. The project is performed by simple vision tasks such as relatively simple predictive coding networks trained on relatively simple synthesised or well-known datasets. This project is an exploratory project without having any concrete hypothesis in advance regarding the internal representations in predictive coding networks in comparison to usual feedforward neural networks, while the author is generally believed that the internal representations of these neural networks probably are more brain-like.

Predictive Coding Networks, Prospective Configuration, Internal Representations, Interpretability Methods, Vision Tasks