Local Decorrelation of Neurons in ANNs through an adversarial Network

Nils Derner(1,2) and Willem Wybo(2)

(1) Systems Neurophysiology, Institute for Zoology, RWTH Aachen University, Aachen, Germany, (2) Dendritic Learning Group, Neuromorphic Software Ecosystems (PGI-15), Peter Grünberg Institut (PGI), Jülich, Germany

How the brain achieves deep learning of rich and robust representations remains an open question. Artificial neural networks (ANNs) provide useful models for studying this problem, as their architectures are often inspired by neuroscience. However, in contrast to state-of-the-art ANNs—which rely on end-to-end backpropagation, labeled data, and batch processing—biological neurons learn locally, largely in an unsupervised and online manner [1].
While there has been substantial progress in developing local learning rules, many of these approaches still require access to non-local information, such as pairwise correlations between neuronal activities [2] or dot products between representations [3].
In this work, we propose a mechanism for decorrelating neuronal activity using a single adversarial layer. This layer is trained to predict the response of a neuron based on the activity of other neurons within the same layer, while each neuron is trained to remain as unpredictable as possible. This adversarial interaction drives the neurons toward statistically independent representations.

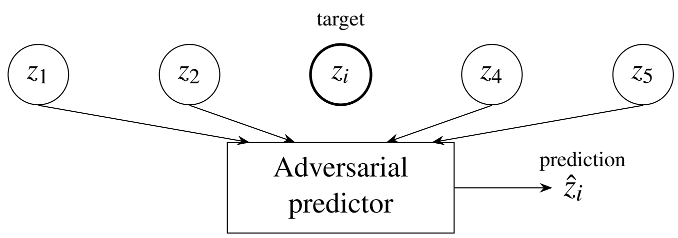

Figure 1: To predict the activity of a target neuron z_i, the predictor uses all other neurons in the same layer, excluding self-input.

The learning rule is further complemented by a predictive objective, where different augmentations of the same input are mapped to invariant representations, and by a regularization mechanism based on SIGReg [4], which enforces any prescribed activity distribution for each neuron. Importantly, all components of the proposed learning rule can be implemented neuron local and are compatible with locality in time.
Preliminary experiments indicate that this approach leads to improved representation diversity while maintaining stable learning dynamics.

[1] G. Tang et al., arXiv, 2110.14092, 2021.

[2] A. Halvagal and F. Zenke, Nat. Neurosci., 26, 2023.

[3] B. Illing, J. Ventura, G. Bellec, and W. Gerstner, NeurIPS, 34, 2021.

[4] Y. LeCun et al., arXiv, 2511.08544, 2025.