Title:

Towards data-driven segregation of class subnetworks in artificial and biological neural networks

Abstract:

In the brain, the information is hypothesized to be represented as a distributed pattern of activity across a pool of processing units. Functional network connectivity studies focus on understanding what features of information are represented and how that information is encoded. The current approaches quantify connectivity as statistical dependence between two spatially distant neurophysiological activities, related to a stimulus or a task. However, functional subnetworks in the brain comprise several nodes and cannot be described simply in terms of node pairs. Moreover, none of the proposed approaches allows a “hypothesis-free” segregation of functional subnetworks, where information categories are not predetermined.

To develop a framework for data-driven subnetwork segregation, we resorted to deep neural networks (DNNs) as proof-of-concept models of distributed processing. We hypothesized that the information-specific subnetworks could be identified based on the predictability of node activity from the other subnetwork nodes. Thus, we used a variable-order Markov model to make predictions of node activity from the activity of other nodes, and segregated the nodes into subnetworks based on the prediction error.

Pruning the subnetworks with lower prediction error impaired the DNN model more than pruning the less accurate subnetworks. Subnetwork size did not correlate with the impairment extent. Subnetworks of the first layers were less class-specific, while that of deeper layers were more class-specific, in agreement with previous reports. The class subnetworks were overlapping, which could be explained by shared class features.

This framework may be transferable to neurophysiological data and contribute to the explainability of DNN models.