Uncovering structure in neural network representations

Neural networks—an increasingly important class of learning algorithms—are trained by finding minima of a function that describes the cost of incorrect predictions. This function depends on the weights of connections in the network, and is minimized numerically. That is, we do not have an equation for the "correct" values of weights in the network architecture. Instead, we iteratively improve the weights until network reaches satisfactory accuracy. Because of this iterative optimization, two networks trained to solve the same task will typically have different weight values, even though their accuracy can be quite similar. This raises a following, important question in neural network research: are these two independently trained networks similar with respect to the way they construct predictions? Or, perhaps, independently trained networks—that solve the same task—have quite different internal "rules" for constructing predictions? This and similar questions are the central focus of research on similarity of neural representations of data. Even though this research area has fundamental implications for applications of neural networks—e.g., in systems that require explanations of automated decisions—initial results toward answering basic questions about similarity of neural representations were obtained only recently.

In this project we want to move analysis of neural network representations beyond comparison of pairs of trained networks. To this end, we plan to pursue two complementary approaches. Firstly, we plan to investigate probabilistic models for neural representations. Specifically, our goal is to develop tractable density models for representations learned by neural networks and then use these models to uncover structure in sets of learned representations. Secondly, we plan to use methods from topological data analysis to develop tools for quantification of representational similarity that focus on relatively invariant properties of the learned representations. Using these two approaches we will propose models and algorithms for a broad family of tasks in research on neural representations. For example, we will work on algorithms that pinpoint inputs with unique network representations, algorithms for uncovering impact of noise in training data on the learned representations, or algorithms that compare neural architectures with respect to the representations they learn across many trained networks. We will also use our methods to characterize effects of various commonly used training-related algorithms on the distributions of learned representations.