Deep artificial neural networks: little brains or big retinas?

Luciano Dyballa[1], Mahmood S Hoseini[3], Greg D Field[4], Michael P Stryker[3], Steven W Zucker[1,2]

[1]Yale Univ., Dept. of Computer Sci, New Haven, CT, [2]Yale Univ., Dept. of Biomed. Engin, New Haven, CT, [3]Univ. of California, San Francisco, Ctr. for Integrative Neurosci., San Francisco, CA, [4]Duke Univ., Dept. of Neurobio., Durham, NC

Deep neural network modeling of biological visual processing is widespread: brains are archetypal pattern analyzers and deep CNNs are currently the best object classifiers. Implicit is the assumption that cortex can be well approximated by CNNs, from which it follows that CNNs are an appropriate foundation for AI. We examine whether this approximation holds using a novel neural manifold obtained with machine learning techniques.

Our approach was developed to understand the visual system in behaving animals responding to a stimulus ensemble. The challenge is to infer the structure and function of neural circuits from neural activity. The usual approach -- embedding trials in 'neural coordinates' -- reveals only population activity and is useful for 'reading out' the state. We turn this around, so that each point on our manifold is a neuron. Nearby neurons on the manifold respond similarly to similar stimuli. When smooth, the manifold coordinates reveal how stimulus selectivity is organized in the neuronal population.

Our manifolds are mathematically related to functional networks. If the underlying network consists of separate, or largely decomposable components, the manifold would be disconnected; this is the case for the retina. If the network were continuous and simple, the manifold would be low-dimensional; this is the case for the ring model of orientation columns. If the circuit consists of many related components, the manifold would be high-dimensional and continuous, as is the case in cortex (V1). Finally, if the network were fully connected, e.g. a clique, the manifold would be degenerate.

The manifold can be computed for artificial neural networks as well, so we tested AlexNet, VGG-16, and Inception-v3. As our 'stimuli', we use a subset of the ImageNet validation set; each trial is a random shift of the input image across position. The average value across trials converts each unit's activity into a 'firing rate'.

The result: deep CNNs yield manifolds that are disconnected like the retina, not continuous like V1. For neuroscience, this could explain the apparent ceiling in modeling cortical data; the limitations of categorical tasks; and suggests future modeling directions. For AI, our manifold can identify co-activations of feature maps across layers, revealing higher-order features, and suggests different approaches to performing 'dropout.' More generally, it illuminates limitations on the categorization problem, and underlines the importance of recurrence in networks for more complex tasks.