Modeling Mid-level Visual Representations through Clustering in a Convolutional Neural Network

Daniel D Leeds (dleeds@fordham.edu)
Shane Hyde (shane.hyde@aol.com)
Fordham University, 441 E Fordham Road
Bronx, NY 10458 USA

Abstract:

The nature of visual properties used in cortical perception is subject to considerable ongoing study. Features of intermediate complexity are particularly uncertain. Convolutional Neural Network (CNN) models, however, have proven to be quite effective in modeling human vision (Yamins et al., 2014) and have performed with great accuracy on image classification tasks (Krizhevsky et al., 2012). Study of representations within layers of CNN models may suggest selectivities in the similarly hierarchical brain. We apply a popular CNN to four diverse stimulus sets. Through clustering, we identify classes of preferred visual patterns for an intermediate model layer (layer 4, out of 8). We find a subset of patterns reflect intuitive visual similarities within and across the datasets, while a broader set of patterns were less accessible to intuitive interpretation. We also observe a heightened correlation between cortical voxel activity and CNN layer 4 responses to a shared dataset, and notably observe increased correlation when weighting model neuron responses based on clustering from each of the four data sets. Our findings suggest behavioral and cortical relevance to visual properties uncovered by clustering on multiple image data sets.

Keywords: Intermediate visual representations; fMRI; Convolutional Neural Networks; K-means clustering

Background

Visual perception employs a hierarchy of cortical regions, encoding increasingly complex properties of the visual input. While the visual properties used in early vision have been well-studied (Hubel & Wiesel, 1968; Kay et al., 2008), properties for intermediate cortical vision remain more elusive. Convolutional neural networks (CNNs) have risen to the forefront for their ability to perform automated image classification tasks (Krizhevsky, 2012), and to predict cortical responses to visual inputs (Yamins, 2014).

Multiple techniques have been pursued to extract candidate intermediate representations for CNNs. Recently, Wang et al. (2016) identified visual concepts by clustering image patches based on responses across CNN units for automobile objects. We adapt the Wang clustering technique across four diverse sets of stimuli to better understand visual properties used in a CNN trained for general object recognition. We explore the utility of the resulting clusters to model fMRI voxel responses for real-world photo stimuli.

Methods

Model network:

We studied a convolutional neural network adapted from AlexNet (Krizhevsky, 2012; Jia et al., 2014). To capture intermediate representations, we use the convolution units in the fourth layer, with CNN weights pre-learned. We study layer responses to image patches taken from distinct locations in each input.

Image datasets:

We consider CNN responses to roughly 20,000 image patches from each of four datasets. The first three sets are object photographs of (1) cars, (2) cows, and (3) guitars from ImageNet (Deng et al., 2009). The fourth set is photographs of objects and scenes from Kay (2008). The first three sets were chosen for within-set homogeneity and cross-set diversity. The fourth set was chosen for comparison with associated fMRI data.

Cluster analysis:

For each dataset, image patches are clustered using K-means based on responses from CNN layer 4. 384 clusters are defined, each associated with a "centroid" specifying the average response of each of 384 CNN units to patches within the cluster.

Neuroimaging analysis:

We compare voxel activity with CNN layer 4 activity responding to photographs from dataset 4. Activity of each layer 4 unit is considered. Activity also is found for each cluster centroid; the response of centroid *k* to image *i*, $r_k^{clust}(i)$ is computed by the weighted sum:

$$r_k^{clust}(i)=\Sigma_j\, w_j^k\, r_j(i)$$

where $r_j(i)$ is the response of CNN unit *j*, and w_j^k is the average response of unit *j* in cluster centroid *k*. We compute the correlation of CNN and voxel responses to the photographs in the fourth dataset.

Results

Cluster analysis:

Ideally, we wished to identify clusters common across multiple datasets, each with a narrow and intuitive set of visual properties revealed by study of the member image patches. All datasets produce an array of clusters with overlapping distributions of "spread" – diversity of layer 4 multi-unit responses (Fig 1). While the more common higher-spread clusters are more challenging to interpret, low-spread clusters typically represent textures (clouds, grass, asphalt) and geometric forms (holes, fences, heads) (Fig 2a,b).

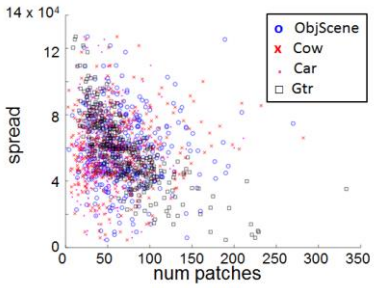

Figure 1: Number of member patches vs. "spread" (mean distance between member patch response and centroid) for each of 384 clusters for each dataset.

A small number of clusters have similar centroids across datasets. These clusters typically have a broad spread of image patches, making visual interpretability more challenging (Fig 2c,d).

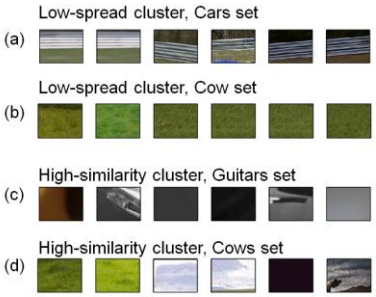

Figure 2: Image patches from (a,b) distinct low-spread clusters; (c,d) cross-set similar clusters.

Neuroimaging analysis:

Comparison with brain data shows significantly higher maximum correlations for cluster centroids than for individual CNN units (Fig. 3a). Notably, there is a positive relation between cluster spread and voxel correlation – clusters with more diverse patches more highly correlate with individual voxel responses (Fig 3b). In all cases, highest correlations are present for voxels in cortical regions associated with mid-level vision (figure not shown).

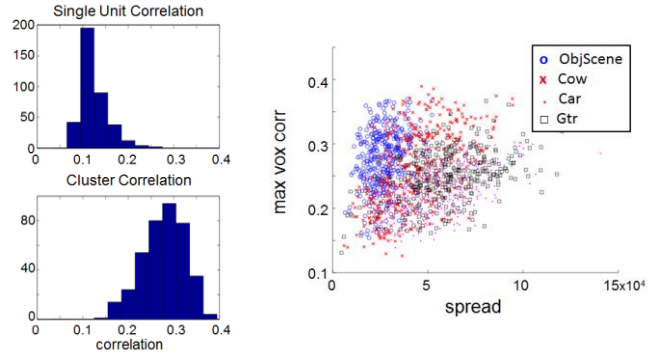

Figure 3: (a) Maximum voxel correlations for single CNN units (top) and cluster centroids (bottom). (b) Spread vs. correlation for each sets' clusters.

Discussion

K-means clustering provides some insights into groupings of intermediate visual properties in CNNs. The most tightly-grouped clusters show texture and geometric properties. More widely spread clusters are harder to interpret visually, but share similarities across datasets and better correlate with mid-level cortical activity, encouraging further study.

Acknowledgments

This work was supported in part by funds from the Fordham Faculty Research Grant to Dr. Daniel Leeds.

References

Hubel D. & Wiesel T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J of Physiology*, 195, 215-243.

Deng, J. et al. (2009). ImageNet: a large scale hierarchical image database. Proc *CVPR*.

Jia, Y. et al. (2014). Caffe: Convolutional architecture for fast feature embedding, *arXiv*: 1408.5093.

Kay, K.N. et al. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352-355.

Krizhevsky, A. et al. (2012). ImageNet classification with deep convolutional neural networks. Proc *NIPS*.

Wang, J. et al. (2016). Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. *arXiv*: 1511.06855v3.

Yamins, D. et al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23), 8619-8624.