Background

Visual perception in the brain is understood to use a network of brain regions selective for increasingly complex properties. While visual properties used in early vision have been well-studied, more complex visual properties used by the brain remain unclear.

Recent studies illustrate Convolutional Neural Networks’ (CNNs), prediction of cortical region responses to visual stimuli (e.g., Yamins, 2014). CNNs’ intermediate representations provide testable hypotheses for properties used in the brain. Wang (2016) recently identified intuitive intermediate properties through clustering of patches from automobile/transit images based on their corresponding CNN encodings.

Methods: Image patch clusters from AlexNet CNN

Four data sets used to study CNN representations
- (1) Cars, (2) Cows, (3) Guitars
- (4) Objects & scenes

Model network

We used Caffe implementation of the AlexNet Convolutional Neural Network (CNN; Krizhevsky 2012, 2a 2014), trained on ImageNet (Deng 2009). AlexNet is composed of 8 layers, each layer finds patterns in outputs from previous layer. Each layer consists of artificial units U1, U2, ..., UK.

CNN layer 4 unit responses extracted for each image input (as an example of Intermediate representation)

Unit responses computed for image patches taken from b x b grid (13x13 at layer 4). For each set, all image patches clustered with K-means clustering (K=384) on layer 4 unit outputs.

Results: Clustering – convergence on visual properties

Intra-cluster variance

Diversity of images in cluster measured by spread:

\( spread = \text{mean squared distance from centroid to member image patches} \)

All sets produce distribution of clusters with wide and narrow spreads. Clusters from Objects & scenes set have smaller spread than Image Net object clusters. Clusters with more patches typically have smaller spread.

Within-set visual properties

Example dense clusters
- Simple textures and shapes
  - Grass, sky, asphalt
  - Edges, curves

Groups of 2 – 10 clusters within same set capture similar properties (inter-cluster square distance < 2x10^6)

Example sparse clusters
- More variable textures
- More variable complex shapes

Cross-set visual properties

Groups of 2 – 20 clusters across sets capture similar properties (inter-cluster square distance < 2x10^6)
- Similar textures and shapes grouped for each set

Discussion

Image-patch clustering provides intuition for intermediate visual representations utilized by artificial CNN model (AlexNet) and by the brain
- Layer 4 AlexNet unit population responses appear organized based on mix of unclear visual patterns and intuitive properties such as shapes, boundaries, and textures
- AlexNet clusters better correlate with voxel responses in mid-level vision than do single layer 4 units
- Additional testing needed on alternative CNNs and alternative image patch sets

Acknowledgements

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

Poster URL:

http://storm.cis.fordham.edu/leeds/LeedsCCN17.pdf