Modeling mid-level visual representations through clustering in a convolutional neural network

Daniel D Leeds; Shane Hyde
Fordham University - Computer and Information Science Department

FORDHAM UNIVERSITY THE JESUIT UNIVERSITY OF NEW YORK

COMPUTER AND INFORMATION SCIENCES DEPARTMENT

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.

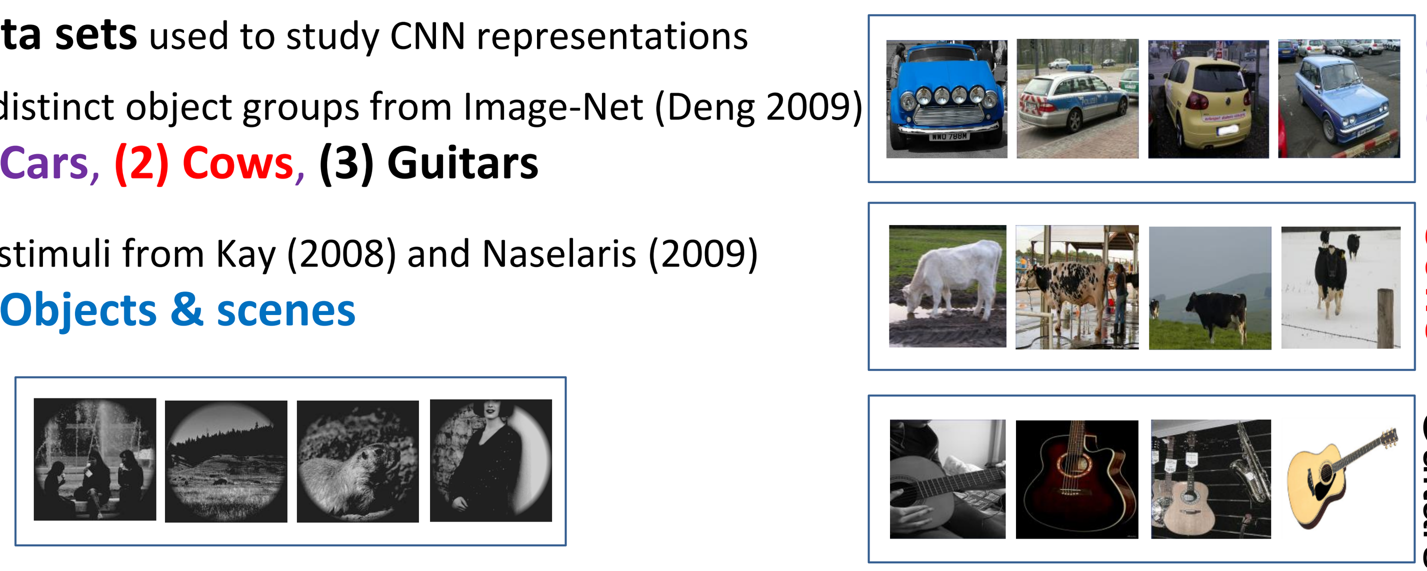
Expanding on Wang, **we cluster image patches from four distinct data sets to identify common properties and assess their relation to cortical encodings.**

Methods: Image patch clusters from AlexNet CNN

Four data sets used to study CNN representations

Three distinct object groups from Image-Net (Deng 2009)
- **(1) Cars, (2) Cows, (3) Guitars**

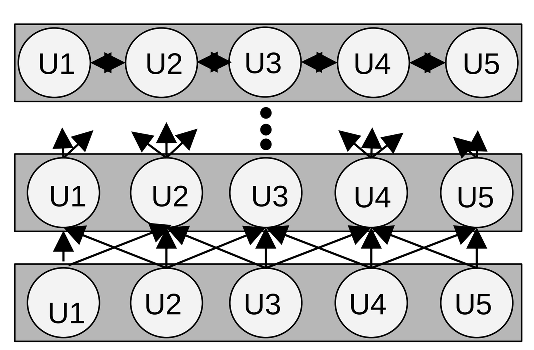
Mixed stimuli from Kay (2008) and Naselaris (2009)
- **(4) Objects & scenes**



Model network

We used Caffe implementation of the AlexNet Convolutional Neural Network (CNN; Krizhevsky 2012, Jia 2014), trained on Image-Net (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)

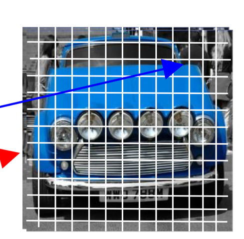
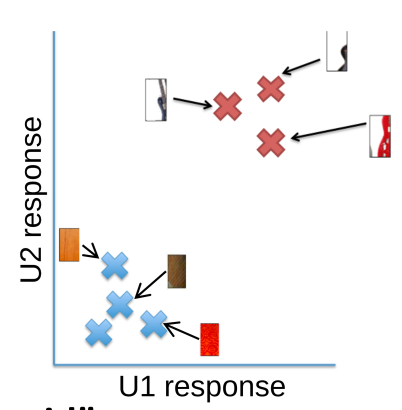


Image patch clustering

For each data set, all image patches clustered with
K-means clustering (K=384) on layer 4 unit outputs.

We record:
- **cluster assignment for each image patch**
- **average response of 384 CNN units for each cluster – "centroid"**

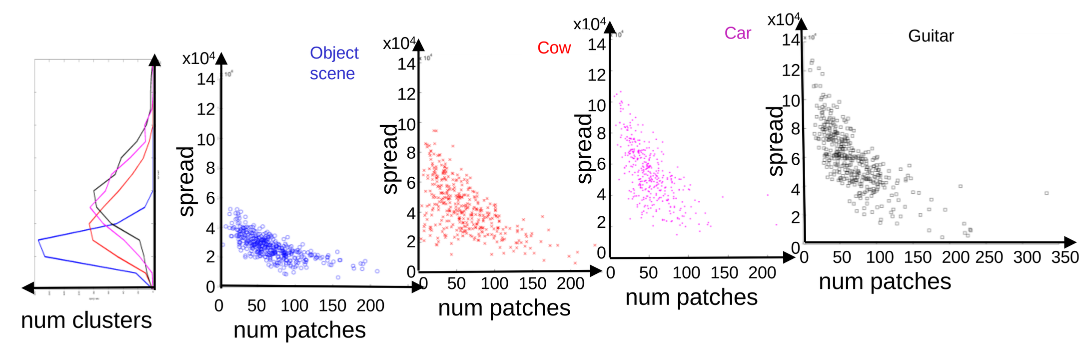


Results: Clustering – convergence on visual properties

Intra-cluster variance

Diversity of images in cluster measured by "spread":

spread = mean squared distance from centroid to member image patches

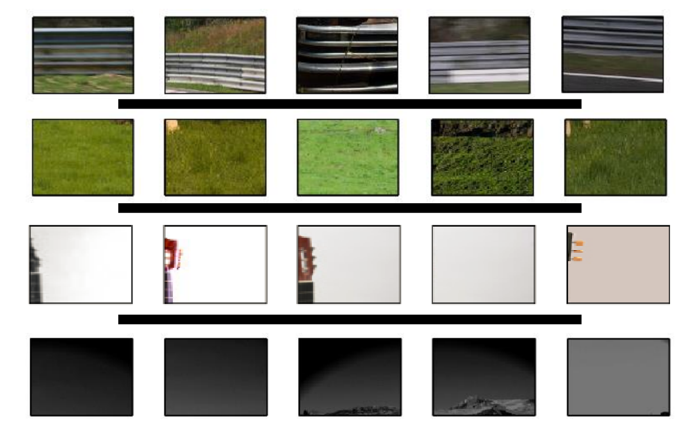


All sets produce distribution of clusters with wide and narrow spreads
Clusters from **Object-Scene** set have smaller spread than **Image-Net** object clusters
Clusters with more patches typically have smaller spread

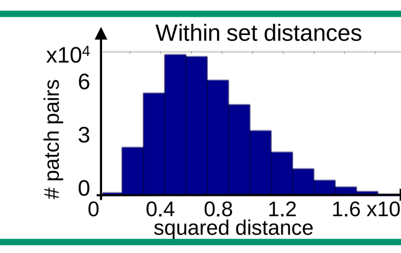
Within-set visual properties

Example dense clusters (spread < 2×10^4):

- **Simple textures and shapes**
  - **Grass, sky, asphalt**
  - **Edges, curves**



Groups of 2 - 10 clusters within same set capture similar properties
(inter-cluster square distance < 2×10^4)

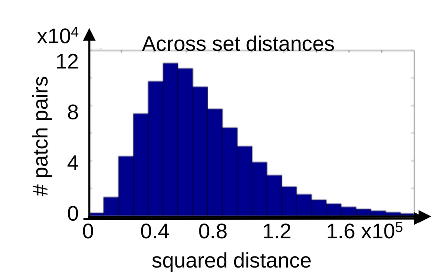


Example sparse clusters (spread > 9×10^4):
- More variable textures
- More variable complex shapes

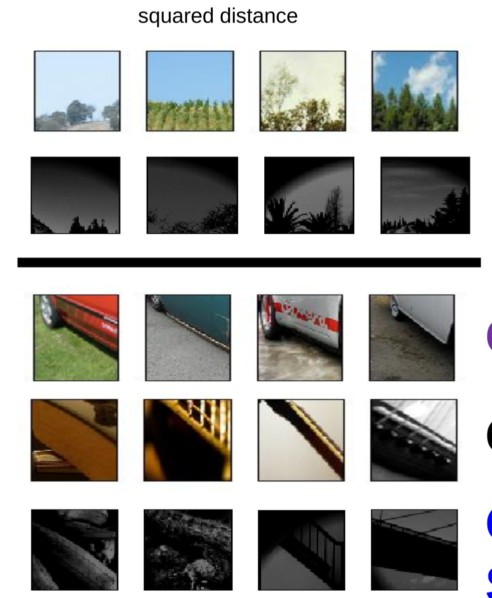
Many hard-to-interpret clusters

Cross-set visual properties

Groups of 2 – 20 clusters across sets capture similar properties
(inter-cluster squared distance < 2×10^4)

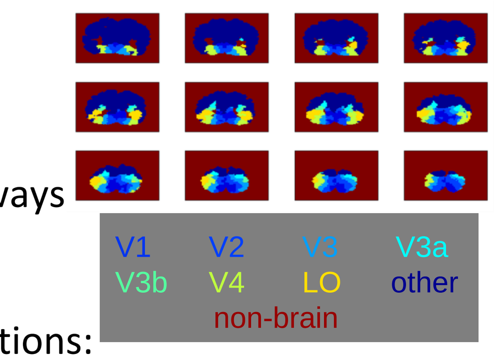


- **Similar textures and shapes grouped for each set**



Methods: Neuroimaging analysis

Neuroimaging data from Kay (2008) and Naselaris (2009)
2 subjects each viewed 1750 Objects and Scene images
Passive viewing, 4s trials
2x2x2.5mm voxels; Coverage of ventral and dorsal visual pathways



CNN layer 4 unit – voxel comparisons

For each image, compute max unit response across all patch locations:

$unit\_resp^i(im) = \mathbf{max}_{x,y}\ unit^i(im_{patch\ (x,y)})$

We find correlation between unit's and voxel's responses to same stimuli.

CNN cluster – voxel comparisons

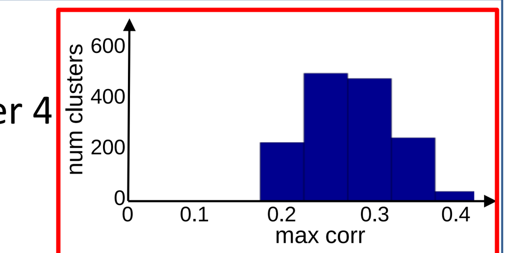
For each CNN cluster, compute weighted response based on centroid

$clust\_resp^i(im) = \sum_n centroid_n^i \times unit\_resp^n(im)$

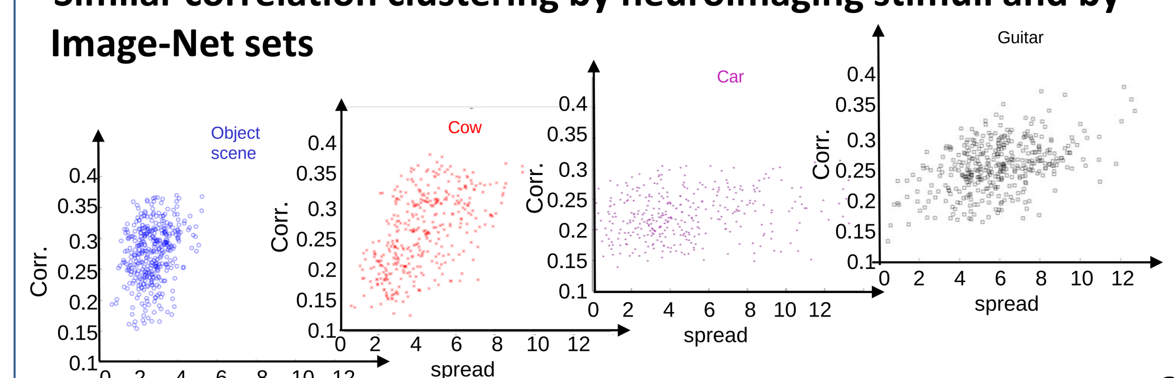
We find correlation between cluster-weighted CNN response and voxel's responses to same stimuli.

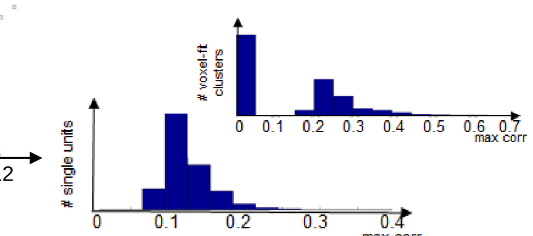
Results: Correlation of voxels and CNN clusters

Cluster-weighted CNN outputs better predict voxel responses to images (r>0.2) than do single CNN units in layer 4
- Similar to CNN units weights learned from clustering of best-fit voxel – CNN unit regression weights
- **Similar correlation clustering by neuroimaging stimuli and by Image-Net sets**

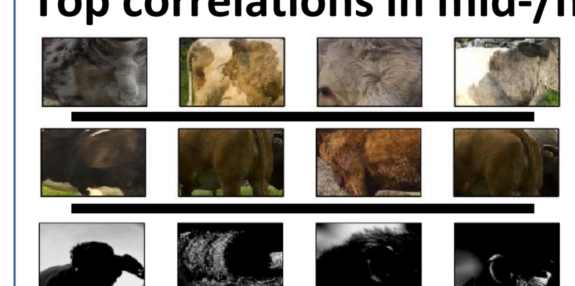


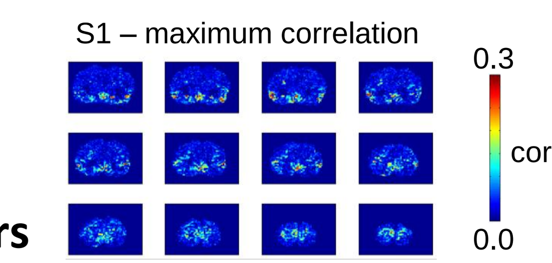
Histogram of highest voxel-cluster correlations for CNN-based clusters across 4x384 learned clusters





Top correlations in mid-/high-level visual cortex





High correlation clusters capture intuitive and unintuitive property groups
- **Including shapes and textures**

Note: all correlations shown here are for Subject S1; patterns for S2 are substantially similar.

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

Contact

Daniel D Leeds
dleeds@fordham.edu
storm.cis.fordham.edu/leeds

Shane Hyde
shane.hyde@aol.com

Computer and Information Sciences
Fordham University

References

1. Deng, J. et al. (2009). ImageNet: a large scale hierarchical image database. Proc CVPR.
2. Jia, Y. et al. (2014). Caffe: Convolutional architecture for fast feature embedding, arXiv: 1408.5093.
3. Kay, K.N. et al. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352-355.
4. Krizhevsky, A. et al. (2012). ImageNet classification with deep convolutional neural networks. Proc NIPS.
5. Naselaris, T. et al. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron, 63(6), 902-915.
6. Wang, J. et al. (2016). Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. arXiv: 1511.06855v3.
7. Yamins, D. et al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS, 111(23), 8619-8624.

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