Modeling voxel visual selectivities through convolutional neural network clustering

Daniel D Leeds; Amy Feng
Fordham University - Computer and Information Science Department

FORDHAM 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., Horikawa 2017). **CNNs' intermediate representations provide testable hypotheses for properties used in the brain.** Wang (2016) and Wu (2015) identified intuitive intermediate properties through clustering of patches, e.g., from automobile/transit images, based on their corresponding CNN encodings.

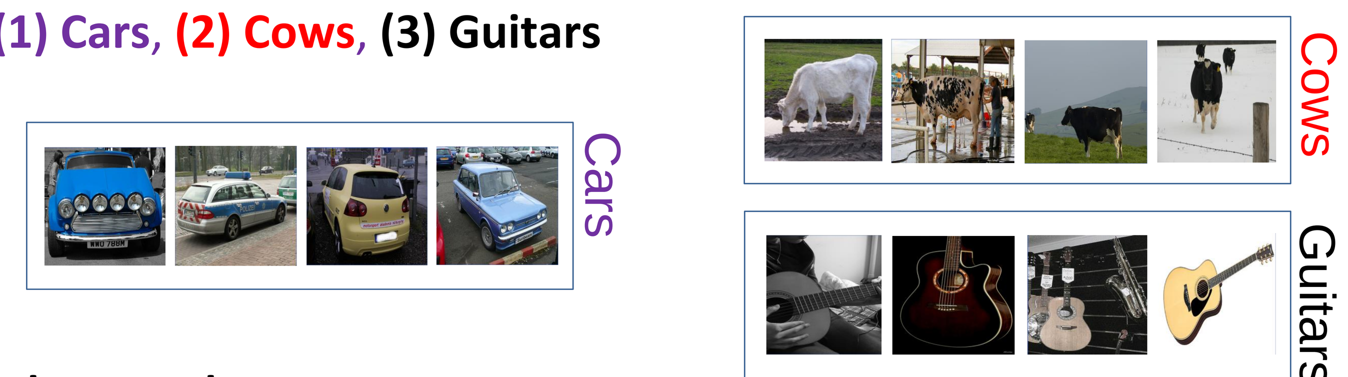
Expanding on Wang and Wu, **we cluster image patches from three data sets to identify common properties and assess their relation to cortical encodings.**

Methods: Image patch clusters from AlexNet CNN

Three data sets used to study CNN representations

Three distinct object groups from Image-Net (Deng 2009)

- **(1) Cars, (2) Cows, (3) Guitars**



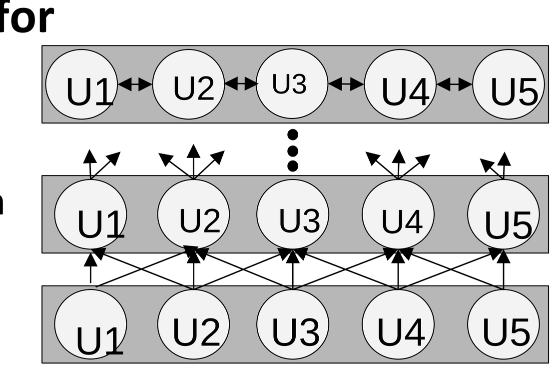
Model network

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 layers 2-5 unit responses extracted for each image input (as examples of low-level to intermediate representations)

Unit responses computed for image patches taken from 13 x 13 grid



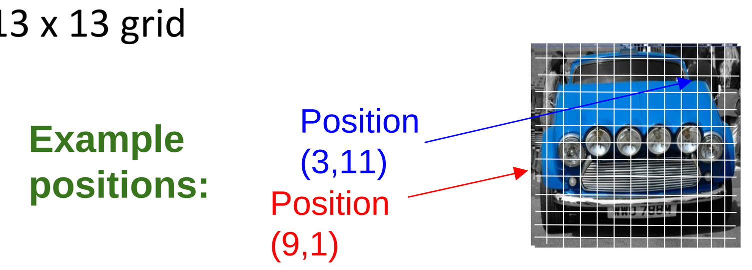
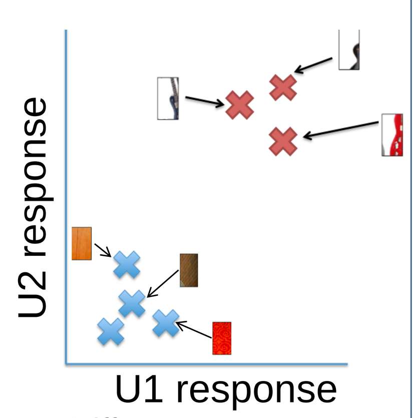


Image patch clustering

For each data set and CNN layer L, all image patches clustered with **K-means clustering (K=384) on outputs from all units in layer L.**

We record:

- **cluster assignment for each image patch**
- **average response of CNN units in layer L for each cluster – "centroid" μ_k**



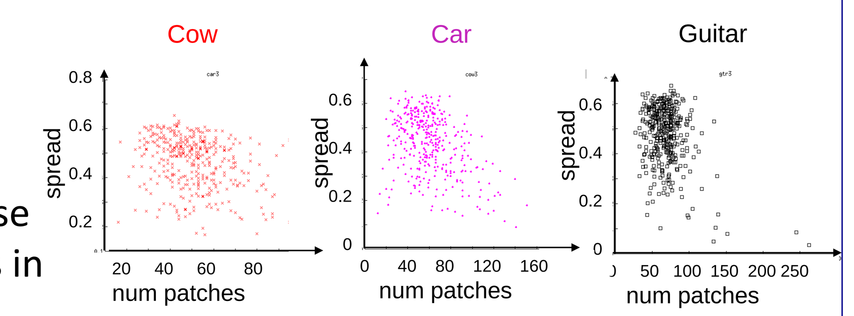
Results: Clustering – convergence of CNN responses

Intra-cluster variance:

Diversity of unit responses in each cluster k :

$$\text{spread} = \frac{1}{N_k}\sum_{i\in k}\left\|x^i - \mu_k\right\|_2^2$$

μ_k is cluster k centroid, x^i is CNN unit response image patch I, N_k is number of image patches in cluster k



Layer	Cars	Cows	Guitars
norm2	0.4016	0.3726	0.3913
conv3	0.4625	0.4337	0.4771
conv4	0.4678	0.4578	0.4993
conv5	0.5083	0.4760	0.5286

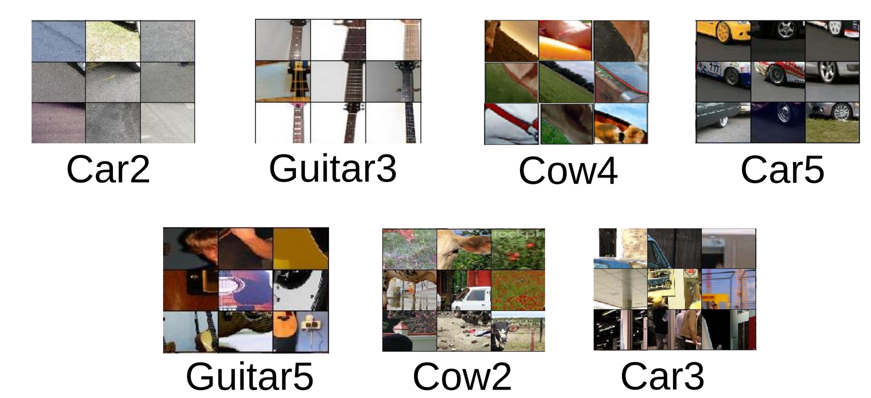
+ Higher layers show more diverse unit responses for each cluster
+ Small number of very high-member, very low-spread clusters

Variability in cluster interpretability

+ Texture + Color
+ Edges + Shapes
+ Object-parts

+ Clusters with unclear themes

Similar themes across layers and image groups



Quantifying clustering interpretability

Methods

Stimuli: Select 30 diverse clusters for CNN layers 2, 3, 5 and for image groups car, cow, guitar (30x3x3 = 270 clusters) with high distance between cluster centroids $\left\|\mu_j - \mu_k\right\|_2^2$

8 subjects through Mechanical Turk

Survey: Record 1-10 "similar properties" across each image cluster

Example cluster response:
1. Cow
2. Grass
3. Land
4. Outside

Example cluster displays

Find consistent subject responses across subjects

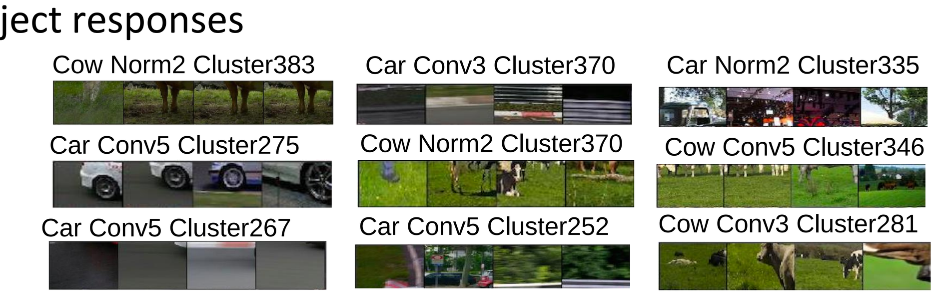
Results

Common responses:
+ Broad category + Sub-part
+ Background + Colors
+ ***Not*** *texture, shape, edge*

k^{th} most-frequent	cow lay2	cow lay3	cow lay5	guitar lay2	guitar lay3	guitar lay5
1	**cow 82**	**cow 60**	**cow 75**	**guitar 80**	**guitar 68**	**guitar 74**
2	**grass 81**	**grass 57**	**grass 70**	**strings 45**	black 52	**strings 45**
3	open 50	blue 52	outside 48	**color 33**	**strings 30**	white 33
4	**green 45**	**animal 39**	**animal 42**	**music 26**	**music 24**	**music 32**
5	**animal 40**	open 38	**green 40**	song 20	wood 23	**color 30**

word (# uses)

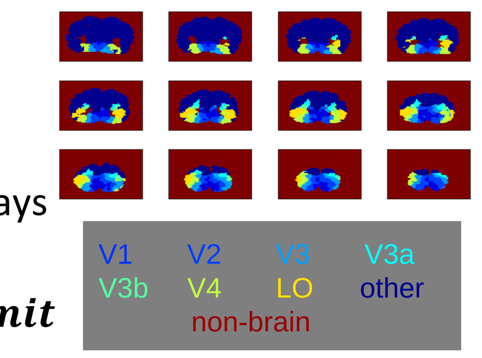
Clusters with most consistent subject responses
+ Textures (grass, asphalt, fence, foliage)
+ Simple shapes (wheel)



Methods: Neuroimaging analysis

Neuroimaging data from Kay (2008) and Naselaris (2009)

We study 1 subject viewing 1750 images of objects and scenes
Passive viewing, 4s trials
2x2x2.5mm voxels; Coverage of ventral and dorsal visual pathways



CNN cluster – voxel comparisons

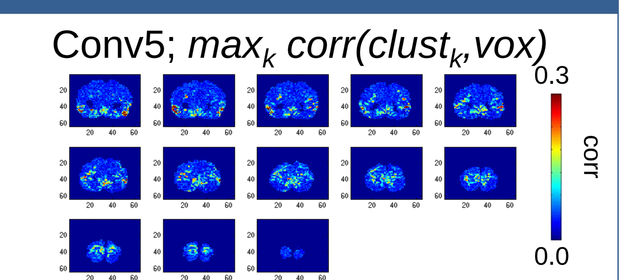
For each cluster, compute weighted-sum of CNN unit responses $\boldsymbol{unit}$ based on each centroid μ_k $\quad \boldsymbol{clust_k} = \boldsymbol{\mu_k^T unit}$

We find correlation between cluster-weighted CNN response and voxel's responses to same stimuli ***corr(clust$_k$,vox)***.

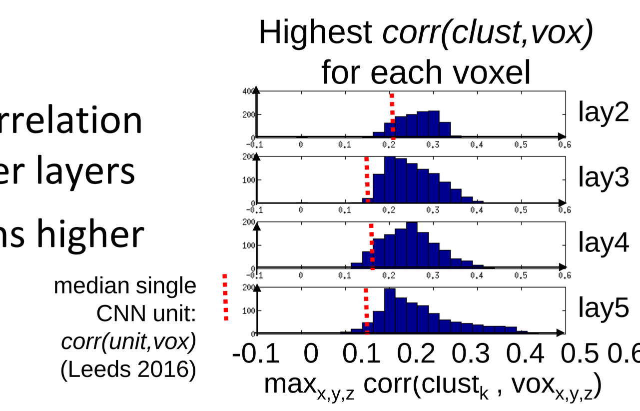
Results: Correlation of voxels and CNN clusters

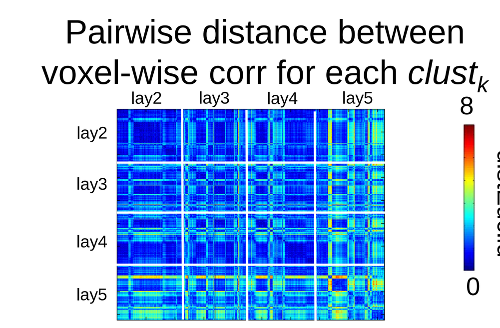
Cluster correlation statistics

+ High correlations in mid-level vision, particularly LO
+ Near-identical distribution of high correlation voxels for Layers 2 – 5

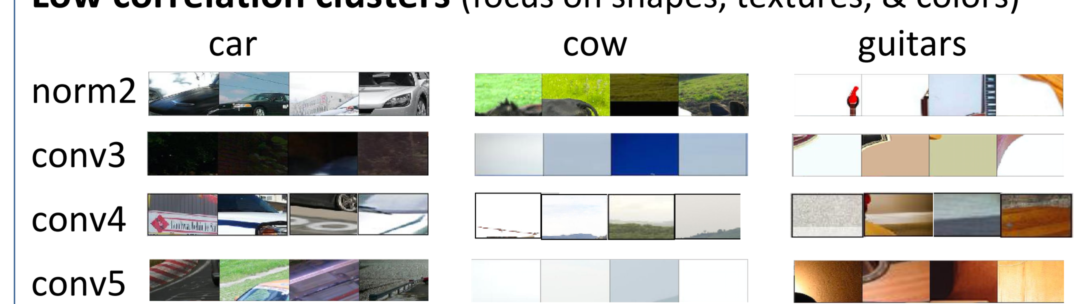


+ Highest cluster correlation get larger at higher layers
+ Cluster correlations higher than single-voxel correlations





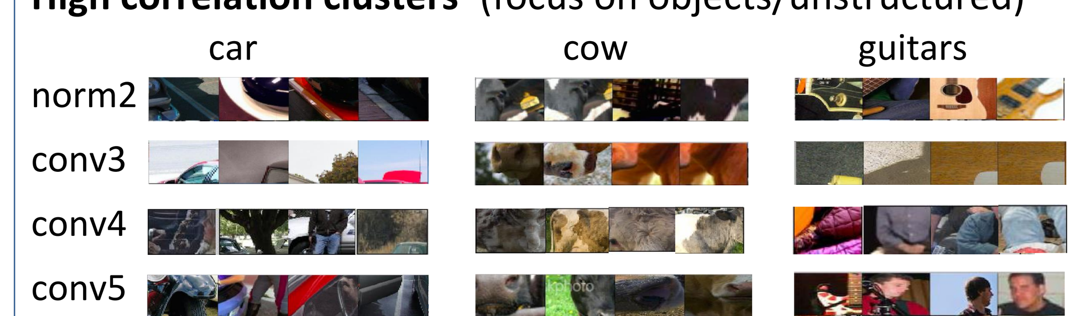
Low correlation clusters (focus on shapes, textures, & colors)



Low Corr Shape/Texture/Color Counts

Layer	Cars	Cows	Gtr	Total
norm2	32	58	48	138
conv3	20	58	53	131
conv5	34	52	34	120
Total	86	168	135	**389**

High correlation clusters (focus on objects/unstructured)



High Corr Shape/Texture/Color Counts

Layer	Cars	Cows	Gtr	Total
norm2	49	37	28	114
conv3	53	42	33	128
conv5	41	30	46	117
Total	143	109	107	**359**

corr(Spread, VoxelCorr)

Layer	Cars	Cows	Gtr
norm2	0.11	0.26	0.47
conv3	-0.07	0.27	0.20
conv4	0.51	0.49	0.41
conv5	0.12	0.28	0.23

+ Same voxel-cluster correlation patterns across CNN layers

Discussion

Image-patch clustering provides intuition for intermediate visual representations utilized by artificial CNN model (AlexNet) and by the brain

- **Layer 2-5 AlexNet** unit population responses **appear organized based on** mix of unclear visual patterns and intuitive **properties such as shapes, textures, and color**
- AlexNet **clusters better correlate with voxel responses in mid-level vision** than do single units
- Highest cluster-voxel correlations tied to most diverse/least simple visual properties

Contact

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

Amy Feng
afeng5@fordham.edu

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. Leeds, DD. and Iotzov, I. (2016). Single kernel models of single-voxel visual selectivities in convolutional neural networks. Cogn Sci Soc.
6. Leeds, DD. and Hyde, S. (2017). Modeling mid-level visual representations through clustering in a convolutional neural network. Cogn Comp Neuro.
7. Naselaris, T. et al. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron, 63(6), 902-915.

References (continued)

8. Wang, J. et al. (2016). Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. arXiv: 1511.06855v3
9. Wu, R. et al. (2015). Harvesting discriminative meta objects with deep CNN features for scene classification. Proc ICCV.
10. Horikawa, T and Kamitani, Y. (2017). Generic decoding of seen and unseen objects using hierarchical visual features. Nat Comms, 8.

Acknowledgements

This work was supported in part by funds to Amy Feng as a Fordham University Clare Boothe Luce Fellow. Thanks to William Charles for assistance with poster edits.

Poster URL:
http://storm.cis.fordham.edu/leeds/LeedsVSS19.pdf