Next 3 topics: modeling conceptual organization

- Modeling* typicality (Blake et al.)
- Can feature models explain behavioral and brain responses? (Mitchell et al., Wen et al. Baroni et al.)
- Modeling Similarity Spaces expressed in human behavior and brain responses (Kriegeskorte et al., Peterson et al.)


Deep Neural Networks Predict Category Typicality Ratings for Images
LAKE ET AL. 2016

Background

- Can deep-learning systems serve as potential cognitive models?
- They try to predict human typicality ratings for a set of naturalistic images
- “for any task that requires relating an item to its category, typicality will influence performance, whether it is the speed of categorization, ease of production, ease of learning, *usefulness for inductive inference*, or word order in language”

Convnets and typicality

- Convents may or may not produce representations that track categorical structure with typicality structure.
  - They learn categorization
  - But perhaps they categorize by learning prototypes?

.

Method: behavioral

- Method:
  - People rate typicality for images drawn from 8 image categories: Typicality ratings were collected for eight categories: banana, bathtub, coffee mug, envelope, pillow, soap dispenser, table lamp, and teapot.
  - Mechanical turk. Each participant rates “how well does this picture fit your idea or image of the category”
  - Mean typicality per image computed across all ss.. Human reliability ratings have good split half . P= 0.92. so two groups of ppl produce similar rankings.

Method: computational

- They use 3 convnet architectures. Describe OverDeat (7 layers). Last layer 1000-way softmax
- top-five error rate of 14.2%, meaning: for over 85 percent of test images, the correct label appeared in the top five guesses.

nections. Convolutional layers take a set of 2D image-like grids as input (called "feature maps"), apply a set of trainable image filters, and output a new set of feature maps. The first two and the last convolutional layers also contain max pooling operations that reduce the resolution of the feature maps. Specifically, the model takes a 231x231 color image as input (three feature maps for RGB channels) and outputs 96 feature maps after applying 11x11 trainable image filters.[1] After three other layers of processing, the last convolutional layer has 1024 feature maps with smaller trainable filters (size 3x3). After the convolutions, the next two layers have 3072 and 4096 fully-connected connectionist units, respectively. Finally, the 1000-way softmax layer produces a probability distribution over the $j = 1, \ldots, 1000$ classes. It does so by first computing the raw class scores y_j from the activity x in the previous layer and weights w_{ij} and then computing the normalized class probabilities z_j, where

$$y_j = \sum_{i=1}^{4096} w_{ij} x_i \quad \text{and} \quad z_j = \frac{e^{y_j}}{\sum_{j=1}^{1000} e^{y_j}}. \qquad (1)$$

Estimating image-typicality

- Their assumption: typicality (human) is related to the strength of the model's classification response (strength) to the category of interest.
- Classification-Strength is estimated in two ways:

Raw typicality

- The raw category score
- Maximize y(j): particular abstract representation of category member that if inserted leads to max activity

$$y_j = \sum_{i=1}^{4096} w_{ij} x_i$$

Contrast typicality

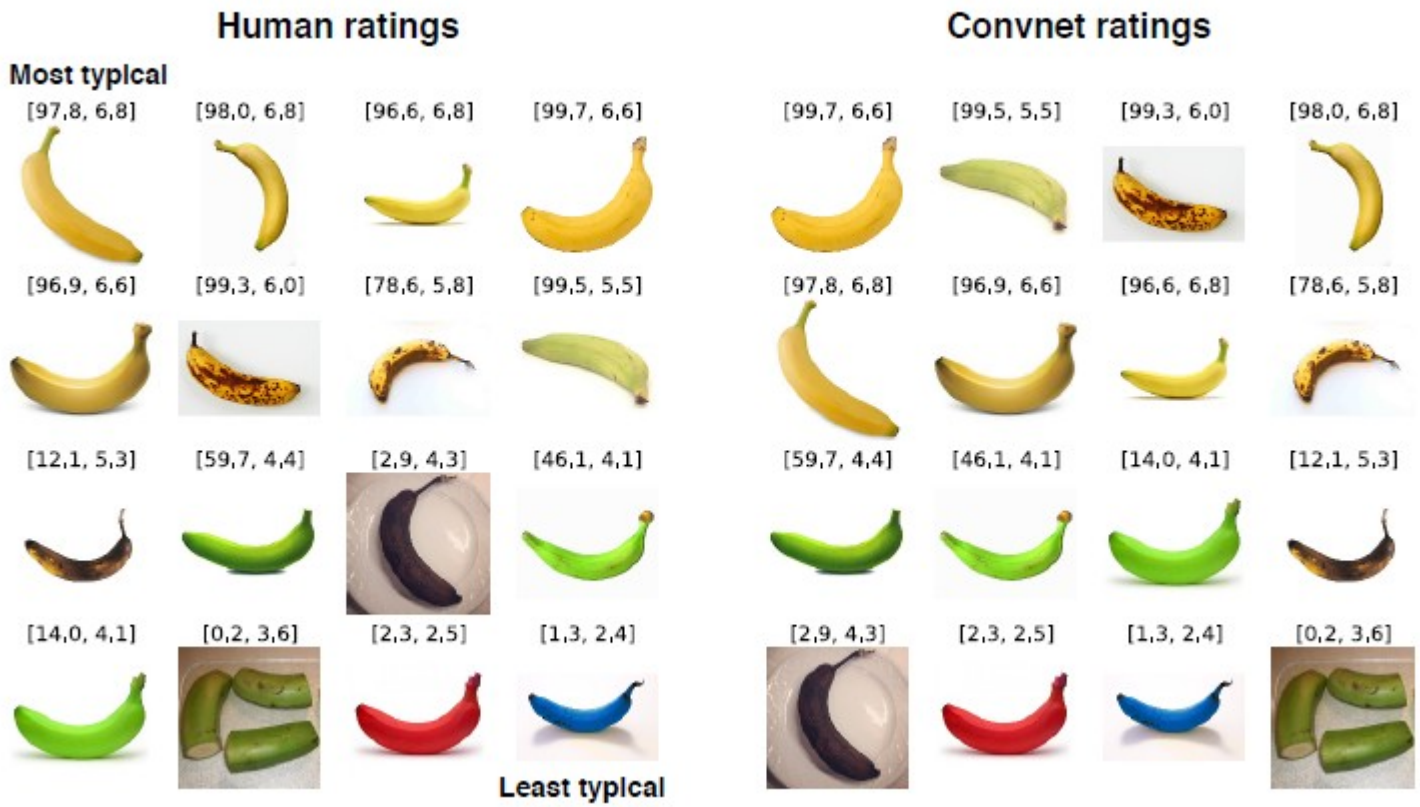
- Benefits images that load on the correct category much more than on other ones.
- Most typical image produces Y(j) that is most differentiated from other categories' response to this image. Indep' from raw value.

$$z_j = \frac{e^{y_j}}{\sum_{j=1}^{1000} e^{y_j}}.$$

Table 1: Rank correlations for human and machine typicality.

Category	OverFeat	AlexNet	GoogLe	Combo	SIFT
Banana	0.82	0.8	0.73	0.84	0.4
Bathtub	0.68	0.74	0.48	0.78	0.39
Coffee mug	0.62	0.84	0.84	0.85	0.63
Envelope	0.79	0.62	0.75	0.78	0.38
Pillow	0.67	0.55	0.69	0.59	0.11
Soap Disp.	0.74	0.79	0.82	0.75	0.09
Table lamp	0.69	0.8	0.7	0.83	0.48
Teapot	0.38	0.21	0.07	0.28	-0.23
Average	**0.67**	**0.67**	**0.63**	**0.71**	**0.28**

Result: Raw and Contrast scores do similarly well


Human ratings
Most typical
[97,8, 6,8]
[98,0, 6,8]
[96,6, 6,8]
[99,7, 6,6]
[96,9, 6,6]
[99,3, 6,0]
[78,6, 5,8]
[99,5, 5,5]
[12,1, 5,3]
[59,7, 4,4]
[2,9, 4,3]
[46,1, 4,1]
[14,0, 4,1]
[0,2, 3,6]
[2,3, 2,5]
[1,3, 2,4]
Least typical
Convnet ratings
[99,7, 6,6]
[99,5, 5,5]
[99,3, 6,0]
[98,0, 6,8]
[97,8, 6,8]
[96,9, 6,6]
[96,6, 6,8]
[78,6, 5,8]
[59,7, 4,4]
[46,1, 4,1]
[14,0, 4,1]
[12,1, 5,3]
[2,9, 4,3]
[2,3, 2,5]
[1,3, 2,4]
[0,2, 3,6]

Interim Conclusion

- “Our results suggest that deep convnets learn graded categories that can predict human typicality ratings, at least for some types of everyday categories”

Examination within layers

- For each layer, the average activation vector from 1300 training images (not experiment images) was computed for each class to serve as the category prototype.
- Typicality was modeled as the cosine distance between the activation vector for a new image and the stored prototype.
- Better prediction in deeper conv layers.

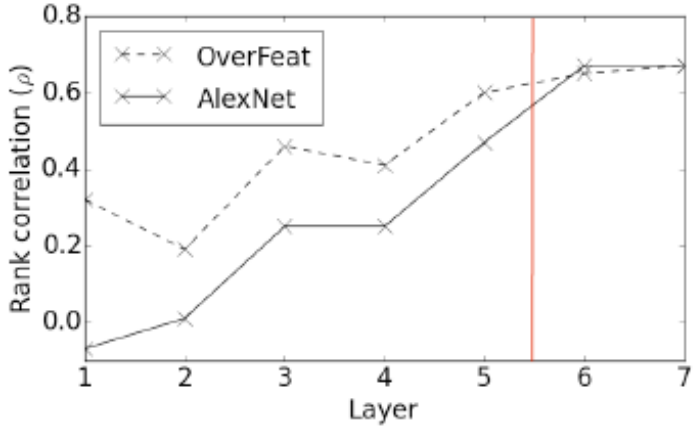


Figure 3: Correlation between human and convnet typicality ratings as a function of network depth. The red line indicates a transition from convolutional (1-5) to standard layers (6-7).