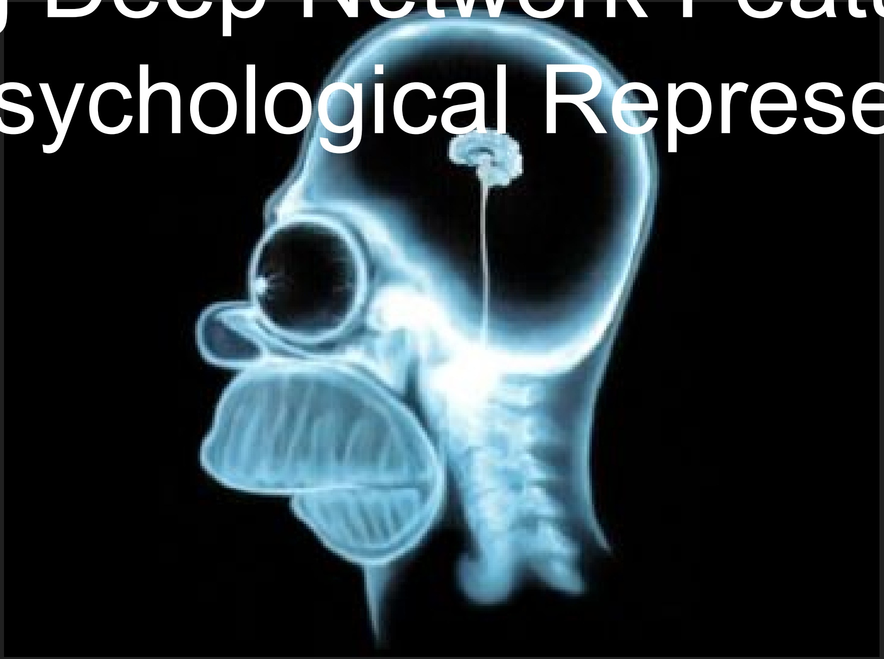
Adapting Deep Network Features to Capture Psychological Representations



Background

- Big question: can representations in deep neural networks be used to predict/understand human psychological representations?

Background

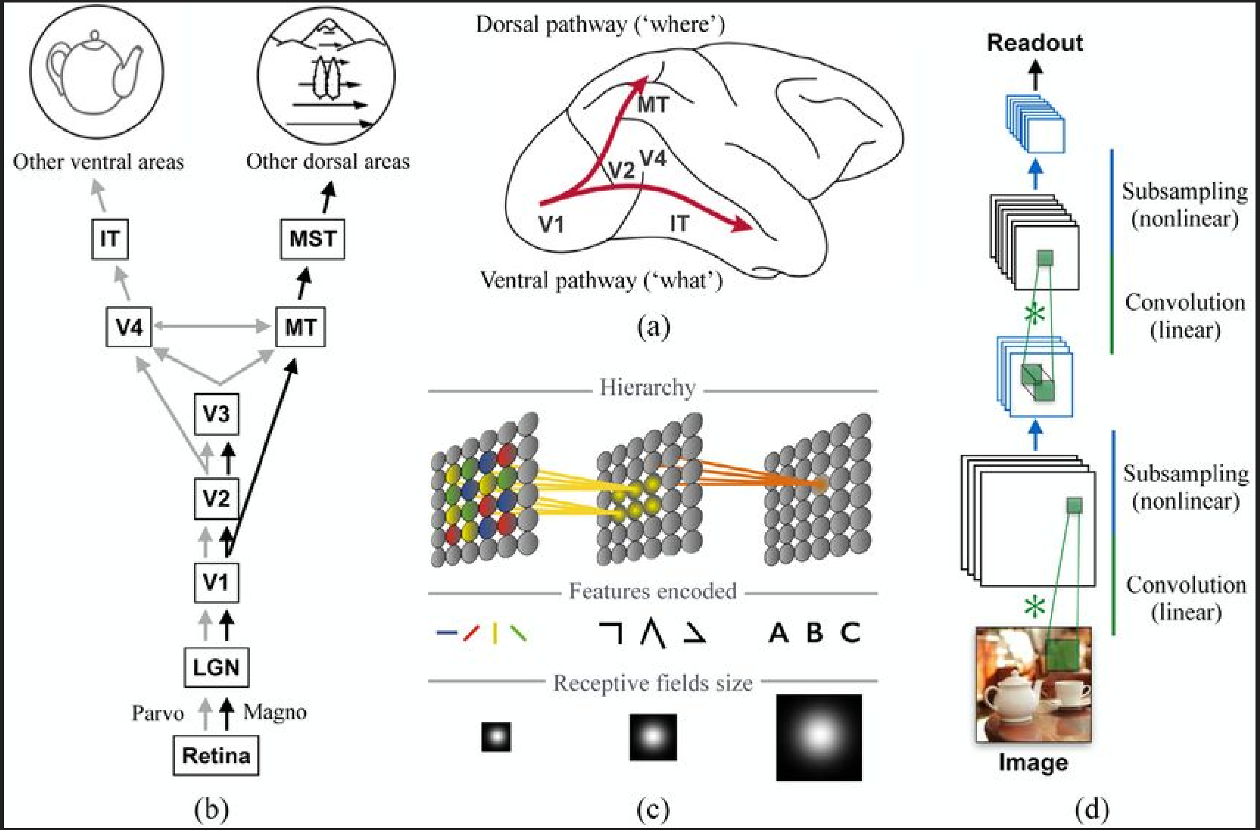
- Big question: can representations in deep neural networks be used to predict/understand human psychological representations?
  - Both are doing (e.g.) “object recognition”, but are they really solving similar problems?

Background

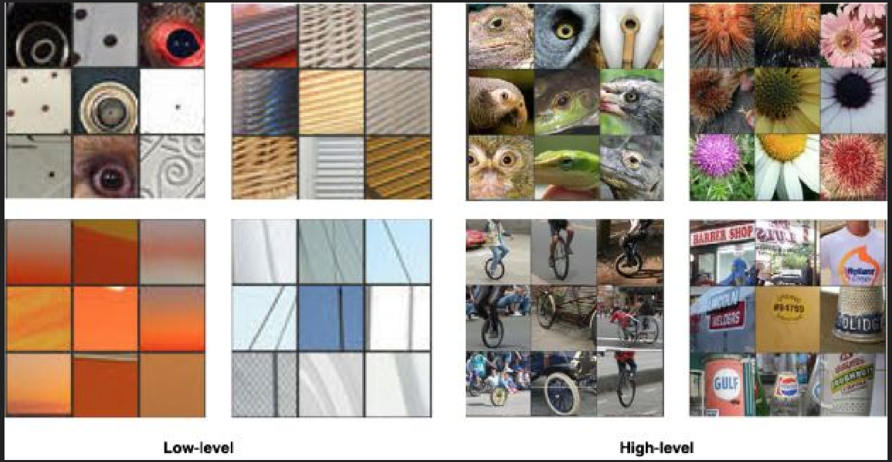
- Big question: can representations in deep neural networks be used to predict/understand human psychological representations?
  - Both are doing (e.g.) “object recognition”, but are they really solving similar problems?
  - If DNNs and humans are solving the same problem, are they doing it in a similar way?

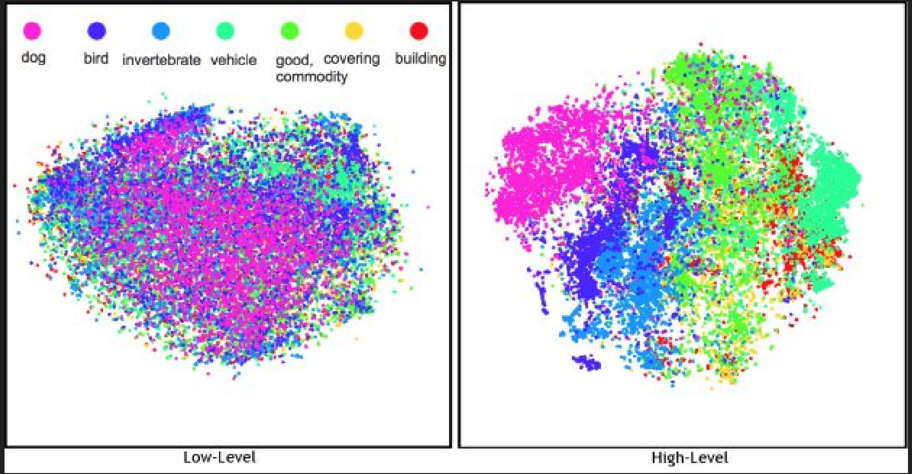
Background

- Big question: can representations in deep neural networks be used to predict/understand human psychological representations?
  - Both are doing (e.g.) “object recognition”, but are they really solving similar problems?
  - If DNNs and humans are solving the same problem, are they doing it in a similar way?
- Reasons to think visual representations may be similar
  - Neural networks have the word “neural” in them
  - Both systems learn to represent useful, potentially abstract features of objects for categorization
  - Human visual system & CNNs are both “feed-forward”, increasing in abstraction


Dorsal pathway ('where')
MT
V2
V4
V1
IT
Ventral pathway ('what')
(a)
Other ventral areas
Other dorsal areas
IT
MST
V4
MT
V3
V2
V1
LGN
Parvo
Magno
Retina
(b)
Hierarchy
Features encoded
A B C
Receptive fields size
(c)
Readout
Subsampling (nonlinear)
Convolution (linear)
Subsampling (nonlinear)
Convolution (linear)
Image
(d)

Low-level -> High-level in CNN



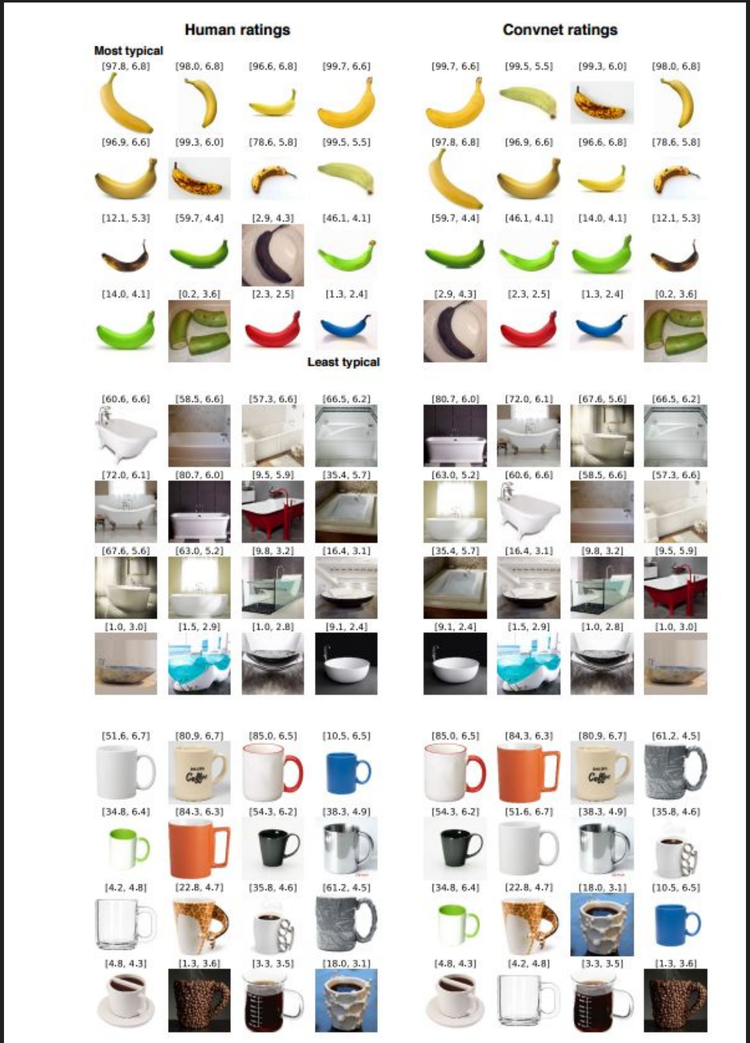


Previous attempts to link CNNs to Human Psychological Representations

- Object typicality
  - How well can DNN representations predict human typicality ratings?
- Object memorability
  - How well can DNN representations predict how well humans remember certain objects?

Typicality

Which is a more “typical” dog?



Lake et al. (2015)

Typicality

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**

Object memorability

- What makes some features more memorable than others?

Dubey et al. (2015)

Object memorability

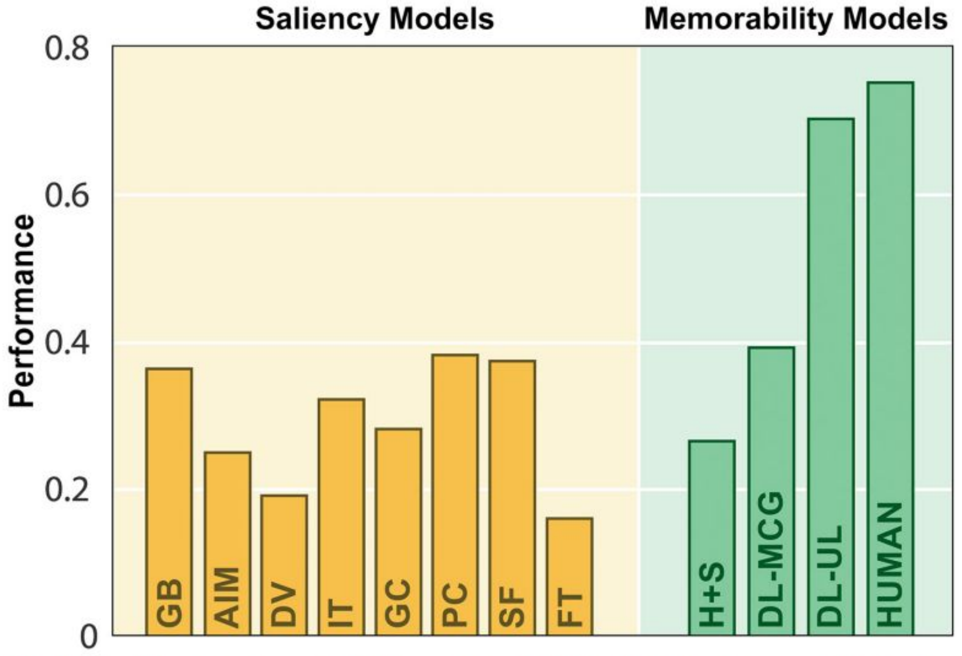


Figure 14: **Rank correlation of predicted object memorability.** Accuracy of the baseline and saliency algorithms on proposed benchmark.

Dubey et al. (2015)

So why are convnets not just people, then?

Tricked by, e.g., images on right in each block, which look identical to humans

(a) (b)

Szegedy et al. (2014)

Methods

- Data set
- Behavioral Experiment
- Deep Network Representations
- Adapted Network Representations
- Representation Comparison

Data Set

120 300 x 300 color photographs of animals

Behavioral Experiment

Constructing Image Similarity Matrix

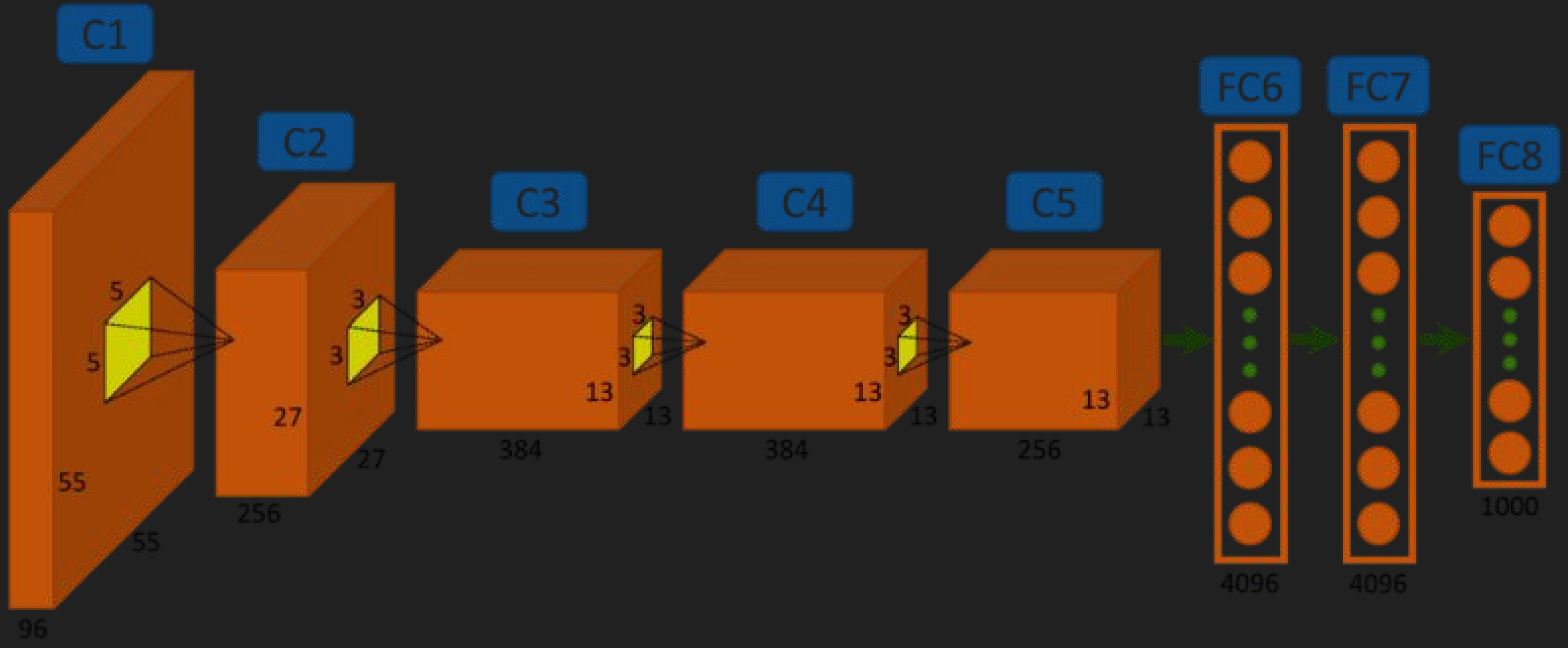
Amazon Mechanical Turk workers shown pairs of images

Each image pair rated from 0-10 similarity by 10 different workers

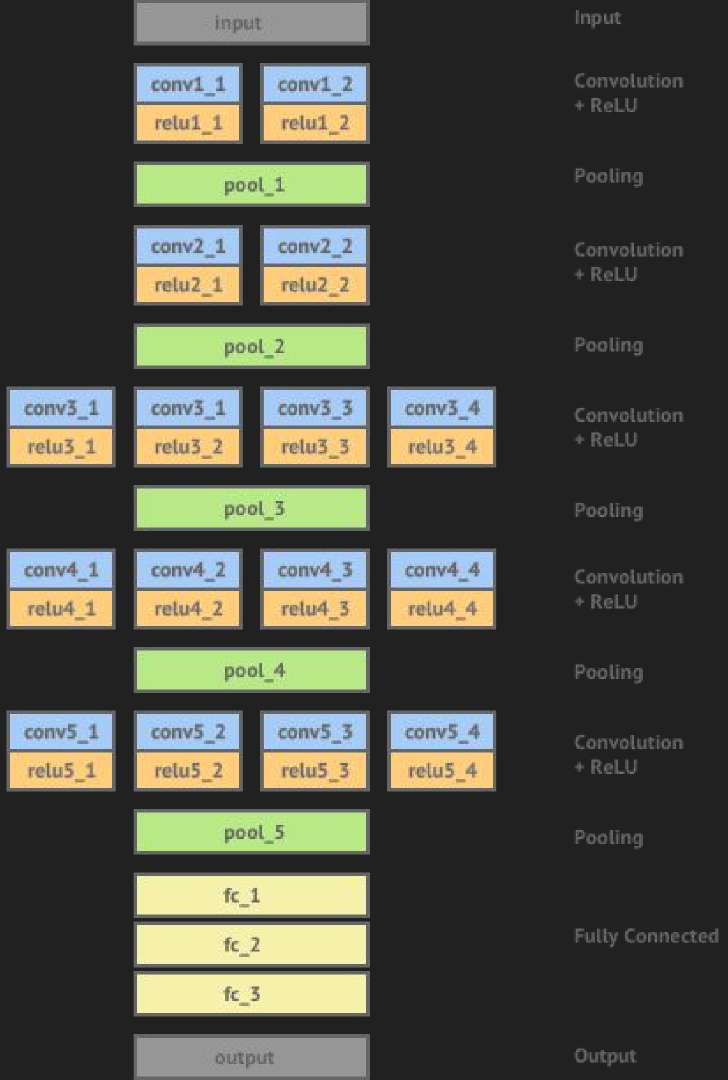
Average similarity rating used in matrix

Deep Network Representations

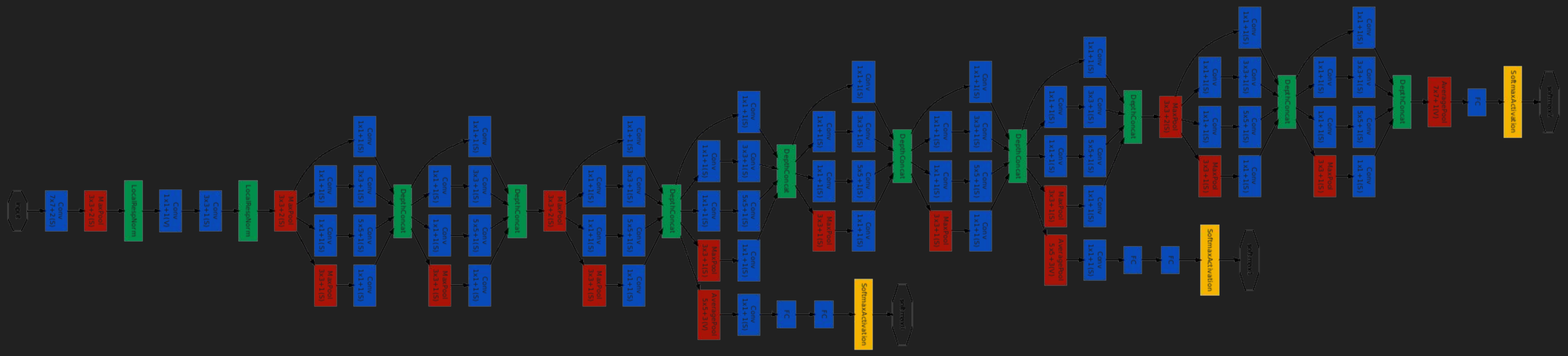
CaffeNet (AlexNet)



VGG-16


input
Input
conv1_1
conv1_2
relu1_1
relu1_2
Convolution + ReLU
pool_1
Pooling
conv2_1
conv2_2
relu2_1
relu2_2
Convolution + ReLU
pool_2
Pooling
conv3_1
conv3_1
conv3_3
conv3_4
relu3_1
relu3_2
relu3_3
relu3_4
Convolution + ReLU
pool_3
Pooling
conv4_1
conv4_2
conv4_3
conv4_4
relu4_1
relu4_2
relu4_3
relu4_4
Convolution + ReLU
pool_4
Pooling
conv5_1
conv5_2
conv5_3
conv5_4
relu5_1
relu5_2
relu5_3
relu5_4
Convolution + ReLU
pool_5
Pooling
fc_1
fc_2
fc_3
Fully Connected
output
Output

GoogLeNet



Convolution

Pooling

Concat

SoftMax

Representation Comparisons

Comparing Neural Network output with human generated similarity matrix

Inner product of image representation vectors is used as a measure for similarity

Correlation between these inner products and human generated similarities

Results:

Table 1: Correlations between human and deep similarities.

	CaffeNet	Google	VGG	HOG+SIFT
R^2	.32	.35	.43	.008

BASELINE

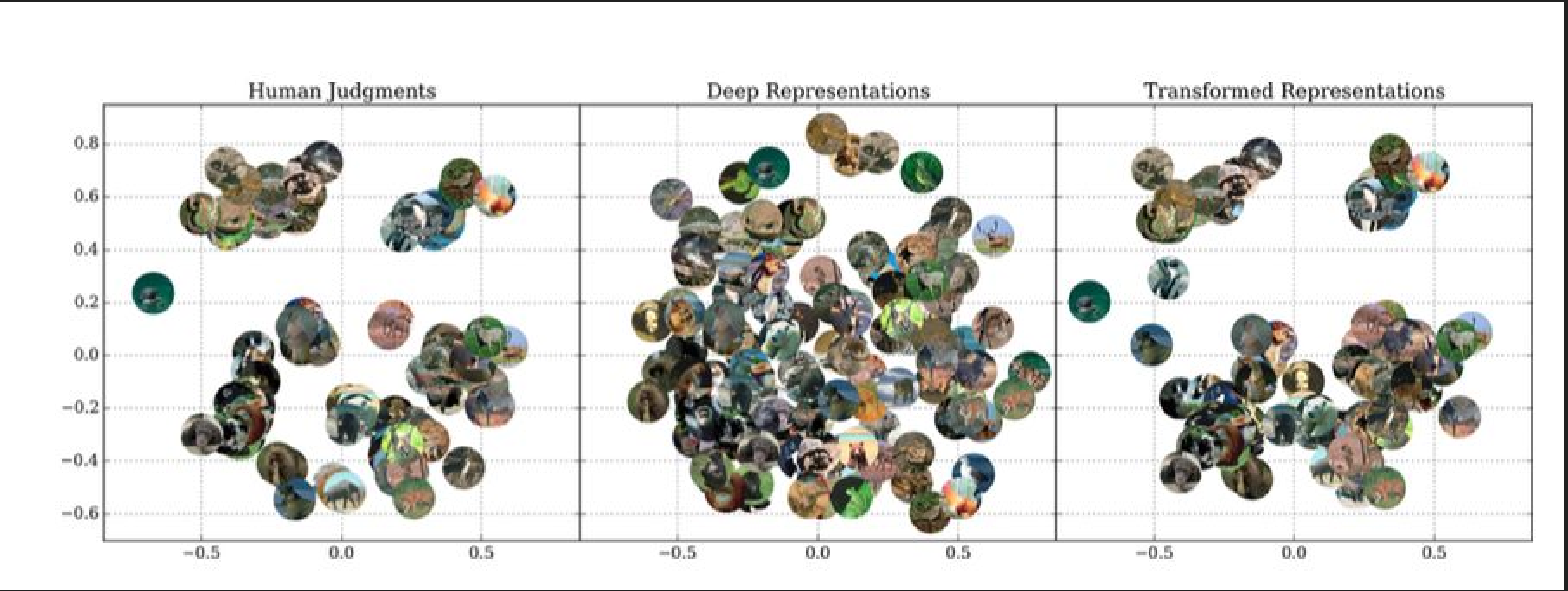
- In general, deeper CNNs perform better.
- HOG + SHIFT:
    - features used produce high classification accuracy in machine vision tasks
    - differ from those that humans use for judging animal similarity

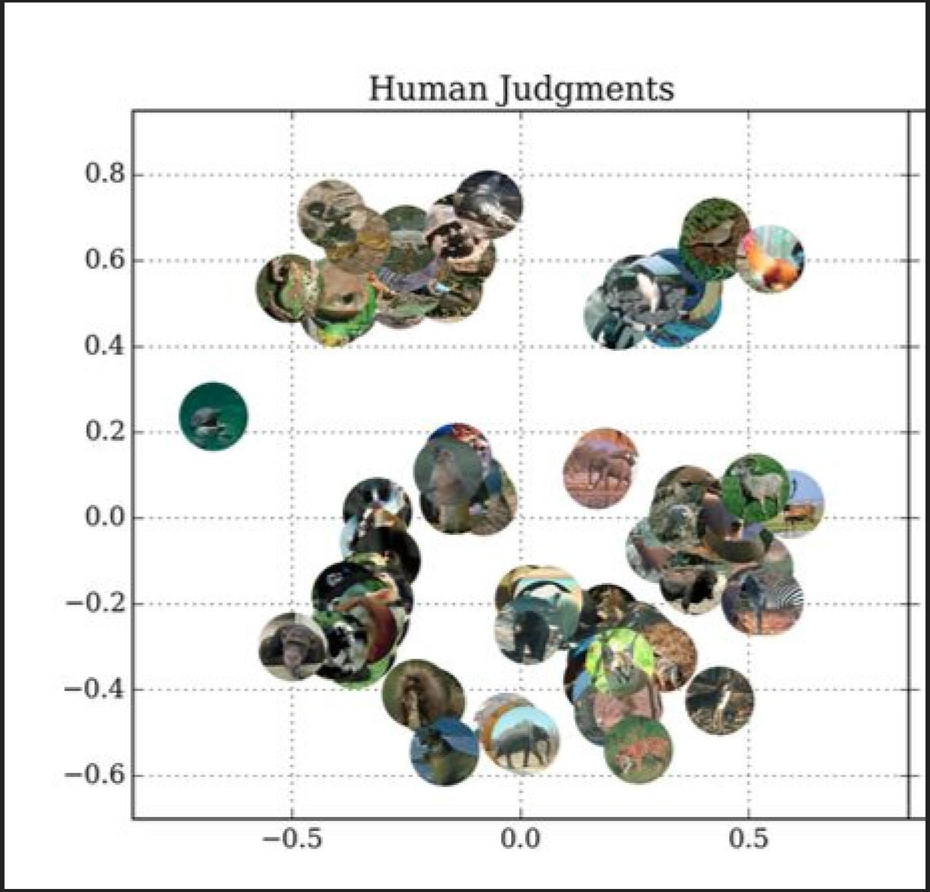
Results: VGG

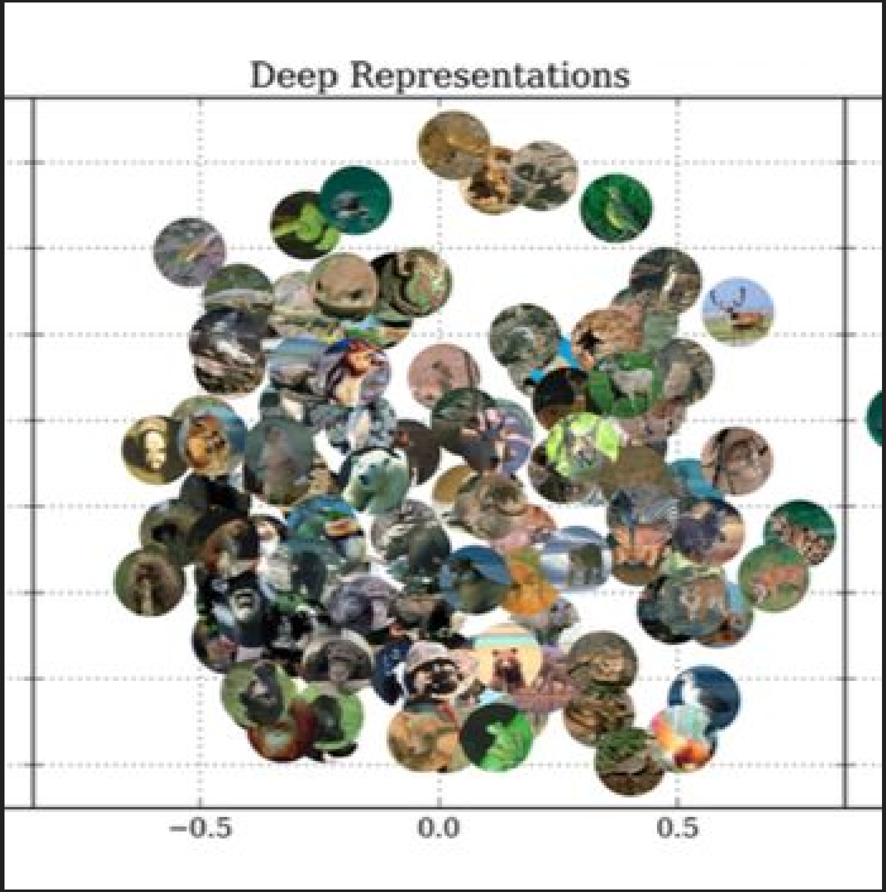
- VGG: Although much of the variance is accounted for, structural aspects of human representations were not preserved

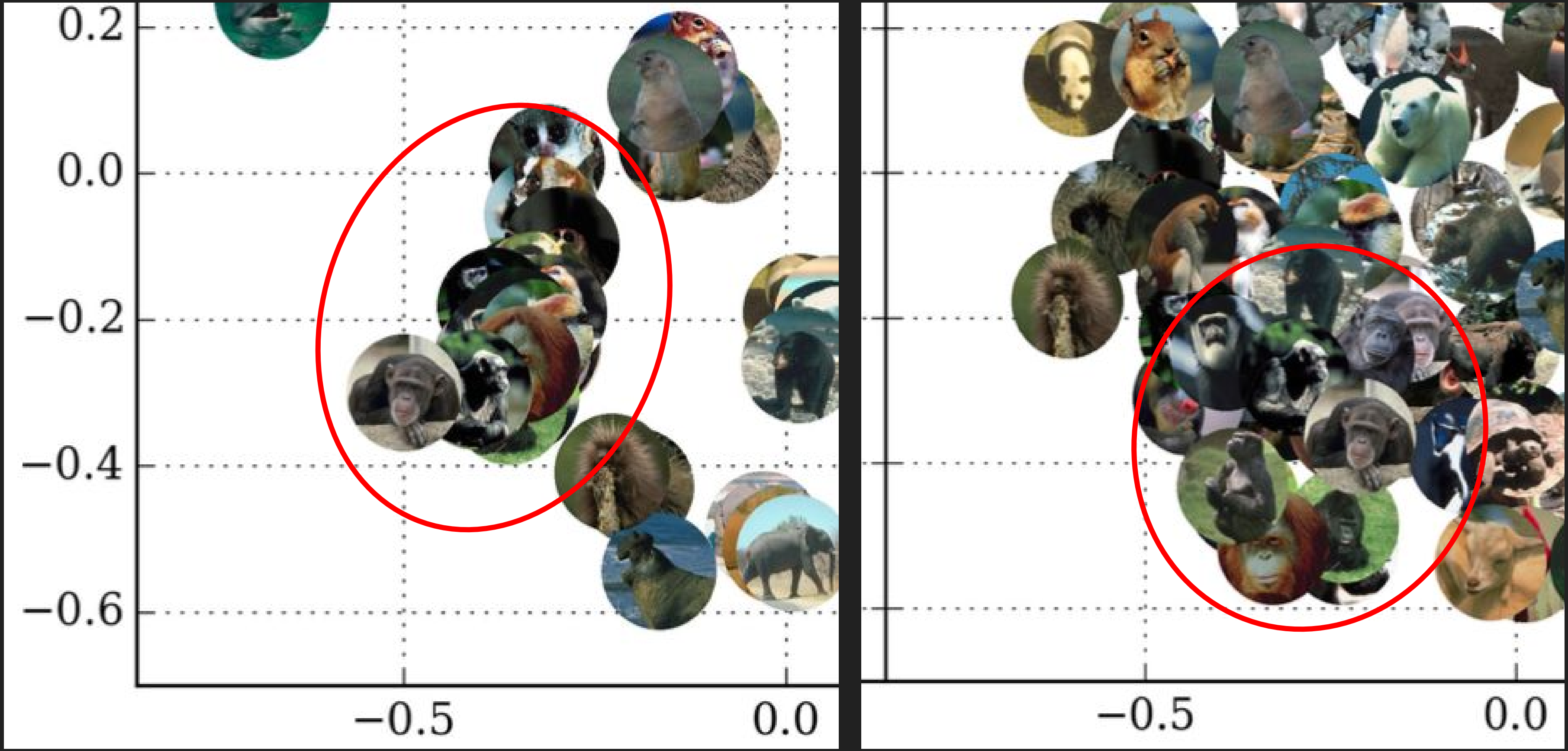
Results: VGG

- Human judgements exhibit several major categorical divisions
- This structure is lost in the predicted data



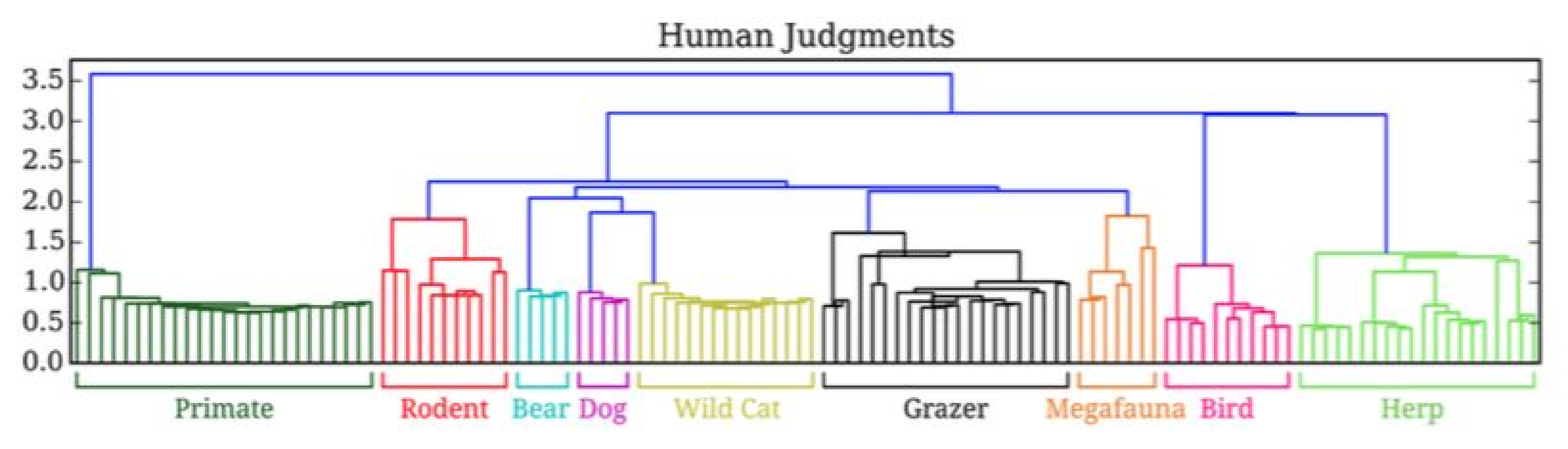

Human Judgments
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.5
0.0
0.5


Deep Representations
−0.5
0.0
0.5

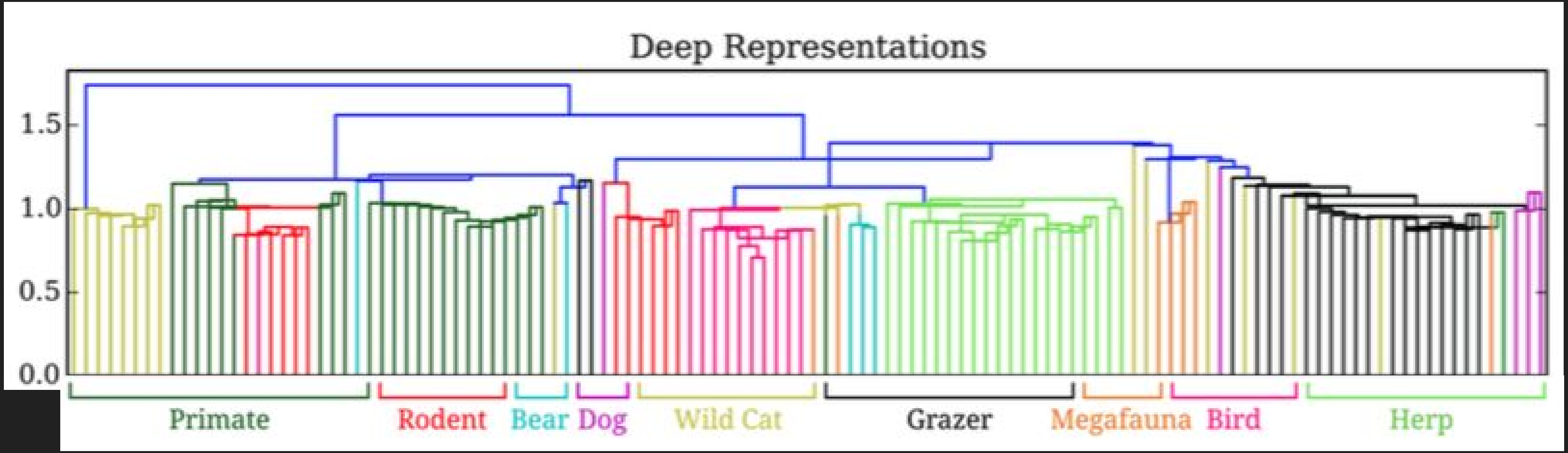


0.2
0.0
−0.2
−0.4
−0.6
−0.5
0.0
−0.5
0.0

Results: Hierarchical Clustering



Results: Hierarchical Clustering



Human representations

Adapted Network Representations

Last layer to classification is generally a linear transform

Solve for the transformation from network to user similarity with linear regression

This is done with L2 regularization and cross-validation to avoid overfitting

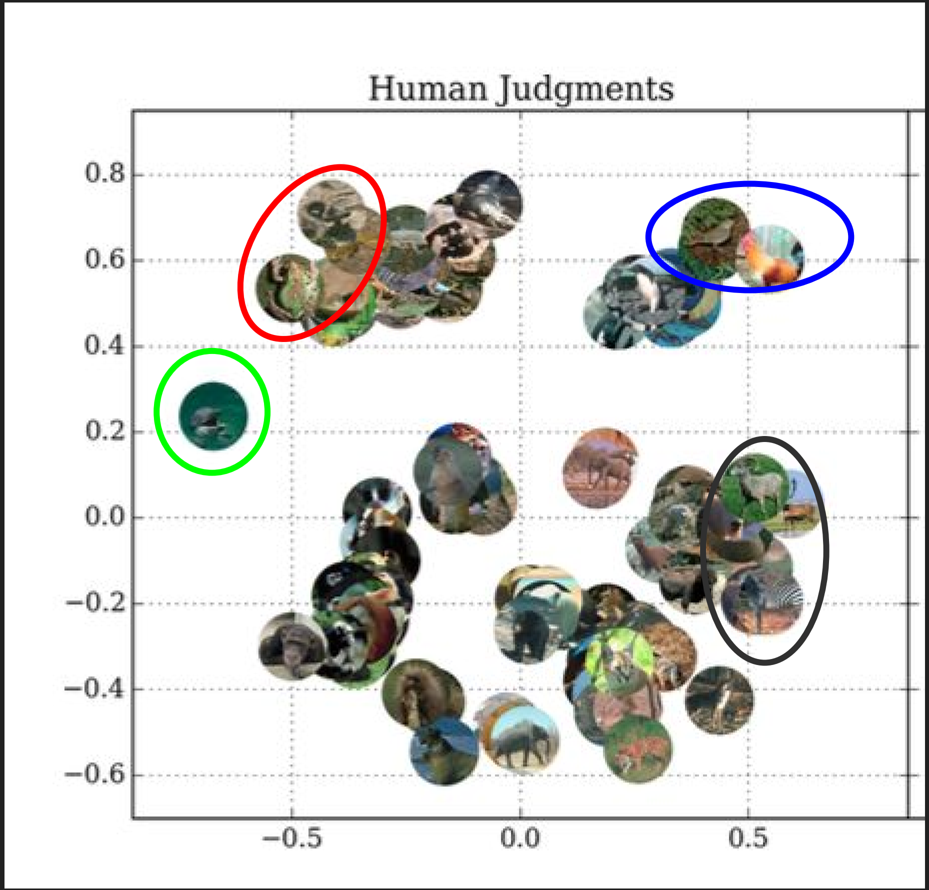
Results: Adapted Network Representations

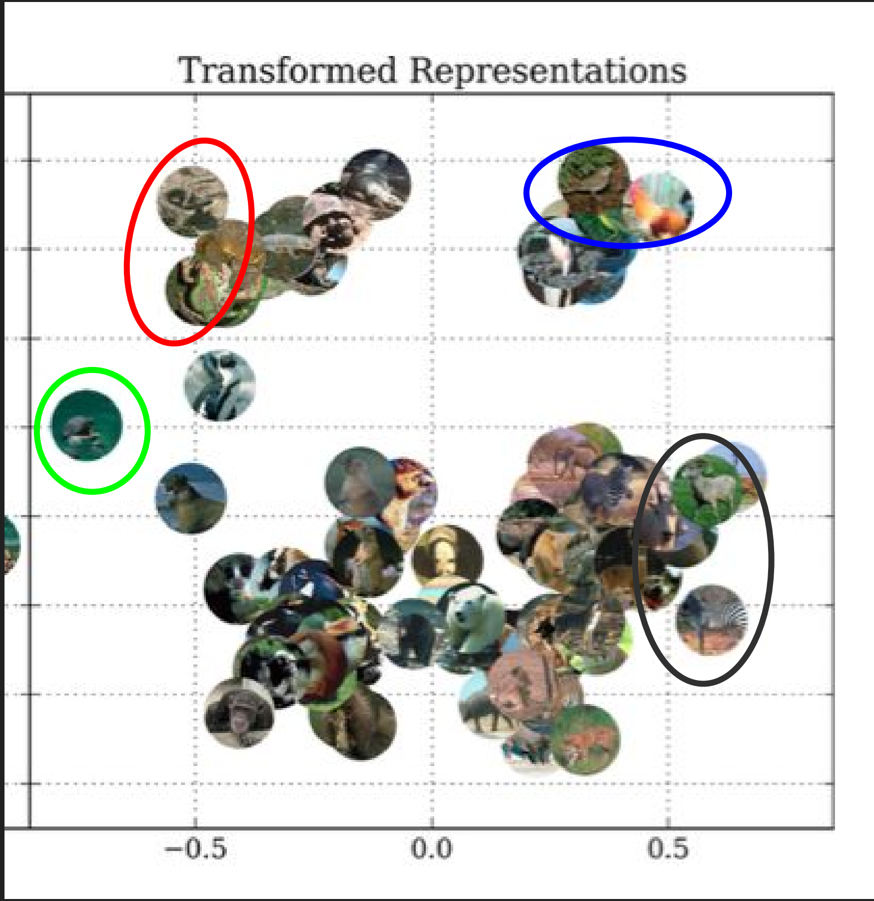
Table 2: Model performance using adjusted CNN features.

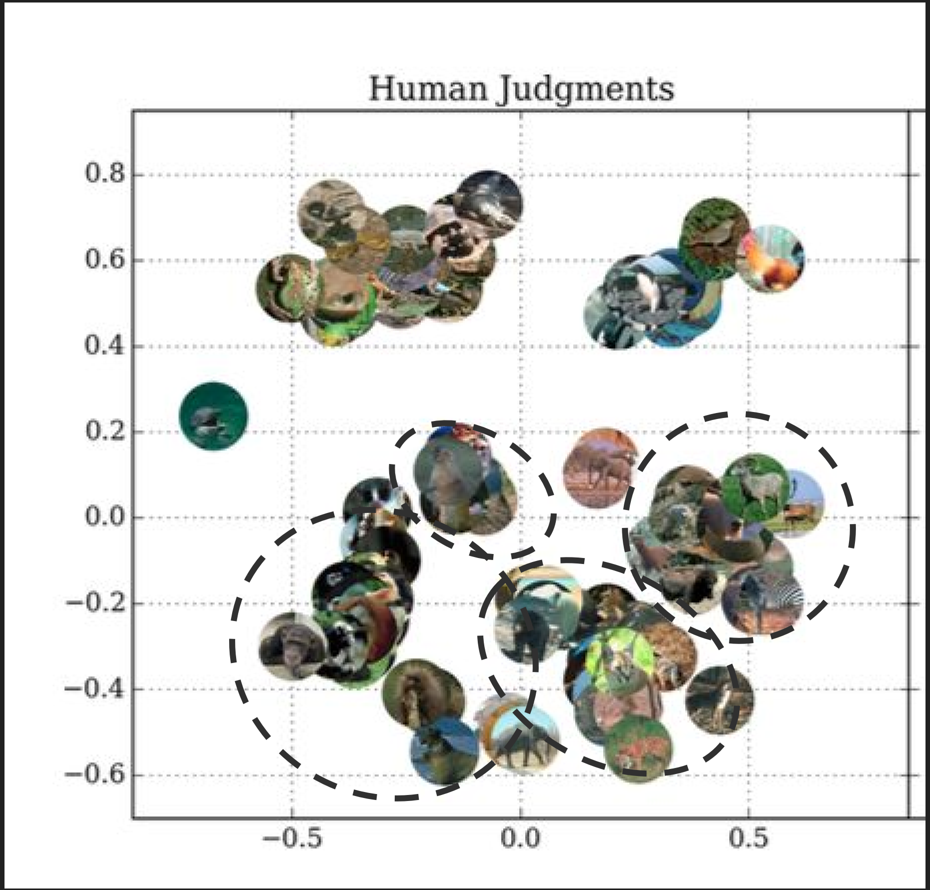
	CaffeNet	Google	VGG	SIFT
R^2	.69	.72	.84	.09

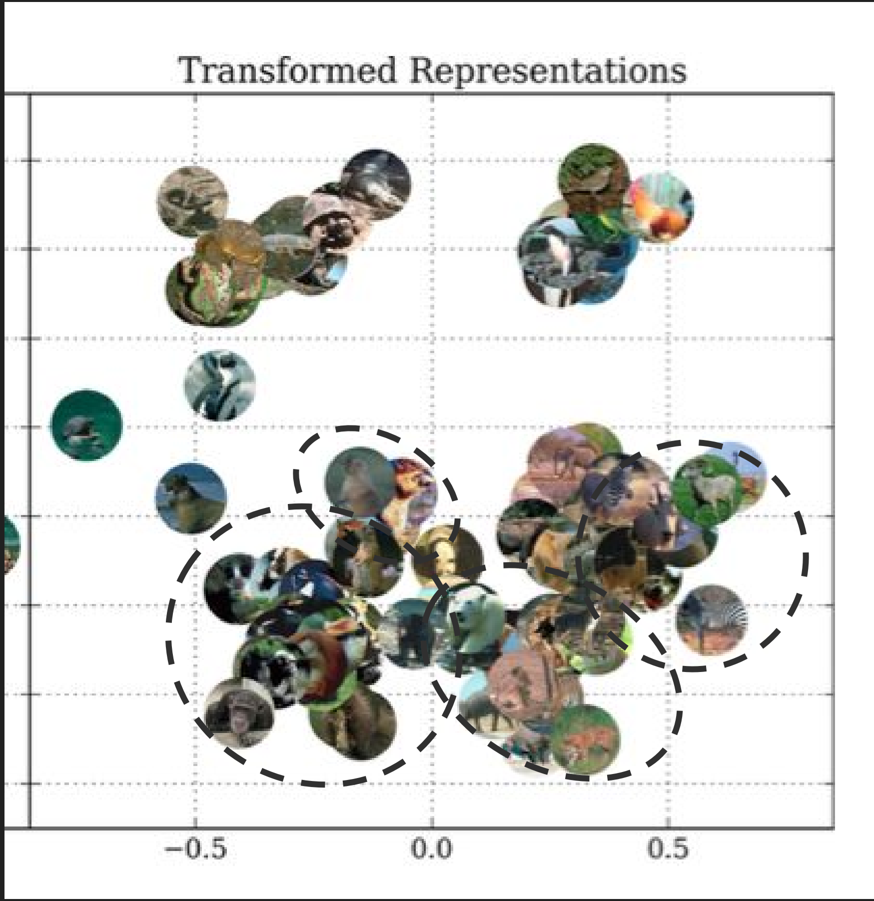
BASELINE

- Average of 6-fold cross validation
- VGG: almost identical to human spatial representation

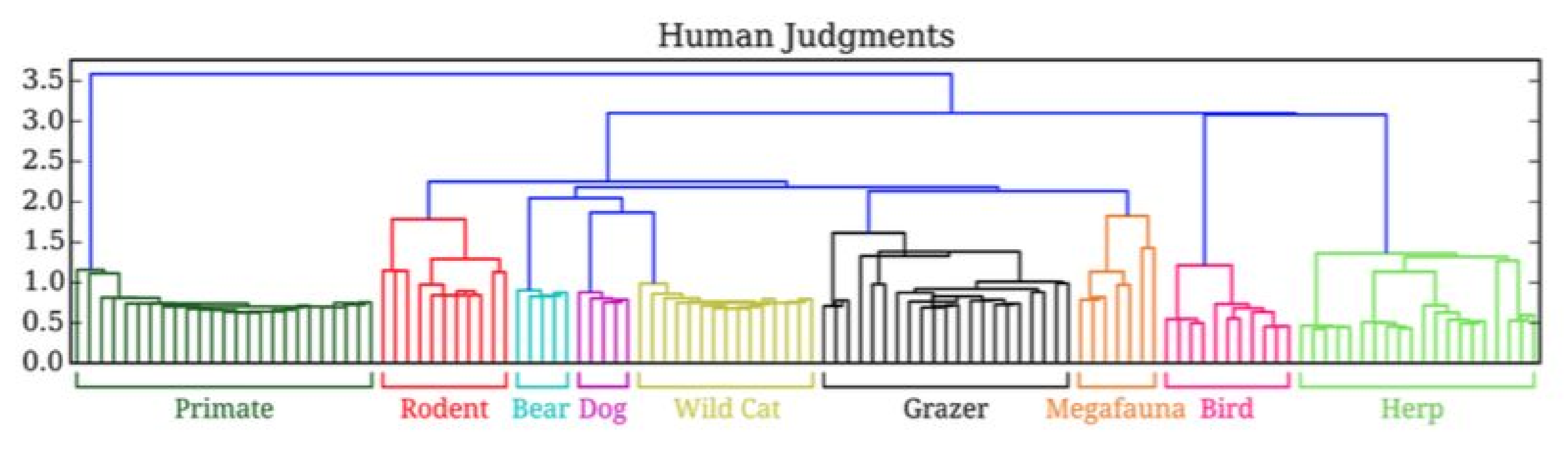

Human Judgments
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.5
0.0
0.5


Transformed Representations
−0.5
0.0
0.5

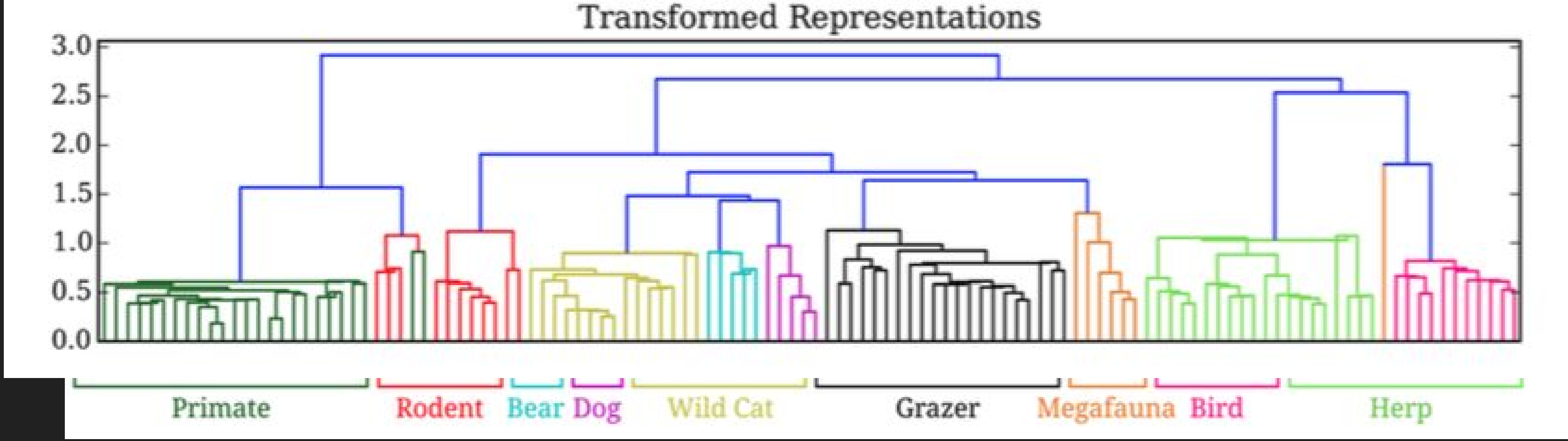

Human Judgments
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.5
0.0
0.5


Transformed Representations
−0.5
0.0
0.5

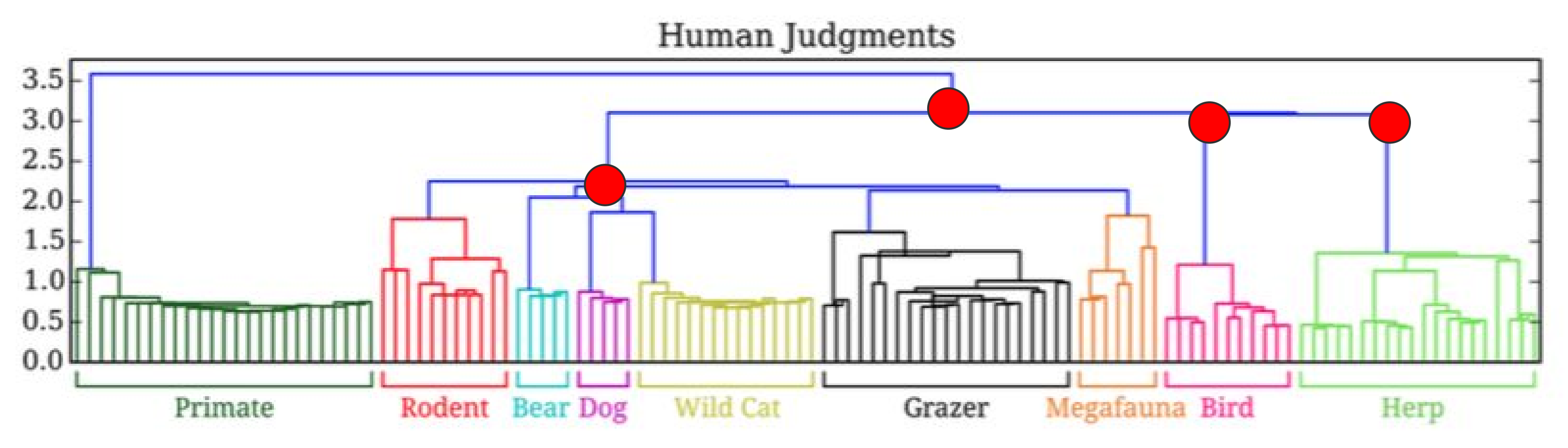
Results: Hierarchical Clustering



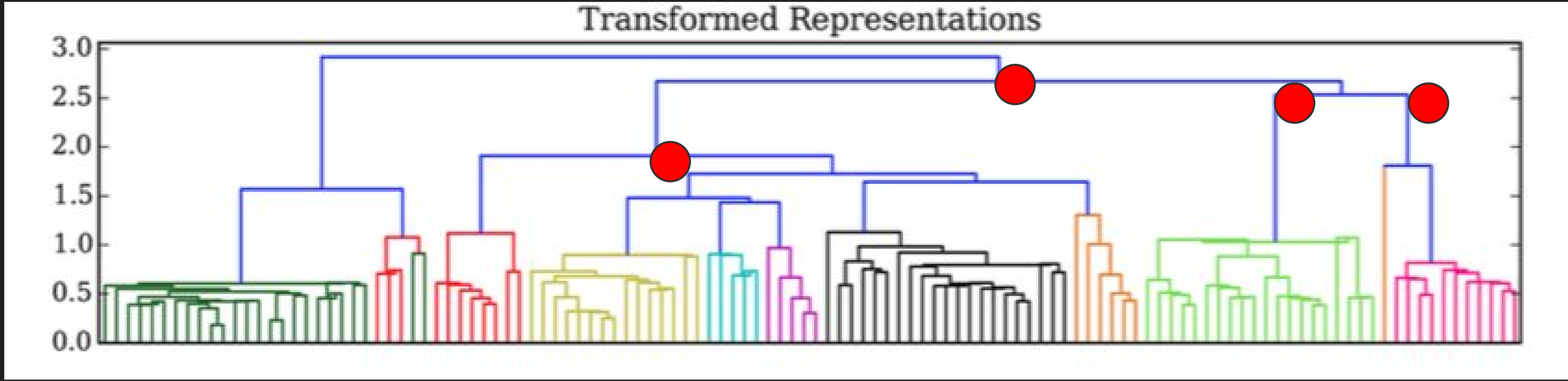
Results: Hierarchical Clustering



Human representations



Human Judgments
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Primate
Rodent
Bear
Dog
Wild Cat
Grazer
Megafauna
Bird
Herp



Transformed Representations
3.0
2.5
2.0
1.5
1.0
0.5
0.0

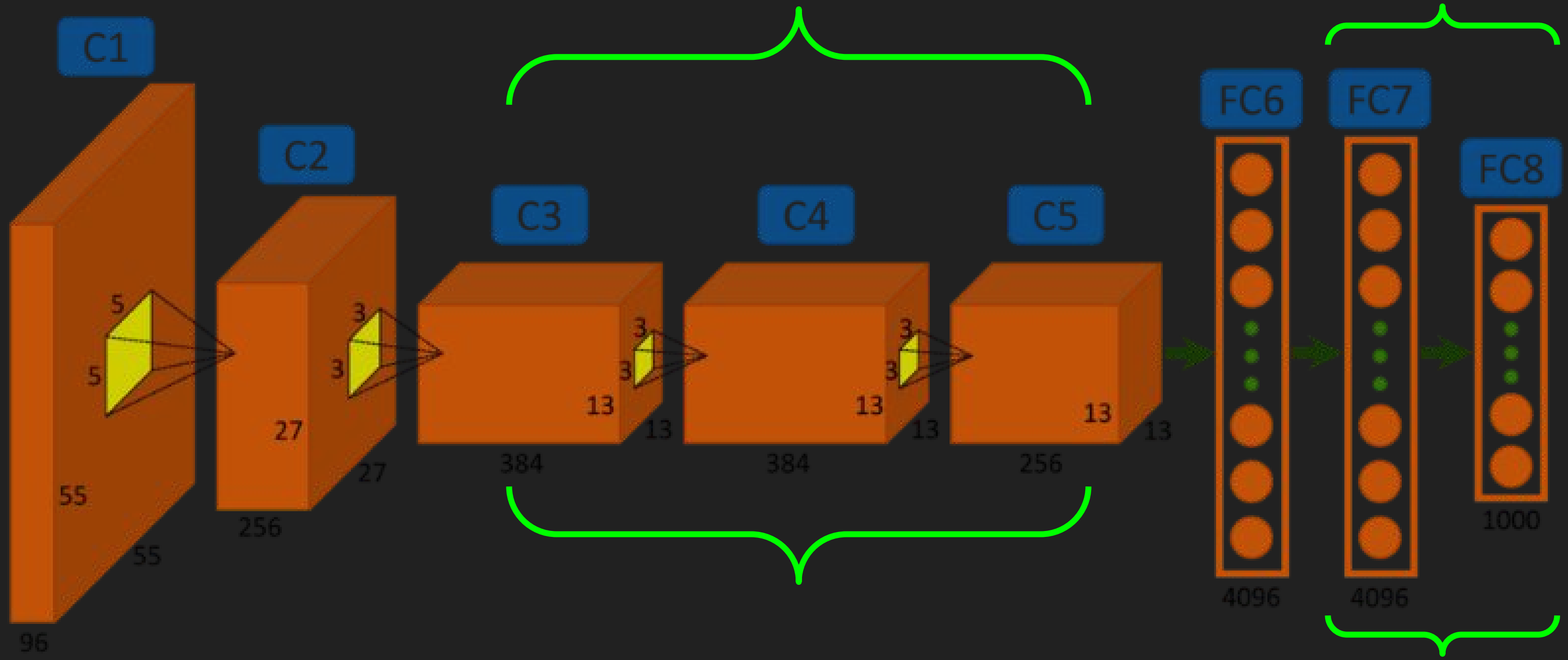
Feature Analysis

- Higher levels in CNN yield the most generic features
- Allowing for domain transfer, but the feature depth depends on the task
- Thus, implying that layer responses at different depths may explain different types of human similarity judgements
  - Conceptual information vs. visual information

Feature Analysis

- Evaluated model performance on predicting similarity judgments
- CaffeNet (Alex Net)

Feature Analysis: CaffeNet



Feature Analysis: CaffeNet

- Performance appears to correspond to layer depth
- Fully connected layers perform better than convolutional layers
    - Human similarity judgements may not be explained well by simpler image features

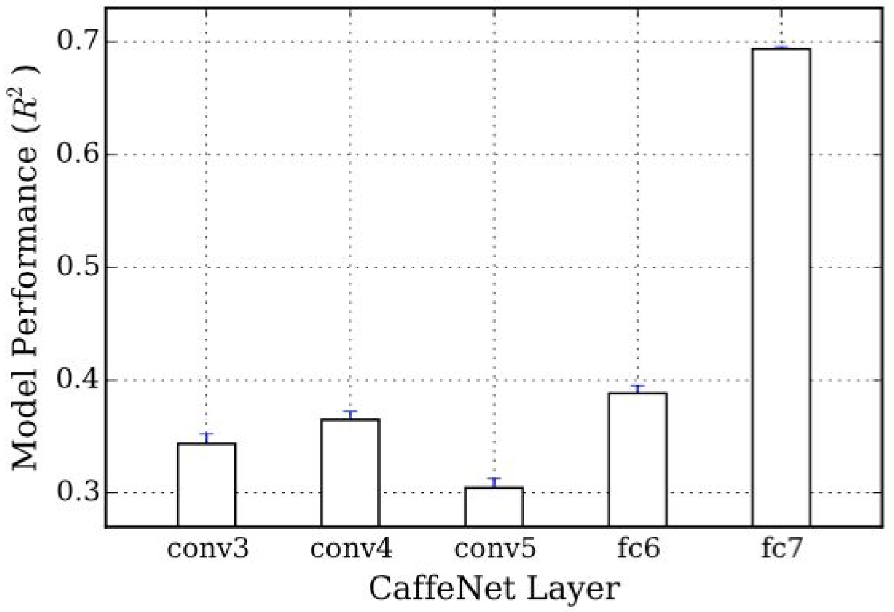
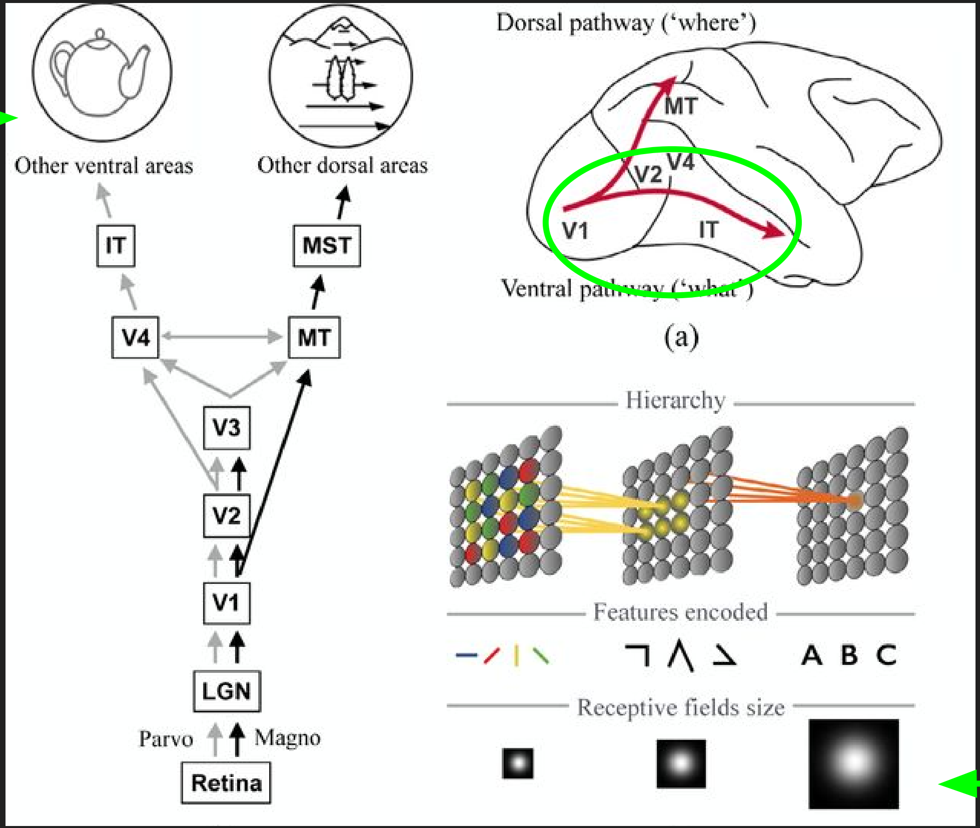
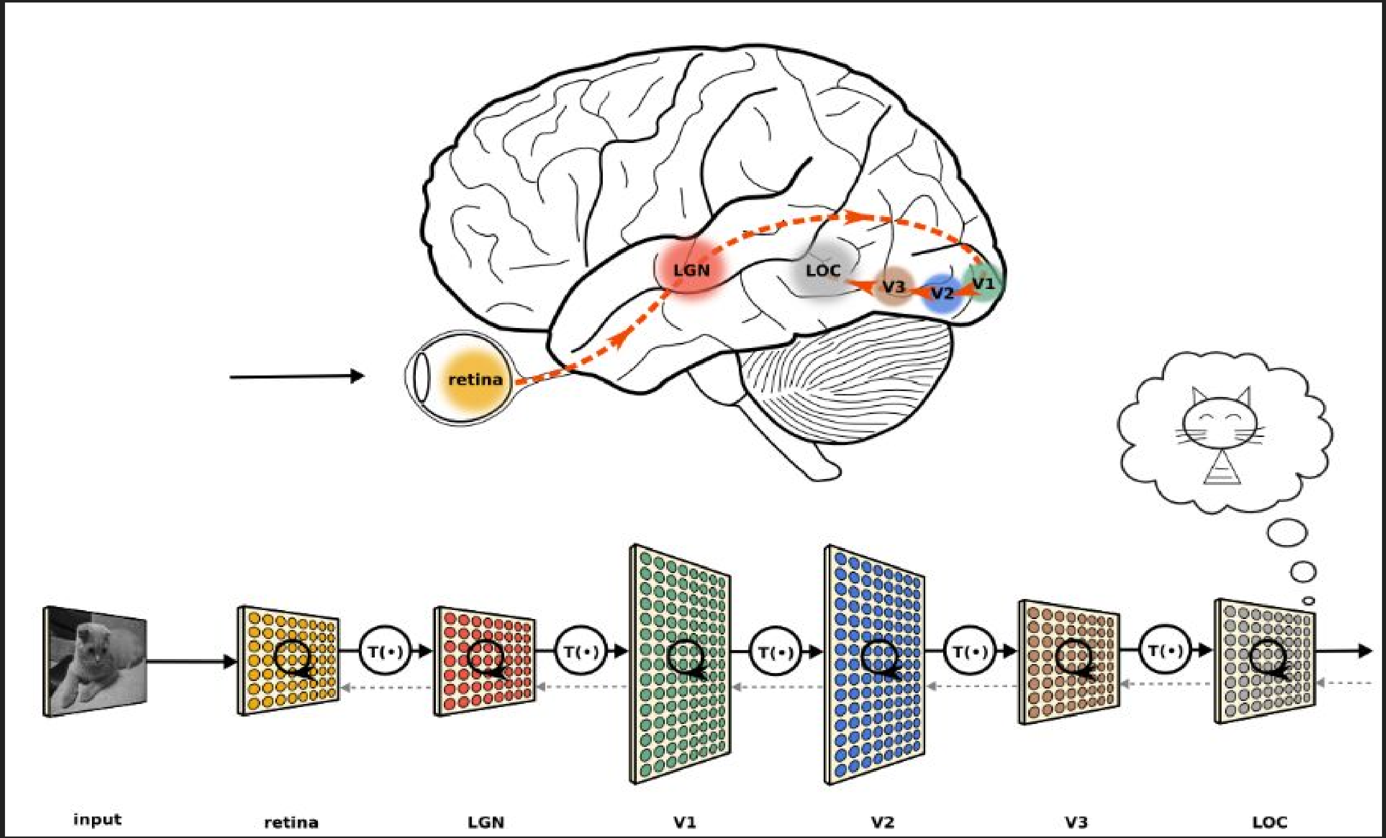


Figure 4: Model performance as a function of feature layer depth in CaffeNet.

Feature Analysis: Visual Processing



Feature Analysis: Visual Processing



Multinomial classification:

- Using human representations as predictors
- Multinomial logistic regression with 6-fold cross validation
- BASELINE: Original VGG16 representations

Multinomial classification:

VGG16	Fine-tuned
$R^2 = 0.94$	$R^2 = 0.89$

Limitations

- Although structure was preserved in the animal classification task, it may not generalize well across domains
- Human categorization behavior exhibits complex patterns like overlapping class assignments which cannot be captured when training data is assigned as single label
- There is a distinction between the computational problems solved by humans and those solved by CNNs
- Others??

Concluding Remarks:

- Adjustment of feature representation through a similarity model successfully preserves the structure of human psychological representations in deep networks
- Fully connected layers outperform convolutional layers in predicticting human similarity judgments
- Using human representations as predictors does not improve accuracies in one-versus-all classification problems
- Beginning to interface cognitive science and A.I.