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?
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
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)
Table 1: Rank correlations for human and machine typicality.

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

- 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

Table 1: Correlations between human and deep similarities.

<table>
<thead>
<tr>
<th></th>
<th>CaffeNet</th>
<th>Google</th>
<th>VGG</th>
<th>HOG+SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>.32</td>
<td>.35</td>
<td>.43</td>
<td>.008</td>
</tr>
</tbody>
</table>
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
Results: Hierarchical Clustering
Results: Hierarchical Clustering
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

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

Human representations
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:

<table>
<thead>
<tr>
<th></th>
<th>VGG16</th>
<th>Fine-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.94</td>
<td>0.89</td>
</tr>
</tbody>
</table>
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 predicting 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.