Modeling categorical search guidance using a convolutional neural network designed after the ventral visual pathway

Gregory J. Zelinsky¹ and Chen-Ping Yu²

¹Department of Psychology, Stony Brook University; ²Department of Psychology, Harvard University

Introduction

- Most of our everyday searches are for categories of things, and a growing body of evidence now exists that attention is guided to target object categories in the context of a visual search task (e.g., [1,3]). But computational models of this categorical guidance of attention are still in their infancy. In previous work we showed that a simple generative model was able to predict this guidance by learning category-consistent features (CCFs)—those features that occur both frequently and consistently across the exemplars of an object category [3]. However, this model’s prediction was limited to a single general relationship; more time is needed to first fixate a target as this target climbs levels in a subordinate-basic-superordinate category hierarchy.

- This restricted scope was likely due to our use of outdated features and methods (SIFT, Bag-of-Words) in this CCF model. Here we extend this work by modeling attentional guidance to individual target categories. We do this again by using CCFs, but now extract these features using a Convolutional Neural Network (CNN) that is modeled after the primate ventral stream—VsNet.

Selecting CCFs using a CNN

- As in our previous work [3], we use the number of CCFs extracted for each category to predict categorical guidance behavior, measured as the time until fixation on a target.

VsNet Architecture

- Each convolutional layer corresponds to a specific area in the primate ventral visual pathway. Note that the ‘V4-like’ layer combines V4 with LO18 & [6] into a single layer.
- Filter sizes at each layer reflect estimates of receptive field sizes in human [4,5].
- The relative numbers of filters across layers are based on estimates of average brain surface areas in humans [6-7].
- Bypass connections between layers reflect known connectivity between brain areas in primate, specifically, V1→V4, V2→TEO, and V4→TE [8,9].
- We compared VsNet to the original BoW-CCF model [3], AlexNet [11], and a convolutional version of the HMAX model [10] that we designed and implemented (details upon request).

Model Comparison to Behavior

As in our previous work [3], we use the number of CCFs extracted for each category to predict categorical guidance behavior, measured as the time until fixation on a target.

- All four CCF models replicate the subordinate-level advantage found in across-level categorical guidance.
- However, VsNet was the best at predicting gaze time-to-target for individual categories at all three hierarchical levels (48, 16, and 4, respectively). It’s performance is also at the subject noise ceiling, defined as one standard deviation from the subject model mean.
- VsNet also makes reasonable predictions, that early layers drive subordinate and basic-level guidance and that higher layers drive superordinate guidance. Layer-specific predictions from AlexNet and Deep-HMAX are less clear.

Object Classification and CCF Visualization

- VsNet beats other CNNs in image classification!

<table>
<thead>
<tr>
<th>ImageNet</th>
<th>AlexNet</th>
<th>Deep-HMAX</th>
<th>VsNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 Accuracy</td>
<td>57.7%</td>
<td>59.6%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Top-5 Accuracy</td>
<td>80.6%</td>
<td>82.4%</td>
<td>83.9%</td>
</tr>
</tbody>
</table>

VsNet outperformed the other CNNs in large-scale image classification, despite having the least convolutional filters and not designed to optimize classification accuracy.

- CNN-CCFs do object detection for free!

Specific object categories can be localized by combining their CCF activation maps.

- What do VsNet CCFs look like? Visualized are the 5 most responsive CCF filters at each layer for the taxi and passenger airplane categories; CCFs look like object parts!

Conclusions

- Look to the brain when building CNN models of behavior. VsNet, a CNN designed after the ventral visual stream, outperformed less biologically-inspired models in image classification, as well as predicting guidance to individual target categories better than other CNN-CCF models.
- CNN-CCFs do object detection for free. CNN-CCFs, learned without object location information, enabled object categories to be localized in images.

References


[7] Yu, C.


