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

Gregory J. Zelinsky[1] and Chen-Ping Yu[2]

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

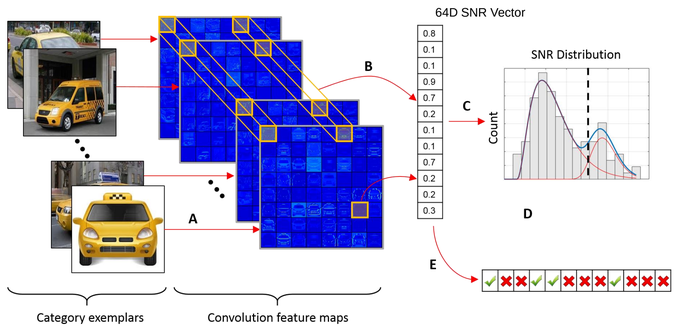
EYE MOVEMENTS AND VISUAL COGNITON

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., [1,2]). 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- This restricted scope was likely due to our use of outdated features and methods (SIFT, Bag-of-Words) in this CCF model.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



A – Input a given set of category exemplars through a CNN.

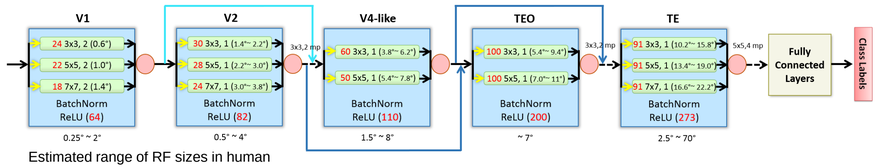
B – For every convolutional filter at each layer, compute a signal-to-noise ratio (SNR) across the category exemplars based on the filter's response frequency and variability.

C – Fit a two-component Gamma mixture model to the distribution of SNRs computed for the filters; the cross-over point indicates the CCF threshold.

D – Apply this threshold to the SNR values to select CCFs.

E – Filters having above threshold SNRs are retained as the CCFs for a given category; below threshold features are pruned away.

VsNet Architecture

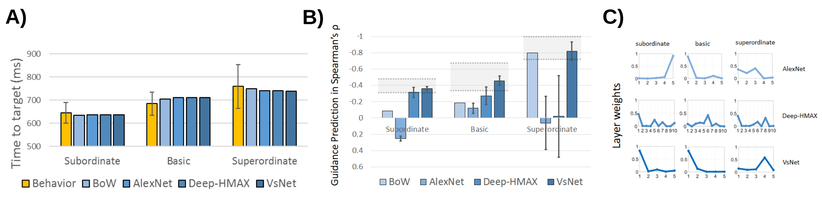


Estimated range of RF sizes in human

- Each convolutional layer corresponds to a specific area in the primate ventral visual pathway. Note that the "V4-like" layer combines hV4 with LO1&2 [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 first fixation on a target.



A – All four CCF models replicate the subordinate-level advantage found in across-level categorical guidance.

B – 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.

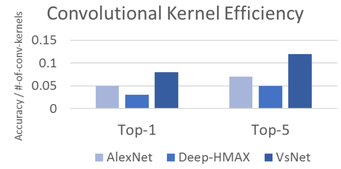
C – VsNet also makes reasonable predictions, that early layers drive subordinate and basic-level guidance and that higher layers drive superordinate guidance. Layer-specific predictions from AlexNet and Deep-HMAX are less clear.

Object Classification and CCF Visualization

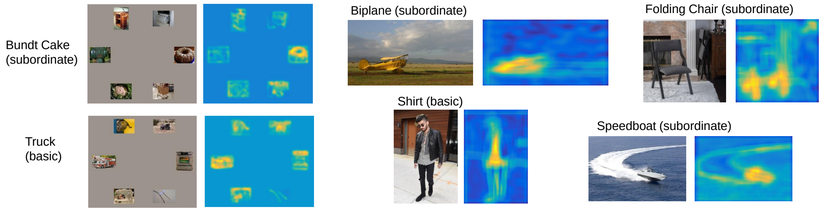
- **VsNet beats other CNNs in image classification!**

ImageNet	AlexNet	Deep-HMAX	VsNet
Top-1 Accuracy	57.7%	59.6%	**61.5%**
Top-5 Accuracy	80.6%	82.4%	**83.9%**

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

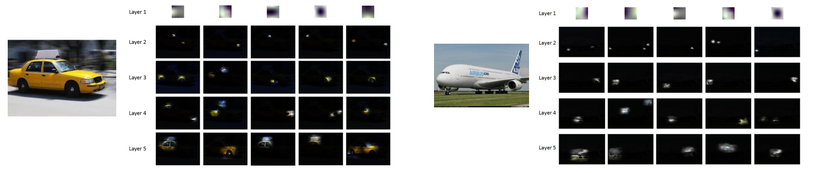


- **CNN-CCFs do object detection for free!**



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

[1] Zelinsky, G. J., Peng, Y., Berg, A. C., & Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. Journal of Vision, 13(3):30, 1-20.
[2] Nako, R., Wu, R., & Eimer, M. (2014a). Rapid guidance of visual search by object categories. Journal of Experimental Psychology: Human Perception and Performance, 40, 50-60.
[3] Yu, C.-P., Maxfield., J.T., Zelinsky, G.J. (2016). Searching for category-consistent features: a computational approach to understanding visual category representation. *Psychological Science*, 27(6):870-884.
[4] Kastner S., et al. (2001). Modulation of sensory suppression: implications for receptive field sizes in the human visual cortex. *Journal of Neurophysiology*, 86(3):1398-1411.
[5] Harvey, B. and Dumoulin, S. (2011). The relationship between cortical magnification factor and population receptive field size in human visual cortex: constancies in cortical architecture. *Journal of Neuroscience*, 31(38):13604-13612.
[6] Larsson, J. and Heeger, D. (2006). Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, 26(51):13128-13142.
[7] Van Essen, D., et al. (2001). Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision Research*, 41(10-11):1359-1378.
[8] Nakamura, H., Gattass, R., Desimone, R., and Ungerleider, L. (1993). The modular organization of projections from areas v1 and v2 to areas v4 and teo in macaques. *Journal of Neuroscience*, 13(9):3681-3691.
[9] Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Current opinion in neurobiology*, 7(4):523-529.
[10] Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424-6429.
[11] Krizhevsky, A., Sutskever, I. & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.