A Goal-driven Model of Visual Search in Natural Scenes Replicates Human Behavior While Relying on Similar Neural Representations

Motahareh Pourrahimi (motahareh.pourahimi@mail.mcgill.ca)
Integrated Program in Neuroscience, 1033 Pine Ave. W.
Montreal, Quebec, H3A 1A1, Canada

Irina Rish (irina.rish@mila.quebec)
6666 St-Urbain Street, 200
Montreal, Quebec, H2S 3H1, Canada

Pouya Bashivan (pouya.bashivan@mcgill.ca)
Physiology, McGill University, 3655 Promenade Sir William Osler
Montreal, Quebec, H3G 1Y6, Canada

Abstract

Visual search, the process of locating a specific item among multiple objects, is a key paradigm in studying visual attention. Due to eccentricity-dependent visual acuity, many animals constantly selectively sample from their environment by moving their gaze location, leading to the formation of search scanpaths, a hallmark of visual search behavior. While much is known about the brain networks involved in visual search, our understanding of the neural computations driving this behavior is limited, leading to challenges in simulating such behavior in-silico. To address this gap, we trained an image-computable artificial neural network to perform visual search from pixels in natural scenes. Model's search scanpaths (spatiotemporal sequence of fixations) were highly consistent with those of humans. It captured the human information integration behavior and relied on neural representations similar to those observed in the primate fronto-parietal attentional control network. Examining the model's latent space revealed how it uses its internal state to construct and update a priority map of the visual space, enabling efficient visual search. Our model provides concrete predictions about the neural computations underlying visual search in the primate brain.

Keywords: visual search; saccadic behavior; artificial neural networks

Methods

Natural scene visual search datasets. We generated a large-scale natural scene search dataset by applying a state-of-the-art object detection model (He, Gkioxari, Dollár, & Girshick, 2017) on the Places image dataset (Zhou, Khosla, Lapedriza, Torralba, & Oliva, 2016), extracting masks for 80 object categories. Each search trial consisted of a cue frame showing an object from the cue category, the search image containing an example of the cued category, and the ground-truth target mask. We discretized the space by overlaying a 10x10 grid on the image (Fig. 1A). We tested the model's performance and behavioral alignment with humans on the COCO Search 18 dataset (Chen et al., 2021) (Fig. 1B). **Model architecture.** General-architecture Visual Search Model (GVSM) consists of a convolutional neural network (CNN) pretrained to perform object recognition on Imagenet (Deng et al., 2009), which has been shown to closely simulate neural activity in the visual cortex (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014), and a transformer assuming the role of the fronto-parietal attentional control network in guiding fixations (Fig. 1C). CNN receives a retinal image to incorporate the eccentricity-dependent visual acuity implemented by a multi-resolution crop module (Mnih, Heess, Graves, & Kavukcuoglu, 2014; Ba, Mnih, & Kavukcuoglu, 2015). **Model training.** GVSM was trained to perform visual search following a 2-stage training paradigm inspired by prior work on saccade-augmented visual categorization (Elsayed, Kornblith, & Le, 2019). 1) We train the model to predict the target area given a random sequence of fixations by backpropagation (Fig. 1D). 2) With fixed transformer parameters, we train an MLP policy network with RL to optimally select fixations (Fig. 1E).

Results

GVSM replicates human saccadic behavior. The model successfully generalized to the COCO Search 18 dataset (Fig. 2A-B). Its search scanpaths were highly consistent with those of humans as measured by the fixation by fixation scanpath prediction method (Kümmerer & Bethge, 2021), where the fixation probability maps conditioned on the saccadic history are compared between model and the humans using various metrics (Fig. 2C-G). Moreover, the model also replicates other human saccadic properties such as saccade sizes, directions, and their joint distribution (not shown).

Replicating humans' information integration during search. GVSM learns a human-like inhibition of return mechanism from being solely trained on the task. Each location's probability of being fixated consistently increases until fixation (the peak), suddenly decreasing afterwards. However, unlike the baselines with hand-engineered inhibition of return mechanisms, the probability does not drop to zero and stays at a nonzero lower level, resembling human behavior in which return fixations happen but with lower probability (Fig.

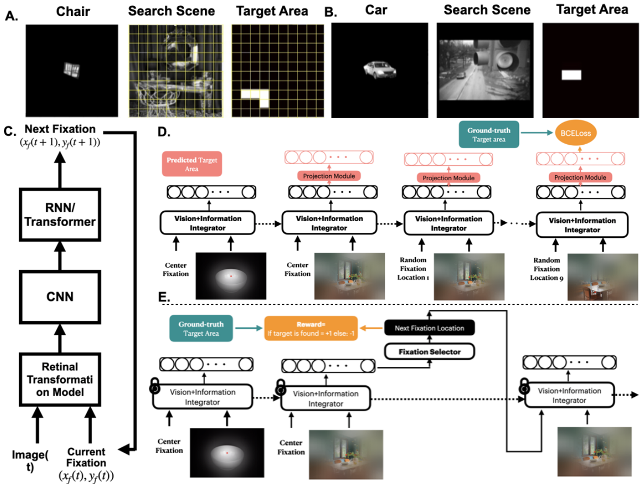

Figure 1: **A.** Example Places Search trial data with chair as the cued category. **B.** Example COCO Search 18 trial data with car as the cued category. **C.** Model schematic. **D.** First stage of training: learning to predict the target area given a random sequence of fixations (supervised). **E.** Second stage of training: learning to optimally select fixations (RL).

3A). **Cue-similarity map emerges in GVSM latent space.** We identified robust representations of cue-similarity maps in GVSM's latent space, akin to those previously found in the primate's fronto-parietal attentional control network (N. Bichot, Heard, DeGennaro, & Desimone, 2015; Bisley & Mirpour, 2019; Machner et al., 2020; Colby & Goldberg, 1999). To do this, we fitted linear decoders to predict the cue similarity across space in both retinocentric (relative to gaze location) and allocentric (relative to image) reference frames (Fig. 3C-D). Cue similarities were computed by comparing the visual representation (final feature maps of the CNN) of the cue frame with that of 1) the fixated retinal image, resulting in a 3x3 retinocentric cue-similarity map, and 2) the full image, resulting in a 10x10 allocentric cue-similarity map.

Geometry of neural representations underlying visual search. The cue-similarity at different locations across the visual field could be encoded by a) separate neural subspaces regardless of their location (a discontinuous representation) or; b) neural subspaces which their pairwise distances (in the latent space) reflect the distances between their corresponding locations in the visual field (a continuous representation) (Fig. 3D). We found a *continuous representation* for both retinocentric and allocentric cue-similarity maps in GVSM latent space (Fig. 3E). Further, the generalization accuracy of the decoders trained on one location and tested on others decreased with increasing distance between their corresponding locations in the visual field, confirming a continuous topographical geometry (Fig. 3F). The allocentric cue-simialrity map was stably encoded in the same subspace across time, judging by high generalization accuracy of the cue-similarity decoders across time (Fig. 3G). However, the retinocentric

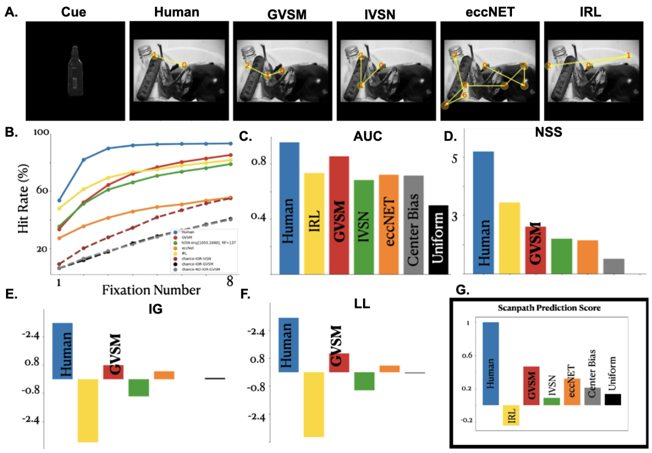

Figure 2: **A.** Example search scanpaths. **B.** Cumulative performance curve. **C.** Area Under the Curve. **D.** Normalized Scanpath Similarity. **E.** Information Gain. **F.** Log Liklihood. **G.** Aggregate scanpath prediction score (Mean of C-F).

cue-similarity map was encoded in a time-varying subspace, particularly changing from the first to the second time step (Fig. 3H) with its encoding subspace rotating across time (Fig. 3I-J; orthogonal Procrustes analysis).

Discussion

We show that an image-computable neural network model trained to perform visual search in natural scenes closely replicates human behavior and relies on hidden state representations closely resembling prior observations from the primate's fronto-parietal cortical network. We believe this model provides an opportunity for the community to test hypotheses about the neural computations underlying visual search, e.g. the fixation selection strategy as well as predicting neural responses of primate brain areas like the Ventral pre-arcuate (VPA), Lateral intraparietal cortex (LIP), and Frontal eye fields (FEF) during visual search.

Acknowledgments

This research was supported by the Healthy-Brains-Healthy-Lives startup supplement grant, the NSERC Discovery grant RGPIN-2021-03035, Calcul Québec, and Digital Research Alliance of Canada. M.P. was supported by the FRQNT Strategic Clusters Program (Centre UNIQUE). P.B. was supported by FRQ-S Research Scholars Junior 1 grant 310924, and the William Dawson Scholar award. M.P. was supported by the UNIQUE Fellowship for M.Sc. and Ph.D. and the Stichting Formation.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2015). *Multiple Object Recognition with Visual Attention.* arXiv.

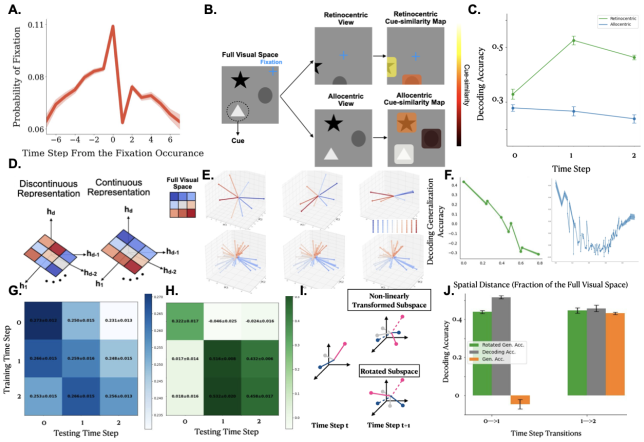

Figure 3: **A.** Probability of fixating at location. **B.** Schematic cue-similarity maps. **C.** Average cue-similarity map decoding accuracy (only the first 3 time steps are included i the following analysis as the performance plateaus after 3 fixations.). **D.** Schematic of the two possible geometries of the cue-similarity representations. **E.** Cue-similarity encoding axes the PC space. **F.** The generalization accuracy from one location to another decreases with increasing distance in the space, indicating a continuous representation. **G.** The allocentric cue-similarity map is consistently encoded in the same subspace. **H.** Retinocentric cue-similarity map is encoded in a variable subspace. **I.** Schematic of a rotational transformation vs. non-linear transformation of a variable subspace. **J.** Retinocentric cue-similarity map encoding subspace rotates across time as shown by the high accuracy of the reconstructed decoders at the target time step using the rotation identified by Procrustes analysis (cross-validated on decoders).

Bichot, N., Heard, M., DeGennaro, E., & Desimone, R. (2015). A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron*, *88*, 832–844.

Bichot, N. P., Xu, R., Ghadooshahy, A., Williams, M. L., & Desimone, R. (2019). The role of prefrontal cortex in the control of feature attention in area V4. *Nature Communications*, *10*, 5727.

Bisley, J. W., & Mirpour, K. (2019). The neural instantiation of a priority map. *Current Opinion in Psychology*, *29*, 108–112.

Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., & Zelinsky, G. (2021). Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, *11*(1), 8776.

Colby, C. L., & Goldberg, M. E. (1999). SPACE AND ATTENTION IN PARIETAL CORTEXfn1. *Annual Review of Neuroscience*, *22*, 319–349.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*, 692–700.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Elsayed, G. F., Kornblith, S., & Le, Q. V. (2019). *Saccader: Improving Accuracy of Hard Attention Models for Vision.* arXiv.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 2961–2969).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*, e1003915.

Kümmerer, M., & Bethge, M. (2021). State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*.

Machner, B., Lencer, M. C., Möller, L., von der Gablentz, J., Heide, W., Helmchen, C., & Sprenger, A. (2020). Unbalancing the Attentional Priority Map via Gaze-Contingent Displays Induces Neglect-Like Visual Exploration. *Frontiers in Human Neuroscience*, *14*, 41.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). *Recurrent Models of Visual Attention.* arXiv.

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* arXiv.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications*, *9*, 3730.

Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., & Oliva, A. (2016). Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.