Human-like Behavior and Neural Representations Emerge in a Goal-driven Model of Overt Visual Search for Natural Objects

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

Like many other animals, humans direct their gaze to selectively sample the visual space based on task demands. Visual search, the process of locating a specific item among several visually presented objects, is a key paradigm in studying visual attention. While much is known about the brain networks underlying 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 naturalistic visual search. After training, the model demonstrated strong generalization in search performance to novel object categories while exhibiting high behavioral consistency with human subjects. Further analysis of the model's population activity revealed an egocentric representation of the priority map, akin to those described in macaques, that persisted in time and was updated with each saccade alongside encoding of the cued object category in a separate subspace. Our model provides a computational framework for further studying the neural circuits underlying visual search in the primate's fronto-parietal cortical network.

Keywords: visual search; saccadic behavior; gaze control; visual attention; priority map; artificial neural network

Methods

Behavioral task. We followed a standard search paradigm from Zhang et al. (2018) (Fig. 1A) that involved viewing a target object (cue) followed by a search array in which one of the objects categorically matched the target.

Model architecture. The model consists of A) a retinal transformation simulating the eccentricity-dependent visual acuity (Mnih, Heess, Graves, & Kavukcuoglu, 2014). B) the retinal transformed image is processed by a convolutional neural network (CNN), simulating the neural processes in the ventral visual pathway (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). The CNN model (Tan & Le, 2020) was trained to perform visual categorization from retinal transformed Imagenet images (Deng et al., 2009). The visual representation is combined with the fixation location in the glimpse network as in Mnih et al. (2014); Ba, Mnih, and Kavukcuoglu (2015). C) The CNN output along with gaze location are fed into a causal transformer that produces the next fixation location. These steps are repeated for as many as 6 fixations in each trial (Fig. 1B).

Model training. The model was trained to do the visual search task following a 2-stage training paradigm inspired by prior work on saccade-augmented visual categorization (Elsayed, Kornblith, & Le, 2019). 1) the model is trained to predict the target location at each step given a sequence of random fixations on the array. The transformer and linear readout parameters are learned using backpropagation by minimizing the Cross-Entropy loss between the last output location and the ground-truth target location (Fig 1B). 2) we fix the transformer's parameters and train a new MLP policy using reinforcement learning to produce the next fixation given the current hidden state of the transformer at every step while incentivized to reach the target as early as it can. Combining supervised learning with reinforcement learning substantially improved the training time of our model compared to other tested alternatives.

Results

Visual search ANN replicates the human saccadic behavior. We behaviorally tested the model on the Natural Object Search Task dataset (Zhang et al., 2018). The model's performance curves followed a similar trend as those of human subjects (Fig 2A). Its scanpaths (spatiotemporal sequence of fixations made during visual search) were highly consistent with those of humans, surpassing the current best model of visual search, IVSN (Zhang et al., 2018) (Fig 2B). Despite this, the model consistently outperformed human subjects in absolute hit rate possibly due to its extensive training on the search trials generated from the same limited set of objects as the test trials. To test this, we further evaluated the model on trials generated from novel object categories (not in the Imagenet) (Fig 2C). The model's behavior was highly consistent with that of humans indicated by similar performance curves (Fig 2D) and high scanpath consistency (Fig 2E). Scanpath similarity score was measured using ScanMatch (Cristino, Mathôt,

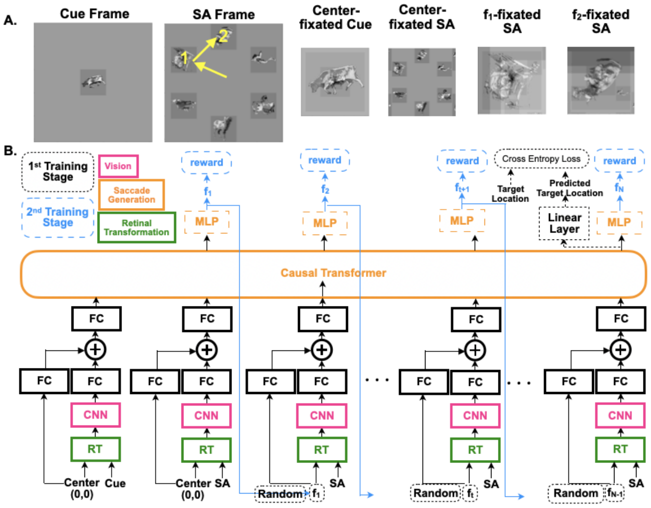

Figure 1: **A.** Example Search Trial. Example cue, search array (SA), and retinal views in different steps. **B.** Model Architecture and Training. The model consists of a retinal transformation (RT), a vision module (a pre-trained CNN), and a fixation generation module (causal transformer + MLP readout). First, the backbone of the fixation generation module (causal transformer) and a linear readout are trained to output the target location using backpropagation. Second, transformer parameters are fixed and an MLP component is trained using reinforcement learning to generate the next fixation location (f) from the latest transformer's output token.

Theeuwes, & Gilchrist, 2010) as in Zhang et al. (2018).

An egocentric representation of priority map in the model's latent space. We investigated whether the model encodes a priority map similar to that previously observed in primates (N. Bichot, Heard, DeGennaro, & Desimone, 2015; Bisley & Mirpour, 2019; Machner et al., 2020; Colby & Goldberg, 1999; Bisley & Mirpour, 2019). To do this we fitted regression models to predict the cue similarity in both egocentric (relative to gaze location) and allocentric (relative to image) frameworks. Cue similarities with each fixated input were computed by comparing (dot product) the center position of the CNN embedding of the cue and 1) that of each fixation for the egocentric map, resulting in a 3x3 priority map (Fig. 3B) and 2) that from fixations on each of the objects on the search array, resulting in a 1x6 priority map (Fig. 3C).

The egocentric priority map was much stronger represented in the model's latent space versus the allocentric one (Fig 3D). Moreover, we also considered the possibility that the allocentric priority map may have been encoded with an inhibition of return (IOR) mechanism. To test this, we repeated the decoding analyses of the allocentric priority map with two adjustments: 1) we considered IOR when generating the ground truth cue similarities in the allocentric priority map (i.e. zeroed out the priority of the locations visited in the past n steps for n-IOR) and; 2) to prevent IOR from artificially boosting our decoding results, we only considered the trials where each decoded location was not visited in the last n steps. From these analyses, we did not find any evidence of an allocentric priority map with IOR in the model's latent space (Fig 3D).

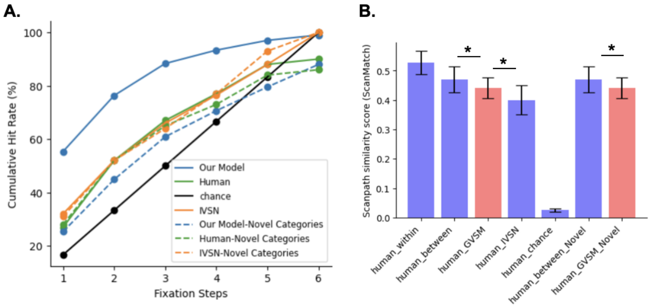

Figure 2: **A.** Cumulative Performance as a function of fixation steps. **B.** Image-by-image consistency in the spatiotemporal pattern of fixation sequences.

We further tested whether the the model's egocentric priority map was consistently encoded in the same subspace across time, by validating priority map decoders fitted on each time step on other time steps. The model's egocentric representation of the priority map was stable across time steps except for the first one (generalization accuracy across locations: for time steps after the first one: mean=0.77, standard deviation (SD)=0.05; for the first time step: mean=0.28, SD=0.22). Moreover, the model's priority map was continuously encoded across the model's latent space, i.e. priorities at nearby locations in the visual space were encoded along more aligned axes in the model's latent space (Fig 3E-G).

Finally, we found that the cue category was also consistently represented in the same subspace across time points, as indicated by its high decodability across time (mean=0.97, and SD=0.01) and high generalizability of the decoding across time steps (mean=0.93, SD=0.07).

Discussion

We showed that an image-computable neural network model trained to perform the visual search task closely replicates humans' behavioral response patterns during this task while relying 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 and the NSERC Discovery grant RGPIN-2021-03035. P.B. was supported by FRQ-S Re-

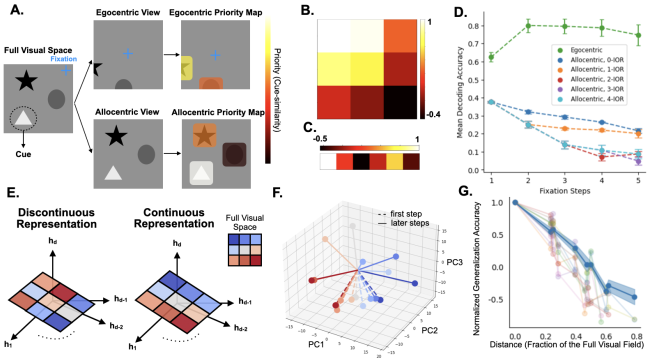

Figure 3: **A.** Schematic of egocentric vs. allocentric priority map. **B.** An example egocentric priority map. The center position of the CNN embedding of the center-fixated cue was compared (dot product) with that of the gaze-location-fixated search array, giving a cue-similarity map indicating the goal-driven priority of each location in the egocentric reference frame. **C.** An example allocentric priority map. The center position of the CNN embedding of the center-fixated cue was compared (dot product) with that from fixations on each object on the search array, resulting in a 1x6 priority map. **D.** The egocentric priority map and not the allocentric priority map was stably decodable (Ridge regression) from the transformer's hidden space. The mean validation accuracy of the priority decoding averaged over all locations on the priority map across time is shown. **E.** Schematic of the two possible geometries of the priority map representation. **F.** Priority decoders' axes for nearby locations in the visual space are more aligned in the model's hidden space. **G.** The normalized generalization accuracy (generalization accuracy across the locations normalized by the maximum over the visual space) decreases with increasing distance in the space, indicating a continuous topographical representation of the priority map in the model's hidden space. Colors as in F.

search Scholars Junior 1 grant 310924, and the William Dawson Scholar award.

References

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

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.

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.

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

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.