Modeling Human Eye Movements with Neural Networks in a Maze-Solving Task

Modeling Human Eye Movements with Neural Networks in a Maze-Solving Task

Jason Li, Nicholas Watters, Hansem Sohn, Mehrdad Jazayeri
{jasli, nwatters, hansem, mjaz}@mit.edu.com
McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, MIT,
43 Vassar St, Cambridge, MA 02139

Abstract

From smoothly pursuing moving objects to rapidly shifting gazes during visual search, humans employ a wide variety of eye movement strategies in different contexts. While eye movements provide a rich window into mental processes, building generative models of eye movements is notoriously difficult, and to date the computational objectives guiding eye movements remain largely a mystery. In this work, we tackled these problems in the context of a canonical spatial planning task, maze-solving. We collected eye movement data from human subjects and built deep generative models of eye movements using a novel differentiable architecture for gaze fixations and gaze shifts. We found that human eye movements are best predicted by a model that is optimized not to perform the task as efficiently as possible but instead to run an internal simulation of an object traversing the maze. This not only provides a generative model of eye movements in this task but also suggests a computational theory for how humans solve the task, namely that humans use mental simulation.

Keywords: mental simulation; saccades; neural networks

Introduction

Throughout the history of cognitive science, eye movements have been appreciated as a window into the workings of the mind and brain (Helmholtz, 1924; Liversedge & Findlay, 2000). However, human eye movements are so rich and varied that characterizing them is difficult even in simple tasks (Beller. et al., 2022; Gerstenberg et al., 2017). Building generative models of eye movements is an even greater challenge (Chen et al., 2017; Zoran et al., 2020), and to date most such work focuses only on static statistics of eye movements (e.g., saliency maps for average fixation positions), ignoring temporal dynamics such as saccade sequences (Kummerer et al., 2016).

To tackle the problem of modeling saccade sequences, we designed a maze-solving task, in which subjects must find the exit location of a path in a maze given a starting point of the path (Figure 1). This task provides an ideal platform for building generative models of eye movements because it offers a near-limitless variety of spatial plans, yet eye movements are largely consistent across humans (Crowe et al., 2000) making them tractable to model. Furthermore, this task may be solved using mental simulation of an object traveling through the maze, so allows us to test mental simulation as a computational theory guiding eye movements (Gerstenberg et al., 2017; Rajalingham et al., 2021; Ullman et al., 2017).

In this work, we develop a novel general-purpose method for incorporating eccentricity-dependent visual acuity and discrete saccades (features of human vision) into a task-optimized, end-to-end differentiable recurrent network. Using this method we construct a space of models with and without mental simulation constraints, and train these on the maze-solving task. We collect eye movement data from human subjects playing the task, and compare this data to eye movements generated by the models to arrive at a hypothesis for how humans solve the task.

Methods

Task and Maze Dataset

In the maze-solving task, a subject is presented with a square maze and an entrance point somewhere on its perimeter. This entrance point is one end of an unbroken, non-branching path through the maze, which exits the maze at some other uniformly sampled perimeter point. The subject is tasked to find this exit point. We create a test set for human and model comparison comprising 200 unique procedurally generated mazes.

Human Data Collection

Using an optical eye tracker (1 ms resolution; EyeLink 1000 Plus, SR Research), we collected eye position data from three human subjects, each of whom completed 400 trials. Subjects were shown the entrance point and were instructed to locate and fixate on the exit point. We extracted saccades from the eye position data by first filtering with a 4ms Gaussian kernel and then thresholding eye velocity at 50 visual degrees per second.

Saccading Recurrent Neural Network (RNN) Models

To model eye movements with recurrent convolutional neural networks, we construct a fovea that provides the network with high visual acuity near the center of the field of view and low acuity in the periphery, like the human eye. We also allow the network to control the position of this fovea and integrate visual information through it in a differentiable manner, in order to pass gradients through the fovea controller via backpropagation.

We model the fovea by applying a circular exponential mask $e^{-d/\tau}$ to the visual input, where d is eccentricity relative to the network's current eye position and τ is a scaling parameter (which we choose to be 8 pixels). After applying the mask, we add noise sampled from a normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 0.05)$. This noise washes out faint information

in the peripheral tail of the foveal mask, analogous to the decreased peripheral photoreceptor density in the human retina.

This approach allows the network to produce an eye position in Cartesian coordinates at each step and differentiably transform that into a fovea-masked input for the next step. This method is general-purpose, and in theory can be incorporated into any RNN that takes visual input.

For our maze task, we incorporate this differentiable saccading mechanism into a convolutional RNN that at each timestep produces both a Cartesian eye position and a Cartesian predicted ball position (Figure 1). The predicted ball position may be regressed to the true position of a dynamic invisible ball traveling through the maze from start point to exit point at constant velocity, as a loss function to the model. We call this a "simulation loss", because it pressures the model to simulate an object traversing the maze.

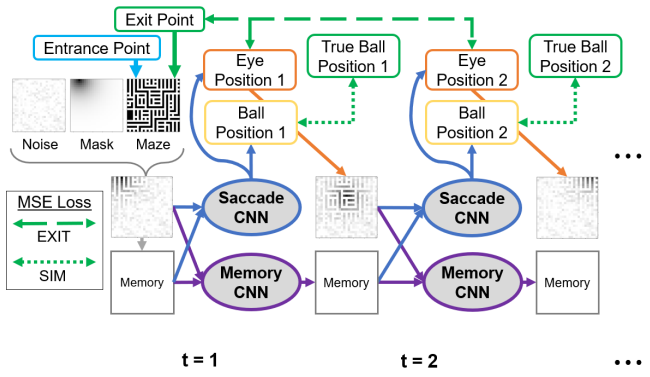

Figure 1: Saccading RNN model. Memory CNN is a 3-layer CNN. Saccade CNN is a strided 3-layer CNN with two 3-layer MLP heads for Cartesian eye position and ball position vectors. Eye position and rendered maze are processed by our differential fovea mechanism to produce visual input for the next step. This architecture can be unrolled through time for an arbitrary number of steps or saccades.

We implement three specific models, all of which have identical architecture and differ only in their objective functions:

1. Exit The EXIT model is trained with a Mean Squared Error (MSE) loss between the eye position at each step and the maze exit point. There is no loss on the model's ball position output. This model represents an optimal exit-finding strategy where the model moves its eyes to the exit in as few saccades as possible.

2. Simulation The SIM model is trained with the simulation loss described above (MSE between each ball position and the position of an invisible "true ball" traveling at 10 pixels per timestep). Eye position is not explicitly constrained, but still plays a critical role in advancing the model's visual field.

3. Hybrid The HYBRID model is trained with a weighted sum of the loss functions for the EXIT and SIM models. The model's eye position must reach the exit quickly *and* allow the model's ball positions to align with the true ball positions. The ratio of SIM to EXIT loss weight is controlled by a coefficient $\beta = 10$, chosen so that the two loss terms have similar magnitudes in a fully trained model.

All models are trained with 8 recurrent steps per maze, batch size 16, through 500,000 iterations using Adam optimizer (Kingma & Ba, 2014) with learning rate 0.001. The training dataset was generated online using a custom procedural maze generator with the same statistics (though not the same samples) as the test set.

Results

See Figure 2 for the behavior of the models and two human subjects on two random mazes in the test set. The human and the models all roughly follow the path through the maze and successfully find the exit point. Humans also display a tendency to cut corners of the maze path. Although the EXIT model finds the exit point faster than the other models, it does not always follow the path closely like humans do and sometimes makes very large saccades. On the other hand, the more uniform saccades generated by the SIM and HYBRID models appear to better match human saccades.

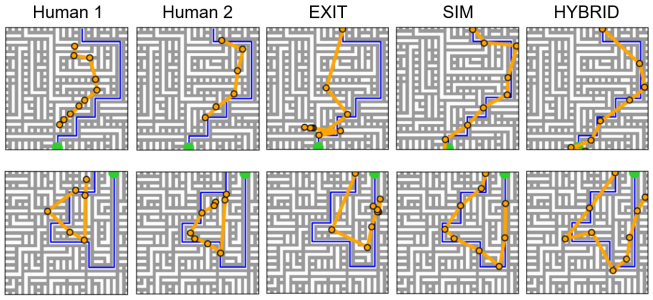

Figure 2: Human and model behaviors on two sample mazes. Orange = eye position, Blue = maze path, Green = exit point.

To quantify these results, we use two metrics for comparing eye movement paths:

- **Nearest neighbors distance** is computed as the mean of the nearest point in path A to each point in path B and the nearest point in path B to each point in path A.
- **Area between paths** is computed as the total plane area of the polygon(s) formed between paths A and B.

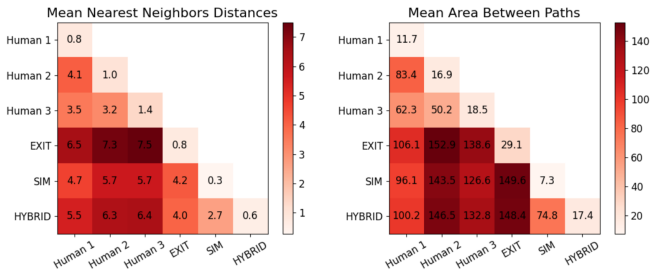

Figure 3: Distance matrices for human and model eye movement paths computed on the test set. The SIM model is the best-performing model under both metrics. Mean model-human scores for each model: nearest neighbors: EXIT 7.1, SIM 5.4, HYBRID 6.1; area: EXIT 132.5, SIM 122.1, HYBRID 126.5]

The SIM model exhibits the lowest mean model-human distances under both distance metrics (Figure 3), while the EXIT model produces the least human-like eye movement paths. Although the SIM model is not quite as similar to humans as humans are to one another, it is the best of our three generative models.

Conclusion

We find that in the maze-solving task, a saccading RNN trained to run an internal simulation of a ball moving through a maze generates eye movements more similar to those of human subjects than a model trained only to solve the task as optimally as possible. This suggests that humans may employ a similar mental simulation when performing this maze-solving task. Further work is needed to explore the relationship between the biological plausibility of the model fovea hyparparameters and model behavior. Future work also includes incorporating our differential saccading method into RNNs trained on other tasks to study the principles of human eye movements in domains beyond maze-solving.

Acknowledgments

M.J. is supported by NIH (NIMH-MH122025), the Simons Foundation, the McKnight Foundation, and the McGovern Institute. H.S. is supported by a NARSAD young investigator grant from the Brain & Behavior Research Foundation.

References

Beller., A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past: Eye-tracking mental simulation in physical inference. *Cognitive Science Proceedings*.

Chen, F. X., Roig, G., Isik, L., Boix, X., & Poggio, T. (2017). Eccentricity dependent deep neural networks: Modeling invariance in human vision. In *2017 aaai spring symposium series.*

Crowe, D. A., Averbeck, B. B., Chafee, M. V., Anderson, J. H., & Georgopoulos, A. P. (2000). Mental maze solving. *Journal of Cognitive Neuroscience*, *12*(5), 813–827.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731-1744. doi: 10.1177/0956797617713053

Helmholtz, H. V. (1924). *Helmholtz's treatise on physiological optics. translated from the third german edition.* (P. C. Southall, Ed.). The Hatton Press, Ltd.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kummerer, M., Wallis, T. S., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences*, *4*, 6–14.

Rajalingham, R., Piccato, A., & Jazayeri, M. (2021). The role of mental simulation in primate physical inference abilities. *bioRxiv*.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Zoran, D., Chrzanowski, M., Huang, P.-S., Gowal, S., Mott, A., & Kohli, P. (2020). Towards robust image classification using sequential attention models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9483–9492).