Glimpse prediction fosters graph-oriented scene representations aligned with the ventral visual cortex

Sushrut Thorat (sushrut.thorat94@gmail.com)
Institute of Cognitive Science, Osnabrück University
Osnabrück, Germany

Adrien Doerig (adrien.doerig@fu-berlin.de)
Department of Psychology and Education, Freie Universität Berlin
Berlin, Germany

Alexander Kroner (alexander.kroner@uni-osnabrueck.de)
Institute of Cognitive Science, Osnabrück University
Osnabrück, Germany

Carmen Amme (camme@uni-osnabrueck.de)
Institute of Cognitive Science, Osnabrück University
Osnabrück, Germany

Tim C Kietzmann (tim.kietzmann@uni-osnabrueck.de)
Institute of Cognitive Science, Osnabrück University
Osnabrück, Germany

Abstract

Understanding how the visual system responds to natural scenes remains a central challenge in vision science. Research shows that the ventral visual cortex (VVC) encodes objects, textures, and the spatial and semantic relationships between them—forming a structured scene representation, akin to a scene graph. However, inferring such representations from images in an interpretable, image-computable way is still an open problem. We propose glimpse prediction—predicting the upcoming visual input given an eye movement (saccadic efference copy)—as a training objective that encourages the emergence of representations with graph properties in artificial neural networks. A recurrent neural network trained with this objective learns spatial covariance between glimpses, across scenes (in-weight learning) and in novel scenes (in-context learning). Importantly, the model's internal representations align closely with VVC responses to natural scenes (Natural Scenes Dataset), despite never observing the full scene or receiving explicit semantic labels. Thus, glimpse prediction offers a principled route to building graph-oriented representations mirroring those in the human ventral visual stream. Combining interpretable concepts from cognitive neuroscience and image-computable neuroconnectionist models, this work advances a comprehensive understanding of the visual system's response to natural scenes.

Keywords: scene representation, scene graph, eye movements, prediction, interpretability, recurrent neural networks

Motivation

How does the visual ventral cortex (VVC) respond to natural scenes? VVC represents scene chunks such as object parts, objects, textures, surfaces (DiCarlo, Zoccolan, & Rust, 2012; Grill-Spector & Weiner, 2014). It also represents relationships between these chunks (Kaiser, Quek, Cichy, & Peelen, 2019), both spatial (e.g. an egg appears above an egg cup) and semantic (e.g. toilet paper appears next to a toilet seat rather than a dishwasher). These chunks and their relationships, constituting a scene graph, comprehensively describe a scene as a combination of its parts (Johnson, Gupta, & Fei-Fei, 2018; Vo, 2021). Although we have a good understanding of the components VVC represents, building an interpretable, image-computable model that can predict VVC responses to natural scenes is challenging.

In building image-computable models that can predict VVC responses, the most successful approaches have taken the form of training artificial neural networks on large-scale image and text datasets (Doerig et al., 2022; Conwell, Prince, Kay, Alvarez, & Konkle, 2024). However, in those approaches it is unclear what the format of the internal representation is - whether it captures graph-like structure. To circumvent this issue of interpretability, we build a 'Glimpse Predictor' model to explicitly encourage a vector representation with graph properties, to predict ventral stream responses to natural scenes. We show that our model learns to encode spatial and semantic relationships between scene chunks, and its internal representation of the scene aligns with VVC responses to natural scenes.

Setup

The Glimpse Predictor (GP) model

To encode graph-like structure, a model needs to represent scene chunks and their relationships.

Human fixation traces were used to define scene chunks - meaningful glimpses that need to be integrated for scene understanding (Henderson, Hayes, Peacock, & Rehrig, 2019). We modeled these traces with DeepGaze3 (DG3; Kümmerer, Bethge, and Wallis (2022)). These glimpses were represented by the AvgPool activations of ResNet50 pretrained on Imagenet (RN50; v1, Torchvision).

To encourage a graph-oriented representation, we turned to sequence prediction. Prediction encourages networks to learn the structure of the generating function (Elman, 1990; Radford et al., 2019). Thus, predicting the next glimpse, given the upcoming saccade, would encourage the model to infer the scene graph that the fixations implicitly traverse.

As shown in Figure 1B, in GP, projections of the current glimpse representation and a cartesian saccadic copy are inputs to a 3-layer LSTM, a projection from which is trained to predict the next glimpse representation, with a contrastive objective: the cosine similarity of the prediction is expected to be higher to the next glimpse and lower to the other glimpses from the current scene and from other scenes.

Datasets

Images of scenes from the MS-COCO dataset (Lin et al., 2014) were used for training and testing the GP: all COCO images from the 2017 train/val splits that were not the special-515 images shown to all participants, with 3 repeats, in the 7T fMRI Natural Scenes Dataset (NSD; Allen et al. (2022)) were split into train (~ 121k images) and val (~ 2k images). The special-515 served as the test set which is used in analyzing the GP. 91 px crops, centered on the DG3 fixations, from the COCO images scaled to 256 px (smaller axis), were taken as glimpses (corresponding to ~ 3 DVA in NSD). If a portion of the glimpse fell outside the scene, that portion was padded black. 10 fixation traces, with 7 fixations each, starting at the center (Figure 1A), were sampled for all the images.

From NSD, GLM beta weights were averaged, across the 3 repetitions, for each of the special-515 images and the 8 participants. We extracted bilateral visual ventral cortex (VVC) responses to these 515 scenes via the ventral stream mask in the NSD 'streams' ROI definitions (RH shown in Figure 1F).

Analysis

As seen in Figure 1C, post-training, the GP predictions were more similar to the target glimpse representations than to the other glimpses' representations, suggesting that the GP learned to predict the next glimpse. Moreover, the contrastive

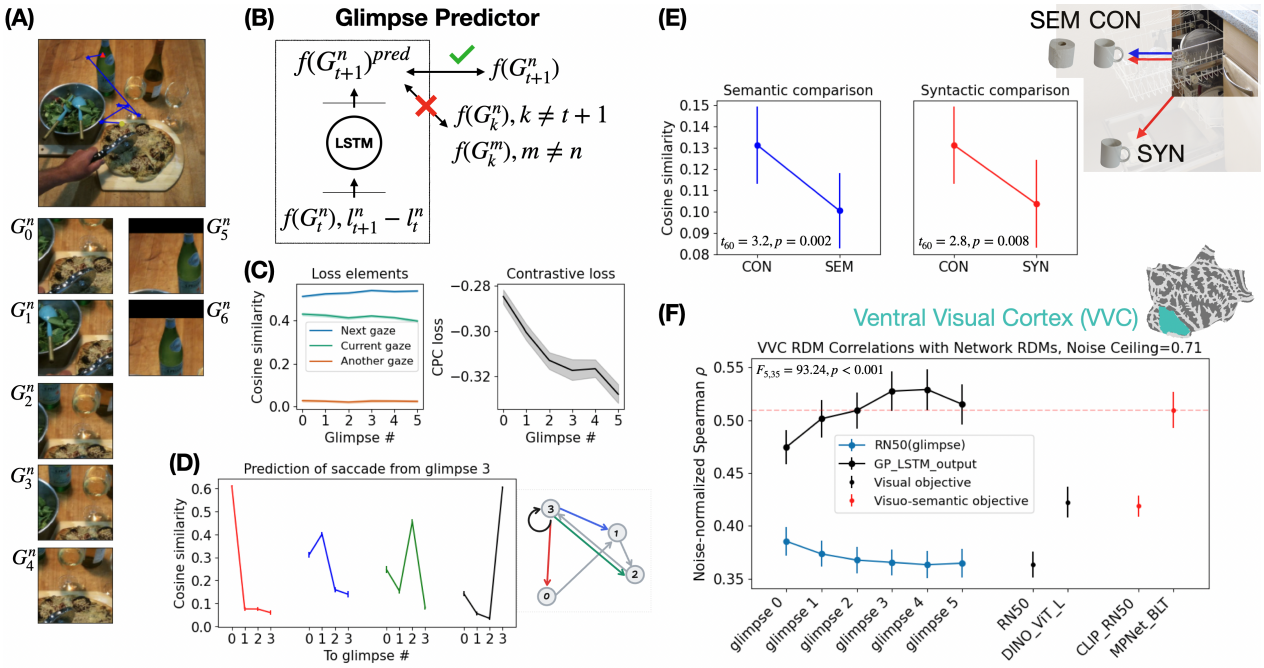

Figure 1: The Glimpse Predictor setup and analysis. In (B), G_t^n refers to the t^{th} glimpse from scene n, l_t^n refers to the location of that glimpse, and $f(G)$ refers to the ResNet50 glimpse representations. 95% confidence intervals of the means are shown.

loss decreased with increasing glimpses, suggesting the existence of integrative processes in the GP.

One indication that glimpses are being integrated into a graph-oriented representation, is the model's ability to predict novel re-fixations. We sampled 1000 fixation traces, taking the first 4 fixations (where no re-fixations happened; min. saccade length ~ 30px), from the test set traces. For each fixation, we sampled the glimpse representation randomly from the set of all glimpses from the test set. Then we simulated re-fixations from the 4^{th} fixation to all the 4 fixations. As seen in Figure 1D, the GP predictions aligned with the glimpse representations corresponding to the locations suggested by the re-fixations, suggesting that the GP indeed utilizes a latent graph to integrate across glimpses. Critically, the weights do not change while learning these arbitrary scene graphs, akin to in-context learning (Olsson et al., 2022).

While graphs can be learned in-context, what information about spatial covariance, specifically object-scene covariance, across scenes did the GP learn with its weights? We took scenes from the SCEGRAM dataset (Öhlschläger & Võ, 2017), which contain semantic and syntactic (in)consistencies at the object level (e.g. a cup or a toilet paper roll in the dishwasher rack [CON/SEM], the cup in the dishwasher rack or on the dishwasher door [CON/SYN]; Figure 1E). With similar preprocessing as the COCO images, we took the central glimpse (without the objects of interest) and simulated saccades to the location of the object in each of the CON/SEM/SYN conditions, for each of the 61 scenes. We compared the GP predictions to RN50 representations of the corresponding isolated objects in upright position (to exclude background and orientation contributions). As seen in Figure 1E, the GP predictions were closer to the semantically-congruent object and matched the congruent object better in its syntactically-appropriate position, suggesting the GP learned spatial object-scene covariance through its experience with the COCO training set.

While the GP contains information about spatial covariance structure of scene chunks (including objects) and can learn such structure in new scenes, do its internal activations resemble ventral stream responses? For each of 515 test set images, we took the activations of the final LSTM layer for each glimpse during one fixation trace and extracted representational dissimilarity matrices (RDMs; correlation distance) per glimpse. We compared these RDMs to RDMs constructed from ventral visual cortex (VVC) responses. As seen in Figure 1F, the GP-VVC correlation increased with added glimpses and peaked at 0.53, which surpasses strong models of VVC response to natural scenes - ResNet50 (full scene representation), DINO (Oquab et al., 2023), and CLIP (Radford et al., 2021) - and is comparable to a state-of-the-art model, MPNet-BLT (Doerig et al., 2022). This result suggests a strong alignment between the graph-oriented GP representations and VVC responses to natural scenes.

Conclusion

Glimpse prediction is a powerful approach to build graph-oriented representations in artificial neural networks, aligning with ventral stream responses to natural scenes, rivaling state-of-the-art models, despite never observing the full scene or receiving any explicit semantic supervision.

Acknowledgments

The project was partially funded by the European Union (ERC, TIME, Project 101039524).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, *15*(1), 9383.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Visual representations in the human brain are aligned with large language models. *arXiv preprint arXiv:2209.11737*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), 19.

Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1219–1228).

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in cognitive sciences*, *23*(8), 672–685.

Kümmerer, M., Bethge, M., & Wallis, T. S. (2022). Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, *22*(5), 7–7.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Öhlschläger, S., & Võ, M. L.-H. (2017). Scegram: An image database for semantic and syntactic inconsistencies in scenes. *Behavior research methods*, *49*(5), 1780–1791.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... others (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... others (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Vo, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20.