Conceptual priorities shape individual gaze patterns during naturalistic visual attention

Amanda J. Haskins (ajhaskins@ucsd.edu), Katherine O. Packard (kateopackard@gmail.com), Caroline E. Robertson (cerw@dartmouth.edu)
Psychological & Brain Sciences, Dartmouth College, 3 Maynard Street Hanover, NH 03755 USA

Abstract

In daily life, we readily recognize people, places, and objects (e.g., "soldier," "stadium," "flag"), as well as the conceptual links between them (e.g., "patriotism"). Here, we show that conceptual-level information shapes individuals' gaze patterns in a naturalistic eye tracking paradigm. Participants (N = 61) explored a large set of real-world photospheres (N = 100) in headmounted VR while their gaze was continuously monitored using in-headset eye tracking. To assess the informational priorities guiding individual differences in gaze, we leveraged the embedding spaces of large vision and language models. We found that individually-specific gaze patterns across diverse real-world photospheres can be captured by a large language model (LLM) that encodes abstract relationships beyond the visual image content. We demonstrate that the embedding spaces of language and vision models explain unique variance in gaze behavior, and that LLM-based models capture individually specific attentional priorities. These results highlight a new dimension of human selective attention: namely, the influence of individuals' unique conceptual-level information seeking priorities.

Keywords: individual differences, naturalistic visual attention, information seeking, concepts

Introduction

What guides individuals' selective attention when viewing real-world scenes? Understanding gaze behavior has been a central goal in psychology, with focus on the roles of "bottom-up" (visual salience) and "top-down" (semantic meaning) scene features (Henderson & Hayes, 2017). However, this dichotomy overlooks the influence of contextual factors on gaze behavior (Awh et al., 2012). While external contexts like tasks and rewards are known to shape attention (Borji & Itti, 2014; Tong et al., 2017), less is understood about how intrinsic factors, like an individual's unique conceptual knowledge and interests, guide their gaze. This gap is important for current theories that forward gaze behavior as a proactive, information-seeking process (Haskins et al., 2020; Hayhoe, 2017). Individuals explore scenes in the real world proactively: they use their conceptual knowledge to form and test hypotheses about object relationships. Here, we explored how conceptual information beyond the visual domain influences gaze patterns during scene viewing. We hypothesized that gaze patterns reflect, in part, individuals' stable, conceptual priorities. To test this, gaze behavior was analyzed from adult participants who explored real-world photospheres in virtual reality (VR). We developed a computational approach using a large language model (LLM; (Devlin et al., 2019)) to approximate conceptual information in scenes and its role in shaping gaze patterns. In brief, we found that individual gaze patterns are strongly predicted by the LLM's conceptual feature space, which accounted for significant unique variance beyond well-established predictors of gaze (i,e., visual, motor features).

Methods

Participants. 66 adults (*n* = 36 female; mean age 20.67 years) participated in this study. Following exclusion, data were analyzed from 61 participants.
Stimuli. We used 100 diverse, information-rich photospheres, selected based on pilot data showing consistent attention to both social (e.g., people) and nonsocial elements. On each 16-second trial, participants were instructed to "look around each scene naturally, like you would in daily life."

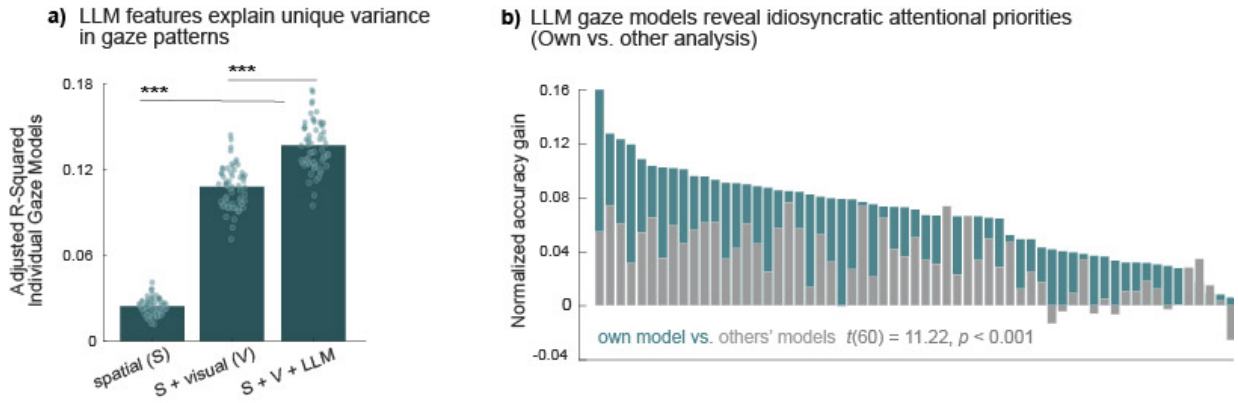

Figure 1: A) LLM embeddings capture unique variance in gaze patterns, relative to control feature spaces. B) LLM gaze models can be used to generate individually-specific gaze predictions on left-out scenes.

Gaze data. Duration-weighted fixation density maps were individually plotted for each subject and scene.

Modeling conceptual information using an LLM. We used the embedding space of an LLM to characterize the conceptual-level information in each photosphere. First, we divided scenes into densely sampled, overlapping "tiles" that were captioned by independent human raters; then, we transformed tiles into language model embeddings. As a control model to the LLM, we used a visual transformer model (ViT; (Dosovitskiy et al., 2020)) trained on image classification, to model the visual content depicted at each image tile. As a second control model to the LLM, we used the tile's equirectangular spatial coordinates (X,Y). These control feature spaces were used to test whether the LLM embedding space captured unique variance in gaze patterns, beyond control spaces.

Results

LLM embeddings explain unique gaze variance. We first asked whether LLM embeddings capture information in gaze patterns that is distinct from established predictors: individual spatial and visual-level feature biases (e.g., center bias, object biases). We performed a variance partitioning analysis on an "omni gaze model" that included LLM features alongside the two control feature spaces. This analysis revealed that LLM features explained significant variance beyond control feature sets (adjusted R^2, all-features model: M = 0.14, SD = 0.02; adjusted R^2, non-language features model: M = 0.11, SD = 0.01; $t(60)$ = 49.6, p < 0.001; Fig 1A).

LLM embeddings predict unique gaze patterns. Next, we built an "LLM gaze model" for each participant by using L2-regularized linear regression to relate their gaze distribution to only the LLM feature space, iteratively training on N-1 scenes, and generating a predicted gaze map for each left-out scene. Then, to test whether LLM gaze models were individually specific, we used the same leave-one-out approach to iteratively assess the accuracy of an individual's own LLM gaze model (correlation, predicted vs. actual map), as compared with the LLM gaze models of all other individuals. Specifically, we computed an "*own-other difference score*" for each participant – subtracting the average accuracy gain (i.e., improvement over the accuracy of a generic baseline model) of others' models from the accuracy gain of the participant's own model. This difference was significant (own: M = 0.07, SD = 0.03; other: M = 0.03, SD = 0.02, $t(60)$ = 11.22, p < 0.001; Fig. 1B), demonstrating clear accuracy gains for one's own LLM gaze model prediction vs. other participants' models in almost every participant.

Discussion

Our results show that gaze patterns offer insight into the conceptual priorities different individuals bring to bear while exploring their visual environment. Specifically, we find that the embedding space of an LLM can be used to capture variance both within individual gaze patterns (variance partitioning) and across individuals (own-other difference). Overall, we find that each individual's gaze reveals their personal conceptual priorities during scene viewing.

Acknowledgments

This work was supported by grants from the Nancy Lurie Marks Family Foundation and the Neukom Institute for Computational Science (C.E.R). A.J.H. was supported by a grant from the National Institute of Mental Health (1F99NS135812).

References

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443.

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*(3), 29–29.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *arXiv [cs.CV]*.

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports*, *10*(1), 14304.

Hayhoe, M. M. (2017). Vision and action. *Annual Review of Vision Science*, *3*(1), 389–413.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747.

Tong, M. H., Zohar, O., & Hayhoe, M. M. (2017). Control of gaze while walking: Task structure, reward, and uncertainty. *Journal of Vision*, *17*(1), 28.