Goal-conditioned world models: Adaptive computation over multi-granular generative models explains human scene perception

Mario Belledonne (mario.belledonne@yale.edu)
Dept. of Psychology, Yale University

Chloë Geller (chloe.geller@yale.edu)
Dept. of Psychology, Yale University

Ilker Yildirim (ilker.yildirim@yale.edu)
Dept. of Psychology, Yale University

Abstract

When we enter a room, perception is challenged to convert sensory data into complex mental representations such as scene geometry. This leaves our percepts both strikingly sparse but also structured so as to support the flexibility of cognition. Here, we apply a new formal model of attention, adaptive computation, to reveal patterns of spatial attention and geometric selectivity in the context of perception of indoor scenes. The model uses the goal of navigating to a visible exit to guide selective processing over a multi-granular scene geometry model, which can represent regions in the room at different levels of resolution. Together the components of goal-driven processing and multi-granular scene states enables the efficient investment of computational resources to resolve geometry relevant to navigational affordances. In the context of a change detection task, we show that both goal-directed attention and multi-granular geometry representations are critical to explaining human responses. Together, adaptive computation and multi-granular geometry representations form powerful computational tools for studying scene networks.

Keywords: goal-conditioned world models; scene perception; multi-granularity; attention; inverse-graphics

Our percepts of everyday scenes are strikingly selective, often drastically impacted by our momentary goals. If in the hall shown in Fig. 1, you will meet with a friend down the hall by the exhibits, you will see the general outline of the room, the obstacle in the middle, and people scattered around, while coarsely summarizing the rest (e.g., the stairs to the left).

This example illustrates three core aspects of perception that so far have eluded a coherent computational account. First, scene percepts are structured with scene geometry and entities, that can support flexible planning in a three-dimensional (3D) world. Second, different representations in a single percept can vary along a ladder of granularity, with some regions more finely represented than others. Third, tasks such as navigation drive the distribution of granularity towards relevant dimensions of the scene. Existing frameworks of perception only address individual aspects of scene perception – with generative approaches emphasizing structure and uncertainty (e.g., Bayesian models of perception and cognition; Chater, Tenenbaum, and Yuille (2006)) and discriminative models emphasizing information loss (e.g., task-optimized deep neural networks; Yamins and DiCarlo (2016)).

Here, we present a new computational architecture to reverse-engineer scene perception. This architecture arises from a principled integration of perception and planning, based on a recently proposed account of attention – adaptive computation (Belledonne, Butkus, Scholl, & Yildirim, 2023). Adaptive computation is a mechanism for bounding and scheduling perceptual processing to construct goal-conditioned structured representations. We use adaptive computation to selectively process a multi-granular generative model of scene geometry, yielding efficient allocation of computational resources to resolve scene geometry relevant to navigational affordances. The resulting model carves a new theoretical landscape in which the result of inference is high-level, structured scene representations similar to structured Bayesian models, while enjoying "intelligent" information loss by conditioning these representations to the goals of the observer, similar to task-optimized DNNs.

We show that this model explains human performance in a change-detection task with indoor scene images. Crucially, this task never overtly mentions anything navigation related; instead, it measures spontaneous, automatic visual processing. In this way, the computations and representations stipulated by our model, adaptive computation and multi-granular, are likely recruited automatically during visual processing.

Figure 1: Scene perception can be strikingly selective, impacted by our goals, including navigational targets.

Computational model

We formalize the perception of goal-conditioned (i.e., selective) scene geometry through a new form of "intelligent" information loss, adaptive computation, that flexibly exploits tasks for the efficiencies they afford. Adaptive computation bounds and schedules computational steps (e.g., Bayesian updates) based on a universal measure of task relevance that integrates the impact of perceptual processing on planning outcomes. This yields a mechanism for automatically carving navigational affordances out of a generative model with multi-granular scene geometry.

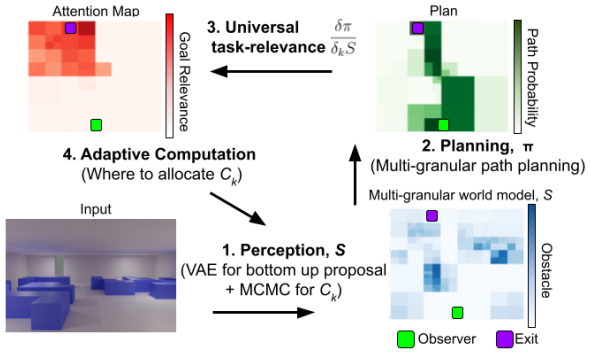

Figure 2: Model architecture that carves goal-conditioned scene percepts in multi-granular generative models. See text.

Multi-granular perception We start by describing the target of the perceptual system, the multi-granular scene state. The scene state, S, consists of spatial "trackers" that represent the occupancy probability for a given region. These trackers are the leaf nodes of a quad-tree (Finkel & Bentley, 1974), supporting multi-granularity by modulating the number of regional subdivisions, with more divisions in a given region leading

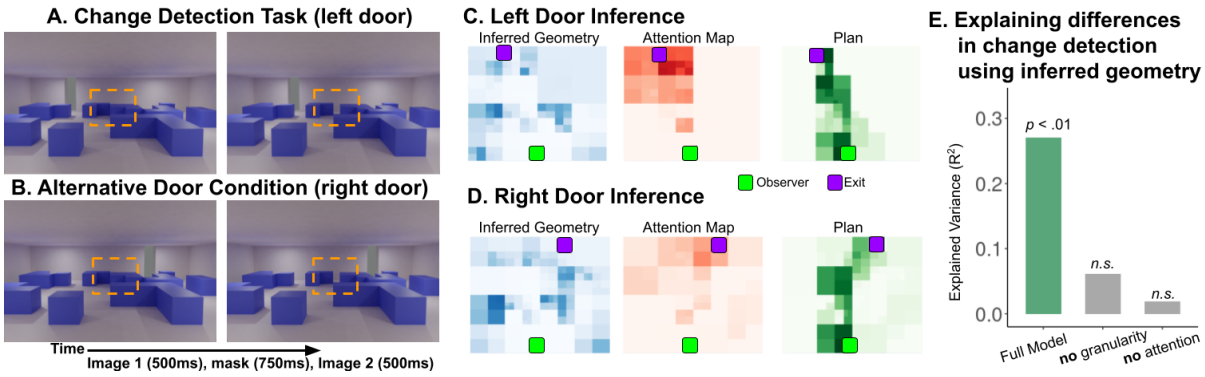

Figure 3: Multi-granularity and adaptive computation are both needed to explain change detection behavior. See text.

to more geometric resolution. The generative model defines a prior over these quad-tree states, $Pr(S)$, via a generative process that iteratively subdivides a region in space into four quadrants (implemented in Gen.jl; Cusumano-Towner, Lew, and Mansinghka (2020)) – up to a maximum of 5 subdivisions (i.e., maximum resolution of 32x32). Each leaf node, $s_k \in S$, determines the Bernoulli weight of obstacle occupancy, drawing from a uniform distribution over $[0,1]$ (0: empty, 1: full).

The generative model makes graphical predictions from S by projecting the leaf node weights $s_k \in S$ to a 3D volume. The transparancy of a region reflects the the geometric uncertainty of its underyling leaf node, with occupancy weights closer to 1 appear more opaque. This volume is then rendered into a 128x128 RGB image using a volumetric renderer (Jakob et al., 2022) with the resulting image, Y_μ serving as the mean of the likelihood over a given observation , $Pr(X \mid S) = \mathcal{N}(X; Y_\mu, Y_\sigma)$, where $Y_\sigma = 0.05$ denotes a diagonal covariance matrix.

The unit of computation, C_k, for the perceptual system denotes a Bayesian update over $s_k \in S$ that collectively approximate the posterior over scene geometry $Pr(S \mid X) \propto Pr(X \mid S)Pr(S)$. These updates have two parts: *(i)* random walk over the occupancy weight in s_k and *(ii)* splitting s_k into 4 children or merging s_k with its siblings. Applying C_k to S updates an individual node $S' = S_{-k} \cup s'_k$ where s'_k is the updated node, and produces a divergence metric that follows from the inverse of the Monte Carlo transition kernel: $\delta_k S = \frac{Pr(S|X)Pr(s'_k|S,C_k)}{Pr(S'|X)Pr(s_k|S',C_k)}$. Given an image X, the nodes, $s_k \in S$, are coarsely initialized by decoding occupancy weights from the embedding of a variational autoencoder (Burgess et al., 2018) that was trained on draws from the generative model.

Multi-granular path planning For planning, the model projects S into a lattice graph connecting neighboring terminal. To compute the shortest path to the exit node, π, the model uses the A* algorithm (Russell, 2010; Fairbanks et al., 2021), weighting the cost of traversing a node in terms of its size (coarser nodes are larger) and occupancy probabilities (i.e., it is more costly to navigate through a region likely to have obstacles). With an updated scene percept S' and its associated plan π', the divergence in planning is the L2 difference in the cost of these shortest paths, $\delta\pi = \|c(\pi) - c(\pi')\|$.

Integration with adaptive computation The model interfaces these perception and planning systems via adaptive computation to enable selective processing of scene geometry along dimensions relevant to navigation. Given an initial S, the task relevance $\frac{\delta\pi}{\delta_k S}$ of each terminal node s_k is approximated when applying C_k. Adaptive computation then allocates the following computational steps according to these task-relevance scores – via $\texttt{softmax}(\frac{\vec{\delta\pi}}{\delta_k S})$. As S is updated, so too are task-relevance scores, adapting the allocation of the remaining computational steps, for a total of 150 steps.

Results

Model infers navigational affordances The model yields goal-conditioned scene geometry inferences: given input images that differ only in their door location (Fig. 3A, B) the model's attention maps and plans substantially diverge, which culminate with differences in the inferred geometries between these computation traces (Fig. 3C,D). The model thus generates a novel representational quantity: navigation-relevant geometry representations.

Explaining human change detection rates We evaluated this model on a dataset of human change detection task (N=45) that previously reported an impact of navigational affordances during the spontaneous processing of indoor scenes (Belledonne, Bao, & Yildirim, 2022). On each trial, subjects briefly viewed images of two rooms in succession (Fig. 3A,B). These scenes had a visible exit across the room (either to the left or right) and obstacles in between. Between the two images of a change detection trial (e.g., Fig. 3A), obstacle placement was modified (with door location constant). Across 30 trial-pairs that shared the same underlying obstacle placements (and modifications) and only differed in terms of door location, we compared the differences in human detection rates with the differences in model geometry inferences.

We find that the model explains significant variance in detection rates ($R^2 = .27, p < .01$, Fig. 3E). Critically, both multi-granularity and attention are necessary for this outcome as either ablation of the model fail to match behavior ($p = .35$, $p = .90$, respectively; Fig. 3E).

Conclusion

Our results offers a new perspective of perception – as a process of carving goal-conditioned world models in multi-granular generative models. Future work can use these traces to explain high-resolution behavioral (e.g., scene reconstruction) and neural (ECoG) measurements.

Acknowledgements

We'd like to thank the Yale Center for Computing Research and the members of the Cognitive and Neural Computation as well as the Perception and Cognition labs at Yale. This work was supported by AFOSR grant # FA9550-22-1-0041.

References

Belledonne, M., Bao, Y., & Yildirim, I. (2022). Navigational affordances are automatically computed during scene perception: Evidence from behavioral change blindness and a computational model of active attention. *Journal of Vision*, *22*(14), 4128–4128.

Belledonne, M., Butkus, E., Scholl, B. J., & Yildirim, I. (2023). Adaptive computation as a new mechanism of human attention. *in submission*.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in betavae. *arXiv preprint arXiv:1804.03599*.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). *Probabilistic models of cognition: Conceptual foundations* (Vol. 10) (No. 7). Elsevier.

Cusumano-Towner, M., Lew, A. K., & Mansinghka, V. K. (2020). Automating involutive mcmc using probabilistic and differentiable programming. *arXiv preprint arXiv:2007.09871*.

Fairbanks, J., Besançon, M., Simon, S., Hoffiman, J., Eubank, N., & Karpinski, S. (2021). *Juliagraphs/graphs.jl: an optimized graphs package for the julia programming language*. Retrieved from `https://github.com/JuliaGraphs/Graphs.jl/`

Finkel, R. A., & Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, *4*, 1–9.

Jakob, W., Speierer, S., Roussel, N., Nimier-David, M., Vicini, D., Zeltner, T., ... Zhang, Z. (2022). *Mitsuba 3 renderer.* (https://mitsuba-renderer.org)

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.