A Feedback Model of Flexible Context Guided Sensory Processing

Lakshmi Narasimhan Govindarajan*, Abhiram Iyer*, Ila Fiete
{lakshmin, abiyer, fiete}@mit.edu

McGovern Institute for Brain Research
K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center
MIT, 43 Vassar Street, Cambridge, MA 02139, USA

Abstract

Visual representations become progressively more abstract along the cortical hierarchy. These abstractions allow us to define notions like objects and shapes, and more generally organize sensory experience. Low-level regions, by contrast, represent simple local features of their inputs. How do the abstract, spatially non-specific, low-dimensional summaries of sensory information in high-level areas flexibly modulate the spatially specific and local low-level sensory representations in appropriate ways to guide attention, context-driven, and goal-directed behaviors across a range of tasks? We build a biologically motivated and trainable neural network model of dynamics in the visual pathway, incorporating lateral, feedforward, and local feedback synaptic connections, and excitatory and inhibitory neurons, together with long-range top-down inputs conceptualized as low-rank modulations of the input-driven sensory responses by high-level areas. We study this model in a visual counting task with images containing several novel 3D objects, each composed of new shape, size, and color combinations. First cued by a visual input depicting one object of a particular color or shape, the model uses its remembered representations of the cue to then modulate the perceptual and counting process for the subsequent image to report the number of objects with the cued color or shape. We show that this model is able to accurately and generalizably count novel combinations of novel objects with the cued attribute. We examine the neural representations that make this possible, shedding light on the nature of top-down contextual modulation of sensory processing and generating predictions for experiments.

Keywords: context; attention; convolutional RNNs; feedback; cued-visual search;

Introduction

We readily use abstract rules and cues to modulate our sensory perception. These forms of modulation include high-level feature-based attention (find Waldo; count the number of hoop shots, etc.), priming, cueing, and other contextual modulations. Such modulation allows us to locate items of interest more rapidly or accurately, and to follow directions or perform goal-directed computations. Understanding how and where sensory-driven neural responses and top-down processes interact has been a longstanding goal in computational cognitive neuroscience (Lamme & Roelfsema, 2000; Friston, 2005; Summerfield & De Lange, 2014; Bar et al., 2006). Extensive psychophysical experiments (Wolfe & Horowitz, 2004) and studies of neural gain modulation (Gilbert & Li, 2013; Gilbert & Wiesel, 1992) shed light on the phenomenon, however a fundamental mechanistic and conceptual problem remains open: what is the nature of the "modulatory homonculus" that decides, given a general and abstract cue, goal, or context, which low-level representations to modulate, and in what combination and in which topographic part of the input space, to sharpen the desired aspects of perception? We lack a cohesive computational framework to link these two levels of representation.

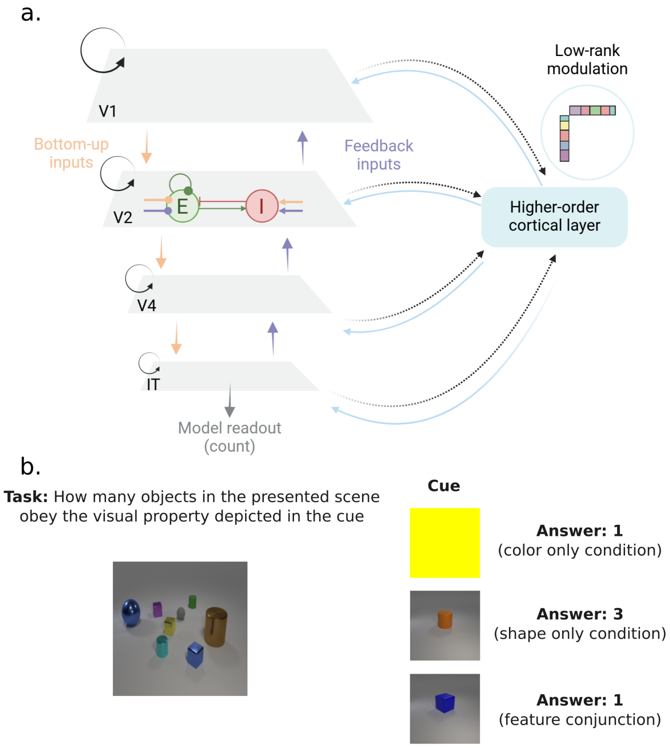

Figure 1: **Low-rank modulations drive context aware processing.** (a) We present a biologically motivated feedback model. Layers in the model are parameterized by recurrent Excitatory (E) and Inhibitory (I) neural populations that interact bidirectionally with a higher-order layer in a low-rank manner. (b) We train and evaluate this model within a cued-delayed-visual search paradigm. The model is tasked with extracting a "rule" from a visual cue input, presented first. The cue is removed, and the visual scene provided next, in which the model must count the number of objects that possess the cued property.

In this work, we combine the known architectures of visual cortex with advances in machine learning to introduce a biophysically-inspired, stimulus-computable model of the modulation of sensory representations by abstract task representations. We consider this model in the context of a challenging cued-delayed-visual search task. We start with a model endowed with several relevant details from biological circuits, including separate (tuned) excitatory and (weakly-tuned) inhibitory populations, lateral inhibition, inter-area feedback, and neuron types with distinct learnable time constants per type. We train our model with gradient descent to solve a (parametrically generated) cued visual search task in which the model has to count and report the number of geometric objects in a scene which possess a cued visual property (color, shape, or a conjunction of the two). Our model

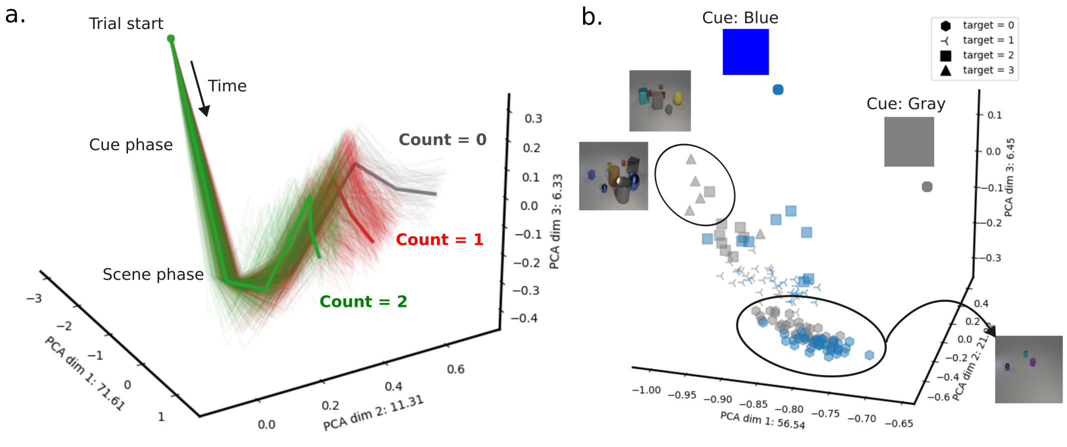

Figure 2: **Interpreting model dynamics in color cue-guided trials** For both (a) and (b) we perform dimensionality reduction on the recurrent states of the last layer in our model. (a) Trial-averaged dynamics reveal the emergence of numerosity information over time. For illustration purposes we pick 256 random trials for the first three classes. (b) Specifically analyzing the effect of disparate cues (here, Blue and Gray) on a variety of scenes. Despite widely varying bottom-up inputs from the different scenes, model dynamics are modulated in such a way that the final network states reflect consistent numerosity information.

learns to solve this task, outperforms state-of-the-art standard DNNs and LLMs, while being interpretable and having orders of magnitude fewer parameters. We believe that our approach holds great promise for generating several testable hypotheses and predictions for neuroscience.

Methods

Task and Stimuli. We study context-guided sensory processing within a cued visual search paradigm. We build input images based on the CLEVR (Johnson et al., 2017) dataset, a visual reasoning benchmark, and with it parametrically generate a counting task. Inputs are resized to a resolution of 128×128px. We render "cues" of three varieties (colors, 3D shapes with neutral colors, and colored 3D shapes) uniformly at random (Figure 1b). During model training and evaluation, we pair scenes with cues and task the model with counting the number of objects in the scene that obey the properties of the cue. By construction, there can be at most 9 objects in the scene consistent with the cue (0 is a valid answer when the scene does not contain any cue-consistent object).

Model. Figure. 1a illustrates our sensory perception backbone. Sensory layers incorporate relevant biological circuit blueprints, including Dale's law, lateral and top-down projections, as well as cell-type specific learnable neuronal time constants. Sensory layers additionally feed into a higher-order layer which computes a low-rank modulatory input (outer product of two rank one vectors) back into the sensory stream. The model is trained end-to-end on the visual counting task via gradient descent. The excitatory neurons of the last layer are transformed by a readout layer that is supervised with the true count (from 0-9) of cue-consistent objects in the scene.

Results & Discussion

Implicit models of context-guided visual processing fail to generalize. We evaluate a simplified version of our model without the low-rank modulations to understand if explicit modulation dynamics are necessary for this particular task. While such a baseline model learns to perform the task during training, evaluation on a held-out set of scenes was 73% accurate, and evaluation for held-out cues dropped significantly to 46%. We note that these experiments were only performed for color cues. Baseline models failed to learn the other cue conditions.

LLMs struggle with zero-shot visual cueing If smaller models struggle to generalize on this task, what about larger models? In an in-context manner (with a number of sample cues, images, and correct answers provided, followed by a new cue and image), we evaluated GPT-4's ability to solve this cued-visual task. Out of 30 random samples from our held-out dataset, GPT-4 was 36% accurate.

Low-rank modulations guide network dynamics and performance appropriately. We test and verify the ability of our proposed low-rank modulations and model to learn and perform well on all cue conditions. Specifically, our models were 89% accurate on color cues, 74% accurate on shape cues, and 66% accurate on conjunctive cues. We compute the chance performance to be around 30%. Visualization of our best performing model's trial-by-trial dynamics is shown in Figure. 2.

In conclusion, we introduce a feedback model of flexible context guided sensory processing. We show that models of this variety yield higher performance on a challenging visual counting task and also present a conceptual understanding for how bottom-up inputs are modulated by top-down processes.

Acknowledgments

We thank the McGovern Institute and the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center for supporting and funding this research.

References

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., ... others (2006). Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, *103*(2), 449–454.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Gilbert, C. D., & Wiesel, T. N. (1992). Receptive field dynamics in adult primary visual cortex. *Nature*, *356*(6365), 150–152.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2901–2910).

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, *23*(11), 571–579.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, *5*(6), 495–501.