Cognition as inference: a unifying account of some neural effects associated with mental imagery and attention

Ghislain St-Yves (stayves@musc.edu)
Dept. of Neurosciences, Neuroimaging Division, 135 Cannon St.
Charleston, SC 29425 USA

Thomas Naselaris (tnaselar@musc.edu)
Dept. of Neurosciences, Neuroimaging Division, 135 Cannon St.
Charleston, SC 29425 USA

Abstract

Mental imagery and attention are difficult to disentangle, suggesting a shared computational mechanism. We propose that imagery and attention are both inferences about the visual world conditioned on retinal input and a high-level anticipatory representation. Neural effects that have been previously associated with imagery occur when the anticipatory representation is determined by a memory and the retinal input is uninformative. Neural effects associated with spatial attention arise when the anticipatory representation is biased toward a location that is experimentally manipulated into or out of alignment with the presented stimuli. We tested the feasibility of this proposal with *in silico* experiments in a deep generative model that roughly analogizes the hierarchy of functionally distinct visual areas. We show that such a model jointly characterizes some of the ways that imagery and attention modulate activity levels, tuning to visual features, and the location and size of receptive fields. Based on these results, we consider the possibility that top-down volitional spatial attention is essentially equivalent to imagining the stimuli at the attended location, and imagining a stimuli involves reinstating the activity evoked by that stimulus near the top of a representational hierarchy.

Keywords: Spatial attention, mental imagery, unified generative model

Introduction

The subjective experiences of imagining and attending to a particular place or object are quite similar, and there is evidence that both imagery and attention depend on feedback from a high-level source (Pearson, Naselaris, Holmes, & Kosslyn, 2015; Dijkstra, Bosch, & van Gerven, 2019). These similarities raise the possibility that imagery and attention may be mediated by a shared computational mechanism.

Recently, we proposed a unifying account of imagery and vision (Breedlove, St-Yves, Olman, & Naselaris, 2018) that generalizes a set of ideas relating vision to probabilistic inference (Rao & Ballard, 1999; T. S. Lee & Mumford, 2003; Friston, 2005). Within this framework neural activity states encode visual features that can be efficiently combined to generate the kinds of stimuli we see in the natural world (Olshausen & Field, 1996). Vision is the process of inferring which of these features are likely to be out in the world, given retinal input. In our generalization of these ideas, imagery is visual inference conditioned not only on the (typically uninformative) current retinal input but also on a representation of a remembered or anticipated retinal input that is reinstated, or “clamped” in a high-level visual area.

In keeping with previous work (Chikkerur, Serre, Tan, & Poggio, 2010; Anderson, 2011), we reasoned that attention might also treated, like imagery, as a form of inference conditioned on both retinal input and a clamped high-level representation. If so, modulations of activation and tuning observed during attention experiments (Klein, Harvey, & Dumoulin, 2014; Hansen, Kay, & Gallant, 2007; Kay, Weiner, & Grill-Spector, 2015) might arise from the way that clamped representations are manipulated to align or misalign with retinal input in a network that performs inference.

To test the feasibility of this idea we constructed a deep generative network (DGN) and subjected it to experiments that permit comparisons between vision, imagery, and different attentional states. The architecture of the DGN is reminiscent of the predictive coding network of (Rao & Ballard, 1999). We further developed methods to ensure that it represents a hierarchy of visual features that roughly analogize some aspects of the features encoded by human brain activity.

We first reproduced our result for the expected responses during an analogue of the imagery experiments of Breedlove et al. (2018), in which a human subject was asked to imagine remembered stimuli (small natural image patches depicting a recognizable object) at different locations in the visual field. In our simulations retinal input was set to 0, while activity at the highest level of the DGN was clamped to a state that would have occurred had the imagined stimulus been seen. We compared activation and tuning to imagined vs. seen stimuli.

We then modeled how attentional effects might play out within the same experimental setting. Here, attention was held at a fixed location as visible images patches varied in content and location. We modeled spatial attention as “clamping” in which a high-level area is clamped to the activation state that would have occurred if the currently visible stimulus had been presented at the attended location. Similar to the case of imagery, we then compared the relative activation and tuning properties of receptive fields, this time across different attentional conditions.

Methods

We developed a learning algorithm to infer the weights connecting distinct processing levels of a linear-Gaussian deep generative network (DGN). The levels of the DGN correspond coarsely to different visual field maps in the brain. The lowest level corresponds to the retina; the highest level to an unspecified high-level visual area. In addition to maximizing a standard log-likelihood objective, our algorithm selected solutions in which units at each of the $L = 5$ levels of the network obeyed a roughly brain-like responses from a distribution of receptive field (RF) size, eccentricity and spatial frequency tuning (Fig. 1B). Importantly, this "desired" distribution reflected the RF and tuning attributes expected during vision only. The algorithm enforced no expectations about tuning or RF distributions during imagery or attention, which are assessed independently from the activity patterns generated by the model.

The DGN is described by its associated joint distribution:

$$p(\vec{r}) = \mathcal{N}(r_L; 0, \Sigma_L) \prod_{l=0}^{L-1} \mathcal{N}(r_l; U_l r_{l+1}, \Sigma_l) \qquad (1)$$

for which the parameters $\theta = \{U_0, \ldots U_{L-1}, \Sigma_0, \ldots \Sigma_L\}$ need to be estimated such that the expected responses under the posterior distribution $p(r_{l\setminus(0)}|r_0 = s; \theta^*)$ approximates the set of desired responses. All subsequent inferences refer to posteriors $p(r_{l\setminus(0,k)}|r_0 = r_0^*, r_k = r_k^*; \theta^*)$ over the activities of a subset of units, given activities of the remaining units. These latter units are called "clamped", since they are fixed to certain specific values. The stimuli $r_0 = s$ is always clamped to some value and we are interested in what happens when some higher-level units are likewise clamped to specific values e.g. $r_k = r_k^*$. We always assume that all units at this higher level are clamped since, though it is not a necessary requirement, it greatly simplifies the manipulations. The activities $a = (r_+^2 + r_-^2)^{1/2}$, where (r_-, r_+) is a pair of "simple cell" units with simple-cell like RF differing only by a phase shift of $\pi/2$, are estimated through a deterministic nonlinear read-out model after sampling the posterior distribution over the linear responses.

The training and validation (here consisting of sampling activities from the model) stimulus set consisted of tiny 32×32 images from the CIFAR-10 dataset, grayscaled and gaussian masked to smoothly remove boundaries (Fig. 1C, left). The resulting masked images are presented at 8 locations surrounding the center 5 pixels away in either direction (Fig. 1C, right).

In our simulated experiments vision is modeled by clamping the retinal input to the DGN to an image patch, leaving units at all higher levels unclamped (Fig. 1A, top). Imagery is modeled by clamping the retinal input to a blank image, while clamping the highest level to the expected activity state associated with seeing a particular image patch (Fig. 1A, middle). Attention is modeled by clamping the retinal inputs to a small image patch, while clamping the highest level to the expected activity state associated with seeing that patch at a location that remains fixed as the retinal input varies (Fig. 1A, bottom). In all cases, once the clamping configuration is specified we then perform exact Bayesian inference to obtain the expected activity states of all unclamped units (Bishop, 2006).

Given the activities sampled in this way, RFs were then independently estimated by a suitable encoding model (Gabor-fwRF, see St-Yves and Naselaris (2018)) under various states of vision, mental imagery and attention (Fig. 1A). We characterized the shift in RF tuning properties across conditions by their change in position, size and spatial frequency preference. All size and distances are expressed in units of "stimulus size" and all frequencies are in "cycle/stimulus".

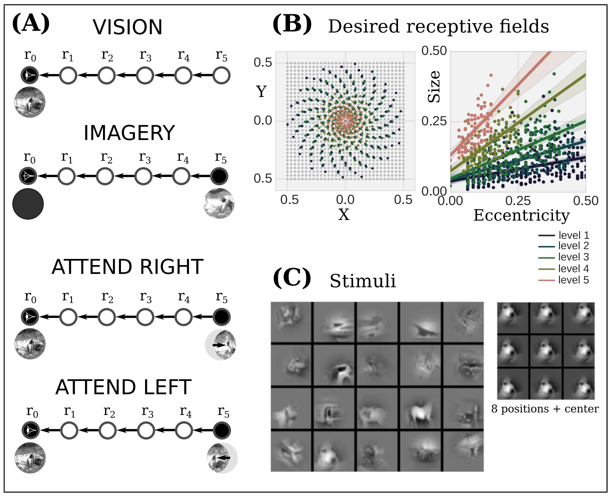
(A)
VISION
r_0 r_1 r_2 r_3 r_4 r_5
IMAGERY
r_0 r_1 r_2 r_3 r_4 r_5
ATTEND RIGHT
r_0 r_1 r_2 r_3 r_4 r_5
ATTEND LEFT
r_0 r_1 r_2 r_3 r_4 r_5
(B) Desired receptive fields
Y
0.5
0.0
0.5
0.5
0.0
0.5
X
Size
0.50
0.25
0.00
0.00
0.25
0.50
Eccentricity
level 1
level 2
level 3
level 4
level 5
(C) Stimuli
8 positions + center

Figure 1: Deep generative network and simulated experiments. (**A**) Probabilistic graphs and inference models of the deep generative network (DGN) used here to simulate responses during vision, imagery and attention. Each level corresponds to a distinct functional visual area labeled r_l. Vision (top) is modeled as pure posterior inference $p(r_{l\setminus(0)}|r_0 = s)$ while imagery (middle) and attention (bottom) are modeled as $p(r_{l\setminus(0,k)}|r_0 = r_0^*, r_k = r_k^*)$. The activity patterns in the conditioning sets are referred to as "clamped" in the main text; random variables are referred to as "unclamped". Imagery and attention differ in the content of the clamped activity patterns. (**B**) The desired distribution of receptive field properties that the DGN is trained to approximate during vision. (**C**) Examples of the masked and shifted stimulus set (left) with the 9 positions the stimulus could have been shown (right).

Results

Imagery

As in Breedlove et al. (2018), we show that previously observed neural effects associated with mental imagery arise as a consequence of conditional inference in the DGN described above. Signal (defined as the median (over a layer) of the variance of the activity given a subset of the stimuli) attenuation in low-level but not high-level areas results from distance-dependent decay from the source (Fig. 2A). Reduction of noise (Fig. 2A) results from clamping, which removes a source of variance (and therefore uncertainty) from the network. Changes to spatial frequency tuning to imagined relative to seen visual features (Fig. 2B) in low-level areas results from the biasing effect of feedback signals emanating from high-level areas that represent seen stimuli with lower spatial frequencies. The more foveal (Fig. 2C) and larger RFs (Fig. 2D) of low-level units during imagery also reflect the biasing effect of units in the level where clamping occurs.

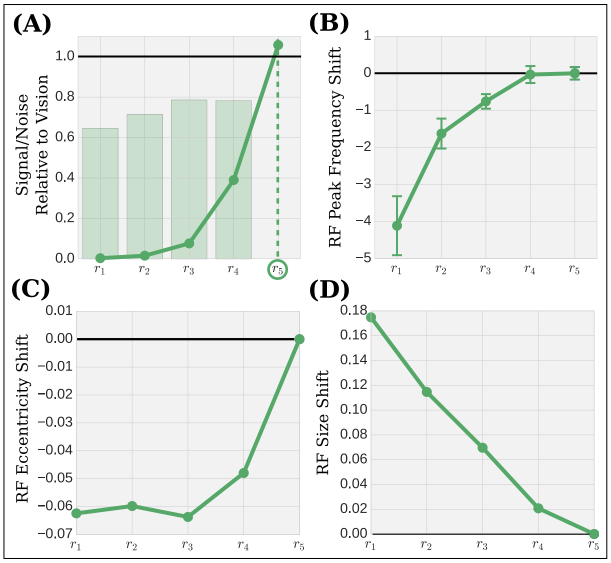

Figure 2: Signal and tuning properties for imagery compared to vision. (**A**) Variance of the signal (curves) and noise (bars) during imagery relative to vision across the 5 levels of representation in the DGN. Clamping occurs at level 5. (**B**) Shift in peak spatial frequency for imagined vs. seen stimuli. Negative values indicate a shift toward lower spatial frequency preference. (**C**) Change in average RF eccentricity for units during imagery and vision. (**D**) Change in average RF size. All tuning properties tend to increase in magnitude with distance from the clamped level.

Attention

To simulate attentional effects we modified the simulated experiment performed for imagery as well as the content of the clamped level. In this case the stimulus is always displayed as it varies in content and location. The clamped level represents the content of the stimulus currently on display but at an attended location that is fixed even as the stimulus varies. As has been observed in several previous studies (Cohen & Maunsell, 2009; Kay et al., 2015), units enjoy a relative boost in signal when their RFs align with both the currently displayed stimulus location and the attended location (Fig. 3A). When the attended location is fixed as stimuli vary in content and location, RF centers shift toward the attended location (Fig. 3B). In agreement with several other studies, the strength of this effect increases with ascent toward the clamped layer; because high-level areas typically have large RFs, the effect strength also shows a dependence on RF size (Fig. 3C). The patterns of attention-related changes in RF location are qualitatively consistent with effects reported in (Klein et al., 2014; Hansen et al., 2007).

Discussion

Our model reproduces many previously observed effects associated with mental imagery (Fig. 2) and, at this early stage, a handful of the most robust and well-known effects associated with attention (Fig. 3). It will be interesting to see what other phenomena are captured under this simple description, and how this overall design can be leveraged in systems with clearer behavioral significance. In the meantime, we here discuss some of the more intriguing interpretations of these preliminary results.

Our results reveal a “counter-gradient” of effects for imagery and attention. Imagery is marked by a monotonic decay of signal away from the clamped level, since no other source of variation is present. In contrast, during attention signals become more and more unaffected by the input supplied by the clamped level with distance from it. This basic pattern also explains the counter-gradient in RF tuning changes between imagery and attention. In low-level areas, the difference in tuning to seen and imagined stimuli is substantial; in high-level areas there is no difference (S. H. Lee, Kravitz, & Baker, 2012; Breedlove et al., 2018). In contrast, low-level areas show negligible tuning shifts across attentional conditions, while high-level areas do (Sprague & Serences, n.d.; Klein et al., 2014; Sheremata & Silver, 2015).

In our account, the specific changes in the tuning and RF attributes we observe results from the (linear) combination of two inputs: the stimulus and a coarse representation of a stimuli in the form of a clamped state. The exact ratio of this combination, as well as the intensity of the inputs, determines the magnitude of the effect. This change in interplay explains the reversal of the observed effect between imagery—the limit of vanishing stimuli—and vision with attentional state. Our account is thus at least conceptually consistent with the framework developed in Albright (2012).

In general, the otherwise puzzling tendency of RFs to be modulated by the task (e.g. spatial and feature attention, but also imagery) raises the question of how changes to RFs subserve the computational goals of vision (Kay et al., 2015; Carrasco, 2011; Klein et al., 2014; Vo, Sprague, & Serences, 2017)). Simple models such as this one may help to clarify the question. Our interpretation is that all these effects arise due to the specific way we query the generative model of the world i.e. that attention *is* a change in the computational goal. The effect that this change has on the representations, and on the RFs, is largely incidental to the computational goal. This is another reason why it is so hard to discern generality in the modulation of RFs, since RFs reflect functional connectivity and their modulations themselves are not the relevant causal factor—a point discussed in more detail in Anderson (2011).

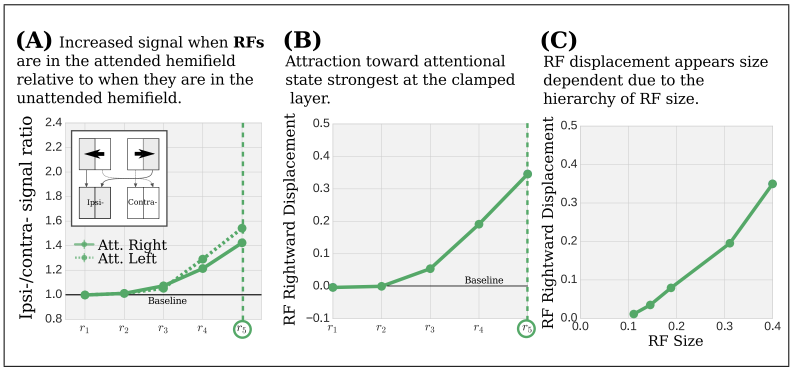

Figure 3: Signal and tuning properties for attention. (**A**) Relative signal across attentional states for units with RFs that, during vision, are located in the same visual hemifield as the numerator attended location. Clamping occurs in level 5. (**B**) Amount of rightward displacement of RFs between left and right attentional conditions. Displacement is the largest change in tuning and is always maximal at the clamped level. (**C**) Displacement vs. size of RFs. Note that, unlike attention, the smallest RFs furthest from the clamped level are most affected during imagery.

Finally, our results suggest an appealingly parsimonious interpretation of the subjective experiences associated with attention and imagery. In our account the neural effects of both imagery and attention derive from the same source of high-level feedback. This suggests that attention may be interpreted as mental imagery when a stimulus is present, while imagery may be interpreted as attention when the stimulus is not.

Acknowledgments

This work was supported by grant NIH R01 EY023384 to TN.

References

Albright, T. D. (2012). On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception. *Neuron*, *74*(2), 227–245.

Anderson, B. (2011). There is no such thing as attention. *Frontiers in Psychology*, *2*, 246.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breedlove, J., St-Yves, G., Olman, C., & Naselaris, T. (2018). Human brain activity during mental imagery exhibits signatures of inference in a hierarchical generative model. *bioRxiv*. doi: 10.1101/462226

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484 - 1525. (Vision Research 50th Anniversary Issue: Part 2)

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A bayesian inference theory of attention. *Vision Research*, *50*(22), 2233 - 2247. (Mathematical Models of Visual Coding)

Cohen, M., & Maunsell, J. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, *12*(12), 1594–1600.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *360*(1456), 815–36.

Hansen, K. A., Kay, K. N., & Gallant, J. L. (2007). Topographic Organization in and near Human Visual Area V4. *Journal of Neuroscience*, *27*(44), 11896–11911.

Kay, K. N., Weiner, K. S., & Grill-Spector, K. (2015, feb). Attention Reduces Spatial Uncertainty in Human Ventral Temporal Cortex. *Current biology : CB*, *25*(5), 595–600.

Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, *84*(1), 227–237.

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, *59*(4), 4064–4073.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America*, *20*(7), 1434–48.

Olshausen, B. A., & Field, D. J. (1996, jun). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends Cogn Sci*, *19*(10), 590–602.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Sheremata, S. L., & Silver, M. A. (2015). Hemisphere-dependent attentional modulation of human parietal visual field representations. *Journal of Neuroscience*, *35*(2), 508–517.

Sprague, T. C., & Serences, J. T. (n.d.). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, *16*, 1879.

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188 - 202. (New advances in encoding and decoding of brain signals)

Vo, V. A., Sprague, T. C., & Serences, J. T. (2017). Spatial Tuning Shifts Increase the Discriminability and Fidelity of Population Codes in Visual Cortex. *The Journal of Neuroscience*, *37*(12), 3386–3401.