Hypothesis-neutral models of higher-order visual cortex reveal strong semantic selectivity

Modeling neural responses to naturalistic stimuli has been instrumental in advancing our understanding of the visual system. Dominant computational modeling efforts in this direction have been deeply rooted in preconceived hypotheses (Yamins et al. 2014; Naselaris et al. 2011). Here, we develop a hypothesis-neutral computational methodology which brings neuroscience data directly to bear on the model development process. We demonstrate the effectiveness of this technique in modeling as well as systematically characterizing voxel tuning properties.

We leverage the unprecedented scale of the Natural Scenes Dataset (Allen et al. 2021) to constrain parametrized neural models of higher-order visual systems with brain response measurements and achieve novel predictive precision, outperforming the predictive success of state-of-the-art models. Next, we ask what kinds of functional properties emerge spontaneously in these response-optimized models? We examine trained networks through structural and functional analysis by running 'virtual' fMRI experiments on large-scale probe datasets.

Strikingly, despite no category-level supervision, since the models are optimized for brain response prediction from scratch, the units in the networks after optimization act strongly as detectors for semantic concepts like 'faces' or 'words', thereby providing one of the strongest evidences for categorical selectivity in these areas. The observed selectivity raises another question: are these units simply functioning as detectors for their preferred category or are they a by-product of a non-category-specific processing mechanism? To investigate this, we create selective deprivations in the visual diet of these models and study semantic selectivity in the 'deprived' networks, thereby also elucidating the role of specific visual experiences in shaping neuronal tuning.

Beyond characterizing tuning properties, we study the transferability of representations in response-optimized networks on different perceptual tasks. We find that the sole objective of reproducing neural targets, without any task-specific supervision, grants these networks intriguing functionalities. Together, this new class of response-optimized models combined with novel interpretability techniques reveal themselves as a powerful framework for probing the nature of representations and computations in the brain.

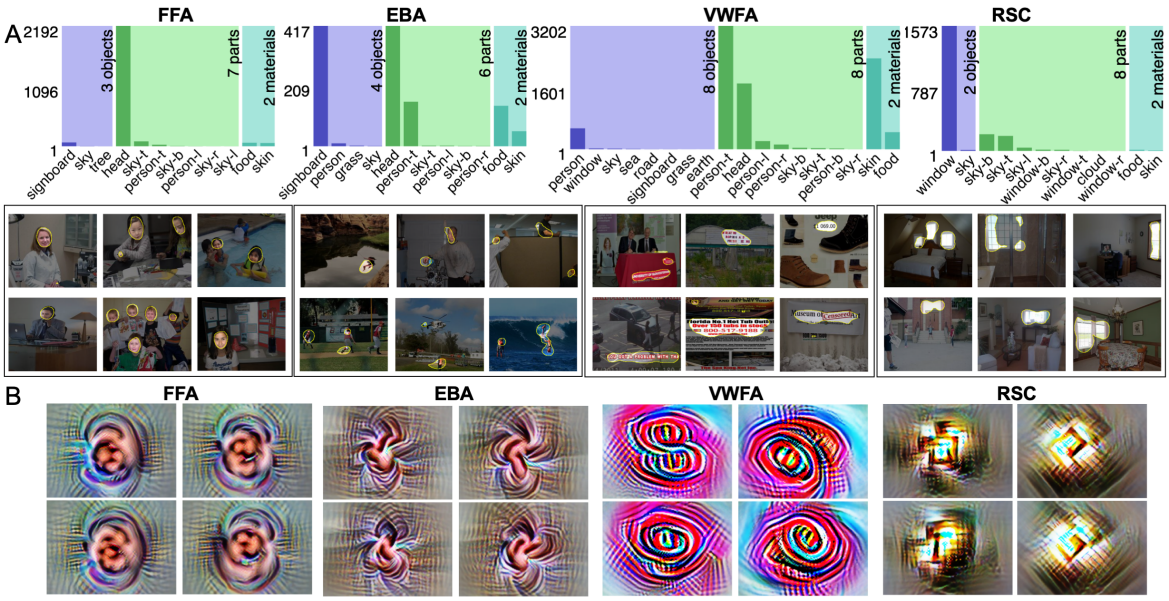

Figure 1: **A. Dissection.** Top: Matched concepts in the model for every region, i.e., the number of units showing high alignment with a human-interpretable concept. Contrary to object recognition networks which are trained with explicit label supervision and show a broad diversity of detectors, concepts in response-optimized networks are highly specific. Bottom: Activation of single units within response-optimized networks in response to an image is visualized as the region in the input space that elicits the top 1% activation in the corresponding filter output. **B. Structural analysis** through MEIs. Each column is a random voxel from a given ROI and each row a different initialization. Face-like visual forms, features akin to orthographic symbols, skin-colored shapes reminiscent of body parts and window-like features emerge spontaneously for FFA, VWFA, EBA and RSC respectively.

We optimize a biologically-motivated model starting from scratch to directly predict the recorded activity in the voxels of a given ROI, thereby learning the dimensions of variance that are important for these ROIs. We focus on four visual, category-selective ROIs: the fusiform face area (FFA), the extrastriate body area (EBA), the visual word form area (VWFA), and the retrosplenial cortex (RSC, a visual place selective area). We utilize a rotation-

equivariant core convolutional neural network architecture and a factorized linear readout that decouples spatial tuning from feature tuning (Klindt et al. 2017). Sharing the core network across subjects, sharing convolutional filters across visual field locations and orientations (translation and rotation equivariance) and a factorized readout jointly enable sample-efficient training of the response-optimized models. Quantitative comparisons showed that on the same subjects, response-optimized networks achieve parity with object-categorization networks in predicting neural responses to novel stimuli, offering a promising alternative tool to study brain representations. We further tested the generalization performance of all models to independent stimuli from new subjects through sample complexity analyses and observed that response-optimized models generalize much more efficiently to novel subjects than task-optimized or explicit categorical models. Response-optimized models further attained approximately 62-66% of the noise ceiling, yielding the most computationally precise models of these regions.

Selectivity, Tolerance and Clutter-invariance. We probe the learned features to shed light on the characteristics of optimized computational models and consequently, the cortical populations they model. We adapt the recent technique of network dissection (Bau et al. 2017) to generate verbal explanations for the behavior of model neurons in response-optimized networks. The goal is to measure the degree of alignment between a neuron's response properties against a large 'concept' dictionary, that spans not just objects but also other fine-grained categories like parts of objects, colors, materials and textures. We find that purely by optimizing for neural response prediction in respective category-selective regions (VWFA and FFA), we can get high-quality word and face detectors in the resulting networks (Figure 1; 'signboard' and 'head' are good proxies for 'words' and 'faces'). Importantly, the response-optimized model had no access to category labels during training; despite this, the model neurons selectively activate in response to high-level concepts like faces/words/body parts across a range of background clutter conditions and variations in appearance. For instance, modeled VWFA voxels respond to text in different backgrounds, fonts, colors, orientations etc., highlighting the invariant orthographic processing in VWFA. For RSC, we observe many 'window' detectors which may be linked to navigational affordances and functional scene understanding. To test whether developing a selectivity requires experience with the unique form of the preferred stimuli, we also train response-optimized models with the same architecture as before, but with a visual diet completely deprived of faces or bodies. Despite this selective deprivation, units in models of FFA & EBA retain their strong selectivity, suggesting that semantic selectivity could arise independent of domain-specific experience and merely as a by-product from tuning by experience with generic non-specific natural images.

Maximally exciting images (MEIs) reveal structured, high-level features We also perform an unconstrained optimization over input noise to discover the input that would result in maximal excitation of individual voxels and intriguingly find that the optimized images still contain human-recognizable patterns. For instance, full 'face'-like features and skin-colored complex shapes, loosely similar in form to body parts, emerge for FFA and EBA voxels respectively; MEIs for the VWFA voxels resemble orthographic units comprising curves and lines of different stroke widths and those for voxels within RSC are reminiscent of windows in different reference frames, which may be linked to RSC's role in spatial cognition.

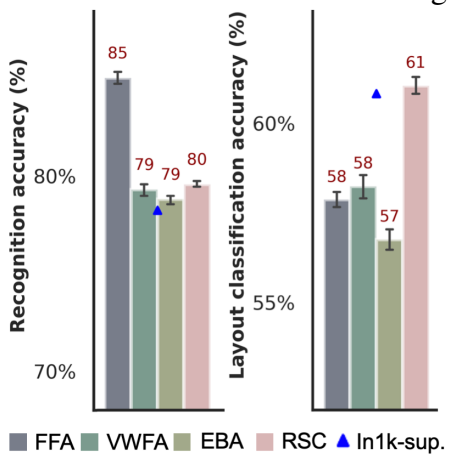

Figure 2: Transfer performance of response-optimized models. FFA and RSC representations achieve the best face recognition and scene layout estimation accuracy respectively, providing novel evidence for their hypothesized specialization.

Models reveal functional distinctions between visual regions We test existing functional specialization accounts which implicate FFA in face perception and RSC in spatial cognition (Mitchell et al. 2018) by simulating face discrimination and spatial layout prediction tasks with independent stimuli in response-optimized models of all brain regions. We find that FFA representations significantly outperform the representations in all other neural models in discriminability for faces, even outperforming the highly transferable representation of ImageNet trained networks, providing novel evidence for the role of FFA in face identification (Figure 2). In the layout estimation task, we find an opposite trend, with the RSC representations outperforming the rest of the neural models, highly suggestive of the role of RSC in scene understanding, particularly aspects relevant to spatial navigation.

In summary, we exploit data-driven computational modeling to place decades worth of prior work on functional specialization using reductionist approaches on firmer grounds of ethological relevance and broad generalization.