Dissecting visual population codes with brain-guided feature accentuation

Jacob S. Prince (jacob.samuel.prince@gmail.com)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Jeongho Park (jpark3@g.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Christopher Hamblin (chrishamblin@fas.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

George A. Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Talia Konkle (talia_konkle@harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA.

Abstract:

A typical view of the world contains diverse objects and people, evoking distributed patterns of activity across visual cortex. How do different functional subregions work together in parallel to process a complex natural scene? Here we introduce brain-guided feature accentuation, which can be applied to encoding models to highlight the specific image content responsible for driving different groups of fMRI voxels. As a proof of concept, we first show that we can attribute the activation of face-selective voxels to human faces, and of scene-selective voxels to the surrounding scene context, all within the same image. Next, we show that these accentuated stimuli can raise (and lower) model-predicted activation levels in category-selective regions of a held-out test subject. Finally, we show that feature accentuation may provide a means to decompose how different scene-selective regions (PPA, RSC) contribute to the representation of individual images. These approaches may eventually help interpret subsets of visual cortex with less-well-understood tuning, and could provide a new method to non-invasively exert control over population activity in human fMRI.

Introduction

A longstanding approach to studying visual representation has involved identifying the group of stimuli that elicit high activation in a particular subset of cortex (Kanwisher, 2010). While such paradigms have been highly productive, they cannot directly speak to the question of how an individual view of a complex scene is processed. In addition, some swaths of high-level visual cortex have tuning preferences that remain poorly understood. Here, we introduce *brain-guided feature accentuation,* a new method aimed at dissecting how different subregions of visual cortex operate jointly to process a given input.

Feature accentuation is an efficient computer vision interpretability method describing *what* and *where* in an image contributes to a given feature's response (Hamblin et al., 2024). Given the inputs of a seed image and a feature vector specified within a deep neural network (DNN) layer, gradient-based procedures update the seed image to: (a) increase the value of the target feature; and (b) regularize the updates by keeping the accentuated image similar to the seed image, in the latent space of an earlier layer. The accentuation thus emphasizes the location of the target feature with minimal distortion of the original image. Suppressing the target feature can be also achieved by flipping the sign of the feature loss. Here, we propose that this technique can be applied in a fruitful way to DNN encoding models of category-selective regions within the Natural Scenes Dataset (NSD; Allen et al., 2022). Our initial analyses suggest that image accentuations can be used in future fMRI studies to test how different image components underlie observed fMRI patterns.

Results

Our NSD encoding models were derived from layer relu7 of an AlexNet model trained on ImageNet using the self-supervised Barlow Twins objective. The DNN backbone was fixed for all analyses. We established the mean encoding axes of face-selective region FFA, and scene-selective regions PPA and RSC, by fitting a set of sparse positive-weighted linear regression models (Prince et al., 2024) using 500 subject-specific training images from NSD, separately for subj01 and subj02. These "encoding axes" (the 4096-dim vectors of relu7 regression weights, averaged over voxels) served as the target features input to the accentuation pipeline.

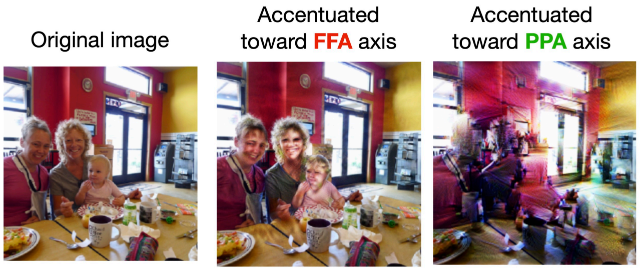
Original image
Accentuated toward **FFA** axis
Accentuated toward **PPA** axis

Figure 1: Accentuating an NSD stimulus (left), to increase predicted activation along either the FFA (center) or PPA (right) encoding axis.

We first accentuated a single NSD stimulus toward the encoding axes of FFA and PPA in NSD subj01. We chose a test image that contained several humans and various objects situated within an indoor scene, and which activated both FFA and PPA to a moderate degree in subj01. Then, we accentuated the image to drive higher predicted activity along the encoding axis of FFA. We observed that the content of the faces in the synthesized image became highly exaggerated, and that there was minimal impact on the background context. Strikingly, when instead accentuating along the encoding axis of PPA, we observed that the humans had been effaced from the image, with further emphasis of rectilinear content such as walls (**Fig 1**). Accentuating toward the encoding axes of different category-selective ROIs thus had intuitive consequences.

We next measured whether images accentuated toward one subject's encoding axes could also drive (or suppress) predicted activity in the encoding model of a separate subject (Tuckute et al., 2024). Examining a diverse set of 50 subject-overlapping NSD test images, we performed separate "drive" and "suppress" accentuations for each image toward subj01's FFA and PPA encoding axes. Then, we measured the predicted activity of these raw and accentuated stimuli in the encoding models of a validation subject (subj02). For the "drive" accentuations, we observed consistent

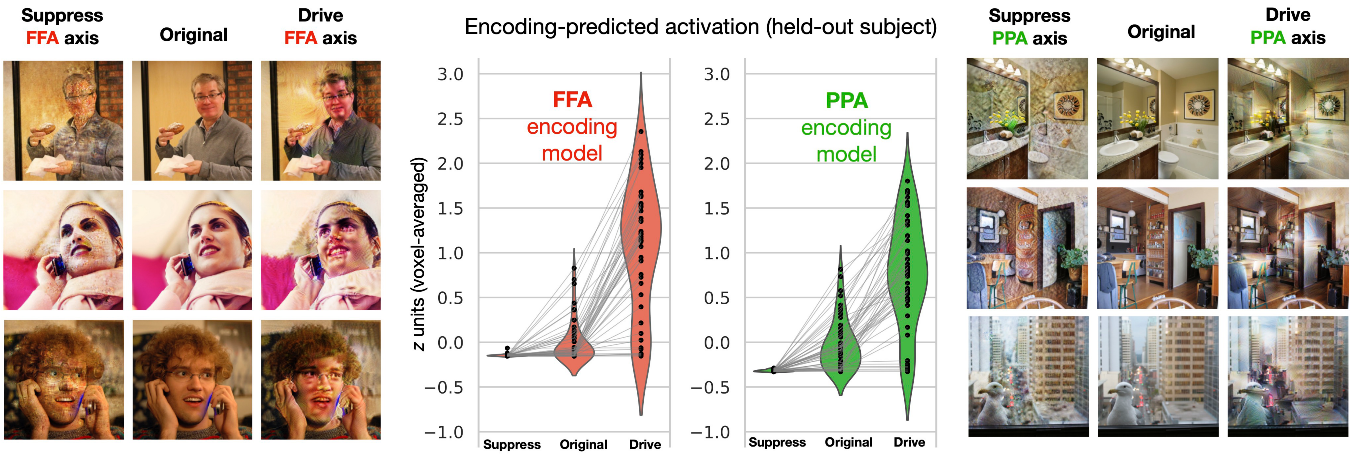

Figure 2: Accentuation of n=50 images to either drive or suppress predicted activity along FFA (left) and PPA (right) encoding axes. Accentuations are derived using encoding axes from subj01 (mean over ROI voxels' encoding weights). Synthesized drive/suppress images are then presented to encoding models from subj02. Violin plots (center) show predicted activation levels (*z* units) for each group of images (dots). 3 representative example images are plotted for each brain region, with corresponding feature accentuations.

increases in predicted activation for the validation subject, and for the "suppress" condition, we observed consistent decreases (**Fig 2**). Thus, similar features were underlying both subjects' encoding axes.

Finally, we attempted to synthesize stimuli that would drive one scene-selective region's encoding axis (PPA) while simultaneously suppressing that of a different scene-selective region (RSC). These areas are hypothesized to perform distinct functions within the scene processing system, but they have highly similar encoding axes ($r = 0.92$ for subj01, $r = 0.82$ for subj02). To test whether we could isolate the features that were uniquely represented by PPA, we accentuated the test stimuli toward the *difference* in vector space between the PPA and RSC axes. On average, we observed that the new stimuli drove higher predicted activity within subj02's PPA encoding model, and lower predicted activity in subj02's RSC encoding model (**Fig 3**), though the effect size was relatively small. Feature accentuation may thus help decompose how voxels with similar selectivity differentially contribute to the representation of individual images.

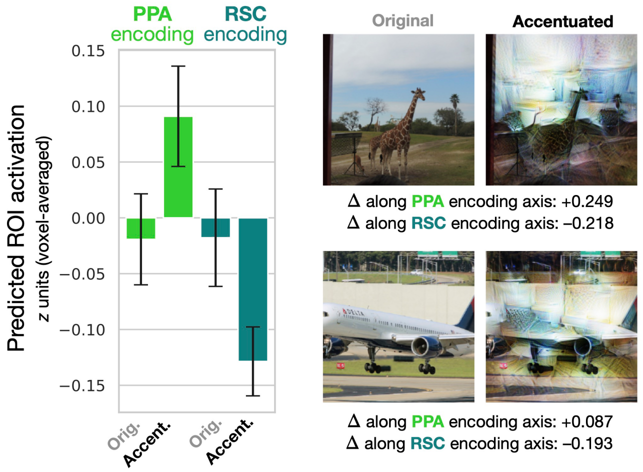

Figure 3: Accentuating toward the difference between PPA and RSC encoding axes. Bar graphs (left) show mean (+/- SEM) predicted activation to 50 original and accentuated images in a held-out subject. Two image examples are shown (right), with differences in encoding model predictions for each ROI.

Discussion

We have demonstrated that accentuating images toward ROI encoding axes can provide both qualitative insight into the nature of visual population codes, and perhaps, a quantitative means to control their activity levels. Our work joins a growing literature aimed at probing human visual representations using brain-guided image synthesis (Gu et al., 2022; Luo et al, 2024), and related work applying network "dissection" to deep encoding models (Khosla and Wehbe, 2022; Sarch et al., 2023).

A critical next step is to validate this paradigm by presenting raw and accentuated stimuli to the same fMRI subjects. The results here are a proof of concept: since there is a correlation between different subjects' FFA encoding axes, it is relatively unsurprising that images accentuated for one subject would predict high activity in another. The images themselves amount to a strong hypothesis–that the highlighted features should activate the target region when scanned. Validating this prediction would mark significant progress in describing the relevant features, while failure would strongly refute the DNN encoding setup and highlight the need for better models. Overall, these approaches may help provide refined models of how complex visual inputs are processed.

Acknowledgements

This work was supported by an NDSEG Fellowship (to J.S.P.) and an NSF CAREER grant (BCS-1942438; to T.K.)

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., . . . others (2022). A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1), 116–126.

Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., . . . Kuceyeski, A. (2022). Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247 , 118812.

Hamblin, C., Fel, T., Saha, S., Konkle, T., & Alvarez, G. (2024). Feature accentuation: Revealing 'what' features respond to in natural images. *arXiv* preprint arXiv:2402.10039.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107 (25), 11163–11170.

Khosla, M., & Wehbe, L. (2022). High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv* , 2022–03.

Luo, A., Henderson, M., Wehbe, L., & Tarr, M. (2024). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, 36.

Prince, J. S., Conwell, C., Alvarez, G. A., & Konkle, T. (2024). A case for sparse positive alignment of neural systems. In ICLR 2024 workshop on representational alignment.

Sarch, G. H., Tarr, M. J., Fragkiadaki, K., & Wehbe, L. (2023). Brain dissection: fMRI-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv* , 2023–05.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., . . . Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behavior,* 1–18.