Brain-Aligned Category-Selective Features from Contrastive Learning

Daniel Janini (janinidp@gmail.com)
Department of Education and Psychology, Freie Universität Berlin
Habelschwerdter Allee 45, 14195 Berlin, Germany

Radoslaw Cichy (rmcichy@zedat.fu-berlin.de)
Department of Education and Psychology, Freie Universität Berlin
Habelschwerdter Allee 45, 14195 Berlin, Germany

Abstract

Researchers have long debated the origins of category-selective visual cortex. Recently, some have argued that face- and scene-selective cortex can naturally emerge from contrastive self-supervised learning instead of domain-specific learning objectives. Here, we aggregated an image set for testing classic effects of the FFA and PPA. We ran replication fMRI experiments for these effects, characterizing the FFA and PPA's distinct feature tuning. We then applied this test battery to a self-supervised vision model, finding that its face- and scene-selective features naturally exhibit many of these effects as well. Our findings support the argument that properties of human category-selective cortex can emerge from contrastive learning objectives, though our test battery also revealed specific shortcomings that could be improved in future models.

Keywords: contrastive learning; faces; scenes

Introduction

What learning processes create face- and scene-selective visual cortex? One possibility is that these brain regions are formed by separate domain-specific learning objectives (e.g., face identification vs scene navigation) (Dobs et al., 2022). Recent work provides computational plausibility for an alternative account, showing that contrastive self-supervised learning can produce face- and scene-selective feature populations (Margalit et al., 2024; Prince, Alvarez, and Konkle, 2024). Here, we explore feature tuning in SimCLR-ResNet50, a self-supervised learning model trained on ImageNet. Our aim was to determine the degree to which this model exhibits the same detailed representational signatures as face- and scene-selective cortex.

Results

Identifying face- and scene-selective features.

In a block-design fMRI experiment, we used a functional localizer approach to identify the FFA and PPA in human participants (n=16). An analogous process was used to identify face- and scene-selective features in the model. These model feature populations will be referred to as model FFA and PPA. See Fig 1A for more information on the localizer approach.

Testing signatures of category-selective cortex.

We created an image set sampling 40 conditions from fMRI papers on the FFA and PPA (5152 images, see Fig 1B). These conditions probe a variety of visual properties including retinal size, curvature, animacy, real world size, face shapes, and scene structure. In a block-design fMRI experiment, we measured the mean activation to each of these 40 conditions in the FFA and PPA. Analogously, we measured the mean activation to each condition in model FFA and PPA. When considering all the conditions, activations in the model FFA and PPA correlated strongly with activations in human FFA and PPA (rho = 0.75 for FFA, 0.92 for PPA, Fig 1C). Next, we test specific contrasts, focusing on those that successfully replicated in our fMRI data (Fig 1C). All contrasts depend on paired t-tests at a threshold of $p<0.05$.

Animacy and Size. Category-selective regions overlap regions preferring animals, big objects, or small objects (Konkle and Caramazza, 2013). Likewise in our study, human and model FFA both preferred animals, while human and model PPA both preferred big objects. Mid-level visual features, as operationalized by Long, Yu, and Konkle (2018), were sufficient to elicit these preferences in both humans and models.

Curvature preferences. We replicated previous findings that human FFA and PPA exhibit opposite preferences for curvature properties (Yue, Robert, Ungerleider, 2020). Human and model FFA preferred curvy textures and objects, while human and model PPA preferred rectilinear textures and objects. Thus, category-selectivity covaries with mid-level curvature preferences.

Faces with texture variation. We replicated the finding that the human FFA generalizes to face shapes across a broad variety of textures, preferentially responding to animal faces, objects that look like faces (pareidolia), and cartoon faces (Tong et al., 2000; Wardle et al., 2020). In comparison to objects, Model FFA exhibited preferences for animal faces and for objects that look like faces, but not for cartoon faces. Thus, model FFA generalizes across some textures but not to simple contours alone.

Scene-specific effects. Activity in the human PPA depends on spatial scale, responding minimally to single objects, an intermediate amount to reachable surfaces, and most to navigable spaces (Josephs and Konkle, 2020). This finding was replicated in human PPA and was also found in model PPA. Next, we replicated findings that human PPA preferentially responds to empty rooms rather than randomly rearranged room surfaces, or to multiple object arrays (Kamps et al., 2016). However, model PPA did not show these effects, indicating that it lacks sensitivity to coherent room surface configurations.

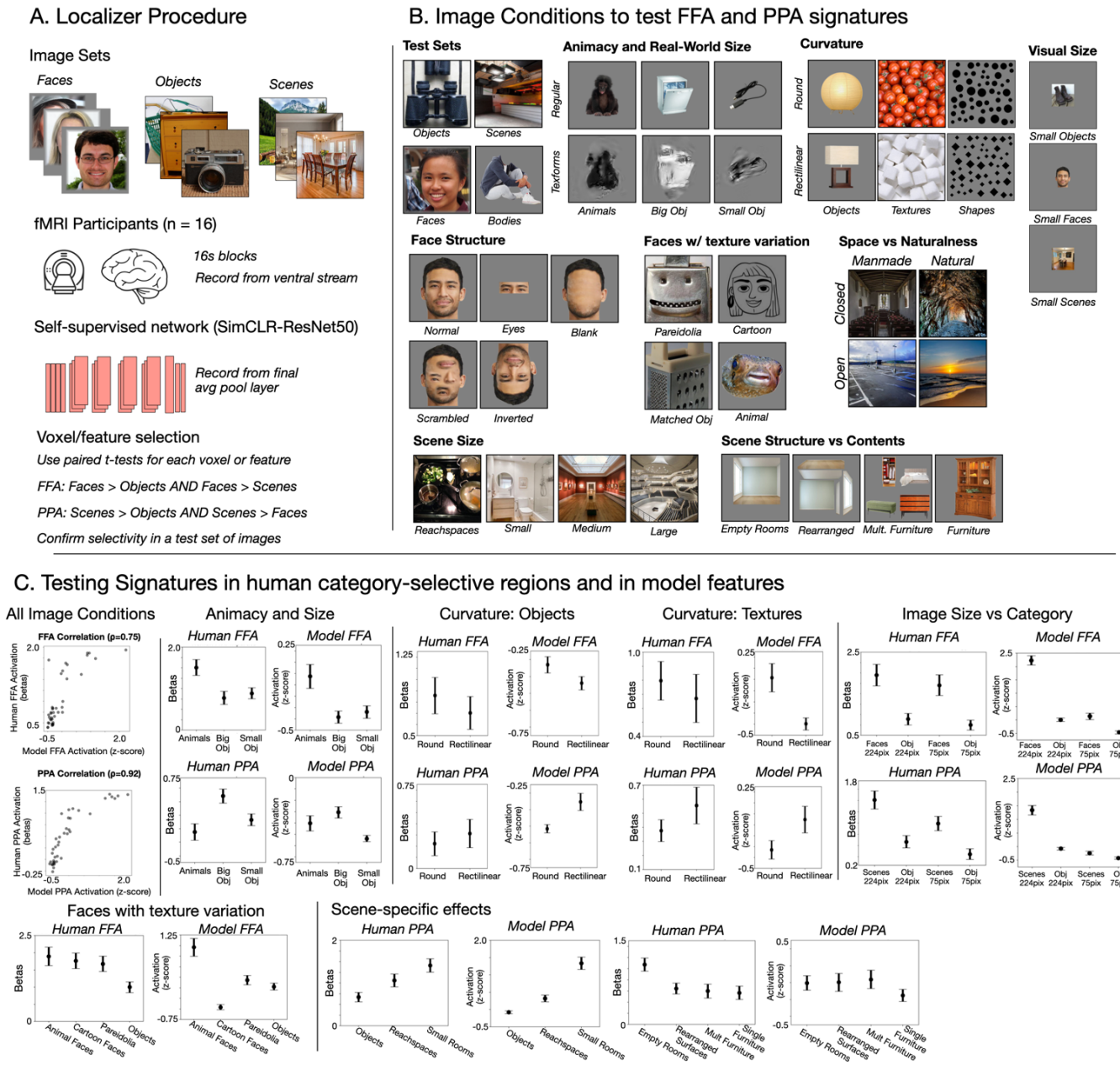

Figure 1. A. Depiction of the localizer procedure. B. Example images from the 40-condition test set. C. Specific effects in human participants and model features.

Comparing retinal size vs category effects. Activity in category-selective regions is influenced by both retinotopy and image category (Groen, Silson, and Baker, 2017). Here, we measure activations to faces, objects, and scenes at two sizes: 224x224 pixels or 75x75 pixels. Human FFA and PPA were more affected by category variation than image size variation. However, model features showed the opposite result, indicating insufficient tolerance to visual size variation, despite large amounts of visual size variation being included in the training augmentations.

Discussion

Contrastive self-supervised learning produced many aspects of category-selective feature tuning without innate biases for domain-specific learning. However, model features were overly influenced by visual size, model FFA did not generalize to line contours, and model PPA was not sensitive to room surface structure. Future experiments could investigate whether different learning factors narrow this gap, including naturalistic image diets, varied augmentation schemes, or fine tuning on domain-specific goals. The battery of tests compiled here provide a thorough benchmark for models of face- and scene-selective cortex.

Acknowledgements

This work was supported by a Humboldt Foundation Postdoctoral Research Fellowship (D.J.), German Research Council (DFG) grants (CI 241/1-3, CI 241/1-7, INST 272/297-2) (R.M.C.), and European Research Council (ERC) Consolidator grant (ERC-CoG-2024101123101) (R.M.C.). We thank the Center for Cognitive Neuroscience Berlin for their resources in conducting our fMRI experiment.

References

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances*, *8*(11), eabl8913.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102.

Josephs, E. L., & Konkle, T. (2020). Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, *117*(47), 29354-29362.

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2016). The occipital place area represents the local elements of scenes. *Neuroimage*, *132*, 417-424.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235-10242.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015-E9024.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, *112*(14), 2435-2451.

Prince, J. S., Alvarez, G. A., & Konkle, T. (2024). Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances*, *10*(39), eadl1776.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive neuropsychology*, *17*(1-3), 257-280.

Wardle, S. G., Taubert, J., Teichmann, L., & Baker, C. I. (2020). Rapid and dynamic processing of face pareidolia in the human brain. *Nature communications*, *11*(1), 4518.

Yue, X., Robert, S., & Ungerleider, L. G. (2020). Curvature processing in human visual cortical areas. *NeuroImage*, *222*, 117295.