A single computational objective may not be sufficient for human-like face discrimination

Ammar Marvi (amarvi@mit.edu) †
Department of Brain & Cognitive Sciences
McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Chengxu Zhuang (chenxuz@mit.edu) †
Department of Brain & Cognitive Sciences
McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Katharina Dobs (katharina.dobs@psychol.uni-giessen.de) ‡
Justus-Liebig University Giessen
Giessen, 35394, Germany

Nancy Kanwisher (ngk@mit.edu) ‡
Department of Brain & Cognitive Sciences
McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

† ‡ Denotes equal contribution

Abstract

Category-selective cortical regions are key to the diverse capabilities supported by the ventral visual pathway (VVP). How might these regions arise in development? One hypothesis suggests that category selectivity emerges from domain-general learning mechanisms. Supporting this idea, artificial neural networks (ANN) trained with self-supervised learning objectives on natural images recapitulate many features of the functional organization of the VVP and show human-like visual classification abilities. However, an adequate model of VVP development should account for all the behavioral abilities it supports, including face recognition. We therefore trained and tested a set of self-supervised networks on datasets composed of object or face images. When testing models via a naturalistic, zero-shot task we find that object-trained models achieve human-level accuracy in the object recognition task, yet face-trained models fail catastrophically at face discrimination. However, when provided labels after training and assessed via linear readout, all models - including those trained on faces - yield high discrimination accuracy, approaching performance of their supervised counterparts. Thus human-like face recognition may not develop from domain-general learning mechanisms. Instead, a single computational objective may only suffice if given a prior on the number, grain, or identity of output categories.

Keywords: face discrimination; self-supervised learning objectives; neural organization; deep neural networks

Introduction

Extensive evidence from behavioral and neural investigations supports the existence of multiple, domain-specific components of the mind and brain, each of which processes a specific kind of information, from perceptual properties of faces, scenes, and speech to abstract, distinctively-human domains of music, language, and other people's thoughts (Kanwisher, 2010; Fedorenko, Behr, & Kanwisher, 2011; Saxe & Kanwisher, 2003; Norman-Haignere, Kanwisher, & McDermott, 2015). How might these systems arise over development? According to one classical view, domain specificity develops from innate, core knowledge (Chomsky, 1965; Spelke & Kinzler, 2007; Fodor, 1983) and tailored learning mechanisms (Gallistel, 1990). An alternative view is that a single, domain-general mechanism shapes existing neural architecture (Arcaro & Livingstone, 2024), taking advantage of the structured statistics within natural input to produce specialized systems. Which of these hypotheses best accounts for the ventral visual pathway and the perceptual abilities it supports?

A powerful lever for answering this question comes from machine learning methods that ask: what kinds of abilities and neural organization can arise in principle under various learning objectives? Specifically, artificial neural networks (ANN) trained on labeled sets of images exhibit internal network activity (Yamins et al., 2014), category selectivity (Blauch, Behrmann, & Plaut, 2022; Dobs, Martinez, Kell, & Kanwisher, 2022), and behavioral performance (Rajalingham et al., 2018; Dobs, Yuan, Martinez, & Kanwisher, 2023) similar to that of humans and non-human primates. However, these networks have been criticized for their reliance on labeled data and non-trivial supervision. Recent self-supervised learning objectives (T. Chen, Kornblith, Norouzi, & Hinton, 2020) attempt to solve this problem using a domain-general contrastive loss to form representations of unlabeled natural images. This training program has produced ANNs with comparable fit to brain data (Zhuang et al., 2021; Konkle & Alvarez, 2022), including category-selective regions (Finzi, Margalit, Kay, Yamins, & Grill-Spector, n.d.; Margalit et al., 2023; Lee et al., 2020; Prince, Alvarez, & Konkle, 2023), as well as near human performance on object recognition (Zhuang et al., 2021). But adequate models of the visual system should account for the full extent of human ability. Does this single computational objective–without domain-specific priors–account for human face recognition abilities?

To find out, we compared the performance of ANNs trained with either supervised or self-supervised learning objectives on datasets containing either objects or faces. We find that–while models perform comparably on object classification when trained with or without supervision, the performance of self-supervised models falls far below their supervised counterparts in fine-grained, zero-shot face discrimination. These findings suggest that although domain-general learning mechanisms may suffice to account for neural and behavioral properties of human object recognition, they are not (yet) sufficient to account for human face recognition, and that perhaps the development of precise face-recognition may require some form of domain-specific prior.

Methods

We pre-trained ANNs using the same Resnet-50 architecture on datasets of varying size (400,000 or 1,200,000) and content (faces or objects) under either a supervised or self-supervised learning objective. We implemented a variety of established self-supervised algorithms: SimCLR (T. Chen et al., 2020), MOCOv2 (X. Chen, Fan, Girshick, & He, 2020), & BYOL (Grill et al., 2020).

We tested object-trained models on object recognition and face-trained models on face discrimination using three different measures. First, we evaluated each model's ability to discriminate stimuli via a zero-shot match-to-sample task (Dobs et al., 2023), where one of two images were matched to a target category based on which image produced activations that better resembled those of the target image. This is a naturalistic task more similar to real-world face recognition. Thus we compared model accuracy against human performance (n=1200) collected via Prolific and previously reported data in Dobs et al. (2023). Second, we trained a linear classification model following Dobs et al. (2023) to perform 100-way discrimination on held-out stimuli from 100 categories. Layer

activations from the model were used to predict an exemplar's category using an SVM with 10 folds of leave-one-out cross-validation. Third, we tested our models on a linear transfer learning task commonly used in machine learning evaluations (Zhuang et al., 2021; T. Chen et al., 2020). This metric utilizes a face dataset that is nearly 3,000 times larger than the previous SVM, and models are trained to discriminate between 8,000 rather than 100 face identities. Similar increases in dataset size were applied to the object dataset. These latter two experiments aim to exhaustively test the utility of representations for downstream categorization tasks. Note that none of the tested categories were included in the smaller dataset used for training, but we cannot exclude some overlap for the larger training dataset.

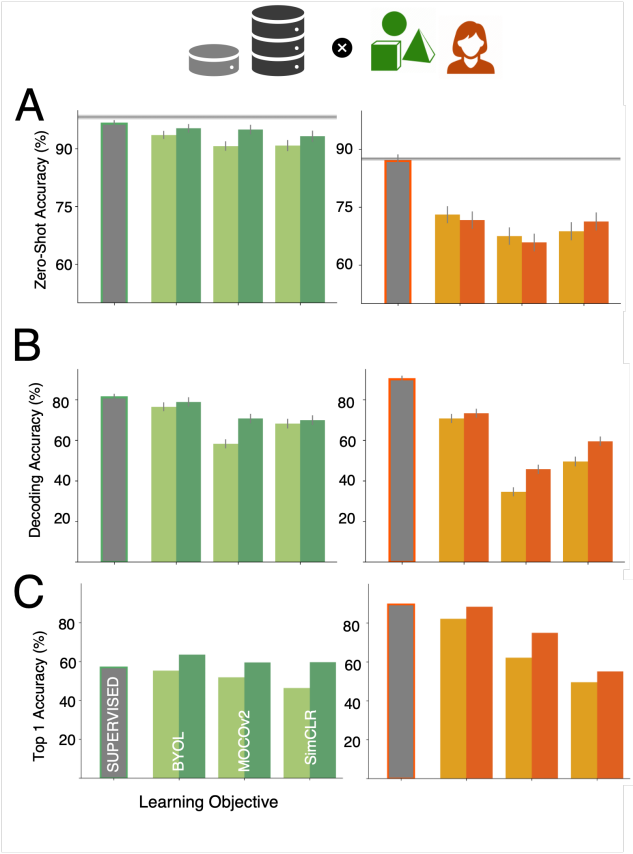

Figure 1: Model and human performance on face discrimination and object categorization tasks. Green bars represent object-trained models, orange face-trained, and gray supervised models. Darker shades indicate models trained on a larger dataset (1,200,000) and lighter shades on a smaller dataset (400,000). (A) Zero-shot performance on the match-to-sample task. Horizontal gray lines indicate human performance. Error bars indicate 95% bootstrapped CI (B) Decoding accuracy of the SVM trained for 100-way discrimination. Error bars represent SEM. (C) Top-1 accuracy on the linear transfer learning task. Object-trained models recapitulate established ML results.

Results

The accuracy of each model on each of the three measures is shown in Figure 1. Results were collected from multiple model layers; only the best-performing layers are shown here.

Zero-Shot Match-to-Sample

We first evaluated discrimination accuracy in a zero-shot match-to-sample task (chance $= 50\%$). Humans have high discrimination accuracy of 98.2% for objects and 87.5% for faces, and supervised models show near-human performance in both tasks, achieving 96.5% and 87.0% in the object and face tasks. However, the self-supervised models exhibit a strikingly different pattern: performance remains nearly as high for objects but drops precipitously for faces.

Linear Decoding

SVM Performance In contrast to the findings for the zero-shot matching task, SVM discrimination performance using layer activations was nearly as high for the BYOL self-supervised models as for the supervised models, for both faces and objects. Within the self-supervised models, performance of MOCOv2 and SimCLR was lower than for BYOL.

Linear Transfer Learning Performance as measured by the linear transfer learning task (chance $< 0.1\%$) showed a similar pattern to the results for SVM, although the drop in performance for MOCOv2 relative to BYOL was greater for faces.

Discussion

Our main finding is that although self-supervised models perform nearly as accurately as supervised models and humans on the zero-shot object matching task, the same learning objectives perform dramatically worse on a face matching task. This result points to an important limitation in the ability of a domain-general learning mechanism: face recognition performance is low when labeled output categories are not provided. However, we also find that if output categories are provided after the model is trained, as in our SVM and Transfer Learning measures, the selective drop on face discrimination in self-supervised models is reduced (and is almost gone for BYOL). Thus a single learning objective on its own, without the provision of labels, cannot account for human performance on both object and face recognition. Instead, it may be that human-like face recognition cannot be learned without a prior on the number, grain, or identity of the categories within this space.

Acknowledgments

This work was supported by the National Institutes of Health funded by grant EY033843 to NK, and the Center for Brains, Minds, and Machines (CBMM) funded by the NSF STC award CC- 1231216

References

Arcaro, M., & Livingstone, M. (2024). A Whole-Brain Topographic Ontology. Retrieved from

https://doi.org/10.1146/annurev-neuro-082823- doi: 10.1146/annurev-neuro-082823

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022, 1). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, *119*(3), e2112566119. Retrieved from https://doi.org/10.1073/pnas.2112566119 doi: 10.1073/pnas.2112566119

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, 2). A Simple Framework for Contrastive Learning of Visual Representations.

Chen, X., Fan, H., Girshick, R., & He, K. (2020, 3). Improved Baselines with Momentum Contrastive Learning.

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (50th ed.). The MIT Press. Retrieved from http://www.jstor.org/stable/j.ctt17kk81z

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). *Brain-like functional specialization emerges spontaneously in deep neural networks* (Vol. 8; Tech. Rep.). Retrieved from https://www.science.org

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023, 8). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(32). doi: 10.1073/pnas.2220642120

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011, 9). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16428–16433. doi: 10.1073/pnas.1112937108

Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K., & Grill-Spector, K. (n.d.). A single computational objective drives 1 specialization of streams in visual cortex 2. Retrieved from https://doi.org/10.1101/2023.12.19.572460 doi: 10.1101/2023.12.19.572460

Fodor, J. A. (1983). *The Modularity of Mind*. The MIT Press. doi: 10.7551/mitpress/4737.001.0001

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA, US: The MIT Press.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... Valko, M. (2020, 6). Bootstrap your own latent: A new approach to self-supervised Learning.

Kanwisher, N. (2010, 6). 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. Retrieved from https://doi.org/10.1073/pnas.1005062107 doi: 10.1073/pnas.1005062107

Konkle, T., & Alvarez, G. A. (2022, 12). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1). doi: 10.1038/s41467-022-28091-4

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & DiCarlo, J. J. (2020, 1). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020.07.09.185116. Retrieved from http://biorxiv.org/content/early/2020/07/10/2020.07.09.18 doi: 10.1101/2020.07.09.185116

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2023, 1). A Unifying Principle for the Functional Organization of Visual Cortex. *bioRxiv*, 2023.05.18.541361. Retrieved from http://biorxiv.org/content/early/2023/05/18/2023.05.18.54 doi: 10.1101/2023.05.18.541361

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281–1296. doi: 10.1016/j.neuron.2015.11.035

Prince, J. S., Alvarez, G. A., & Konkle, T. (2023, 1). A contrastive coding account of category selectivity in the ventral visual stream. *bioRxiv*, 2023.08.04.551888. Retrieved from http://biorxiv.org/content/early/2023/08/07/2023.08.04.55 doi: 10.1101/2023.08.04.551888

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018, 8). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, *38*(33), 7255. Retrieved from http://www.jneurosci.org/content/38/33/7255.abstract doi: 10.1523/JNEUROSCI.0388-18.2018

Saxe, R., & Kanwisher, N. (2003, 8). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835–1842. doi: 10.1016/S1053-8119(03)00230-1

Spelke, E. S., & Kinzler, K. D. (2007, 1). *Core knowledge* (Vol. 10) (No. 1). doi: 10.1111/j.1467-7687.2007.00569.x

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, 6). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021, 1). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3). doi: 10.1073/pnas.2014196118