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

Ammar I. Marvi[1,2 †], Chengxu Zhuang[1,2 †], Katharina Dobs[3 ‡], Nancy G. Kanwisher[1,2 ‡]

[1]Massachusetts Institute of Technology, Cambridge MA, USA
[2]McGovern Institute for Brain Research, Cambridge MA, USA
[3]Justus-Liebig University Giessen, Giessen, 35394, Germany
†,‡Equal contribution
Contact: amarvi@mit.edu

Original design by Vivian C. Paulun

QUESTION

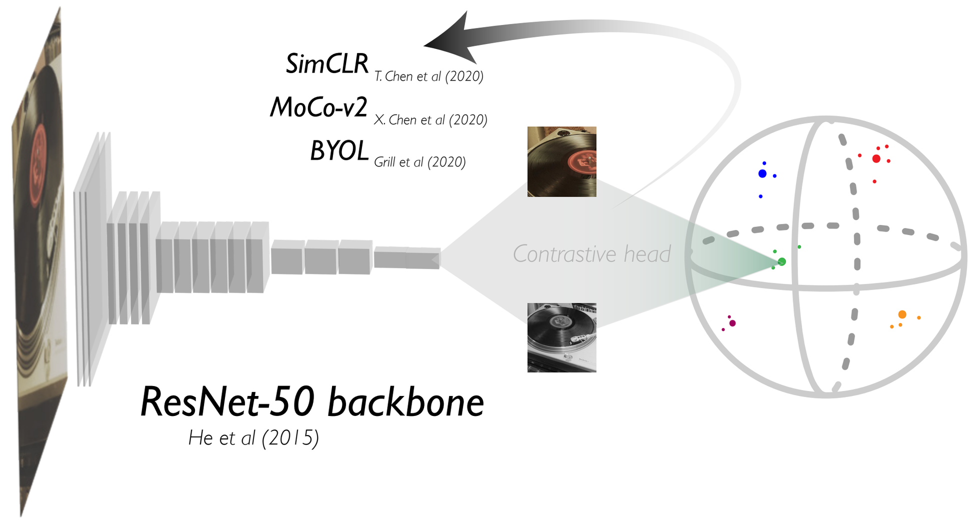
Recent evidence suggests that domain specificity develops from a ***single, domain-general computational objective***

Can such an objective account for human-like face discrimination?

Margalit et al (2024)
Finzi et al (2023)
Prince et al (2023)

TRAINING

We trained deep neural networks (DNNs) with unlabeled images and a self-supervised loss function



DNNs were trained on diets consisting *only* of faces ***or*** objects...

...and were subsequently tested on their respective category

Object classes manually selected to exclude face & face-like images (Dobs et al 2022)

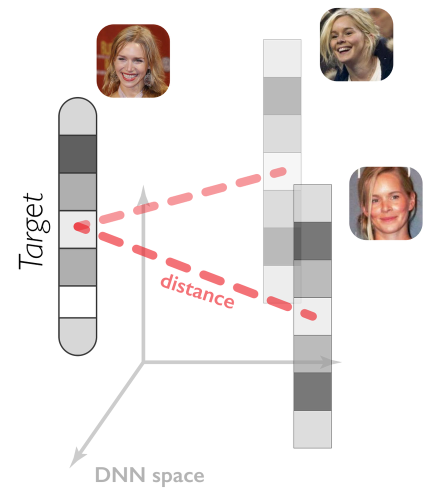
MATCH-TO-SAMPLE

Zero-shot discrimination

In the manner of Dobs et al (2023)

Evalute discriminaton performance on novel face & object classes

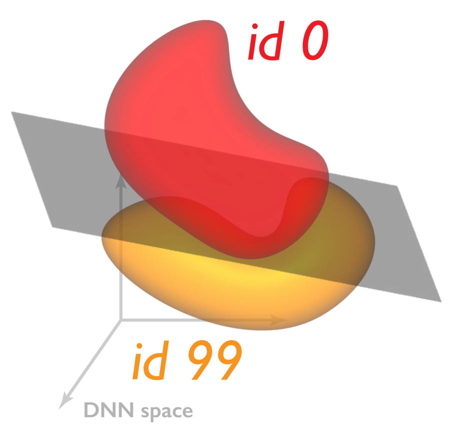
How do DNNs compare to human performance *(n=3,732 subjects)*



LINEAR DECODING

Support vector machine (SVM)

In the manner of Dobs et al (2022)



Train a linear SVM to decode 100 face & object classes *held-out* during training

Perform ten-fold leave-one-out cross validation over1,000 total images

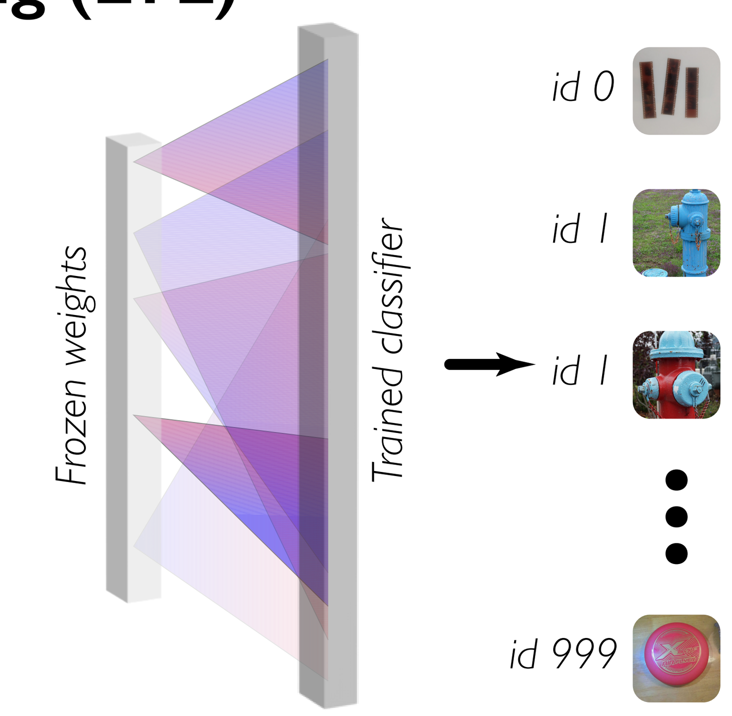
Increase training & the number of classes...

Linear transfer learning (LTL)

In the manner of Zhuang et al (2021)

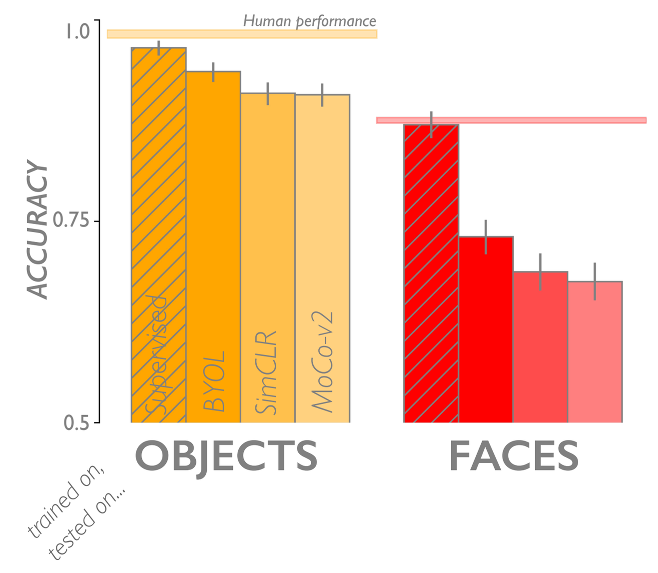
Train a linear classifer to decode the full set of classes in the original datasets

Perform gradient descent over millions of images to classify thousands of classes



RESULTS

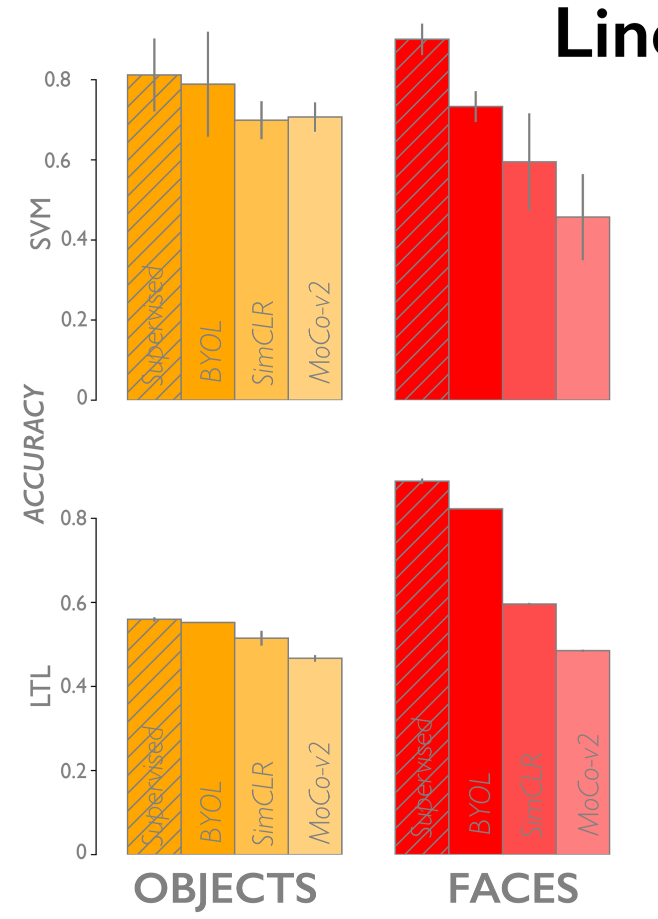
Match-to-sample *(zero-shot)*



Supervised networks perform comparably to humans

Face but not object discrimination drops precipitously in self-supervised networks

Linear decoding



Object-trained networks exhibit robust discrimination, with and without supervision

Self-supervised face discrimination continues to lag. Performance slightly recovers under some learning objectives, *provided labels and a linear readout*

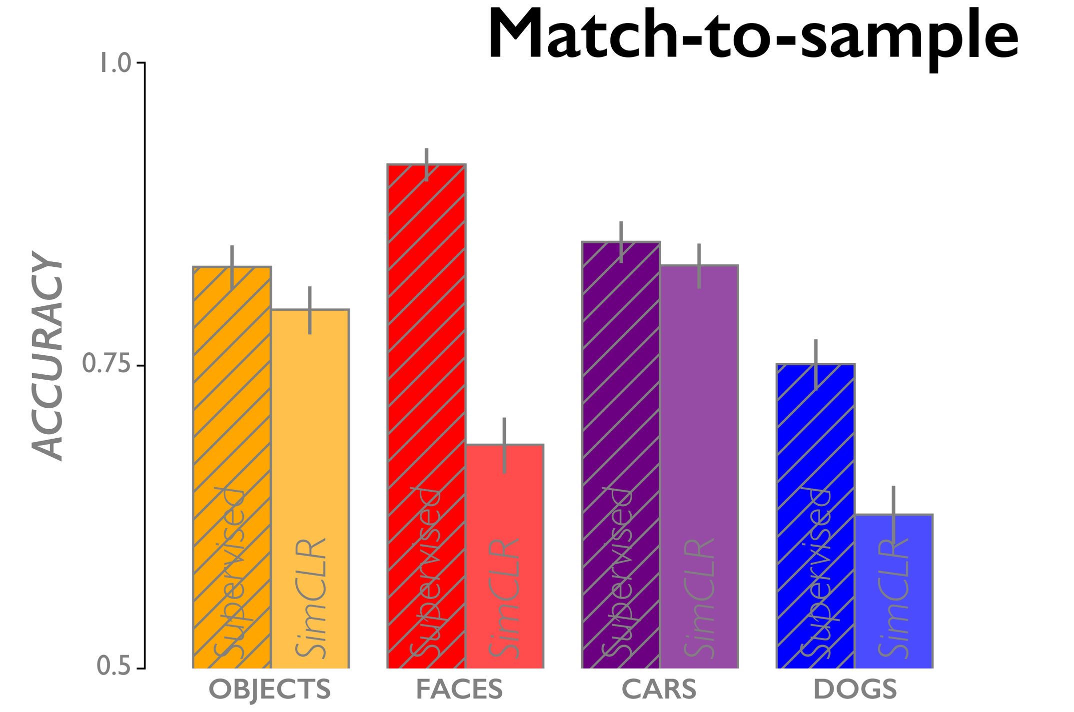
...FACES?

Are faces special, or do these results extend to fine-grained discrimination of *any* category?

To find out, we trained new networks on diets consisting only of objects, faces, *cars*, or *dogs*

n = 100,000 images per category

Match-to-sample



Preliminary results show a performance drop for *some*, but *not all*, fine-grained categories

DISCUSSION

A single computational objective, in the absence of labeled data, does not account for both object and face discrimination

Human-like face discrimination may require a prior on the number, grain, or category of labels in the representation space

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age.
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations.
Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved Baselines with Momentum Contrastive Learning.
Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. In Sci. Adv (Vol. 8). https://www.science.org
Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). 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). https://doi.org/10.1073/pnas.2220642120
Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K., & Grill-Spector, K. (2023). A single computational objective drivescspecialization of streams in visual cortex. https://doi.org/10.1101/2023.12.19.572460
Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised Learning.
He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition.
Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. Neuron. https://doi.org/10.1016/j.neuron.2024.04.018
Prince, J. S., Alvarez, G. A., & Konkle, T. (2023). A contrastive coding account of category selectivity in the ventral visual stream. BioRxiv, 2023.08.04.551888. https://doi.org/10.1101/2023.08.04.551888
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge.
Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences, 118(3). https://doi.org/10.1073/pnas.2014196118