Disentangled Face Representations in Deep Generative Models and the Human Brain

Paul Soulos (psoulos1@jh.edu)
Department of Cognitive Science, 3400 N Charles St
Baltimore, MD 21218

Leyla Isik (lisik3@jh.edu)
Department of Cognitive Science, 3400 N Charles St
Baltimore, MD 21218

Abstract:

Despite decades of research, much is still unknown about the computations carried out in the human face processing network. Recently deep networks have been proposed as a computational account of human visual processing. While they provide a good match to neural data throughout visual cortex, they lack interpretability. Here we use a new class of deep generative models, disentangled representation learning models, which learn a low-dimensional latent space that "disentangles" different interpretable dimensions of faces, such as rotation, lighting, or hairstyle. We show that these disentangled networks are a good encoding model for human fMRI data and allow us to investigate how semantically meaningful face features are represented in the brain. We find that several interpretable dimensions, including both identity-specific and identity-invariant dimensions, are distributed widely across the face processing system. The remaining "entangled" representations may be the basis of identity recognition in the brain. These disentangled encoding models provide an exciting alternative to standard "black box" deep learning models and have the potential to change the way we understand face processing in the human brain.

Keywords: Neural Networks; Generative Models; Disentangled Representations; fMRI; Face Processing; Encoding models

Humans are very good at recognizing faces despite variability in both identity preserving transformations (such as 3D rotation and lighting), and identity-relevant, stable features (such as facial features or skin tone). The complexity of face processing can be seen in the relatively poor decoding of face identity from fMRI data compared to other visual categories (Kriegeskorte et al., 2007). Recently, deep generative models have been shown to provide a good match to human fMRI data as well as high decoding accuracy of individual faces (VanRullen & Reddy, 2019). However, like most deep learning models, they lack interpretability.

Figure 1: Images generated by dVAE when single dimensions corresponding to smile (top) and 3D rotation (bottom) are varied.

Results

Disentangled Generative Models

We test a new class of neural networks, disentangled generative models (specifically a disentangled variational autoencoder or dVAE), as a model of face processing in the brain. These models isolate semantically meaningful factors of variation in individual latent dimensions. Based on hyperparameter search (Duan et al., 2019), we use FactorVAE (Kim & Mnih, 2018) with 24 latent dimensions. The dimensions capture changes in interpretable factors like smile and 3D rotation. When each latent dimension is changed, images generated by the network smoothly change along these single dimensions (Fig 1). Out of the 24 latent dimensions, human annotators agreed on semantic labels for 16, while the other 8 were considered entangled or did not control a clear transformation.

We compare our disentangled model against a standard entangled generative VAE (Kingma & Welling, 2014) matched in terms of training and parameters, as well as the discriminatively trained VGG-Face (Parkhi et al., 2015). To match model dimensions, we reduce the dimensionality of the VGG-Face representations to the first 24 principal components. The dVAE and VAE latent dimensions are highly correlated (CCA r = 0.92), but the dVAE and VGG were only moderately correlated (CCA r = .52).

Encoding Model

We use the publicly available fMRI dataset from VanRullen & Reddy (2019), which includes four subjects viewing roughly 8000 face images each a single time. Each subject also saw 20 face test images between 40-60 times. We fit an encoding model between each models' latent space and the voxel activity for the training face images. We then use the learned beta weights to predict fMRI responses to the test images.

Despite the additional disentanglement constraint,

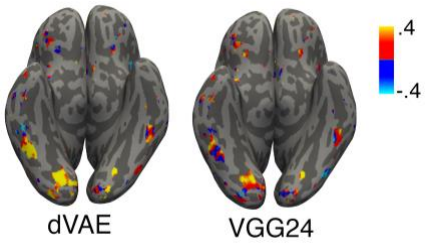

Figure 2: Encoding model performance in face-selective voxels. Ventral view of one representative subject's encoding model performance for dVAE (left) and VGG-face, dimensionality reduced to 24 PCs (right).

the dVAE model achieves similar encoding performance to the standard VAE (data not shown) and VGG24 in face-selective voxels (one representative subject shown in Fig. 2). In an ROI analysis, we find that dVAE provides as good encoding performance as a standard VAE and VGG, but none of the models are predictive of pSTS face responses (Fig. 3).

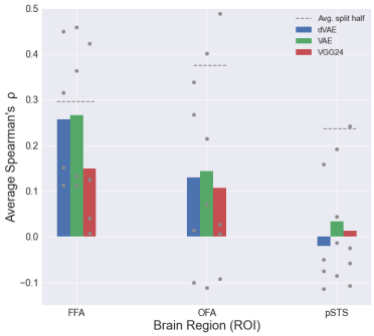

Figure 3: Average encoding model performance in face ROIs for dVAE (blue), VAE (green) and VGG (red). Gray dots represent individual subject performance.

Feature Tuning Across the Face Network

The main advantage of disentangled models is the ability to examine how voxels respond to semantically meaningful dimensions. To do this, we predict test responses based on each individual latent dimension (Fig. 4). In the FFA, we find that background, smile, skin tone, and entangled dimensions are predictive of voxel

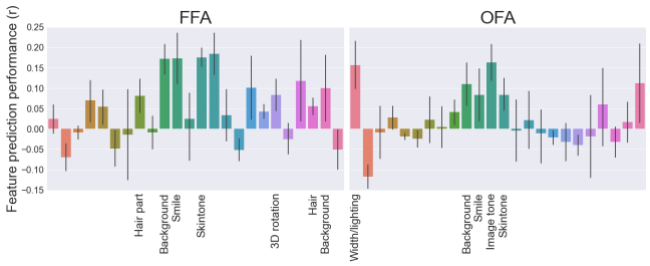

Figure 4: Encoding performance by dimension.

responses. In the OFA, lower-level visual dimensions like lighting and image tone are strongly predictive.

To understand how dimensions are represented across the brain, we can visualize their predictivity in a winner take all manner on the surface of the brain (one representative subject shown in Fig 5). While there are some interpretable dimensions like smile that load heavily on ventral face-selective voxels, most voxels are best predicted by entangled dimensions, particularly in the FFA.

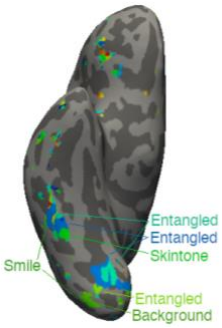

Figure 5: Predictive features of face-selective voxels. Ventral view of right hemisphere for the same representative subject shown in Fig 2. Each face-selective voxel is color coded by the feature which best predicts its voxel activity.

Conclusions

We find that disentangled models are good encoding models of human face responses, with similar performance to standard VAEs and VGG. The benefit of dVAEs is that they provide interpretable dimensions in an unsupervised manner. Similar approaches have been used to analyze macaque physiology data (Higgins et al., 2021), but this is the first comprehensive whole brain investigation of disentangled representations in humans. While we find several face-specific, semantically meaningful dimensions in FFA, many voxels are best predicted by entangled dimensions. While this lack of interpretability may seem troubling, these entangled dimensions may work together to encode identity, which would account for their prominent activity in FFA. By factoring out identity-preserving transformations, this disentangled modeling approach provides a relevant, low dimensional space to investigate identity coding across the face network in future work.

References

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C., Lerchner, A., & Higgins, I. (2019, September 25). *Unsupervised Model Selection for Variational Disentangled Representation Learning*. International Conference on Learning Representations. https://openreview.net/forum?id=SyxL2TNtvr

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), 6456. https://doi.org/10.1038/s41467-021-26751-5

Kim, H., & Mnih, A. (2018). Disentangling by Factorising. *Proceedings of the 35th International Conference on Machine Learning*, 2649–2658. https://proceedings.mlr.press/v80/kim18b.html

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20600–20605. https://doi.org/10.1073/pnas.0705654104

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *BMVC*.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, *2*(1). https://doi.org/10.1038/s42003-019-0438-y