Reconstructing faces from fMRI patterns using Generative Adversarial Networks

Rufin VanRullen (rufin.vanrullen@cnrs.fr)
CerCo, CNRS, Université de Toulouse
Toulouse, 31055 (France)

Leila Reddy (leila.reddy@cnrs.fr)
CerCo, CNRS, Université de Toulouse
Toulouse, 31055 (France)

Abstract:

Recent years have seen steady improvements in our ability to read out sensory inputs from fMRI brain response patterns – so-called "fMRI mind-reading". While visually distinct inputs, such as objects from different categories, can be accurately decoded and partially reconstructed from fMRI patterns, it has proved more difficult to distinguish visually similar inputs, such as different instances of the same category. Here, we apply a recently developed deep learning system to the reconstruction of face images from human fMRI patterns. We trained a variational auto-encoder (VAE) neural network using a GAN (Generative Adversarial Network) unsupervised training procedure over a dataset of > 200K celebrity faces. The auto-encoder latent space provided a meaningful (topologically organized) 1024-dimensional description of each image. We then presented > 4000 face images to a human subject in the scanner, and learned a simple linear mapping between the multi-voxel fMRI activation patterns and the 1024 latent dimensions. Then we applied this mapping to novel test images, turning the obtained fMRI patterns into VAE latent codes, and ultimately the codes into face reconstructions. Qualitative and quantitative evaluation of the reconstructions reveal robust pairwise decoding (>90% correct), and a strong improvement relative to a baseline model relying on PCA decomposition.

Keywords: faces, fMRI, deep learning, GAN, VAE, brain decoding, mind-reading

Introduction

Decoding sensory inputs from brain activity is both a modern technical challenge and a fundamental neuroscience enterprise. Multi-voxel fMRI pattern analysis, inspired by machine learning methods, has produced impressive "mind-reading" feats over the last 15 years (Haxby et al., 2001; Carlson, Schrater, & He, 2003; Kamitani & Tong, 2005; Kay, Naselaris, Prenger, & Gallant, 2008). A notoriously difficult problem, however, is to distinguish brain activity patterns for visually similar inputs, such as objects from the same category, or distinct human faces (Kriegeskorte, Formisano, Sorger, & Goebel, 2007; Kaul, Rees, & Ishai, 2011; Axelrod & Yovel, 2015). Here, we propose to take advantage of recent developments in the field of deep learning. Specifically, we use a variational auto-encoder or VAE (Kingma & Welling, 2014), trained with a Generative Adversarial Network (GAN) procedure (Goodfellow et al., 2014; Larsen, Sønderby, Larochelle, & Winther, 2015), as illustrated in Figure 1. The "face latent space" of the resulting network provides a description of numerous facial features that could approximate face representations in the human brain. In this latent space, faces and face features (e.g., maleness) can be represented as linear combinations of each other, and different concepts (e.g., male, smile) can be manipulated using simple linear operations (Figure 2). We reasoned that it could prove advantageous, when decoding brain activity, to learn a mapping between the space of fMRI patterns and this kind of latent space, rather than the space of image pixels.

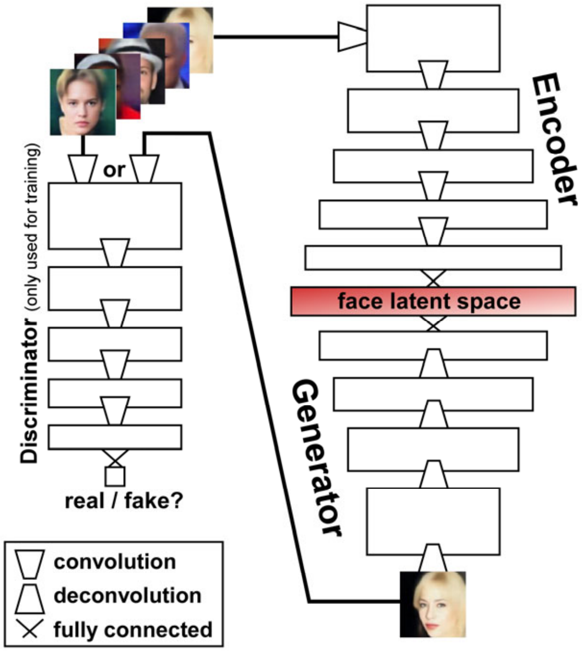

Figure 1. **VAE-GAN Network Architecture**. Three networks learn complementary tasks. The Encoder network maps a face image onto a latent representation (1024-dimensional), shown in red, which the Generator network converts into a novel face image. The Discriminator network (only used during the training phase) outputs a binary decision for each given image, either from the original dataset, or from the Generator output: is the image real or fake? Training is called "adversarial" because the Discriminator and Generator have opposite objective functions. (For simplicity, this diagram does not reflect the fact that the VAE latent space is actually a variational layer, which samples latent vectors stochastically from a probability distribution).

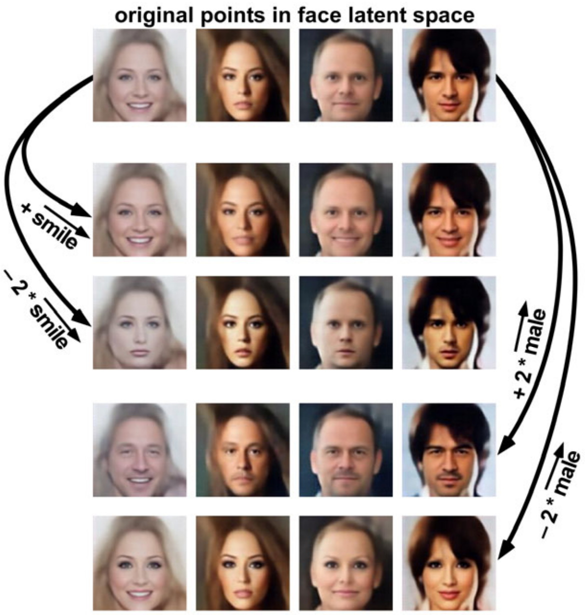

Figure 2. Latent space properties. The VAE latent space can be sampled and manipulated with simple linear arithmetic. The top row shows four original faces. The lower rows show the result of linear operations on the sample faces. For example, adding or subtracting a "smile vector" $\overrightarrow{smile}$ (computed by subtracting the average latent description of 1000 faces having a "no-smile" label from the average latent description of 1000 faces having a "smile" label) creates images of the original faces smiling or frowning (2nd and 3rd rows). The same operation can be done by adding or subtracting (a scaled version of) the average vector $\overrightarrow{male}$ (4th and 5th rows), making the original face more masculine or more feminine. In short, the network manipulates concepts, which it can extract from and render to pixel-based representations.

Methods

VAE architecture and GAN training

We trained a "variational auto-encoder" (VAE) deep network (10 layers) using an unsupervised "generative adversarial network" procedure (GAN) for 15 epochs on a labeled database of 202,599 celebrity faces (CelebA dataset). During GAN training, 3 sub-networks learn complementary tasks (Figure 1). The Encoder network learns to map a face image onto a 1024-dimensional latent representation (red in Figure 1), which the Generator network can use to produce a novel face image; the Encoder is rewarded for making the output face image as close as possible to the original image. The Generator network learns to convert latent 1024-D vectors from the latent space into plausible face images. The Discriminator network (only used during the training phase) learns to produce a binary decision for each given image (either from the original dataset, or from the Generator output): is the image real or fake? The Discriminator and Generator have opposite objective functions: the Discriminator is rewarded if it can reliably determine which images come from the Generator (fake) rather than from the dataset (real); the Generator is rewarded if it can produce images that the Discriminator network will not correctly classify. At the end of training, the Discriminator network was discarded, and the Encoder/Generator networks were used as a standard auto-encoder. Specifically, we used the Encoder to produce 1024-D latent codes for each input face image shown to our human subject, and these codes served as the design matrix for the fMRI GLM (General Linear Model) analysis (see "Brain decoding" section below). We used the Generator to reconstruct face images based on the output of our "brain decoding" system.

PCA model

Principal Component Analysis (PCA) was used as a baseline (linear) model for face decomposition and reconstruction. Retaining only the first 1024 principal components (PCs), each image could be turned into a 1024-D code to train our brain decoding system (as detailed below), and output codes could be turned back into face images for visualization using the inverse PCA transform.

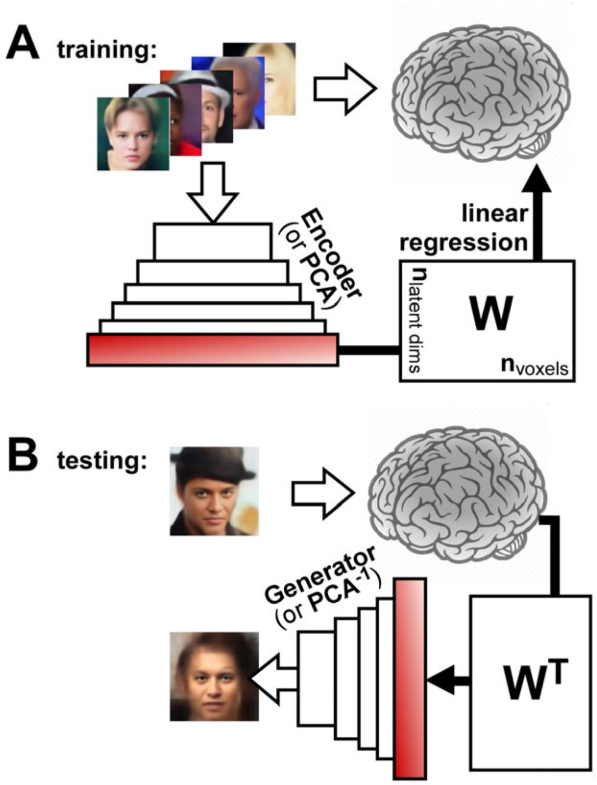

Figure 3. Brain decoding of face images based on VAE-GAN latent representations. A. The subject saw more than 4,300 faces (one presentation each) in a rapid event-related design. The same face images were also run through the "Encoder" network (as described in Figure 1) or a PCA decomposition, to obtain a 1024-dimensional latent face description. The "brain decoder" was a simple linear regression, trained to associate the 1024-dimensional latent vector with the corresponding brain response pattern. This linear regression, with 1024 parametric regressors for the BOLD signal, produced a weight matrix W (1024 by n_{voxels} dimensions) optimized to predict brain patterns in response to face stimuli. **B.** In the "testing phase", we presented 20 novel faces (at least 14 randomly interleaved presentations each) to the subject. The resulting brain activity patterns were simply multiplied by the transposed weight matrix W^T (n_{voxels} by 1024 dimensions) and its inverse covariance matrix to produce a linear estimate of the 1024 latent face dimensions. The Generator network (Figure 1) or an inverse PCA transform was then applied to translate the predicted latent vector into a reconstructed face image.

Brain decoding

We trained a simple brain decoder (linear regression) to associate the 1024-D latent representation of face images (obtained by running the image through the "Encoder", as described in Figure 1, or using a PCA transform as described above) with the corresponding brain response pattern, recorded when a human subject viewed the same faces in the scanner. This procedure is illustrated in Figure 3A. The subject saw more than 4,300 faces (one presentation each) in a rapid event-related design, and we used the VAE-GAN latent dimensions (or the image projection onto the first 1024 PCs) as 1024 parametric regressors for the BOLD signal. The linear regression performed by the GLM analysis thus produced a weight matrix W (1024 by n_{voxels} dimensions, where n_{voxels} is the number of voxels in the brain region-of-interest) optimized to predict brain patterns in response to face stimuli.

To use this brain decoder in the "testing phase", we simply inverted the linear system, as illustrated in Figure 3B. We presented 20 novel faces to the same subject, which had not been seen in the training phase. Each test face was presented at least 14 times (randomly interleaved) to increase signal-to-noise ratio. The resulting brain activity patterns were simply multiplied by the transposed weight matrix W^T (n_{voxels} by 1024 dimensions) and its inverse covariance matrix to produce an estimate of the 1024 latent face dimensions. We then used the Generator network (as illustrated in Figure 1) to translate the predicted latent vector into a reconstructed face image. For the baseline PCA model, the same logic was applied, but the face reconstruction was obtained via inverse PCA of the decoded 1024-D vector.

Results

We used the VAE-GAN model described in Figure 1 to train a brain decoding system. During training (Figure 3A), the system learned the correspondence between brain activity patterns in response to numerous face images (more than 4300 examples, involving 6 hours of scanning over 4 separate sessions) and the corresponding 1024-D latent representation of the same faces within the VAE network. The learning procedure assumed that each brain voxel's activation could be described as a weighted sum of the 1024 latent parameters, and we simply estimated the corresponding weights via linear regression (GLM). After training (Figure 3B), we inverted the linear system, such that the decoder was given the brain pattern of the subject viewing a specific, novel face image as input (a face that was not included in the training set), and its output was an estimate of the 1024-dimensional latent feature vector for that face. The image of the face was then generated (or "reconstructed") through the generative (VAE-GAN) neural network.

We contrasted the results obtained from this deep neural network model with those produced by another, simpler model of face image decomposition: principal components analysis (PCA, retaining only the first 1024 principal components from the face celebrity dataset). The PCA model also describes every face by a vector in a 1024-dimensional latent space, and can also be used to reconstruct faces based on an estimate of this 1024-D feature vector.

For both the deep neural network and PCA-based models, we defined a subset of the gray matter voxels as our "region-of-interest". Indeed, many parts of the brain perform computations that are not related to face processing or recognition; entering such regions in our analysis would adversely affect signal-to-noise. We included in our selection all 75,004 voxels that showed a significant response to visual stimulation (t-value>2.28 in a t-test for stimulation present vs. absent). In addition, we reasoned that certain voxels may only be activated by a small subset of our visual stimuli (think of a neuron population selectively responding to the "mustache" feature, which is only present in ~4% of faces); such voxels would most likely not be included based on the previous criterion. We therefore defined another voxel selection criterion, based on a comparison of the residual variances of (i) a baseline GLM model with only a face present/absent regressor, and (ii) the full GLM model encompassing the 1024 face regressors. 88,121 additional voxels were included based on this criterion, for a total of 163,125 voxels in the region-of-interest (for the PCA model, the total was similar: 169,407 voxels). The selected voxels are depicted in Figure 4. It is important to highlight that the above voxel selection criteria were applied only to the 4,300 training face images, but not to the 20 test images; therefore, the decoding analysis does not suffer from "circular reasoning" issues (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009).

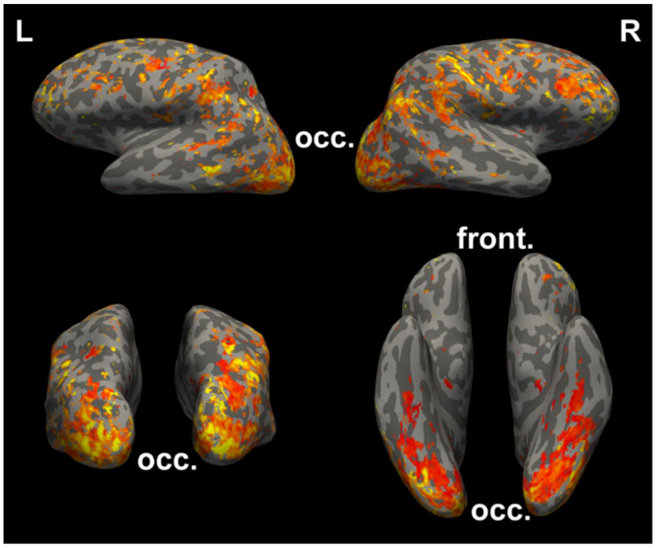

Figure 4. Voxels selected for brain decoding. 163,125 voxels were selected based on either their visual responsiveness, or their GLM goodness-of-fit during the brain decoder training stage (Figure 3A). The color code (red to yellow) reflects the goodness-of-fit. Occipital and frontal poles are marked "occ." and "front." respectively.

Examples of the reconstructed face images from the test image set are shown in Figure 5. Only the VAE-GAN model but not the PCA model could reconstruct an acceptable likeness of the original faces. We quantified the performance of our brain decoding system by correlating the brain-estimated latent vectors of 20 faces with the 20 actual vectors, and used the pairwise correlation values to measure the percentage of correct classification. For each of the 20 test faces, we compared the decoded 1024-D vector to the ground-truth vector from the actual test image, and to that of another test image: brain decoding was "correct" if the correlation with the actual vector was higher than with the other vector. This was repeated for all (20*19) pairs of test images, and the average performance compared to chance (50%) with a binomial test. Reconstructions from the GAN model achieved 92% classification ($p<10^{-10}$), while the PCA model only reached 77.5% (still highly above chance, but much below the GAN model, $p<10^{-10}$).

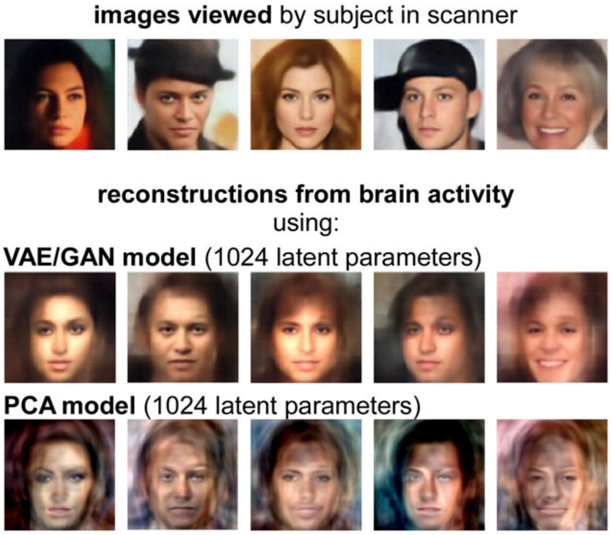
images viewed by subject in scanner
reconstructions from brain activity
using:
VAE/GAN model (1024 latent parameters)
PCA model (1024 latent parameters)

Figure 5. Examples of face reconstructions. Top row: 5 examples of images (among the 20 test images) and their reconstructions from a brain decoder based on the VAE-GAN model (middle row), or based on a PCA decomposition of the face dataset (bottom row). The quality of brain decoding was quantified with a pairwise pattern classification, and the average performance compared to chance (50%) with a binomial test. Brain decoding from the VAE-GAN model achieved 92% correct performance ($p<10^{-10}$), the PCA model only 77.5% ($p<10^{-10}$).

Discussion and Conclusion

We found that we could take advantage of the expressive power of deep neural networks (in particular, VAEs and GANs) to provide a better image space for linear brain decoding. Compared to PCA, which operates in pixel space, our approach produced qualitatively and quantitatively superior results. In particular, we could reliably distinguish the fMRI pattern evoked by one face from another, an outcome which had so far proved elusive (Kriegeskorte et al., 2007; Kaul et al., 2011; Axelrod & Yovel, 2015).

One explanation for our method's performance could be that the topology of the VAE-GAN latent space is ideally suited for brain decoding. We already know that this space supports linear operations on faces and facial features (Figure 2). We also know that, by construction (due to the variational training objective of the VAE, and the generative objective of the GAN), nearby points in this space map onto similar-looking but always visually plausible faces. This latent space therefore makes the brain decoding more robust to small mapping errors. In addition to these technical considerations, it might simply be that the VAE-GAN latent space is topologically similar to the space of face representations in the human brain. This speculation could easily be tested in the future, for example using Representational Similarity Analysis or RSA (Kriegeskorte et al., 2008).

Our approach could readily be extended from faces to other stimulus domains where GANs have proved relevant: for example, flowers, shoes, natural scenes or indoor scenes. Within the realm of faces, the current brain decoding model could be applied to the visualization of the facial feature selectivity of any voxel or ROI in the brain (simply by running corresponding columns of the W matrix into the face Generator network). The approach could also serve to investigate the brain representation of behaviorally important facial features, such as gender, race, emotion or age (by deriving the corresponding 1024-D latent vector, e.g. $\overrightarrow{smile}$ in Figure 2, and correlating it with each column of the W matrix). Finally, the decoding system could be extended to the reconstruction of faces that are not seen but imagined: this would be a true "mind-reading" achievement.

Acknowledgments

This work was funded by an ERC Consolidator Grant P-CYCLES number 614244.

References

Axelrod, V., & Yovel, G. (2015). Successful decoding of famous faces in the fusiform face area. *PLoS ONE, 10*(2), e0117126. doi: 10.1371/journal.pone.0117126

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *J Cogn Neurosci, 15*(5), 704-717.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative adversarial nets.* Paper presented at the Advances in Neural Information Processing Systems.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat Neurosci, 8*(5), 679-685.

Kaul, C., Rees, G., & Ishai, A. (2011). The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain. *Front Hum Neurosci, 4*, 238. doi: 10.3389/fnhum.2010.00238

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*(7185), 352-355. doi: 10.1038/nature06713

Kingma, D. P., & Welling, M. (2014). *Auto-encoding variational Bayes.* Paper presented at the International Conference on Learning Representations.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci U S A, 104*(51), 20600-20605. doi: 10.1073/pnas.0705654104

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141. doi: 10.1016/j.neuron.2008.10.043

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience, 12*(5), 535.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.