Feature-disentangled reconstruction of perception from multi-unit recordings

Thirza Dado (thirza.dado@donders.ru.nl)[1]

Paolo Papale (p.papale@nin.knaw.nl)[2]

Antonio Lozano (a.lozano@nin.knaw.nl)[2]

Lynn Le (l.le@donders.ru.nl)[1]

Feng Wang (f.wang@nin.knaw.nl)[2]

Marcel van Gerven (marcel.vangerven@donders.ru.nl)[1]

Pieter Roelfsema (p.roelfsema@nin.knaw.nl)[2,3,4,5]

Yağmur Güçlütürk (y.gucluturk@donders.ru.nl)[1]

Umut Güçlü (u.guclu@donders.ru.nl)[1]

[1] Donders Centre for Cognition, Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6500 HE, Nijmegen, The Netherlands. [2] Department of Vision and Cognition, Netherlands Institute for Neuroscience, 1105 BA, Amsterdam, The Netherlands. [3] Laboratory of Visual Brain Therapy, Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Institut de la Vision, Paris F-75012, France. [4] Department of Integrative Neurophysiology, Centre for Neurogenomics and Cognitive Research, Vrije Universiteit, 1081 HV, Amsterdam, The Netherlands. [5] Department of Psychiatry, Amsterdam UMC, University of Amsterdam, 1105 AZ, Amsterdam, The Netherlands

Abstract

Here, we aimed to explain neural representations of perception, for which we analyzed the relationship between multi-unit activity (MUA) recorded from the primate brain and various feature representations of visual stimuli. Our encoding analysis revealed that the feature-disentangled latent representations of generative adversarial networks (GANs) were the most effective candidate for predicting neural responses to images. Importantly, the usage of synthesized yet photorealistic images allowed for superior control over these data as their underlying latent representations were known a priori rather than approximated post-hoc. As such, we leveraged this property in neural reconstruction of the perceived images. Taken together with the fact that the (unsupervised) generative models themselves were never optimized on neural data, these results highlight the importance of feature disentanglement and unsupervised training as driving factors in shaping neural representations.

Keywords: feature disentanglement; generative modeling; neural coding; multi-unit activity

Introduction

The brain is adept at recognizing a virtually unlimited variety of different visual inputs. Every stimulus creates a unique pattern of brain activity that carries information about that stimulus in some shape or form - but this stimulus-response transformation remains largely unsolved due to the complexity of multi-layered visual processing in the brain. The field of neural coding aims to characterize this relationship where *neural encoding* seeks to find how properties of external phenomena are stored in the brain (van Gerven, 2017), and vice versa, *neural decoding* aims to find what information about the original stimulus is present in and can be retrieved from the recorded brain activity by classification (Haxby et al., 2001; Kamitani & Tong, 2005; Horikawa & Kamitani, 2017), identification (Mitchell et al., 2008; Kay, Naselaris, Prenger, & Gallant, 2008) or reconstruction (Nishimoto et al., 2011; Du, Du, & He, 2017; Güçlütürk et al., 2017; Shen, Horikawa, Majima, & Kamitani, 2019; VanRullen & Reddy, 2019; Dado et al., 2022). Note that the latter problem is considerably harder as its solution exists in an infinitely large set of possibilities whereas those of classification and identification can be selected from a finite set.

Although neural representations are constructed from experience, an infinite amount of visual phenomena can be represented by the brain to successfully interact with the environment. That is, novel yet plausible situations that respect the regularities of the natural environment can also be mentally simulated or *imagined* (Dijkstra, Bosch, & van Gerven, 2019). From a machine learning perspective, generative models achieve the same objective by capturing the probability density underlying a huge set of observations. Generative adversarial networks (GANs) (Goodfellow et al., 2014) are among the most impressive generative models to date which can synthesize novel yet realistic-looking images (Brock, Donahue, & Simonyan, 2018; Karras, Aila, Laine, & Lehtinen, 2017; Karras, Laine, & Aila, 2019; Karras et al., 2021) from z-latent vectors, which represent the visual information of their corresponding images in their low-dimensional code. In particular, feature-disentangled GANs have been designed to separate the factors of variation in the generated images. One member of the family of feature-disentangled GANs is StyleGAN (Karras et al., 2019) - which maps the conventional z-latent via an 8-layered MLP to an intermediate and less entangled w-latent space.

We hypothesized that feature-disentangled w-latents represent the visual information similar to the brain such that we could utilize it for neural encoding and -decoding. We also evaluated the latents of Contrastive Language-Image Pre-training (CLIP) which represent both images and text in a shared representational space that captures their semantic relationships (Radford et al., 2021). We found that w-latents indeed outperformed z- and CLIP latents in predicting neural responses to images. We then used w-latents for neural reconstruction by the adoption and improvement of the experimental paradigm of (Dado et al., 2022) as follows: visual stimuli were synthesized by a feature-disentangled GAN and presented to a macaque with cortical implants in a passive fixation task. A neural decoder was fit based on the recorded brain activity and the ground-truth latents of the stimuli. Reconstructions were created by feeding the predicted w-latents from brain activity from a held-out test set to the feature-disentangled GAN. This work is the first to perform neural coding of photorealistic images using intracranial recordings.

Methodology

We recorded multi-unit activity (Super & Roelfsema, 2005) with 15 chronically implanted electrode arrays (64 channels each) in one macaque (male, 7 years old) upon presentation of images in a passive fixation experiment. We used two datasets of visual stimuli: (i) face images synthesized by StyleGAN3 (pretrained on FFHQ) and (ii) high-variety natural images synthesized by StyleGAN-XL (pretrained on ImageNet). In the analysis, we used linear mapping to evaluate our claim that the feature- and neural representation effectively encode the same stimulus properties, as is standard in neural coding (Naselaris, Kay, Nishimoto, & Gallant, 2011; Güçlü & van Gerven, 2015). A more complex nonlinear transformation would not be valid to support this claim since nonlinearities will fundamentally change the underlying representations.

In encoding, the neural response per electrode was modeled using kernel ridge regression to find how brain activity was linearly dependent on the stimulus features. Regularization was required to avoid overfitting since we predicted from feature space. For decoding, multiple linear regression was used to model how the individual units within feature representations were linearly dependent on the brain activity per electrode.

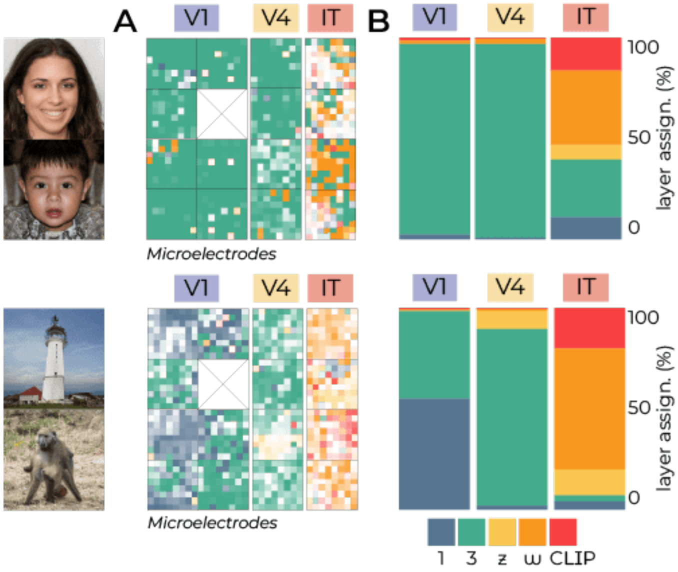

Figure 1: For each individual microelectrode, we fit five encoding models to predict its neural response via low and mid-level feature representations of VGG16 (pretrained for face- or object recognition), z-, w- and CLIP latent representations of the visual stimuli. The representation that resulted in the highest encoding performance (Student's t-test) was assigned to each microelectrode. The distribution of assigned features shows a complexity gradient where the more low-level representations are assigned to early brain regions V1 and V4 and more high-level representations to IT with a preference for w-latents.

Results

Neural encoding

Encoding with low-level, mid-level and latent representations of images resulted in the well-established complexity gradient where early representations are mainly predictive of responses in early visual areas and deeper representations of more downstream visual areas (Güçlü & van Gerven, 2015; Freiwald & Tsao, 2010; Chang & Tsao, 2017). As such, we confirmed that z-, w- and CLIP latent representations indeed encode high-level information about the visual stimuli and thus explained the neural responses in the inferior temporal (IT) cortex which is located at the end of the visual ventral pathway (Figure 1). Among these latents, w-latents were the best candidate to predict neural activity to visual stimuli.

Neural decoding

Neural decoding of neural activity via feature-disentangled w-latents resulted in highly accurate reconstructions that closely resembled the stimuli in their specific visual characteristics (Figure 2). Table 1 quantitatively demonstrates the similarity between stimuli and their reconstructions. The contribution of each visual area was determined by the occlusion of the electrode recordings in the other two brain areas (rather than fitting three independent decoders on subsets of brain activity). It is reasonable to say that, of the three cortical areas, area IT resulted in the highest similarity and thus contained the most information about that representation.

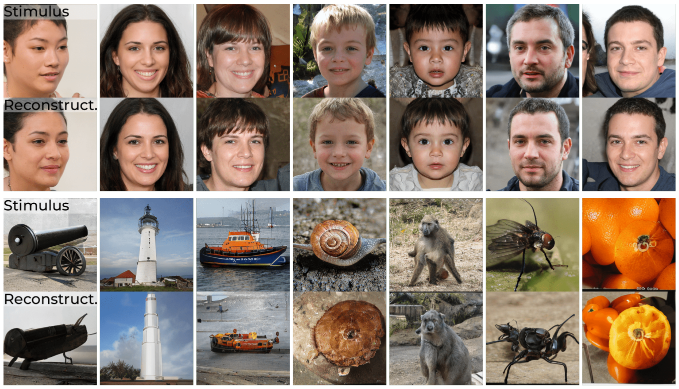

Figure 2: Qualitative decoding results. Seven arbitrary but representative test set stimuli (top row) and their reconstructions from brain activity (bottom row).

Table 1: Quantitative decoding results. The upper and lower block display performance when reconstructing face- and natural images, respectively, in terms of four metrics: cosine similarity using MaxPool layer outputs 1, 3 and 5 of VGG16 for object recognition and between the w-latents of stimuli and of their reconstructions.

	VGG16-1	VGG16-3	VGG16-5	Latent
All	**0.6066**	**0.5192**	**0.6607**	**0.5548**
V1	0.5435	0.4503	0.5351	0.5022
V4	0.5430	0.4531	0.5323	0.5026
IT	*0.5590*	*0.4718*	*0.5638*	*0.5176*
All	**0.4083**	**0.2555**	**0.2497**	**0.8032**
V1	*0.3929*	0.2223	0.1367	0.7336
V4	0.3790	0.2270	0.1617	0.7614
IT	0.3798	*0.3127*	*0.1692*	*0.7653*

Conclusion

This work leveraged feature-disentangled GANs to understand how visual information is represented in the brain, highlighting how unsupervised models can capture the underlying structure and patterns of visual data so it can be aligned with biological processes. The high reconstruction performance achieved by decoding via w-latent space indicates the importance of feature disentanglement to explain neural representations of perception, offering a new way forward for the previously limited yet biologically more plausible unsupervised models of brain function. These findings have implications for the advancements of computational models and the development of clinical applications for people with disabilities. For instance, neuroprosthetics to restore vision in blind patients as well as brain computer interfaces (BCIs) to enable nonmuscular communication with individuals who are locked-in.

References

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013–1028.

Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., van Gerven, M., & Güçlü, U. (2022). Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space. *Scientific reports*, *12*(1), 1–9.

Dijkstra, N., Bosch, S. S. E., & van Gerven, M. A. M. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, *23*(5), 423–434. Retrieved from `https://doi.org/10.1016/j.tics.2019.02.004` doi: 10.1016/j.tics.2019.02.004

Du, C., Du, C., & He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. In *2017 international joint conference on neural networks (ijcnn)* (pp. 1049–1056).

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, *30*.

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.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, *8*(1), 1–15.

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

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4401–4410).

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

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191–1195.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, *21*(19), 1641–1646.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS computational biology*, *15*(1), e1006633.

Super, H., & Roelfsema, P. R. (2005). Chronic multiunit recordings in behaving animals: advantages and limitations. *Progress in brain research*, *147*, 263–282.

van Gerven, M. A. (2017). A primer on encoding models in sensory neuroscience. *Journal of Mathematical Psychology*, *76*, 172–183.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, *2*(1), 1–10.