Brain-grounding of semantic vectors improves neural decoding of visual stimuli

Shirin Vafaei¹, Ryohei Fukuma^{1,2,3}, Huixiang Yang², Haruhiko Kishima¹, Takufumi Yanagisawa^{1,2,3}

Keywords

Abstract

Developing algorithms for accurate and comprehensive neural decoding of mental contents is one of the long-cherished goals in the field of neuroscience and brain-machine interfaces. Previous studies have demonstrated the feasibility of neural decoding by training machine learning models to map brain activity patterns into a semantic vector representation of stimuli. These vectors, hereafter referred as pretrained feature vectors, are usually derived from semantic spaces based solely on image and/or text features and therefore they might have a totally different characteristics than how visual stimuli is represented in the human brain, resulting in limiting the capability of brain decoders to learn this mapping. To address this issue, we propose a representation learning framework, termed brain-grounding of semantic vectors, which fine-tunes pretrained feature vectors to better align with the neural representation of visual stimuli in the human brain. We trained this model this model with functional magnetic resonance imaging (fMRI) of 150 different visual stimuli categories, and then performed zero-shot brain decoding and identification analyses on 1) fMRI and 2) magnetoencephalography (MEG). Interestingly, we observed that by using the brain-grounded vectors, the brain decoding and identification accuracy on brain data from different neuroimaging modalities increases. These findings underscore the potential of incorporating a richer array of brain-derived features to enhance performance of brain decoding algorithms.

Introduction

The development of brain decoding algorithms is essential for advancing Brain-Machine Interfaces (BMIs) (Lebedev & Nicolelis, 2006; Willett et al., 2023; Willsey et al., 2022; Stavisky & Wairagkar, 2023), enabling precise communication and motor control for individuals with speech or motor impairments. Simultaneously, these algorithms offer a unique opportunity to delve into the intricate processes of human brain information processing (Naselaris et al., 2011; Haxby et al., 2001; Yamins et al., 2014; Kellis et al., 2010; Brouwer & Heeger, 2009; Haynes & Rees, 2006), shedding light on the fundamental mechanisms that underlie information encoding process in the human brain. Furthermore, this progress can augment the effectiveness of neurofeedback systems by enabling the precise decoding of cognitive patterns and delivering real-time neurofeedback, thereby assisting patients in the refinement of their cognitive and emotional faculties (Fukuma et al., 2022; Chaudhary et al., 2022; Cortese et al., 2016; Sitaram et al., 2017).

¹Department of Neurosurgery, Graduate School of Medicine, Osaka University, Suita, Japan

²Institute for Advanced Co-Creation Studies, Osaka University, Suita, Japan

³ATR Computational Neuroscience Laboratories, Seika-cho, Japan

Recently it has been shown that the perceived or imagined visual stimuli (i.e., images) can be decoded from the neural activity patterns either as semantic attributes associated with those images (Haynes & Rees, 2005; Kamitani & Tong, 2005; Thirion et al., 2006; Brouwer & Heeger, 2009; Horikawa & Kamitani, 2017) or as category-level classes (Cox & Savoy, 2003; Nakai, Koide-Majima, & Nishimoto, 2021; Haxby et al., 2001; Haynes & Rees, 2006; Horikawa & Kamitani, 2017) or as in the form of reconstructed visual representations of the presented or imagined images (Miyawaki et al., 2008; Shen et al., 2019a; Shen et al., 2019b; Koide-Majima, Nishimoto, & Majima, 2023).

The typical decoding approach that was used in previous studies consists of training a machine learning models to map brain activity patterns to semantic vector representation of the stimuli (Horikawa & Kamitani, 2017; Fukuma et al., 2022; Toneva & Wehbe, 2019). These vectors were usually obtained by methods such as one-hot encoding of the categories in their dataset (Makin et al., 2020) or semantic vectors derived from pretrained large language models (Pereira et al., 2018; Fukuma et al., 2022), as well as those derived from feature spaces of deep neural networks trained for object recognition (Horikawa & Kamitani, 2017) or multimodal text-image processing models (Toneva & Wehbe, 2019).

These vectors, hereafter referred to as pretrained feature vectors, are usually derived from large language models mostly encode the "distributional semantics" (Pennington, Socher, & Manning, 2014) and the ones that are derived from deep neural networks for object recognition mostly encode visual features of an image such as basic contours, colors, or categories (Zeiler & Fergus, 2013). Likewise, the vectors that are derived from multimodal neural networks primarily encode the semantical attributes that are shared within those modalities (Radford et al., 2021).

However, in order to make brain decoders more accurate and generalizable, we hypothesized that the semantic vectors that will be used for brain decoding analysis should also contain some features about how information is represented in the human brain, and not solely contain image-based and/or text-based features.

Furthermore, distinct brain regions are selective to distinct semantic categories (Haxby et al., 2001; Huth et al., 2016), therefore, we hypothesized it is hardly possible that the semantic spaces derived from large neural networks for image, text or image/text processing can fit all the semantic information representation in the human brain (Schrimpf et al., 2020), as the information encoded in the brain activity patterns in each brain regions might be totally from a different nature than the features encoded in neural networks trained for image and/or text processing.

Also, as previous studies have shown, that the choice of neural network model/layer highly impacts the accuracy of brain decoders (Horikawa & Kamitani, 2017; Toneva & Wehbe, 2019; Benchetrit et al., 2023) and different studies choose different types of models to train their brain decoders, suggesting that it is still unknown which model can fit the semantic system in the human brain the best, and therefore to be used in decoding algorithms (Benchetrit et al., 2023).

Therefore, we sought to create a novel semantic representation of concepts, that is based on integrating features from the neural activities of seen objects, into the pretrained feature spaces and investigate whether the brain decoders can improve the zero-shot decoding of 1) brain activity patterns obtained by same neuroimaging technique that is used to fine-tune the pretrained feature spaces and 2) brain activity patterns obtained by a different neuroimaging technique rather the one that pretrained feature vectors are being fine-tuned by. We would like to note that by zero-shot decoding, we mean the test dataset consisted of images from totally different categories than training dataset, and we do not mean new images from categories that were used in the training dataset.

To accomplish this, we developed our framework, "brain-grounding of semantic vectors", which is an autoencoder with a two-term loss function that 1) learns to reconstruct the pretrained feature vectors and 2) constrains the representational similarity matrix (RSM) (Kriegeskorte et al., 2008) of its latent space to be as similar as possible to the RSM of their corresponding brain activity patterns. Therefore, after training, the features extracted from the latent space of autoencoder consists of a compact representation of pretrained feature vectors that are fine-tuned to better align with how visual information are represented in the human brain.

To evaluate how brain-grounding of pretrained feature vectors affect the brain decoding analyses, we extracted the autoencoder's latent features of all categories in the dataset as the "brain-grounded vectors" and performed brain decoding (Horikawa & Kamitani, 2017) and identification (Kay et al., 2008; Horikawa & Kamitani, 2017) analyses on the brain activity patterns of categories that were not used in the autoencoder's training procedure. Our results demonstrated that using brain-grounded vectors, the accuracy of brain decoding, and identification analyses increases, although the pattern of increase varies depending on the choice of pretrained feature space and the type of analysis.

Furthermore, although brain-grounded vectors are derived by neural activity patterns measured by functional magnetic resonance imaging (fMRI), we tested whether these vectors can be used to improve the decoding and identification performance of brain activity patterns measured by other neuroimaging data such as Magnetoencephalography (MEG) and the positive results from all fMRI and MEG decoding suggest that this method is consistent and can be generalized to be used to decode/identify the brain activity data obtained by different neuroimaging modalities and different subjects that were not used during the training procedure.

Results

Brain-grounding of semantic vectors

To develop the proposed framework, we developed a multimodal learning autoencoder framework that works by taking a pretrained feature space and brain activity patterns corresponding to the visual stimuli dataset and train it in a way that its latent space integrates the structure of visual representation in the human brain to the pretrained feature spaces. More specifically, we used an image-based feature space (i.e., features from the image encoder model of CLIP (Radford et al., 2021) and a text-based feature space (i.e., features from the GloVE model (Pennington, Socher, & Manning, 2014). The fMRI dataset that we used here to finetune the pretrained feature vectors is the generic object decoding (GOD) dataset (Horikawa & Kamitani, 2017) which contains

fMRI brain activity patterns of 5 subjects while they were watching 1200 images of 200 distinct object categories selected from ImageNet (Deng et al., 2009). First, we extracted the original semantic vectors for each category in the GOD dataset (see methods) and represented each category by its corresponding pretrained feature vector. Then, to obtain the "brain-grounded semantic space", we trained the autoencoder with a two-term objective function. The first term is a simple mean squared error (MSE) loss between true and predicted pretrained feature vectors (with the goal of reconstructing them). The second term is the MSE loss between the representational similarity matrix (RSM) (Kriegeskorte et al., 2008) of fMRI signals and the autoencoder's latent space in each batch. Mathematically:

$$loss = \frac{1}{m} \sum_{i=1}^{m} \left[(\alpha) \left(y - y' \right)^{2} + (1 - \alpha) (RSM_{l} - RSM_{b})^{2} \right]$$
(1)

where y is the original semantic vector, y' is the reconstructed semantic vector, RSM_l is the RSM of autoencoder's hidden layer, RSM_b is the RSM of corresponding brain activity patterns, and m is the number of samples in each batch. α is the hyperparameter that determines the level of brain-grounding.

Given that our aim here is to decode the visual object categories, we used the fMRI activity patterns of lateral occipital complex (LOC) region of human brain to fine-tune pretrained feature vectors. LOC is a region that is located in the occipital lobe and is primarily responsible for visual processing of objects and shapes (Kourtzi & Kanwisher, 2000). Concurrently, to choose the appropriate pretrained feature vectors, we extracted the category-specific pretrained feature vectors of each category in the training data of GOD dataset.

For each subject in the GOD dataset, brain region, pretrained feature vector and α we trained a different autoencoder. When training the autoencoder for a particular subject (i.e., subject A) in the GOD dataset, we used the averaged RSM of fMRI brain signals of all other subjects in the GOD dataset. We later used this these braingrounded vectors for decoding the fMRI brain signals of subject A, to avoid information leakage. We trained autoencoders on a wide range of α consisting of values from the [0.0001, 0.001, 0.01, 0.1, 1]. We specially included 1 as one of the options to see if we exclude the brain-grounding part from this framework, how it will affect the downstream analyses.

Figure 1 shows an overall procedure of this framework.

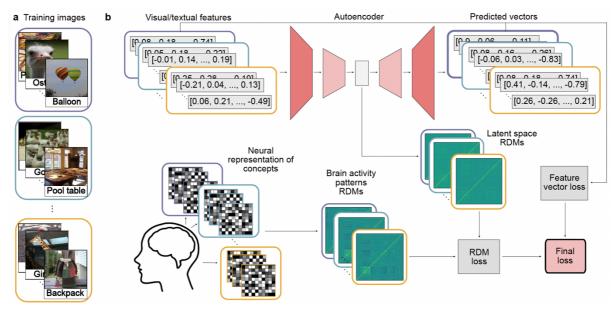


Figure 1| The brain-grounding framework. a) Samples of images in the GOD dataset, each rectangle are indicators of batches of samples that were used in training b) First, the original visual or textual semantic vectors were extracted. Then, the autoencoder is trained in a way that while it is learning to reconstruct the original visual/textual features of the training images, the RSM of its latent space to become as similar as possible to the RSM of corresponding brain activity patterns.

fMRI brain decoding and identification of visual stimuli

As the first step to determine how well the brain-grounded vectors can improve the brain decoding accuracy, we performed decoding analysis by training linear regression models to learn to map brain activity patterns to their corresponding pretrained feature vector or brain-grounded vectors (figure 2). More specifically, for all samples and each unit in the semantic vectors, we trained a separate linear regression model. We evaluated the decoding accuracy by taking Pearson correlation between true and predicted vectors both in sample-wise and dimension-wise manner. In the sample-wise, we calculated the Pearson correlations of true and predicted vectors for each sample and then averaged the results for all samples and subjects. In the dimension-wise manner, we calculated the Pearson correlations of true and predicted vectors for each feature unit and then averaged the results for all feature units and subjects. First, we investigated how well we can decode the brain activity patterns measured by fMRI using brain-grounded vectors, and by comparing them to the case where the is no brain-grounding component (i.e., alpha is equal to one or we have used the original semantic vectors), we can observe the contribution of brain-grounding part to the results.

Next, we were interested if using "brain-grounded" for decoding brain activity can also boost identification accuracy. In identification analysis, we are interested in probing how well we can identify the predicted vector among a large set of candidate vectors that are representing different semantic categories (Kay et al., 2008; Horikawa & Kamitani, 2017) (figure 2). For identification analysis, we computed the Pearson correlation coefficient between the predicted semantic vector and all other candidate vectors. If the correlation coefficient between the true and predicted vectors ranks among the top x%, we deem the identification accuracy of this case

as x%. The final identification accuracy is determined by averaging the identification accuracies across all categories and subjects in the test dataset.

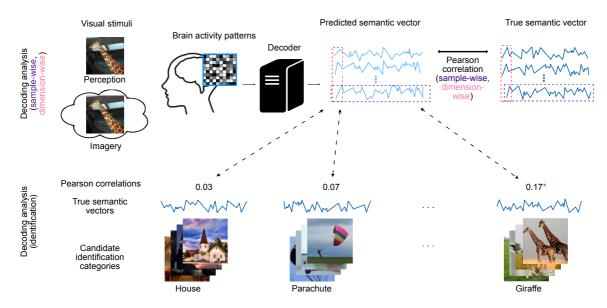


Figure 2| Decoding/identification procedure. For brain decoding, we first measure the brain activity patterns of seen or imagined stimuli and then train the decoders to learn to map the stimuli's corresponding semantic vector (either the pretrained feature vectors or the brain-grounded feature vectors). In the identification analysis, we first consider a large amount of candidate stimuli. Then, we calculate the Pearson correlation coefficient of predicted semantic vector with the semantic vectors of each of the candidate stimuli.

In figure 3, we show the results of fMRI decoding and identification analyses among different semantic spaces on both perception data and imagery data of GOD dataset. We evaluated the decoding accuracy in 3 different ways: Sample-wise correlation, dimension-wise correlation and identification analyses. In evaluating the results in the sample-wise manner, we take the Pearson correlation coefficient between true and predicted vectors for all samples, and then average the results across all samples and subjects. In the dimension-wise manner, we take the correlation coefficient for feature values of all samples in each unit and then average them across the total number of feature units and subjects. In the identification analysis, we are interested in evaluating how well we can identify a predicted vector among a large set of candidate vectors. Therefore, accuracy is defined by the percentage of candidate categories, in which their correlation with the predicted vector were lower than the correlation between the actual true and predicted vector.

In the scene-wise and dimension-wise results, we observed that the accuracy tends to increase when the contribution from brain part also increases (i.e., when alpha is lower in the loss function). In the identification analysis, we observed that the accuracy tends to increase when alpha is set to 10^{-1} and 10^{-2} and then decreases.

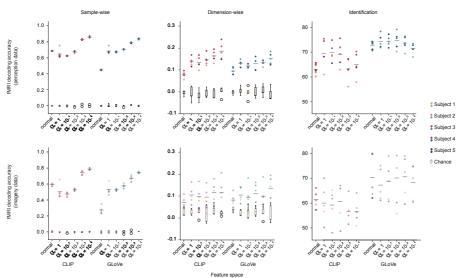


Figure 3| fMRI decoding and identification results. The GOD dataset consisted of fMRI activity patterns when participants were observing the visual stimuli (perception experiment) and when they were imagining the experiment (imagery experiment). The top row shows the decoding and identification results of perception data, and the bottom row shows the results of decoding and identification of imagery data. Each column shows the metric in which we assessed the effects of using brain-grounded vectors compared to their original cases. Each dot represents a subject. We observed when alpha decreases, although not linearly, the sample-wise and dimension-wise decoding accuracies tend to increase. In the identification analysis, when alpha decreases (therefore there is more contribution from the brain part in the autoencoder framework), the accuracy tends to increase but then decreases. This might be due to some biases in the way brain-grounded vectors are created (i.e., they are biased in one space), which can be considered for future work to improve.

Generalization of decoding performance on other modalities: MEG dataset

To further investigate whether these "brain-grounded vectors are effective in decoding brain activity patterns obtained by other neuroimaging modalities, we performed both decoding and identification analysis of the brain activity patterns obtained by MEG while the different subjects watched the same 1200 images of GOD dataset. This is particularly interesting because MEG capture brain activity patterns in a different aspect from fMRI. Here, we took average of all brain-grounded vectors when alpha was set to 0.01 that are based on fMRI brain signals across fMRI subjects, and used the averaged vectors as the brain-grounded vectors to decode and or identify MEG signals. Interestingly, we observed that in sample-wise, dimension-wise and identification accuracies of when we use the brain-grounded spaces increase compared to the case where we use the normal space.

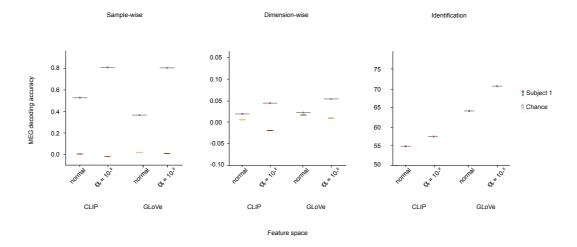


Figure 4 MEG decoding and identification results.

Discussion

Here we proposed a simple framework for creating a new type of semantic space, that vectors in this space contain some features about how visual information is represented in the human brain. This framework works by aligning the geometry of concept representation in human brain into the pretrained feature vectors derived from text or image processing tasks. We show that by using this new semantic space, decoders can better learn to map neural activity patterns into their corresponding semantic vector, and further be identified, even for categories that were not used in autoencoder's training procedure. Furthermore, we show that even though this new semantic space is created by leveraging brain activity patterns measured by fMRI, it can be used to decode/identify category representations in data obtained by other neuroimaging modalities, suggesting that grounding the original vectors using the second-order representation in human brain results in creating vectors that can be robustly used to decode brain activity data obtained by other neuroimaging techniques.

Recently, the spotlight in neuroscience and machine learning research has shifted towards developing multimodal learning models. As an example, in a pioneer study by (Karpathy & Fei-Fei, 2015), they show that by aligning the features extracted from convolutional neural networks trained for classifying the image and recurrent neural networks trained for generating texts, they can highly improve the quality of caption generation for an image. The more recent model, "CLIP", is also a text and image multimodal model in which it learns the visual concepts by natural language supervision and has shown great zero-shot capabilities compared to its similar models (Radford et al., 2021).

In neuroscience, several studies have either directly used human's brain activity data or monkey's brain activity data to make the representations of neural networks more aligned with those in the human brain and/or monkeys and have shown that these neural networks gain performance in object recognition (Federer et al., 2020) or robustness towards adversarial attacks (Muttenthaler et al., 2022).

However, up to our knowledge, fine-tuning the neural representations based on geometry of information representation in the human brain has not been explored. Interestingly, despite using a limited number of brain samples to train our models, we observed that brain decoding, and identification analyses can be increased for categories neuroimaging data and that were not used in training the autoencoder model.

Overall, we believe that our model is a promising approach towards creating more informative semantic spaces. In future work we are interested in upgrading our model using a larger brain dataset and assess its performance in various of downstream tasks such as brain decoding/identification analyses or tasks such object recognition.

Methods

Creating semantic vectors

Semantic vectors are multidimensional representations of data that encode the underlying semantics, relationships, and context within that data. In the context of brain decoding analysis, they have been widely used as a meaningful representation of stimuli to train decoders to learn to map neural activity patterns to their corresponding semantic vector representations. In this study, we used two different types of semantic spaces that have been previously used in brain decoding studies. Namely, we used features from the last layer of CLIP's image encoder and GLoVe model. We created semantic vectors for all categories in the ImageNet dataset fall 2011 release. For image-based semantic (CLIP), we fed one image from each category and used the feature representations its final layer as the representative of that category. In the ImageNet dataset, each category has been nominated by one or several names using crowdsourcing (Deng et al., 2009). For obtaining the GloVe labels of each category, we took average of the available GloVe vectors of all the labels of that category, and if for that category there wasn't any vector in the GloVe dictionary, we excluded that category from the whole analysis.

GloVe

GloVe is a method that can generate 300-dimensional semantic vector representations of a given word, from a normalized version of statistical results of word co-occurrences obtained from a corpus consisting of more than billions of tokens. Words with similar meanings have vectors that are close in the vector space, enabling GLoVe to capture the semantics of words and their contextual associations. Here we used the pretrained word vectors of the 42B tokens file (https://nlp.stanford.edu/data/glove.42B.300d.zip). For each image category in ImageNet dataset, we considered their annotations obtained by crowdsourcing, and calculated the averaged GLoVe representations of all available annotations in GloVe dictionary, as a representation of that category. If none of the annotations of a particular category didn't exist in GLoVe dictionary, that category was excluded from the whole analysis in this study.

CLIP

CLIP is a model that connects vision and language by generating semantic vectors for both images and text. The unique strength of CLIP lies in its ability to map images and textual descriptions into a shared vector space, where the similarity or dissimilarity between vectors accurately reflects the semantic relationships between the two modalities. To create a CLIP semantic vector for each category in ImageNet, we extracted an image from

each category, and then extracted the features from the ViT-B/32 Transformer image encoder of the CLIP model for each image.

fMRI dataset

Dataset description

We made use of one publicly available dataset (Horikawa & Kamitani, 2017), commonly referred as "Generic Object Decoding". Five healthy subjects (one female and four males, aged between 23 and 38 years) with normal or corrected-to-normal vision had participated in their experiments. Experiments consisted of presenting natural object images to subjects and recording their brain activity while they were perceiving the visual stimuli (perception experiment) or imagining them (imagery experiment). Images were selected from the ImageNet (Deng et al., 2009) dataset. Training dataset consisted of neural recording of 1200 images (150 categories, 8 images per category), all performed in perception manner. Test dataset consisted of both perception and imagery neural recordings of 50 images (50 images were selected from 50 categories that were not used in the training dataset, 1 image per category, presented 35, 10 times respectively).

MEG dataset

MEG Experiment

MEG experiments were conducted within our laboratory. Three (3) subjects (3 males, aged between 25 and 34) were seeing images from the GOD dataset, while their brain activity was recorded with MEG. Experiments were conducted in perception manner. Each of the images in either training or test data was repeated 6 times and participants were asked to fixate on a central dot on images. The study adhered to the Declaration of Helsinki and was performed in accordance with the protocols approved by the ethics committee of our University Clinical Trial Center (No. 18027). All participants approved and signed the informed consent.

MEG preprocessing

The process was initiated by importing the raw MEG data into Brainstorm, a specialized software tool for neuroimaging analysis. Subsequently, essential filters were employed, including a high-pass filter at 0.5 Hz and a notch filter at 60Hz and its harmonics to eliminate unwanted frequency components. To address potential artifacts, Independent Component Analysis (ICA) was employed to extract cardiac and blink artifacts. The robustness of the analysis was further enhanced by utilizing room data to compute noise covariance and generating a subject-specific head model from individual MRI data. This enabled the accurate estimation of neural sources in the brain. Following source estimation, individual source activities were projected onto the default brain model (FSAverage) for consistency. To precisely mark the onset of each stimulus, analogue triggers were utilized. After these preprocessing steps, a high-quality dataset was obtained in the form of a matrix with dimensions of 200 samples (equivalent to milliseconds) by 7200 trials by 15002 vertices.

Subsequently, a crucial feature, the sensorimotor cortical potential (SCP), was calculated by averaging signals within specific time windows, resulting in a 7200 by 15002 matrix that served as the foundation for subsequent analyses. These preprocessing steps collectively ensured the integrity and accuracy of the MEG data, providing a solid foundation for investigations into neural activity and responses.

MEG ROI selection

ROI selection consisted of two steps. First the neural sources in the brain that underlie the recorded MEG signals were estimated through a process called source localization. These source localized data were then anatomically registered to the Human Connectome Project's HCP brain parcellation. In our study, we selected five distinct regions for the extraction of neural data: the primary visual cortex (PVC), early visual cortex (EVC), dorsal visual cortex (DSVC), ventral visual cortex (VSVC), and the motor area (MEO). These selection of brain regions represent a comprehensive exploration of both early visual processing and motor functions, enabling us to delve deeper into the neural underpinnings of our research questions with a specific focus on these functionally relevant brain regions. All the above-mentioned processes were performed using brainstorm.

Autoencoder framework

The autoencoder consists of two fully connected layers with ReLU activation functions. The number of dimensions in the autoencoder's latent space is set to be half of the number of dimensions of the original vectors. For each subject, brain region and semantic space type, a separate autoencoder was trained. When training the autoencoder for a particular subject, we used the averaged brain RSMs of all other subjects. After we could finish the training process, we fed all original semantic vectors to the trained model used the intermediate features of the resulting trained autoencoder as the brain-grounded features.

We create the RSM matrices of brain activities or latent space by calculating the pairwise cosine similarity of each of the two data points. During the training process, we used the difference between upper triangle of each of the RSM matrices to constrain the autoencoder to make representations to be more brain-like.

Neural decoding of visual stimuli

We performed brain decoding by constructing linear regression models to predict semantic vectors from brain activity patterns. For predicting each unit of semantic vectors, a separate set of linear regression models were trained. Prior to applying regression analysis, we performed voxel selection (Horikawa & Kamitani, 2017), and brain activity patterns were Z-normalized.

More formally, given brain activity patterns as $x = \{x_1, x_2, ..., x_n\}^T$ representing activity of n neural activity data points (i.e., voxels in the fMRI data, source estimated neural activity patterns from MEG sensors, neural amplitude recorded from each channel in each second in the ECoG) from the region of interest, regression function can be represented as:

$$y(x) = \sum_{i=1}^{n} w_i x_i + w_0$$

where x_i is a scalar value specifying amplitude of the brain data point i, w_i is the weight of voxel i and w_0 is the bias.

For each subject, semantic space type and brain region, we trained a separate set of linear regression function as decoders. When decoding fMRI data of a particular subject to the brain-grounded semantic spaces, we used the brain-grounded space in which that subject was not used to create. When decoding MEG data of a particular subject, we used the averaged brain-grounded semantic spaces of all fMRI subjects.

Identification analysis

In the identification analysis, the predicted vector was identified among a large set of candidate vectors. First we prepared one random image from each class of imagenet dataset. Then, for each semantic space (i.e., GloVe and CLIP's pretrained feature vectors or GLoVe and CLIP's brain-grounded vectors of when alpha was set 0.01), we calculated the corresponding semantic vectors of all images that we had randomly selected from imagenet. If we couldn't obtain the GLoVe embeddings of a category, that category was getting excluded from the whole analysis. When obtaining the brain-grounded vectors of all ImageNet categories, we fed the original GLoVe/CLIP pretrained feature vectors to the corresponding trained autoencoder, and obtained the corresponding brain-grounded vectors. Then, for each category in the GOD dataset, we calculated the Pearson correlation coefficient of true and predicted vector, and the predicted vector and all other candidate vectors, and assigned the identification accuracy as the percentage of candidate categories, in which their correlation with the test predicted vector was lower than the correlation of true and predicted vectors.

References

- Benchetrit, Y., Banville, H., & King, J.-R. (2023). Brain decoding: Toward real-time reconstruction of visual perception. *arXiv Preprint arXiv:2310.19812*.
- Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3–4), 130–174. https://doi.org/10.1080/02643294.2016.1147426
- Brouwer, G. J., & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, 29(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009
- Chaudhary, U., Vlachos, I., Zimmermann, J. B., Espinosa, A., Tonin, A., Jaramillo-Gonzalez, A., Khalili-Ardali, M., Topka, H., Lehmberg, J., Friehs, G. M., Woodtli, A., Donoghue, J. P., & Birbaumer, N. (2022). Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training. *Nature Communications*, *13*(1), 1236. https://doi.org/10.1038/s41467-022-28859-8
- Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, 7(1), 13669. https://doi.org/10.1038/ncomms13669
- Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/10.1016/S1053-8119(03)00049-1

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848
- Federer, C., Xu, H., Fyshe, A., & Zylberberg, J. (2020). Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, *131*, 103–114. https://doi.org/10.1016/j.neunet.2020.07.013
- Fukuma, R., Yanagisawa, T., Nishimoto, S., Sugano, H., Tamura, K., Yamamoto, S., Iimura, Y., Fujita, Y., Oshino, S., Tani, N., Koide–Majima, N., Kamitani, Y., & Kishima, H. (2022). Voluntary control of semantic neural representations by imagery with conflicting visual stimulation. *Communications Biology*, 5(1), 214. https://doi.org/10.1038/s42003-022-03137-x
- 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. https://doi.org/10.1126/science.1063736
- Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5), 686–691. https://doi.org/10.1038/nn1445
- Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523–534. https://doi.org/10.1038/nrn1931
- Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1), 15037. https://doi.org/10.1038/ncomms15037
- Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210–1224.
- Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210–1224.
- Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685. https://doi.org/10.1038/nn1444
- Karpathy, A., & Fei-Fei, L. (2015). *Deep visual-semantic alignments for generating image descriptions*. 3128–3137.

- Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352–355. https://doi.org/10.1038/nature06713
- Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., & Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of Neural Engineering*, 7(5), 056007. https://doi.org/10.1088/1741-2560/7/5/056007
- Koide-Majima, N., Nishimoto, S., & Majima, K. (2023). Mental image reconstruction from human brain activity. *bioRxiv*. https://doi.org/10.1101/2023.01.22.525062
- Kourtzi, Z., & Kanwisher, N. (2000). Cortical Regions Involved in Perceiving Object Shape. *Journal of Neuroscience*, 20(9), 3310–3318. https://doi.org/10.1523/JNEUROSCI.20-09-03310.2000
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 4.
- Lebedev, M. A., & Nicolelis, M. A. L. (2006). Brain-machine interfaces: Past, present and future. *Trends in Neurosciences*, 29(9), 536–546. https://doi.org/10.1016/j.tins.2006.07.004
- Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4), 575–582. https://doi.org/10.1038/s41593-020-0608-8
- Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915–929.
- Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2022). Human alignment of neural network representations. *arXiv Preprint arXiv:2211.01201*.
- Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A. K., & Kornblith, S. (2023). Improving neural network representations using human similarity judgments. *arXiv Preprint* arXiv:2306.04507.
- Nakai, T., Koide-Majima, N., & Nishimoto, S. (2021). Correspondence of categorical and feature-based representations of music in the human brain. *Brain and Behavior*, 11(1), e01936. https://doi.org/10.1002/brb3.1936
- Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2), 400–410.

- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. 1532–1543.
- Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1), 963. https://doi.org/10.1038/s41467-018-03068-4
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). *Learning transferable visual models from natural language supervision*. 8748–8763.
- Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, *617*(7960), 360–368. https://doi.org/10.1038/s41586-023-06031-6
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*. https://doi.org/10.1101/407007
- Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019). End-to-End Deep Image Reconstruction From Human Brain Activity. *Frontiers in Computational Neuroscience*, 13. https://doi.org/10.3389/fncom.2019.00021
- Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633.
- Sitaram, R., Ros, T., Stoeckel, L., Haller, S., Scharnowski, F., Lewis-Peacock, J., Weiskopf, N., Blefari, M. L., Rana, M., Oblak, E., Birbaumer, N., & Sulzer, J. (2017). Closed-loop brain training: The science of neurofeedback. *Nature Reviews Neuroscience*, 18(2), 86–100. https://doi.org/10.1038/nrn.2016.164
- Stavisky, S. D., & Wairagkar, M. (2023). Listening in to perceived speech with contrastive learning. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-023-00742-1
- Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *Neuroimage*, *33*(4), 1104–1116.
- Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32.

- Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., & Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, 620(7976), 1031–1036. https://doi.org/10.1038/s41586-023-06377-x
- Willsey, M. S., Nason-Tomaszewski, S. R., Ensel, S. R., Temmar, H., Mender, M. J., Costello, J. T., Patil, P. G., & Chestek, C. A. (2022). Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder. *Nature Communications*, 13(1), 6899. https://doi.org/10.1038/s41467-022-34452-w
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111
- Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *ArXiv*, *abs/1311.2901*. https://api.semanticscholar.org/CorpusID:3960646