Visual Encoding and Decoding of the Human Brain Based on Shared Features

Chao Li*[1], Baolin Liu*†[2] and **Jianguo Wei[1]**
[1] College of Intelligence and Computing,
Tianjin Key Laboratory of Cognitive Computing and Application,
Tianjin University, Tianjin, China
[2] School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing, China

Abstract

Using a convolutional neural network to build visual encoding and decoding models of the human brain is a good starting point for the study on relationship between deep learning and human visual cognitive mechanism. However, related studies have not fully considered their differences. In this paper, we assume that only a portion of neural network features is directly related to human brain signals, which we call shared features. In the encoding process, we extract shared features from the lower and higher layers of the neural network, and then build a non-negative sparse map to predict brain activities. In the decoding process, we use back-propagation to reconstruct visual stimuli, and use dictionary learning and a deep image prior to improve the robustness and accuracy of the algorithm. Experiments on a public fMRI dataset confirm the rationality of the encoding models, and comparing with a recently proposed method, our reconstruction results obtain significantly higher accuracy.

1 Introduction

In recent years, many studies of the visual information processing mechanism of the human brain, the encoding and decoding of brain signals, and artificial intelligence technology using convolutional neural networks (CNNs) have been mutually enlightening and synergistic [Pei *et al.*, 2019; Rajalingham *et al.*, 2018]. In this context, we built an encoding model for human brain activation based on widely used convolutional neural networks, and then decoded the human brain's visual information by reversing the encoding process. This research can facilitate our understanding of the visual information processing mechanism of the human brain, as well as further explore the relationship between the human brain and deep learning.

The framework proposed in this paper has two parts (see Figure 1), the first of which is "encoding." In this part, we assume that only some of the features of the neural network have a clear relationship with human brain signals, which are called shared features. They are obtained by projecting the features of low and high layers of the neural network into their respective shared spaces. Then a non-negative sparse map is established between the shared features and brain activity to achieve the encoding process. The second part of the framework is "decoding," in which visual stimuli are reconstructed through neural network visualization technology, meanwhile, dictionary learning and sparse representation are introduced to improve the robustness of the algorithm. We tested the proposed framework on an open fMRI dataset [Kay *et al.*, 2008, 2011; Naselaris *et al.*, 2009] and the experimental results verified the rationality of the encoding models and demonstrated that this method can reconstruct the basic contours and textures of natural image stimuli.

The main innovations presented in this article are as follows:

- First, based on known neural cognitive mechanisms and machine learning methods, neural network features are projected onto a shared feature space that can be used to build a better encoding model to predict brain activation. We find that encoding models based on low-level features of a neural network can be more accurate than classic models based on Gabor wavelet features.
- Second, dictionary learning technology is used to train a corresponding dictionary for the shared features, and the sparse representation method is used to estimate features from measured brain signals. This method is robust to noise from fMRI signals and significantly improves the accuracy of feature estimation.
- Third, an unconditional model is used to generate an image, and the unbiasedness of the generated model ensures the interpretability of our method. In our proposed method, low-level visual features determine the contours of the generated image, and its principles are easy to explain. High-level features enrich the details of the image, and although their physical meaning is difficult to explain, related effects are constrained by low-level features, thereby making the results controllable.

* Chao Li and Baolin Liu equally contributed as co-first authors.
† Corresponding Author, email: liubaolin@tsinghua.edu.cn

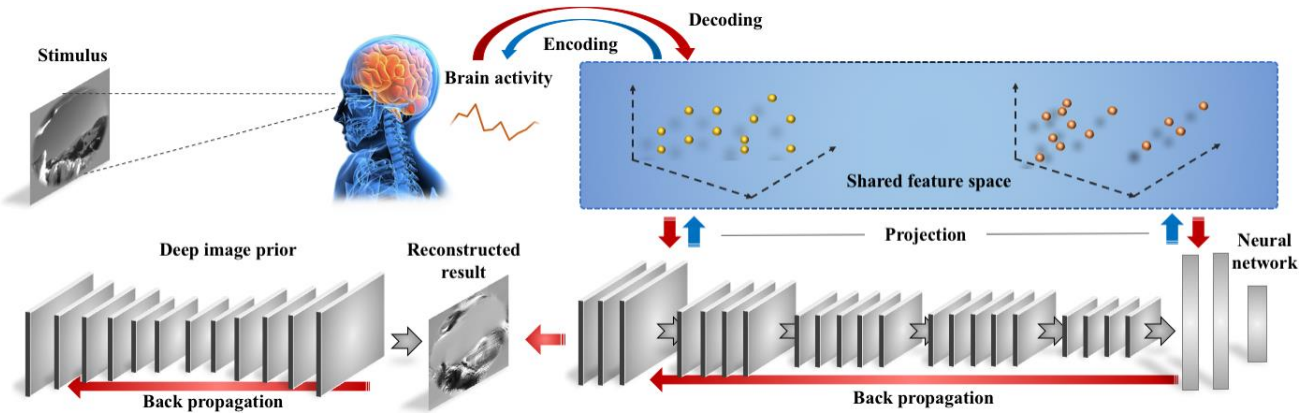

Figure 1: Schematic diagram of encoding and decoding.

2 Related Work

This work includes the encoding and decoding process, as well as implementing image reconstruction by using the recently developed neural network visualization technology. The related works are described as below.

A good encoding model can describe the information processing mechanism of human visual cortex and can accurately predict brain activities. It has been shown that simple cells in V1 are sensitive to stripes at specific spatial locations, spatial frequencies and orientations [Hubel and Wiesel, 1968; Olshausen and Field, 1996]. Based on this characteristic, Kay et al. used Gabor wavelets to extract the visual features in specific receptive fields, and established a voxel-wise encoding model to predict the activation of early visual areas [Kay *et al.*, 2008]. In addition, some recent studies have found that deep neural networks can also be used to extract effective features for modeling. For example, St-Yves and Naselaris used low-level features extracted from deep neural networks as input to obtain higher encoding accuracy [St-Yves and Naselaris, 2017]. Since then, researchers have extracted hierarchical visual features from deep networks and proved that there is a certain correspondence between the low / high-level features of neural networks and the low-level / high-level visual regions of the human brain [Horikawa and Kamitani, 2017; Wen *et al.*, 2018].

The task of decoding is to recover the information of visual stimulus from the measured brain signals, where reconstructing visual stimuli is a challenging branch. According to the visual stimulus materials, related works can be divided into reconstruction of artificial images and reconstruction of natural images. Artificial images include binary images [Miyawaki *et al.*, 2008] and handwritten characters [Du *et al.*, 2017; van Gerven *et al.*, 2010]. Because this type of images is limited to a specific category, its reconstruction task is relatively simple. However, the reconstruction of natural images does not limit the content of the images, so it requires a higher generalization capability of the reconstruction model. The related research can be traced back to the works of the Gallant Lab team. The idea of these works is to select the images / videos that best match to the measured subject's brain responses from a large data set based on the encoding model to approximate the observed images [Naselaris *et al.*, 2009] or videos [Nishimoto *et al.*, 2011]. These studies make good use of relevant results of encoding models. However, these methods are difficult to generate images flexibly. Since then, researchers have been trying to achieve image reconstruction based on deep learning. Seeliger et al. first established a mapping between brain activation and hidden variables of a pre-trained deep generation network, and then used the deep generation network to generate stimulus images [Seeliger *et al.*, 2018]. Shen et al. first built decoding models to estimate the features of each layer of a deep CNN, and then used a deep generation network to perform iterative calculations for the reconstruction of images [Shen *et al.*, 2019]. The results obtained by their methods have high structural and semantic accuracy, but because of relying on deep generation networks as strong prior information, there may be structural and semantic deviations in the generated results due to prior bias.

This paper is also related to the visualization of neural networks. Here we briefly introduce the feature inversion of convolutional neural networks. Mahendran and Vedaldi proposed a TV-norm-based energy function to reconstruct the natural pre-image from the features of different CNN layers [Mahendran and Vedaldi, 2016]. Since then, image reconstruction methods based on CNN and deep generation networks have been developed to further improve the accuracy of reconstructed images [Dosovitskiy and Brox, 2016a, 2016b]. However, these methods rely on training, which may cause deviations in generated results due to prior bias. Recently, Lempitsky et al. proposed deep image prior, which is an unconditional model and can generate image from a single feature by iterating a full convolutional network [Lempitsky *et al.*, 2018]. Considering this method does not require additional training and can effectively recover images, we use it as a prior to achieve feature-to-image reconstruction.

3 Method

In this study, we extract features from pre-trained AlexNet with Caffe version, then build and test encoding models on a public functional MRI dataset. In this section, we first introduce the functional MRI dataset, and then describe the encoding and decoding methods.

3.1 Functional MRI Dataset

In this study, we use the fMRI dataset collected by Kay et al. [Kay *et al.*, 2008, 2011; Naselaris *et al.*, 2009]. In the experiment, a 4T scanner was used to obtain fMRI data (BOLD signal) with a spatial resolution of 2 × 2 × 2.5 mm^3 and a TR of 1s. All stimuli are grayscale natural images with circular masks and fixed points. The image size is 500 px × 500 px, and the field of view is 20 × 20° . Two subjects participated. The data set consists of a training set and a test set. The training set contains 1,750 different images and the test set contains 120 new images. In the preprocessing procedure, the response amplitude (single value) of each stimulus image is estimated by deconvolving the response time course of each voxel. For each voxel response in the training set, calculating the ratio between the absolute value of the response and its standard error, and defining the median of the ratio as the voxel's signal-to-noise ratio (SNR) [Kay *et al.*, 2011]. In our study, we further divide the training set into two parts: training set Trn_1 is used to train the encoding models, which contains 1,575 images; training set Trn_2 is used to evaluate the models, which contains 175 images. A test set of 120 images is used to evaluate the performance of the decoding method.

3.2 Encoding (Feature to Brain Signals)

Recently, related research has confirmed that there is a certain correspondence between the low/high-level features of convolutional neural networks and the low/high-level visual regions of human brain [Wen *et al.*, 2018]. However, the feature space of the neural network and the representation space of brain activity are unlikely to completely overlap. Therefore, we first give the definition of shared feature space: the intersection between the feature space of a particular layer of a neural network and the representation space of brain activation. Then we try to find the shared features from the features of neural network using a linear transform and use them to estimate brain activation. It can improve the encoding accuracy and can facilitate our understanding of the similarities and differences between human brain and neural network.

Encoding with Low-Level Features

Early literatures supports that early visual areas of the human brain are sensitive to low-level features such as Gabor wavelets [Kay *et al.*, 2013, 2008], and AlexNet's first convolutional layer shares similarity with Gabor wavelets, so we use the features of the first convolution layer as the low-level features to build encoding models for early visual areas. In addition, since the BOLD signal of a voxel reflects the population response characteristics of neurons within the voxel, the response characteristics of a voxel are likely to be related to the features of multiple convolution channels [Li *et al.*, 2018]. In general, we use AlexNet's pool1 feature $\Phi_{pool1} \in R^{27\times 27\times 96}$ as the initial feature, and for a single voxel, try to construct an encoding map as follows:

$$\min_{\boldsymbol{w},b} \sum_{\boldsymbol{x}\in \boldsymbol{Trn_1}} \left[< \mathcal{M}_{pool1}\left(\Phi_{pool1}(\boldsymbol{x})\right), \boldsymbol{w} > + b - r(\boldsymbol{x_0}) \right]^2 \quad (1)$$
$$s.t.\ |\boldsymbol{w}|_1 + |b|_1 < \lambda$$
$$\boldsymbol{w},\ b > 0,$$

where x_0 and x are original and down-sampled stimulus, respectively; $\Phi_{pool1}(x)$ is the feature of pool1 layer when AlexNet takes x as input; $\mathcal{M}_{pool1}$ is a linear mapping, which is responsible for projecting the pool1 features into a shared space, and the outputs of it is represented as a vector; $<\cdot,\cdot>$ is product operation, $\boldsymbol{w}$ is regression coefficients for the shared feature, b is a constant, r is the evoked brain activity when the subject observes the image x_0. When $\mathcal{M}_{pool1}$ is known, $\boldsymbol{w}$, b can be calculated by Lasso regression.

Here pool1 feature is obtained by convolution and pooling operations with 96 convolution kernels. Therefore, this feature has 96 channels. The convolution kernel of each channel is shown in Figure 2a. Some convolution kernels are probably not significantly related to brain activity, so we use the following steps to evaluate and select the channels:

1) Initialization: Denoting the selected channel set is S, and the feature's tensor form is $\Phi_{pool1}^{i,j,k}$, where i, j, and k are the indexes of width, height, and channel of feature pool1, respectively. Let $S = \{1,2,\dots,96\}$, $\mathcal{M}_{pool1}(\Phi_{pool1}^{i,j,k}) = \sum_{k\in S} \Phi_{pool1}^{i,j,k} / |S|$, where $|S|$ is the size of S. Calculate the encoding map for all voxels in early visual areas and let V_0 be the number of voxels whose predictive power is greater than 0.4 (The predictive power of a voxel model is defined as the Pearson correlation between predicted and measured voxel responses for the images in Trn_2). Let $ind = 1$.
2) Taking the number of voxels with high predictive power (predictive power greater than 0.4) as the evaluation index, iteratively calculate the value of this index after removing a channel in S, thereby obtaining the most useless channel in S. Remove the most useless channel, let V_{ind} be the number of voxels with high predictive power at this time, and C_{ind} be the index of the removed channel. Let $ind = ind + 1$.
3) Repeat 2) until $ind = 96$.

The curve of V_{ind} is shown in Figure 2b. It can be seen that when $ind = 85$, the highest value is obtained, so we select the 12 channels remaining at this time as the elements in S, and let $\mathcal{M}_{pool1}(\Phi_{pool1}^{i,j,k}) = \sum_{k\in S} \Phi_{pool1}^{i,j,k} / |S|$ to build the final encoding models.

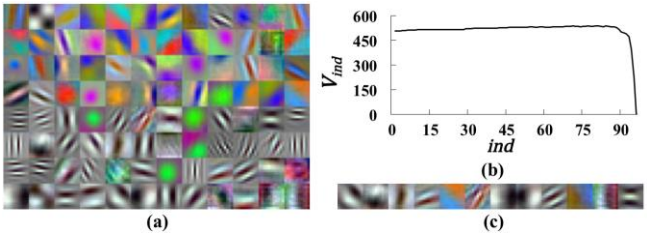

Figure 2: (a) The convolution kernels of AlexNet's first convolution layer. (b) The curve of V_{ind}. (c) The kernels corresponding to the selected channels.

Encoding with High-Level Features

High-level features reflect more semantic information, and the encoding and decoding of high-level features can obtain additional information outside early visual areas, so it is very important for reconstruction. According to the research by Wen et al., AlexNet's first fully connected layer (FC6) has relatively high encoding accuracy for V4, MT, PPA, FFA, LO and other brain regions [Wen *et al.*, 2018]. Therefore, we select the FC6 layer as the key layer and extract high-level shared features from it, and then build encoding models based on the features. Due to the lack of knowledge about the activation rules of mid- and high-level visual areas, we use an unsupervised learning method to estimate the shared feature space. Here, we train a non-negative dictionary, and use this dictionary as a projection matrix to map features into the shared space. There are two advantages in this way: Firstly, under the influence of non-negative properties, the elements in the dictionary tend to represent the local structure of features [Lee and Seung, 1999], which makes the dimensions of shared features more independent; Secondly, projection can reduce the dimension of features, which can avoid overfitting in subsequent modeling.

To avoid overfitting, we extract FC6 features from a public image database (Webvision) and train a non-negative dictionary. Denoting the extracted feature sample set is $\{\boldsymbol{\Phi}_{\boldsymbol{fc6}}^{\boldsymbol{i}} \in \mathbb{R}^{4096\times 1}, i = 1,\dots,N\}$, and the non-negative dictionary $\boldsymbol{D}_{\boldsymbol{fc6}}$ is obtained by calculating the following formula:

$$\begin{aligned} \min_{\boldsymbol{D}\in\mathcal{C},\boldsymbol{\alpha_i}\in\mathbb{R}^{k\times 1}} \ & \sum_{i=1}^{N} \frac{1}{2}\left\|\boldsymbol{\Phi}_{\boldsymbol{fc6}}^{\boldsymbol{i}} - \boldsymbol{D}_{\boldsymbol{fc6}}\boldsymbol{\alpha_i}\right\|_2^2 \\ s.t. \quad & \boldsymbol{D}_{\boldsymbol{fc6}} \geq 0,\ \forall i, \boldsymbol{\alpha_i} \geq 0, \end{aligned} \tag{2}$$

where $\mathcal{C} \equiv \{\boldsymbol{D} \in \mathbb{R}^{m\times k}: \boldsymbol{D} = [\boldsymbol{d_1},\dots,\boldsymbol{d_k}], \|\boldsymbol{d_j}\|_2 \leq 1, \forall j = 1,\dots,k\}$, we set $k = 700$ to reduce the dimension, $\boldsymbol{\alpha_i}$ is the representation vector. Denoting $\mathcal{M}_{fc6}(\boldsymbol{\Phi}_{\boldsymbol{fc6}}) = \boldsymbol{D}_{\boldsymbol{fc6}}^{\boldsymbol{T}}\boldsymbol{\Phi}_{\boldsymbol{fc6}}$, then a voxel-wise encoding model can be established by solving the following formula:

$$\begin{aligned} \min_{\boldsymbol{w},b} \ & \sum_{\boldsymbol{x}\in\boldsymbol{Trn_1}}\left[< \mathcal{M}_{fc6}\left(\Phi_{fc6}(\boldsymbol{x})\right), \boldsymbol{w} > + b - r(\boldsymbol{x_0})\right]^2 \\ & + R(\boldsymbol{w}, b), \end{aligned} \tag{3}$$

where $R(\boldsymbol{w}, b)$ is the L2 regularization term.

3.3 Decoding (Brain Signals to Features to Image)

Estimation of Low-Level Feature

In order to achieve feature estimation, after constructing the encoding map, we first give a measure of encoding loss in the general form as follows:

$$E(\boldsymbol{\mathcal{M}}) = \sum_k o_k\left(\boldsymbol{w_k^T}\boldsymbol{\mathcal{M}} + b_k - r_k\right)^2, \tag{4}$$

where $\boldsymbol{\mathcal{M}}$ is the shared feature, and o_k, r_k, $\boldsymbol{w_k}$, and b_k are the predictive power, measured brain activity, mapping coefficients, and mapping offset of the kth voxel, respectively. The shared feature can be estimated just by finding the minimum value of E, but this method will be affected by the high noise of the fMRI signal and the deviation of the encoding map, making the estimated feature accuracy low. In order to improve the robustness and accuracy of feature estimation, when estimating low-level shared features, we first train a non-negative dictionary for shared feature $\boldsymbol{\mathcal{M}_{pool1}}$, then estimate features using the dictionary. Specifically, we can estimate the low-level shared features by solving the following formula:

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \ & E_{\mathcal{M}pool1}\left(\boldsymbol{D}_{\boldsymbol{\mathcal{M}pool1}}\boldsymbol{\alpha}\right) \\ & + \lambda_\alpha\|\boldsymbol{\alpha}\|_1 + \beta\|\boldsymbol{\alpha}\|_2^2 + \gamma\left\|\boldsymbol{D}_{\boldsymbol{\mathcal{M}pool1}}\boldsymbol{\alpha}\right\|_2^2 \\ s.t. \quad & \boldsymbol{\alpha} > 0 \end{aligned} \tag{5}$$

$$\widehat{\boldsymbol{\mathcal{M}}}_{\boldsymbol{pool1}} = \boldsymbol{D}_{\boldsymbol{\mathcal{M}pool1}}\widehat{\boldsymbol{\alpha}}, \tag{6}$$

where $\boldsymbol{D}_{\boldsymbol{\mathcal{M}pool1}}$ is the dictionary, λ_α and β are constants used to control the sparsity and variance of $\boldsymbol{\alpha}$, respectively. γ is used to control the variance of $\widehat{\boldsymbol{\mathcal{M}}}_{\boldsymbol{pool1}}$.

Low-level features are sensitive to the contours and positions of objects in the image (see Section 4.3). When estimating high-level features, we can use the estimated low-level features as a priori information, and simultaneously generate images and estimate high-level features, so that the estimated high-level features and low-level features are coordinated.

High-Level Feature Estimation and Image Reconstruction

After getting low-level features, back-propagation is used to simultaneously reconstruct the stimulus image and estimate high-level features. Without regularization, it is difficult to reconstruct the image. Here we use the deep image prior as the regular term [Lempitsky *et al.*, 2018], the formula is as follows:

$$\begin{aligned} \hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ & \sigma_p\left\|\mathcal{M}_{pool1}\left(\Phi_{pool1}\left(f_\theta(\boldsymbol{z})\right)\right) - \widehat{\boldsymbol{\mathcal{M}}}_{\boldsymbol{pool1}}\right\|_2^2 \\ & + \sigma_f E_{\mathcal{M}_{fc6}}\left(\mathcal{M}_{fc6}\left(\Phi_{fc6}\left(f_\theta(\boldsymbol{z})\right)\right)\right), \end{aligned} \tag{7}$$

$$\hat{x} = f_{\hat{\theta}}(\boldsymbol{z}), \tag{8}$$

$$\widehat{\boldsymbol{\mathcal{M}}}_{\boldsymbol{fc6}} = \mathcal{M}_{fc6}\left(\Phi_{fc6}(\hat{\boldsymbol{x}})\right), \tag{9}$$

where f_θ is a randomly initialized hourglass network with 6 layers, θ is the parameter, $\hat{\boldsymbol{x}} \in \mathbb{R}^{W\times H\times 1}$ is the estimated image, and $\boldsymbol{z} \in \mathbb{R}^{W\times H\times C}$ is a random code vector subject to the distribution $U[0,1]$, σ_p and σ_f are the coefficients of the two loss terms, respectively. The above method constructs a nonlinear mapping f_θ by optimizing the parameter θ in the hourglass network, thereby mapping the random code vector $\boldsymbol{z}$ to the image we want to reconstruct. (The code for f_θ is available at https://github.com/Nehemiah-Li/Deep-image-prior.) Although the deep image prior does not have a learning process, the experimental results show that as long as the feature estimation is accurate, back propagation based on the prior can effectively reconstruct the observed image.

4 Experimental Results

4.1 Analysis for Encoding Process

In order to know which brain regions the encoding model can be well applied to, we calculated the correlation coefficient between the predicted responses and measured responses for each voxel model on the test set, and counted the proportion

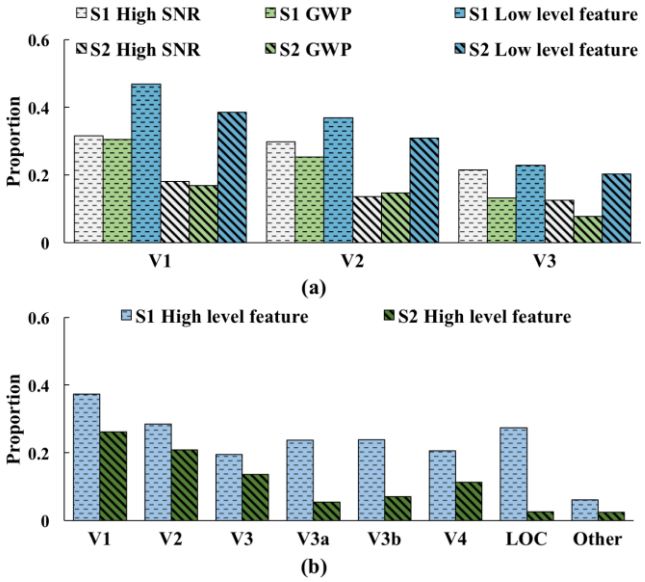

Figure 3: (a) Distribution of low-level-feature-based encoding models in early visual areas. The vertical axis represents the proportion of the counted voxels to all voxels in the region. GWP represents the encoding model based on Gabor wavelet pyramid, and the relevant data comes from the literature [Kay *et al.*, 2008]. (b) The proportion of high-level-feature-based encoding models in each scanned brain region.

of voxels with significant correlation coefficients ($P < 0.01$) in each brain region (for comparison, the significance test method is exactly the same as that of literature [Kay *et al.*, 2008]). Figure 3a shows that the proportion is highest in the V1 region, and the more the number of high SNR voxels, the more the number of significant voxels. The distribution of the two subjects shows consistency. In addition, we also compared our models with Gabor-wavelet-based encoding models [Kay *et al.*, 2008] (see Figure 3a). It can be seen that the encoding models based on the low-level shared features have higher accuracy. Because the function of the first convolutional layer is easy to explain, the models based on the low-level features have strong interpretability. The proportion of encoding models based on the high-level features in each visual region is shown in Figure 3b. High-level features are used to model all brain regions in the data set (the data set gives a total of 8 visual areas). It can be seen that the high-level features can encode mid- and high-level areas such as V3A, V3B, V4 and lateral occipital area (LOC). A considerable number of encoding models have also been built in undivided (other) visual area (it has a large cardinal number).

4.2 Representation Ability of the Dictionaries

In the process of feature estimation, the dictionary plays an important role. After training the dictionaries, we analyze the key attributes of the dictionaries on another dataset. Table 1 shows the dictionary names, feature dimensions, the root mean square error (RMSE) and correlation coefficients between the normalized actual features and the represented features, the average number of non-zero values in the representation vectors and the atoms in the dictionaries. It can be seen that the dictionary $\boldsymbol{D}_{\mathcal{M}pool1}$ have sufficient representation ability for the low-level shared features and can achieve such representation accuracy using less than 200 non-zero elements. In theory, if the shared space does not completely overlap with the feature space, some information will be lost after the projection transformation. Therefore, as a projection matrix, $\boldsymbol{D}_{fc6}$ has no special requirement for its representation ability.

Dictionary	Dimension	RMSE	Correlation	Non-zeros	Atoms
$\boldsymbol{D}_{\mathcal{M}pool1}$	729	0.061	0.959	145.21	4000
$\boldsymbol{D}_{fc6}$	4096	0.029	0.765	206.08	4000

Table 1: The representation ability of dictionaries.

Feature	Dimension	Max Value	Subject	RMSE	Correlation
$\mathcal{M}_{pool1}$	729	123.71	Subject1	**17.033**	**0.803**
			Subject2	18.454	0.764
$\mathcal{M}_{fc6}$	700	72.364	Subject1	11.105	**0.494**
			Subject2	**10.523**	0.468

Table 2: Accuracy of feature estimation.

4.3 Analysis for Decoding Process

Figure 4 shows some reconstructed images in the test set. It is shown that the reconstructed results can capture the contours of the observed images, and the textures in some images are consistent. The reconstruction accuracy of the data of subject 1 is stronger than that of subject 2, which is consistent with the difference between the encoding accuracy of two subjects. Because the estimation accuracy of the shared features directly affects the quality of the reconstructed image, we first analyze the accuracy of feature estimation by comparing the RMSE and correlation between the estimated features and the actual features. Table 2 shows the average values of accuracy indexes of the estimated features in the test set of 120 images. It is shown that the accuracy of estimated features on the data of subject 1 is slightly higher than that on the data of subject 2, and the estimation accuracy of low-level features is higher than that of high-level features.

In addition, in order to qualitatively analyze the impact of low-level features and high-level features on the reconstruction results, we separately use low-level features and high-level features to reconstruct the image, and the results are shown in Figure 5. We can see that the images reconstructed from low-level features have basic contours, and the positions of edges are more accurate. Moreover, by comparing the reconstruction result with the estimated low-level feature map, we can see that the estimated feature map reflects the main information such as the contour and edge of image. The position of the contours of the images reconstructed by high-level features alone may not be accurate, but the topological

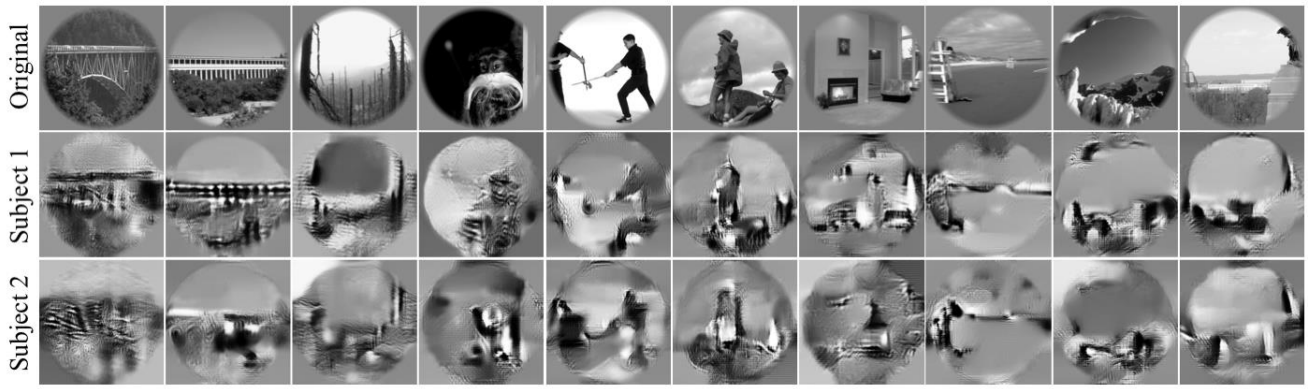

Figure 4: Image reconstruction results.

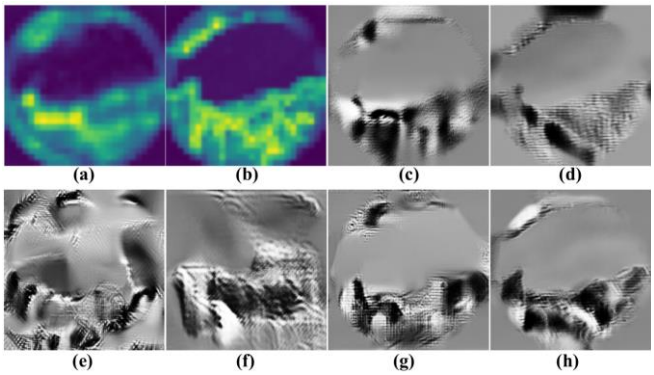

Figure 5: (a, b) Estimated and real feature maps of $\mathcal{M}_{pool1}$. (c, d) Images reconstructed from estimated and real low-level features. (e, f) Images reconstructed from estimated and real high-level features. (g, h) Images reconstructed from estimated and real shared features. The original image is shown in Figure 4 (penultimate column).

structure is consistent with original image. It may be that the high-level features have a wider receptive field and are more abstract. When the high-level features and low-level features are combined to reconstruct the image, the contour and texture accuracy of the result are significantly improved.

Finally, we compared our method with the method proposed by Seeliger et al. [Seeliger *et al.*, 2018]. To our knowledge, their method is the latest method based on this dataset. For comparison, we use the same quantitative evaluation method for the reconstruction results: performing a behavioral perceptual study using Amazon Mechanical Turk (www.mturk.com). Specifically, in each test, the worker is presented with an original image in the test set, and meanwhile he is required to choose between a real reconstruction and a different reconstruction randomly selected from the test set. This process is repeated ten times for each image in the test set, and different reconstructions randomly selected in each test were used as interference terms. More details about the evaluation method can be seen in [Seeliger *et al.*, 2018]. Table 3 shows the statistical results of this test. Decision accuracy represents the total number of correct decisions in all comparisons. Image accuracy indicates the number of images in the test set that can be correctly identified after a majority vote is applied to ten decisions for each image. If there are at least 8 correct decisions out of 10 comparisons, it means that the image is identifiable. Table 3 shows that our lowest accuracy (S2) is slightly lower than the accuracy achieved by Seeliger et al., while the highest accuracy (S1) and average accuracy are significantly higher than the accuracy of their results. (Note that Seeliger et al. used hyperalignment method to average the multiple subjects' data into a single hyperaligned subject data with improved SNR. In theory, the SNR of the data of subject 1 is closer to that of hyperaligned data.)

Method	Decision accuracy	Image accuracy	Identifiable images
Seeliger et al.	66.4%	70.0%	43.3%
Ours (S1)	**76.2%**	**89.2%**	**58.3%**
Ours (S2)	66.1%	71.7%	38.3%
Ours (average)	71.2%	80.5%	48.3%

Table 3: Accuracy of reconstructed images.

5 Conclusion

Based on the concept of shared features, we built encoding models to explore the relationship between the deep CNN and human brain activity, and then proposed a reconstruction method for decoding brain signals. Experimental results show that the low-level shared features of the deep CNN can efficiently encode the early visual areas of the human brain, while the shared features of the fully connected layer can encode the mid- and high-level visual areas of the human brain. The reconstruction method effectively integrates the information of low- and high-level shared features, and nearly half of the reconstructions retain the main features of visual stimuli. Further research on the encoding mechanism of the high-level visual regions of the human brain and its intrinsic relationship with neural networks will be important for understanding the human brain's visual cognitive mechanism and improving the accuracy of decoding brain information.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2018YFB0204304), National Natural Science Foundation of China (No.U1736219 and No.61571327).

References

[Dosovitskiy and Brox, 2016a] Alexey Dosovitskiy, and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks, In *30th Conference on Neural Information Processing Systems*, pages 658–666, 2016.

[Dosovitskiy and Brox, 2016b] Alexey Dosovitskiy, and Thomas Brox. Inverting Visual Representations with Convolutional Networks, In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4829–4837, June 2016.

[Du *et al.*, 2017] Changde Du, Changying Du, and Huiguang He. Sharing deep generative representation for perceived image reconstruction from human brain activity, In *Proceedings of the International Joint Conference on Neural Networks*, pages 1049–1056, May 2017.

[Horikawa and Kamitani, 2017] Tomoyasu Horikawa, and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8: 1–15, 2017.

[Hubel and Wiesel, 1968] D. H. Hubel, and T N Wiesel. Receptive Fields and Functional Architecture of monkey striate cortex. *Journal of Physiology*, 195: 215–243, 1968.

[Kay *et al.*, 2008] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452: 352–355, March 2008.

[Kay *et al.*, 2013] Kendrick N. Kay, Jonathan Winawer, Ariel Rokem, Aviv Mezer, and Brian A. Wandell. A Two-Stage Cascade Model of BOLD Responses in Human Visual Cortex. *PLoS Computational Biology*, 9: e1003079, May 2013.

[Kay *et al.*, 2011] Kendrick N Kay, Thomas Naselaris, and Jack L Gallant. fMRI of human visual areas in response to natural images. *CRCNS.org*, 2011.

[Lee and Seung, 1999] Daniel D. Lee, and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788–791, 1999.

[Lempitsky *et al.*, 2018] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. Deep Image Prior, In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, June 2018.

[Li *et al.*, 2018] Chao Li, Junhai Xu, and Baolin Liu. Decoding natural images from evoked brain activities using encoding models with invertible mapping. *Neural Networks*, 105: 227–235, September 2018.

[Mahendran and Vedaldi, 2016] Aravindh Mahendran, and Andrea Vedaldi. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. *International Journal of Computer Vision*, 120: 233–255, December 2016.

[Miyawaki *et al.*, 2008] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60: 915–929, 2008.

[Naselaris *et al.*, 2009] Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63: 902–915, 2009.

[Nishimoto *et al.*, 2011] Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21: 1641–1646, 2011.

[Olshausen and Field, 1996] Bruno A Olshausen, and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381: 607, 1996.

[Pei *et al.*, 2019] Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, Feng Chen, Ning Deng, Si Wu, Yu Wang, Yujie Wu, Zheyu Yang, Cheng Ma, Guoqi Li, Wentao Han, Huanglong Li, Huaqiang Wu, Rong Zhao, Yuan Xie, and Luping Shi. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*, 572: 106–111, August 2019.

[Rajalingham *et al.*, 2018] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, 38: 7255–7269, August 2018.

[Seeliger *et al.*, 2018] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and M.A.J. van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181: 775–785, November 2018.

[Shen *et al.*, 2019] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15: e1006633, January 2019.

[St-Yves and Naselaris, 2017] Ghislain St-Yves, and Thomas Naselaris. The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, 2017.

[van Gerven *et al.*, 2010] Marcel A.J. van Gerven, Floris P. de Lange, and Tom Heskes. Neural decoding with hierarchical generative models. *Neural Computation*, 22: 3127–3142, December 2010.

[Wen *et al.*, 2018] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28: 4136–4160, December 2018.