DCNN-GAN: Reconstructing Realistic Image from fMRI

anonymous

CONFIDENTIAL. For review only.

Abstract

Visualizing the perceptual content by analyzing human functional magnetic resonance imaging (fMRI) has been an active research area. However, due to its high dimensionality, complex dimensional structure, and small number of samples available, reconstructing realistic images from fMRI remains challenging. Recently with the development of convolutional neural network (CNN) and generative adversarial network (GAN), mapping multi-voxel fMRI data to complex, realistic images has been made possible. In this paper, we propose a model, ***DCNN-GAN****, by combining a reconstruction network and GAN. We utilize the CNN for hierarchical feature extraction and the DCNN-GAN to reconstruct more realistic images. Extensive experiments have been conducted, showing that our method outperforms previous works, regarding reconstruction quality and computational cost.*

1 Introduction

The externalization of the mental content is a fundamental research area in neuroscience. In the last decade, the analysis of multi-voxel fMRI patterns using machine learning techniques allows for the interpretation of visual content. Recent work has progressed from matching seen images to exemplars [1], to introducing deep neural networks (DNN) [2] to extract the hierarchical neural representations of the human visual system and to reconstruct the images seen by the subject.

However, while the brain activity measured by fMRI can be decoded (translated) into DNN features across multiple layers of the network [3], the large size of the features, and the absence of regularization in the regression model, contribute to the low decoding accuracy. As a result, the reconstructed image bears little resemblance to the original one. Some model only optimizes the reconstructed image to be similar to natural ones, without utilizing the categorical information of the image. Other work focuses on the reconstruction of a particular type of images [4] [5], which has improved detail but lacked generality.

In this paper, we propose DCNN-GAN, a new model that reconstructs more realistic images from fMRI. Our model consists of a reconstruction network and a recently proposed GAN, pix2pix [6], that allows for pixel-wise image generation. In our proposal, an encoder network based on VGG-19 [7] extracts the features from the input image. An fMRI decoder learns the mapping from the fMRI data to the extracted features and decodes features from fMRI test data. In the DCNN-GAN, the reconstruction network outputs the coarse image from the decoded features. The GAN generates a more realistic image from the coarse one.

Our proposed method significantly reduces the size of the decoded features and uses Ridge regression in the decoder to improve numerical stability which contributes to the improvement in decoding accuracy. The DCNN-GAN can render images with more semantically plausible details due to the introduction of category-specific prior. Compared to the reconstruction of a particular image category, our work can be applied to reconstructing various categories of images. Moreover, our proposal stands out for its efficiency, in that it achieves real-time reconstruction.

Through both quantitative comparison and human assessment of the images reconstructed using our method, we have observed an enhancement of the reconstructed image quality among various image categories.

2 Related Work

As a technically challenging task, reconstructing realistic images from fMRI has been an active area of research in computational neuroscience over the last decades. Before the introduction of deep neural networks(DNN), previous works have only achieved matching the images to similar ones [1] and reconstruction of contrast-based image [8] that is low in resolution. These methods directly decode the fMRI into the image to be reconstructed, which limit the number of possible outputs and is unfit for reconstructing images with higher resolution. Some model [9] has used variational autoencoder instead of DNN, resulting in generating relatively blurry images.

Recent works have used DNN [2] [3] [10] to obtain the features of the input images. With DNN, the process of reconstruction usually involves two crucial steps, the decoding of fMRI and the reconstruction of the image using the decoded DNN features.

Decoding of fMRI. The decoding of fMRI activity aims to translate the fMRI pattern measured when the human test subject sees an image, into DNN features that can represent the seen image. Recent work has used sparse linear regression (SLR) [11] to learn the relationship between fMRI data and the DNN features of all convolutional layers of a VGG-19 model with the same input image [2]. However, we find such linear regression model to behave inadequately without regularization. Also, it is difficult to accurately decode the features of all convolutional layers because the accumulative feature size is too large compared to the small number of fMRI samples.

Reconstruction of images. Various methods and models have been introduced to reconstruct images from decoded DNN features. Recent work [3] has proposed an iterative algorithm that optimizes the reconstructed image so that the DNN features of the image are similar to those decoded from fMRI activity. While the decoded features capture the hierarchical visual information of the image, the difficulty in decoding all features across multiple layers of the VGG-19 model prohibits the model from generating higher quality images. The reconstructed image contains shape that resembles the original image but presents no identifiable textures. We also find the iterative method converges slowly and is therefore unfit for real-time reconstruction.

3 Approach

3.1 Neuroscience backgrounds

Human visual system. The human visual system processes and interprets the visual input to build a mental image of the surroundings. The visual cortex (VC) located in the cerebral cortex is responsible for processing the visual image. It is composed of subsequent regions, namely V1, V2, V3, V4 [12] [13], lateral occipital complex (LOC) [14], fusiform face area (FFA) [15], and parahippocampal place area (PPA) [16]. In this paper, the entire VC is selected as the brain region of interest (ROI) in the decoding process.

Functional magnetic resonance imaging. The fMRI data we used in this paper records the blood-oxygen-level-dependent (BOLD) signal, which measures the hemodynamic response that reflects the level of brain activity [17]. As depicted in Figure 1., the fMRI data is a four-dimensional sequence consisting of 3-D volumes sampled every few seconds. Each volume consists of over 700,000 voxels, each voxel measuring $2 \times 2 \times 2mm$.

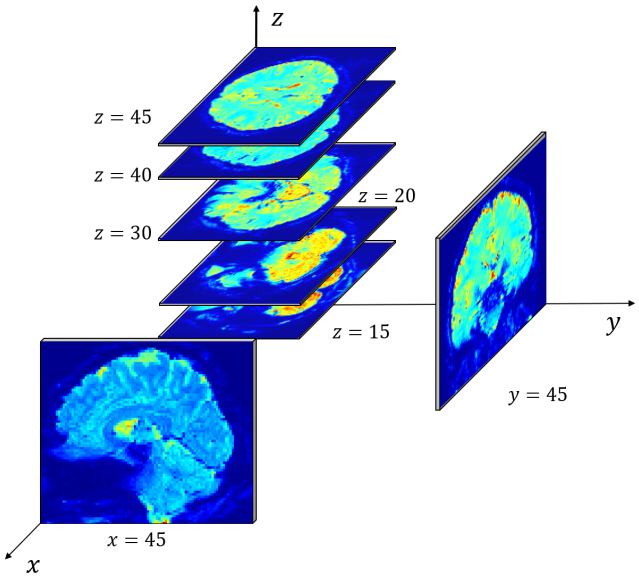

Figure 1. Visualization of raw fMRI data, showing slices of a fMRI volume scanned in three directions: sagittal (x), coronal (y) and axial (z).

3.2 Model Formulation

In this section, we explain in detail the proposed reconstruction model. As shown in Figure 2, the reconstruction model contains three parts:

1. The encoder network E, which extracts feature z from the original image x.
2. The fMRI decoder D_f, which is trained to learn the mapping from fMRI data to z.
3. The DCNN-GAN that performs image reconstruction using the decoded features.

The encoder network. We build the encoding network E using VGG-19 model pre-trained on ILSVRC2012 [18]. By exploiting the feature extracting property of the VGG-19, E maps the original image x to feature vector z, as well as the categorical information c of the image. In order to reduce the computational cost of training the fMRI decoder and to improve the accuracy of decoding, we take the output of the first fully connected layer (fc_7) as the feature vector z. This will reduce the dimension of z to 4096, which is sufficient for preserving the visual information of the original image.

The fMRI decoder. We use a linear least-squares regression model with Tikhonov regularization (Ridge regression) [19] to build the decoder. In previous work, an ordinary least squares model is used, which is unstable due to the absence of regularization. The regularization technique increases the numerical stability

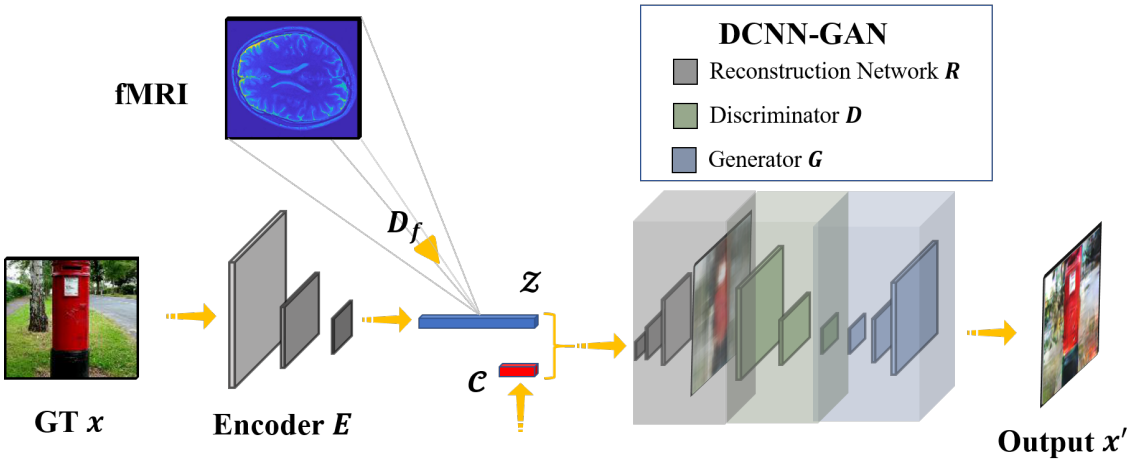

Figure 2. Overview of the reconstruction method. The model consists of: 1) The encoder network; 2) The fMRI decoder D_f; 3) The DCNN-GAN. Please see Section 3.2 for detailed descriptions.

of our model.

Given the feature z of an input image, and the fMRI multi-voxel data X when showing subject the same image, the model computes the weight vector w and the bias b in the regression function.

$$z = w^T X + b \tag{1}$$

By minimizing the objective function

$$\|z - (Xw + b)\|_2^2 + \alpha \|w\|_2^2 \tag{2}$$

where $\|w\|_2^2$ is the L2 regularization term, and α is the regularization strength parameter.

DCNN-GAN. DCNN-GAN is defined as a combination of the reconstruction network R and GAN. R is a deconvolution network that reconstructs the coarse image from decoded feature vector z. The GAN, composed of a generative network G and a discriminative network D, takes a coarse image $R(z)$ and the categorical information c as input and outputs a refined image x'.

The idea behind is that the information decoded from fMRI is insufficient to reconstruct realistic images, and rendering semantically essential details to the image without knowing its category is impractical. Therefore, we use GAN to introduce the image prior, based on the categorical information of the image. This will optimize the reconstructed image to be similar to images of the same category, adding more semantically plausible details to the image.

The objective function of the network is the combination of the reconstruction network loss, the conditional GAN loss, and a traditional L1 loss. The generative network tries to minimize the objective while the discriminative network tries to maximize it. Therefore, the final objective function can be expressed as:

$$\begin{aligned} G = \arg\min_C \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \\ + \theta \mathcal{L}_2(R) \end{aligned} \tag{3}$$

Given w the random noise,

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D) = \log D(R(z), x') \\ + \log(1 - D(R(z), G(R(z), w))) \end{aligned} \tag{4}$$

is the loss function of the conditional GAN, and

$$\mathcal{L}_{L1}(C) = \|x' - G(R(z), w)\|_1 \tag{5}$$

is the L1 loss function, and

$$\mathcal{L}_R(G) = \|x - R(z)\|_2 \tag{6}$$

is the loss function of the reconstruction network. We optimize the loss of reconstruction network and the loss of G, D alternatively in practice.

4 Experiment

In this section, we validate the effectiveness of our reconstruction model by experiments. We have trained and tested our model on the following datasets.

4.1 Datasets

ILSVRC2012. The ILSVRC2012 dataset is a subset of the large hand-labeled ImageNet dataset. The validation and test data consist of 150,000 images, and the training data contains 1.2 million images. All the images are hand labeled with the presence or absence of 1000 object categories which do not overlap with each other.

fMRI on ImageNet. Originally used in [2]. The fMRI data were recorded while subjects were viewing object images (image presentation experiment) or were imagining object images (imagery experiment). In the training image session, a total of 1,200 images from 150 object categories were each presented only once. In the test image session, a total of 50 images from 50 object categories were presented 35 times each. All images were from ImageNet (Fall 2011 release). In this paper, we use the fMRI data of the VC region in the image presentation experiment. The number of voxels in the VC region is 4466. The number of fMRI samples available is 2700.

4.2 Implementation Details

We have implemented our approach using Pytorch [20]. The encoder network E is based on VGG-19 model, pretrained on ILSVRC2012 training set. The input of E is the original image of size $224 \times 224 \times 3$. We take the output of the first fully connected layer of VGG-19 with size 4096×1 as the feature vector used to train the fMRI decoder.

The fMRI decoder is constructed using a linear regression model. We have chosen Ridge regression, which is a linear least squares with L2 normalization. The number of samples in the training set is 2200, and the number of fMRI voxels is 4400. The L2 regularization strength α is empirically set to 0.7.

The reconstruction network R is a deconvolution network consisting of a fully connected layer with input size 4096×1 and output size $512 \times 7 \times 7$, followed by 4 deconvolutional layers with kernel size set to 4×4 and output channel set to $256, 128, 128, 128$, and a convolutional layer with kernel size set to 1×1 and output channels set to 3. The output size of G is $112 \times 112 \times 3$. We used the Adam optimizer, with initial learning rate set to 0.01 and exponential decay. On the ILSVRC2012 training set, we trained our network for 200 epochs with batch size set to 256. We add Gaussian noise to the input to increase the robustness of the network.

We trained the GAN using the output images from the reconstruction network R and the corresponding original images as paired data on a specific category of the ILSVRC2012 training set for 500 epochs. The input is resized to $128 \times 128 \times 3$, the same size as the output, the batch size is 256, and the learning rate is 0.001.

4.3 Visual comparison with previous models

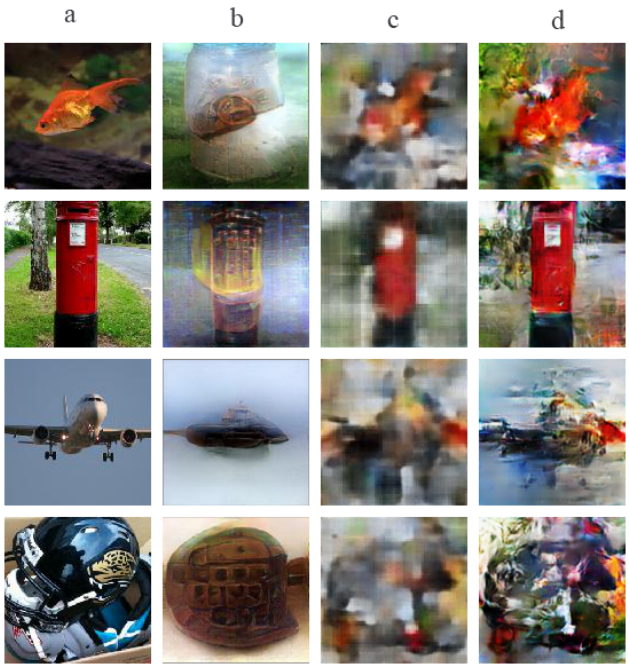

Figure 3. Comparison of reconstructed images from different models. (a) Original images (b) Reconstructed images (previous model) (c) Reconstructed images (the output of R) (d) Reconstructed images (the output of DCNN-GAN)

The results of the existing model are less recognizable because they tend to preserve the shape of the object while sacrificing the texture, while our proposal reaches a balance between the shape and texture. Some of the test instances (i.e. the mailbox) introduce objects that do not belong to the input, indicating that the existing model may over-fit the training set. From the comparison we can conclude that, the outputs of the DCNN-GAN are significantly improved in resolution and details when compared to the intermediate outputs of the deconvolution network. It supports the idea that GAN can optimize the generation of images by introducing category-specified prior knowledge.

4.4 Comparison of decoding accuracy

In this part, we compare the fMRI decoding accuracy between the linear regression model used in previous

work and the ridge regression model in our method. We use two common regression metrics. The coefficient of determination, as known as R-squared, reflects the goodness of fit of a model, where 1 is the best and lower the worse. The root-mean-square error (RMSE) measures the error between the predicted and observed values. The result shown in Table 1 indicates an improvement in decoding accuracy using our model compared to the previous model.

Table 1. Results of decoding accuracy

Metrics	Linear	Ridge	Mean
R-squared	-0.3093	0.3184	-0.0039
RMSE	0.4960	0.3614	0.4402

4.5 Human Perceptual Study

Evaluating the quality of the reconstructed images is an open problem. None of the traditional metrics can effectively measure the similarity between two images with complex structural and textural information. Therefore we conducted a double-blind perceptual study on a group of randomly selected volunteers in the interest of holistic evaluation of our results.

Perceptual study. Over 40 volunteers were surveyed in the study, and each of them was presented a sequence of trials on which the image reconstructed by the existing model was pitted against the image generated by our model with the same input. Given the origin image x, volunteers were asked to select the image which they viewed as a better reconstruction of x and report whether it was a satisfying reconstruction. Each image was presented for 1 second, and volunteers were given unlimited time to decide their response.

The result of the perceptual is listed as follows.

Table 2. Results of perceptual study

Items	Existing model	Our proposal
Reconstruct(%)	44.3	55.7
Satisfaction(%)	12.4	20.1

We can derived the conclusion that our model has achieved improvement in performance compared to existing work.

5 Conclusion

In this paper, we proposed a novel model DCNN-GAN and applied it to the output features of the fMRI decoder to reconstruct realistic images. Compared to previous works, the overall quality of the reconstructed image was considerably enhanced, and our model achieved real-time reconstruction. The future work is to reconstruct realistic images of which categories do not exist in the training set.

References

[1] Shinji Nishimoto, AnT. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and JackL. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641 – 1646, 2011.

[2] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *arXiv e-prints*, page arXiv:1510.06479, October 2015.

[3] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *bioRxiv*, 2017.

[4] Rufin VanRullen and Leila Reddy. Reconstructing Faces from fMRI Patterns using Deep Generative Neural Networks. *arXiv e-prints*, page arXiv:1810.03856, October 2018.

[5] Alan S. Cowen, Marvin M. Chun, and Brice A. Kuhl. Neural portraits of perception: Reconstructing face images from evoked brain activity. *NeuroImage*, 94:12 – 22, 2014.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[8] 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(5):915 – 929, 2008.

[9] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, and Zhongming Liu. Variational autoencoder: An unsupervised model for modeling and decoding fmri activity in visual cortex. *bioRxiv*, 2018.

[10] 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(12):4136–4160, 2018.

[11] Masa-aki Sato. Online model selection based on the variational bayes. *Neural Computation*, 13(7):1649–1681, 2001.

[12] Stephen A Engel, David E Rumelhart, Brian A Wandell, Adrian T Lee, Gary H Glover, Eduardo-Jose Chichilnisky, and Michael N Shadlen. fmri of human visual cortex. *Nature*, 1994.

[13] Martin I Sereno, AM Dale, JB Reppas, KK Kwong, JW Belliveau, TJ Brady, BR Rosen, and RB Tootell. Borders of multiple visual areas in humans revealed

by functional magnetic resonance imaging. *Science*, 268(5212):889–893, 1995.

[14] Zoe Kourtzi and Nancy Kanwisher. Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, 20(9):3310–3318, 2000.

[15] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

[16] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598, 1998.

[17] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain Magnetic Resonance Imaging with Contrast Dependent on Blood Oxygenation. *Proceedings of the National Academy of Science*, 87:9868–9872, December 1990.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[19] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, February 2000.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.