Perceived Image Reconstruction from Human Brain Activity via Time-series Information Guided Generative Adversarial Networks[⋆]

Shuo Huang, Liang Sun, Muhammad Yousefnezhad, Meiling Wang, and Daoqiang Zhang ✉

College of Computer Science and Technology
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
dqzhang@nuaa.edu.cn

Abstract. Understanding how human brain works has attracted increasing attentions in both fields of neuroscience and machine learning. Previous studies have used autoencoder and generative adversarial networks (GAN) to improve the quality of perceived image reconstruction from functional Magnetic Resonance Imaging (fMRI) data. However, these methods mainly focus on acquiring relevant features between stimuli images and fMRI while ignoring the time-series information of fMRI, thus leading to sub-optimal performance. To address this issue, in this paper, we develop a time-series information guided GAN method for reconstructing visual stimuli from human brain activities. In addition, to better measure the *modal difference*, we leverage a pairwise ranking loss to rank the stimuli images and fMRI to ensure strongly associated pairs at the top and weakly related ones at the bottom. Experimental results on real-world datasets suggest that the proposed method achieves better performance in comparison with several state-of-the-art image reconstruction approaches.

Keywords: Perceived image reconstruction · functional Magnetic Resonance Imaging (fMRI) · Long-Short Term Memory (LSTM) · generative adversarial networks (GAN).

1 Introduction

"Reading minds" has been one of the most significant challenges in the field of neuroscience in the past for a long time [1,2]. To this end, an algorithm called the human brain encoding and decoding has been proposed, where the encoding part embeds information into neural activities, while the decoding part extracts

[⋆] This work was supported by the National Key Research and Development Program of China (Nos. 2018YFC2001600, 2018YFC2001602) and the National Natural Science Foundation of China (Nos. 61876082, 61861130366 and 61732006), the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship (No. NAF\ R1\180371).

information from neural activities [3]. Functional Magnetic Resonance Imaging (fMRI) is one of the most popular tools for studying the human brain, using blood oxygen level dependence (BOLD) signals as a proxy for visual neural activity. The main idea is to use these measurements of neural activities to process cognitive state [4,5].

Recent years, several deep neural network based methods have been proposed for decoding the cognitive states in human brains. For instance, some studies use the outputs of DNN to reveal the neural activities in human visual cortex [6,7,8]. However, there are still some challenges for the perceived image reconstruction from human brain activity with fMRI. In particular, 1) fMRI data is usually high-dimensional with a lot of complex noises, which interferes with the mining of real brain activity and influences the reconstruction results; 2) the pairwise samples are treated as time point samples, which ignores the time-series information of the visual task; 3) the limited mapping between the stimuli images and the evoked brain avtivity patterns, which fails to correctly assess the correlation between the two cross-modal data.

To address these issues, in this paper, we propose a novel visual stimuli reconstruction method based on LSTM and GAN. Specifically, there are three components in our method to solve the challenges mentioned before. The first part is the stimuli images encoder, which is used for mapping the stimuli imag.es to a latent space through deep neural network. The second part is a LSTM network, used for fMRI feature mapping to extract time-series information from fMRI. The last part is the discriminator for stimuli image generation, which generates the images as similar as the original input images. We also employ the pairwise ranking loss [9] to encourage the similarity of ground truth caption-image pairs to be greater than that of all other negative ones.

The major contributions of this paper are two folds. First, we propose a novel method to reconstruct the visual images from the evoked fRMI data. A time-series information guided GAN method is proposed to capture the time-series information in fMRI data via LSTM network and complete the task of stimuli image reconstruction through GAN. Second, we introduce a pairwise ranking loss to measure the relationship between the stimuli images and fMRI signals. This loss function ranks the stimuli images and fMRI that ensure strongly associated (corresponding) is at the top and weakly correlated at the bottom.

2 Proposed Method

2.1 Notations

Let N be the number of images which we used in the visual stimuli task, and D denotes the dimensions of stimuli images. Suppose $X = \{x_{pq}\} \in \mathbb{R}^{N \times D}, p = 1 : N, q = 1 : D$ denotes the stimuli images. At the same time, the preprocessed fMRI time series for S subjects is denoted by $Y = \{y_{mn}\} \in \mathbb{R}^{T_f \times V}, m = 1 : T_f, n = 1 : V$, where T_f is the number of time points in units of TRs (Time of Repetition), V is the number of voxels, and y_{mn} denotes the functional activity

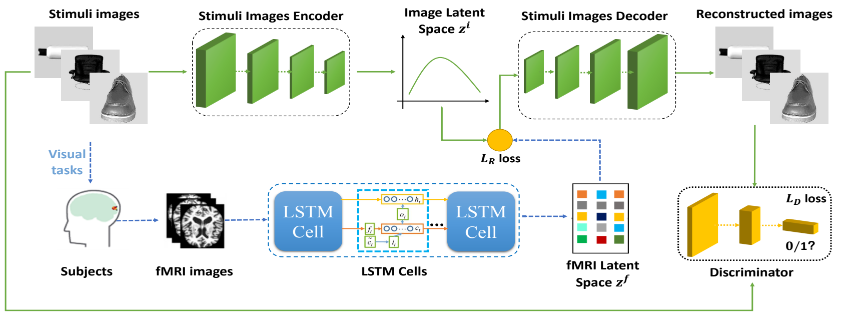

Fig. 1. The schematic diagram of time-series information guided GAN method

for the subject in the m-th time point and the n-th voxel. As proposed method is a cross-modal data reconstruction task, the samples are pairwise, which is saying that the number of the samples is T, and $T = N = T_f$. Here, for convenience, we let (x_t, y_t) be a pairwise sample at time point $t, t = 1, 2, \ldots, T$.

2.2 Time-series Information Guided GAN

We develop a time-series information guided GAN method for modeling the relationships between the visual stimuli (images) and the evoked fMRI activity patterns. Our method generates two different modals from a shared latent space, via the two-view specific generative model and a discriminative model. The schematic diagram of the proposed method is shown in Fig.1. There are three sub-networks in the proposed model, i.e., 1) a image encoder for mapping the stimuli images into latent space, 2) a LSTM generator for fMRI feature mapping, and 3) a discriminator for image reconstruction.

Stimuli Images Autoencoder: Due to only a small number of sample can be used to train the network for stimuli image reconstruction task, we pretrain an autoencoder to improve the performance. The pretrained encoder network is employed to map the stimuli images into a latent representation. Herein, in the cross-modal reconstruction model, the image encoder can map the features of visual stimuli into the image latent space z^i, where the latent feature $z^i_t = f_\theta(x_t)$. Here $f(\cdot)$ is the encoder function, θ is the parameters in the encoder. While the decoder network reconstructs the original input image $\hat{x}_t = g_\phi(z^i_t)$ by using the nonlinear function $g(\cdot)$, where ϕ is the parameters in the decoder. The loss function of the autoencoder can be defined as follows:

$$\min_{\theta,\phi} \frac{1}{T} \sum_{t=1}^{T} \| x_t - g_\phi(f_\theta(x_t)) \|_F^2 \, , \tag{1}$$

LSTM Network for fMRI Feature Mapping: The fMRI generator produces an output image $\hat{y}_t$ given the corresponding neural response in sequential order

$y_t, t \in 1 \cdots T$. Here, the generated image $\hat{y_t}$ is as similar as possible to the reconstructed image in the next step $y_{\hat{t+1}}$. Therefore, the generator should be a sequential LSTM model, which produces the sequentially next image, $\hat{y_t} = \mathcal{L}(y_1, y_2, \cdots, y_t)$, $t = 1, 2, \ldots, T$. The LSTM network maps the fMRI signals into the fMRI latent space z^f, where the latent feature $\hat{y_t} = \mathcal{L}(y_t)$, $t = 1, 2, \ldots, T$. Here, $\mathcal{L}(\cdot)$ defines the LSTM network mapping.

Discriminator for Stimuli Image Generation: We use two loss components to compute the loss between $G(z_t^i)$ and x_t on the basis of features from the trained deep neural networks. For the first component, which we refer to [12], is feature reconstruction loss $\mathcal{L}_f$, which determines whether features are activated above a threshold at all. The feature activation matrices for one fully connected layer $\mathcal{F}$, denoted $\phi_{F(x_t)}$ and $\phi_{G(z_t^i)}$ are transformed to binary representations and by using the threshold α. Here, we follow the same setting as in [12], that $\alpha = 1.0$.

The feature loss $\mathcal{L}_f$ then can be determined as follows:

$$\max_{z_i} -\frac{1}{T}\sum_{t=1}^{T}\sum_{F} \phi_{F,b} x_t \log(\phi_{F,b}(G(z_t^i))) - (1 - \phi_{F,b} x_t) \log(1 - \phi_{F,b}(G(z_t^i))) \quad (2)$$

The second component of the losses is the discriminator loss. Here we use the original discriminator loss from GAN. The discriminator discriminates the real sample. Here, to make the discriminative result close to 1, we let the generated image $\hat{x_t}$ close to $\log D(x)$ to fool the discriminator. $\mathcal{L}_d$ can be defined as follows:

$$\max_{D} V(D, G) = E_{x\sim p_{data}(x)}[\log(D(x))] + E_{z\sim p_z(z)}[\log(D(1 - D(G(z))))] \quad (3)$$

The hybrid loss function $\mathcal{L}_D$ combine the two loss components as below:

$$\mathcal{L}_D = \lambda_f \mathcal{L}_f + \lambda_d \mathcal{L}_d, \quad (4)$$

where λ_f, λ_d are hyper-parameters to balance the effects of the two loss components.

2.3 Ranking Loss for Cross-Modal Data Fusion

One of the most significant challenges for reconstructing the visual stimuli is how to model the relationship between the stimuli images and the evoked fMRI scans. Inspired by [9], we develop the rank loss from the image-textual reveral to visual stimuli reconstruction field for measuring the relationship between the two cross-madal data. We denote $(\hat{x}_t, \hat{y}_t)$ as the pairwise image-fMRI sample at time point t, which generated from the two specific generators of cross-modal data. We further denote the non-corresponding samples by using $\hat{x}'_t$ and $\hat{y}'_t$, where $\hat{x_t}'$ goes over stimuli images independent of $\hat{y_t}$, and $\hat{y_t}'$ goes over brain activities not evoked by $\hat{x_t}$. The objective function ensures that the groundtruth image-activity pairs at the top and weakly related ones at the bottom. Therefore, we optimize the ranking loss below:

$$\mathcal{L}_R = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{Rank}(\hat{x_t}, \hat{y_t}), \tag{5}$$

where the single pairwise sample ranking loss $\mathcal{L}_{Rank}$ is defined as follows:

$$\mathcal{L}_{Rank} = \sum_{\hat{x_t}'}[\alpha - s(\hat{x_t}, \hat{y_t}) + s(\hat{x_t}', \hat{y_t})]_+ + \sum_{\hat{y_t}'}[\alpha - s(\hat{x_t}, \hat{y_t}) + s(\hat{x_t}, \hat{y_t}')]_+, \tag{6}$$

where α is a margin, $s(\hat{x_t}, \hat{y_t}) = - \parallel (\max(0, \hat{x_t} - \hat{y_t})) \parallel^2$ is the order-violation penalty used as a similarity. Futher, $[x]_+$ represents $\max(x, 0)$.

The overall loss function is then given as follows:

$$\mathcal{L}_{loss} = \lambda_D\mathcal{L}_D + \lambda_R\mathcal{L}_R, \tag{7}$$

where λ_D, λ_R are hyper-parameters to balance the effects of the two loss functions. We randomly choose all the parameters from $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. The values were determined via optimizing on the training set. Optimization specifically for each dataset may improve the results further.

3 Experimental Results

Datasets: In this paper, we employ two publicly available datasets to validate the proposed method, including, a) Open NEURO[1] dataset, and b) Handwritten digits dataset. For the Open NEURO dataset, we select the dataset numbered DS105 [2] in the Open NEURO Project. In DS105, 6 subjects were stimulated with grayscale images in 8 categories, and each subject underwent 12 runs of experiments. Among them, subject No.5 miss one run of data record, with only 11 runs of data. In this paper, we use a leave-one-out cross validation strategy to adjust the parameters and evaluate the effectiveness of the method we propose. In each phase, data from five subjects are used for training, while data from one subject is used during the test stage.

For the handwritten digits dataset, we use the same dataset as the experiment in [6]. There are one hundred gray-scale handwritten digit images (50 of digital "6" and the equal numbers of digital "9") in the dataset. The image resolution is 28×28. A 10-fold cross validation was performed (i.e. each category contained 45 training data and 5 test data for each experiment).

Experiment Settings: The proposed method is compared with four well-known methods, including 1) Bayesian canonical correlation analysis (BCCA)[10]: A multiview linear generative model designed for neural encoding and decoding. 2) Deep canonically correlated autoencoder (DCCAE)[11]: A cross-view representation model to learn the deep representations from multiview data. 3) Deep

[1] http://openneuro.org

Table 1. Performances of compared methods on the *DS*105 dataset.

Model	Euc_dis↓	p-value	PCC↑	p-value	MSE↓	p-value
BCCA	0.787±0.153	1.3839*e*-14	0.561±0.159	9.8467*e*-11	0.208±0.062	8.2238*e*-9
DCCAE	0.751±0.196	1.6552*e*-10	0.584±0.193	8.5767*e*-10	0.171±0.104	2.0229*e*-7
DGMM	0.652±0.122	2.4369*e*-7	0.636±0.146	2.8228*e*-7	0.124±0.069	3.4167*e*-4
DCGAN	0.641±0.089	7.9318*e*-5	0.651±0.096	8.0966*e*-6	0.116±0.074	0.0055
Proposed	**0.609±0.061**	——	**0.689±0.063**	——	**0.091±0.051**	——

Table 2. Performances of compared methods on the handwritten digits dataset.

Model	Euc_dis↓	p-value	PCC↑	p-value	MSE↓	p-value
BCCA	0.679±0.155	1.1709*e*-10	0.423±0.139	1.6853*e*-22	0.119±0.023	1.8554*e*-20
DCCAE	0.631±0.064	9.5486*e*-9	0.529±0.047	5.0496*e*-20	0.077±0.018	3.3630*e*-11
DGMM	0.585±0.061	0.0025	0.801±0.061	0.0291	0.037±0.019	0.004
DCGAN	0.581±0.055	0.0238	0.799±0.057	0.0163	0.038±0.022	0.0096
Proposed	**0.568±0.037**	——	**0.812±0.059**	——	**0.033±0.015**	——

generative multiview model (DGMM)[6]: A deep generative multi-view learning model for reconstructing the perceive images from brain fMRI activities. 4) Deep convolutional generative adversarial network (DCGAN)[12]: A GAN framework used to generate arbitrary images from the stimuli domain (i.e., handwritten characters or natural gray scale images).

Three evaluation metrics are used to measure the reconstruction performance of different methods, including 1) Euclidean distance (Euc_dis), the smaller the value is, the more similar reconstructed result we obtained. 2) Pearson's correlation coefficient (PCC), which shows the correlation between the original and reconstructed images. 3) Mean squared error (MSE), which calculates the pixel-level error between the reconstructed image and the original image. The smaller the error, the more similar the reconstructed image is to the real image.

Quantitative Analysis: Performances of compared methods on two datasets were listed in Table 1 and 2. Table 1 shows the experimental results of dataset DS105, several observations can be drawn as follows. First, the proposed method obtains a considerably better performance compared with the other methods. Second, by comparing the proposed method with BCCA, a linear model for stimuli reconstruction, we can see that our method is always out-perform BCCA. These results show that our reconstruction method with deep network is better than linear model by extracting nonlinear features from visual images and fitting images. Third, compared with DCCAE, the proposed method shows significantly better performance. As a nonlinear cross reconstruction model, DCCAE achieves better performance than BCCA, but compared with our method, there is a lack of time-series information mining. Fourth, the performance of DGMM is moderate on both of the two datasets. This may be caused by the performance gap between DGMM's deep network model and GAN's generative discriminant model. The last but not the least, compared with DCGAN, LSTM network in

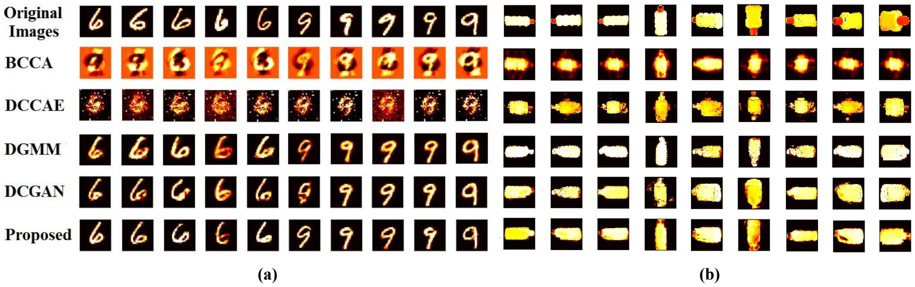

Fig. 2. Image reconstruction results of different methods on two datasets. (a) Handwritten digits dataset (b) $DS105$ dataset (category=bottles).

our method plays an important role in mining the correlation between the stimuli images and the brain activity patterns.

For the handwritten digits dataset, the results are shown in Table 2. The quantitative results on the three evaluation metrics are also at the best level. For the three campared methods of BCCA, DCCAE and DGMM, we refer to the experimental settings in [6], and also refer to their experimental results on MSE. And for Euc_dis and PCC here, we obtained similar results as which on DS105 dataset. The reason is as analyzed above. Compared with DCGAN, our method also takes the better results because of the use of LSTM network and the cross-modal ranking loss. In addition, p-values are also displayed in the tables to verify the significance of our experimental results.

Qualitative Analysis: The reconstructed results on two different datasets are shown in Fig.2(a)-(b), respectively. In each figure, the top row shows the presented visual images, while the following rows show the reconstructed results of all compared methods.

In Fig.2(a), the reconstructed handwritten digits are very similar to the original images. Compared with our method, the performances of BCCA and DCCAE are not acceptable. The complex noises often influence the their reconstruction results and the results also lack of the basic features in the original images. Furthermore, the reconstruction results of DGMM and DCGAN are coarse too. Although their results are better than those of BCCA and DCCAE, they lost some information in details compared with our method, because they did not take the time-series information into account. The reconstructing results of DS105 (categories = “bottles”) are shown in Fig.2(b). As can be seen from the figure, our method produces better reconstruction results than the compared methods. Fig.2(b) also indicates that the effect of our method is obviously better than other methods on the reconstruction of natural images. In particular, BCCA and DCCAE cannot provide acceptable performance in characterizing detailed contours, which may be related to their mapping capabilities. DGMM and DCGAN are better than the first two methods, but they are not as good as our method when describing image details, such as color.

4 Conclusion

In this paper, we present a time-series information guided GAN method for perceived image reconstruction from the human brain activities. Our method is not only a generative model to model the relationship between the stimuli image and the evoked brain activities, but also take the time-series information of fMRI data into account. Furthermore, the pairwise ranking loss is introduced to measure the relationship between the stimuli images and the corresponding fMRI data, which ensures that the strongly associated pairs is at the top and the weakly related ones is at the bottom. Our reconstruction model can also achieve better performance in comparison with state-of-the-art reconstruction methods on both of the two publicly available fMRI datasets.

References

1. K. Smith, "Brain decoding: reading minds," *Nature News*, vol. 502, no. 7472, p. 428, 2013.
2. J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.
3. K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, p. 352, 2008.
4. R. J. Brachman and J. G. Schmolze, "An overview of the KL-ONE knowledge representation system," *Cognitive Science*, vol. 9, no. 2, pp. 171–216, April–June 1985.
5. J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.
6. C. Du, C. Du, L. Huang, and H. He, "Reconstructing perceived images from human brain activities with bayesian deep multiview learning," *IEEE transactions on neural networks and learning systems*, 2018.
7. Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven, "Reconstructing perceived faces from brain activations with deep adversarial neural decoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 4246–4257.
8. G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Frontiers in Computational Neuroscience*, vol. 13, 2019.
9. J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.
10. Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, "Modular encoding and decoding models derived from bayesian canonical correlation analysis," *Neural computation*, vol. 25, no. 4, pp. 979–1005, 2013.
11. W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.
12. K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, "Generative adversarial networks for reconstructing natural images from brain activity," *NeuroImage*, vol. 181, pp. 775–785, 2018.