2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS) | 979-8-3503-7333-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/SCISISIS61014.2024.10759955

A Study on Brain Decoding of Image Stimuli Using a Diffusion Model

Fumi Ishizaki
Graduate School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
g2020505@is.ocha.ac.jp

Ichiro Kobayashi
Graduate School of Humanities and Sciences
Ochanomizu University
Tokyo, Japan
koba@is.ocha.ac.jp

***Abstract*—Humans recognize the external world by processing information received by the eyes and other sensory organs in the brain. Understanding how the human brain processes complex information from the outside world is expected to improve the performance of image and speech recognition technologies, which have made remarkable progress in recent years. In this study, we focus on brain activity decoding of visual experience, aim to read what humans are looking at by predicting image features from brain activity data, and further attempt to develop a method to output high-definition and semantically valid images by generating images from predicted features using a diffusion model. As a result, similarity to the stimulus image was confirmed in the image generated by the training data, but the evaluation data confirmed that there is still room for further study.**

***Index Terms*—Brain decoding, diffusion models, image stimulation**

I. Introduction

In recent years, a number of studies in artificial intelligence research have used deep neural network (DNN) to represent visual experience from human brain activity [1]–[4], showing homology between the hierarchical representation of convolutional neural network (CNN) models processing visual information in DNN and the hierarchical representation of visual information in the brain [5]. In [1], [3], a decoding model [6] was constructed to predict the latent state represented in each hierarchy of the CNN that processes the stimulus image from human brain activity data with visual stimuli, and the visual experience in the human brain was generated as an image. In addition, in [4], a Stable Diffusion model [7] that generates images from text, and in [2], a Versatile Diffusion model [8] using a Very Deep Variational Autoencoder for even lower-order reconstruction is used to reconstructs visual images in the human brain as generating high resolution images by the diffusion model's image generation capability.

In this study, we attempt to decode human visual experience by predicting image features from brain activity data, focusing on brain activity decoding related to visual experience in the same way as these studies. In particular, we attempt to develop a method for outputting high-definition and semantically valid images by utilising the image reproduction function using the Stable Diffusion Model and generating images based on predicted features regressed from human brain activity.

II. Method

An overview of the method proposed in this study for decoding brain activity using a diffusion model is shown in Figure 1. An neural decoding model is constructed to predict the features of the stimulus image from the brain activity data acquired while viewing the image. The neural decoding model learns weights using linear regression so that the predicted image feature values and the stimulus image feature values approach each other. The stimulus image is then reconstructed from the predicted image features using a diffusion model, Stable Diffusion [7].

A. Neural Decoding

Human brain activity patterns change in response to the content of the cognitive state, behaviour and other stimuli experienced. Therefore, it is thought that by decoding brain activity, it is possible to read what humans are actually receiving. The neural model [9] is a model that estimates what humans are experiencing when specific brain activities are measured [10].

In the present study, correspondence was modelled by predicting the content of stimulus images from brain activity.

B. Diffusion Model

Diffusion models are a type of generativemodel [11]. In general, the transformation from noise to data is difficult, but it can be easily achieved by considering it as the inverse transformation of the transformation from data to noise. The data are generated by the denoising process. No learning is required for the diffusion process, and learning is required only when generating data in the denoising process.

As shown in the Figure 2, the diffusion process is a Markov process in which Gaussian noise is gradually added to the data as time progresses from 0 to T. If the data at time t is x_t, the noise following the standard normal distribution added at time t is ϵ_t, and the parameter that determines the strength of the noise addition is $\beta_t \quad (0 < \beta_t < 1)$, the noise addition from time $t-1$ to time t addition from time $t-1$ to time t is defined as follows.

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{1}$$

$$q(x_t|x_{t-1}) = N(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

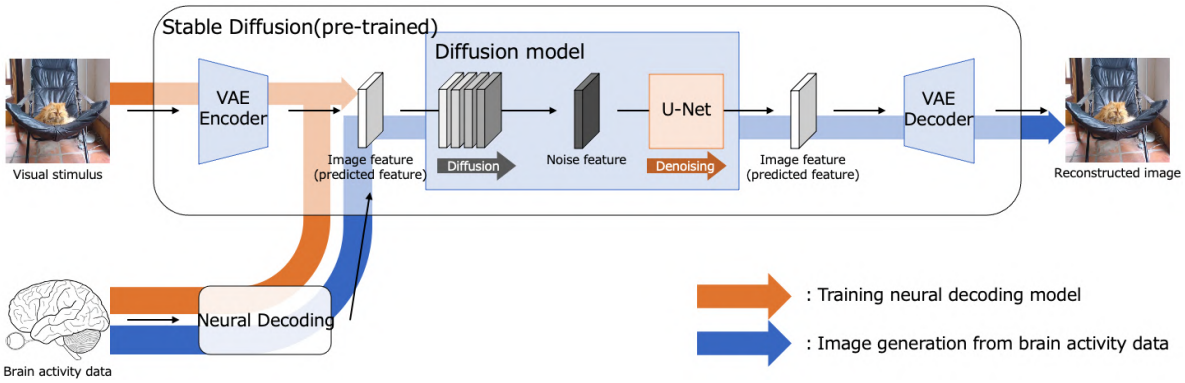

Fig. 1. Schematic diagram of this study: regression prediction of the latent state of VAE of Stable Diffusion to generate images from brain activity states using an neural decoding model.

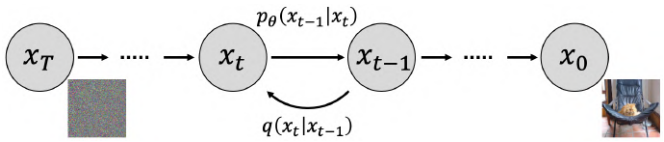

Fig. 2. Diffusion process

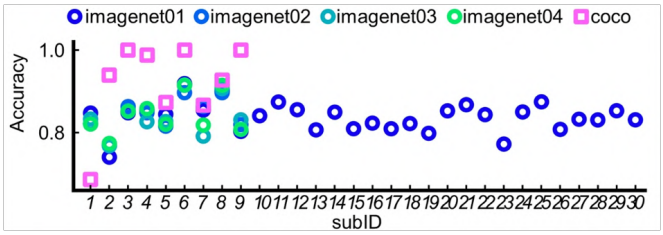

Fig. 3. "The mean recognition accuracy of each subject in each session of ImageNet and COCO experiments," from [12]

It is known that the denoising process, in which noise is gradually removed from x_T to finally become data x_0, can be represented by a normal distribution if the diffusion process is defined by a sufficiently small Gaussian noise load as in Equation 1. When estimating the mean and variance in the model, let θ be the parameter and $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ be the estimated mean and variance, respectively, the distribution of x_{t-1} given x_t is as follows.

$$p_\theta(x_{t-1}|x_t) = N(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Therefore, x_{t-1} is obtained by learning a model to estimate the meanμ_θ and varianceΣ_θ at time t and sampling from a normal distribution. This sampling can be repeated to obtain the final data x_0.

This research is based on the Latent Diffusion Model (LDM) [7], which enables image generation, such as processing part of the image generation using learned weights[1], and uses a Variational Autoencoder (VAE) to convert pixel images into a latent embedded representation, which is then used to generate a low-dimensional latent space. Efficient learning is possible by using a 3D latent space.

III. Experiments

A. Implementation Details

1) Dataset: We used the Natural Object Dataset (NOD) [12], a large fMRI dataset containing 57,120 ImageNet [13] and MS-COCO [14] natural image responses from 30 subjects. The NOD is a large fMRI dataset containing 57,120 ImageNet and MS-COCO natural image responses from 30 subjects.

[1]provided at https://huggingface.co/CompVis/stable-diffusion-v-1-4-original was used.

TABLE I
Experimental Data Setting

Subject Number	6	6&8	1&3–9	1–9
Number of training pictures	3,000	6,000	23,000	26,000
Number of validation pictures	1,000	2,000	8,000	9,000

In the present study, data from nine subjects who participated in all experiments were used for trials with ImageNet images (36,000 images in total). The average recognition accuracy of the most salient object presented in the images across subjects was 83.7% for whether it was an organism or not, with all subjects except sub-02 performing equally well (Figure 3).

2) Feature Extraction: The image features used when training the neural decoding model are extracted from the latent layer of the pre-trained VAE [15] Encoder [16] within Stable Diffusion.

3) Construction of Decoding Model: An encoding model was constructed using brain activity data and image features extracted by VAE Encoder. A model that predicts the time series of image features using the time series of brain activity as an explanatory variable was trained by ridge regression[2]. The model was regressed on the brain activity data observed by fMRI and the features 2, 4 and 6 seconds before the time series, taking into account the reaction time (hemodynamic response) of the increase in blood flow associated with the neural activity (Figure 4). A 10-segment cross-validation was

[2]Ridge regression was used as provided at https://github.com/alexhuth/ridge

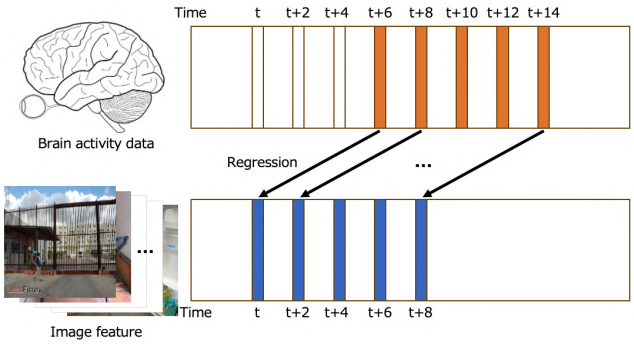

Fig. 4. This hemodynamic response results in brain activity data corresponding to the image presented at time t being observed a few seconds later. This figure shows the state of the paired data with the brain activity data after 6 seconds.

performed by shuffling the training data with 100 chunks, and the regularisation term with the best average correlation coefficient was adopted.

Of the data from the nine subjects, only the data from sub-06, which had the highest average recognition accuracy, was used. Additionally, data from sub-08, which showed similar recognition accuracy, was used for two subjects. Data from eight subjects, excluding sub-02, which had low recognition accuracy, was used for the nine subjects. A total of 12 nueral decoding models were constructed using the data of 9 persons each.

4) Image Generation: Images were generated from predicted image features obtained by the neural decoding model using the Stable Diffusion pre-trained diffusion model and the VAE Decoder.

B. Results

Examples of images generated from brain activity data are shown below. Pearson's product-rate correlation coefficients were calculated using the image features of the stimulus images and the predicted image features, and the prediction accuracy of each model was evaluated accordingly.

Figure 5 shows examples of images generated using the brain activity of each validation data. It was confirmed that the images became blurred as the number of subjects increased. Table II shows that the correlation coefficient is less than 0.1 for any of the validation data, indicating that there is no correlation.

Figure 6 shows examples of images generated using the brain activity of each training data set. It was found that the results with a smaller number of subjects produced images that better represented the general shape, and that the stimulus images and the generated images were similar. The results of the correlation coefficients are shown in Table II: a strong positive correlation of 0.88 was observed for the training data of sub-06 only, and 0.75 for sub-06 and sub-08. The correlation coefficients were also 0.45 and 0.44 when training on data from all but sub-02 and the nine subjects, indicating a high degree of similarity between the stimulus image and the generated image in the correlation coefficients.

Fig. 5. Image generated from the evaluation data and the image being viewed at the time. Figure shows an example of a regression with features from 6 seconds ago.

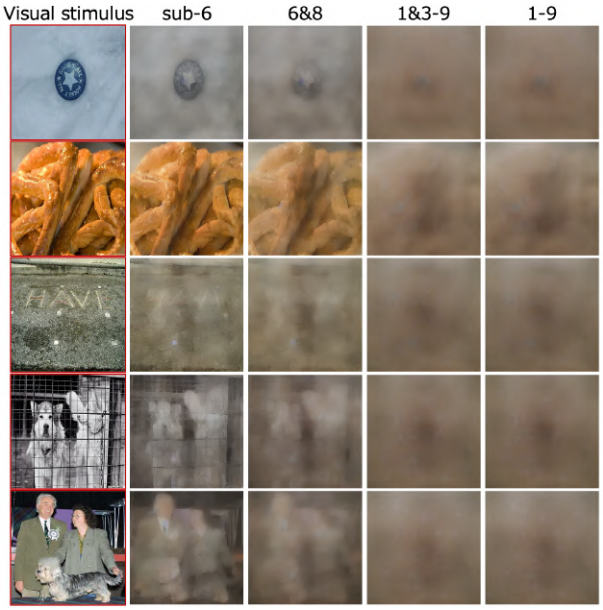

Fig. 6. Images generated from the training data and the image being viewed at the time. Figure shows an example of a regression with features from 6 seconds ago.

TABLE II
PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT OF THE PROPOSED PREDICTION MODEL.

Subject Number	6	6&8	1&3–9	1–9
training data	0.88	0.75	0.45	0.44
validation data	0.017	0.020	0.027	0.027

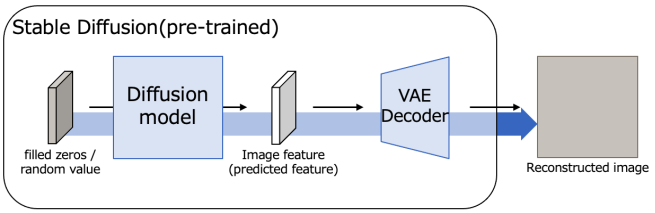

Fig. 7. Schematic diagram of additional experiments: image generation was performed using each data instead of image features predicted from brain activity data.

Fig. 8. Image output generated using data where all dimensions are filled with zeros.

Fig. 9. Image output generated using data where all dimensions are filled with random values.

C. Discussion

The fact that the accuracy worsens as the number of subjects increases is thought to be mainly due to individual differences in brain activity data. Since there are differences in ROI in the brain for each subject, it is necessary to fill in the differences for each individual. In qualitative evaluation, it has been found that color is more likely to be perceived as more expressive [3]. In this study, many of the outputs were similar in hue, suggesting that shape is easier to represent than color.

An additional experiment (Figure 7) was conducted to confirm that the image generation was not dependent on the generation capability of the diffusion model, but referred to the brain activity data. Instead of image features predicted from brain activity data, image generation was performed using data in which all dimensions were filled with zeros and data in which all dimensions were filled with random values.

The output results of an additional experiment are shown in Figures 8 and 9, respectively. Almost the each same outputs were observed for all the data as shown in Figures 8 and 9, indicating that the outputs in Figures 5 and 6 are related to the brain activity data. The above confirmed that the proposed model generates images with reference to the brain activity data.

IV. Conclusion

In this study, an neural decoding model was constructed to predict the image features of the stimulus image from the brain activity, and image reconstruction was performed from the brain activity using Stable Diffusion, which enables the generation of high-resolution images.

As a result, similarity to the stimulus image was confirmed in the image generated by the training data, but it was difficult to say that the evaluation data generated an image similar to the stimulus image. In the future, we would like to make use of brain activity features as conditions in the denoising process, so that highly accurate image reconstruction can be carried out even with evaluation data.

References

[1] Horikawa Tomoyasu and Kamitani Yukiyasu. “Generic decoding of seen and imagined objects using hierarchical visual features,” in Nature Communications, vol. 8, No. 1, p. 15037, 2017.
[2] Furkan Ozcelik and Rufin VanRullen. “Natural scene reconstruction from fMRI signals using generative latent diffusion,” 2023.
[3] Shen Guohua, Horikawa Tomoyasu, Majima Kei and Kamitani Yukiyasu. “Deep image reconstruction from human brain activity,” in PLOS Computational Biology: Public Library of Science, vol. 15, No. 1, pp. 1–23, January 2019.
[4] Takagi Yu and Nishimoto Shinji. “High-resolution image reconstruction with latent diffusion models from human brain activity,” in bioRxiv: Cold Spring Harbor Laboratory, 2022.
[5] Yamins Daniel L. K., Hong Ha, Cadieu Charles F., Solomon Ethan A., Seibert Darren and DiCarlo James J. “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” in Proceedings of the National Academy of Sciences: National Academy of Sciences, vol. 111, No. 23, pp. 8619–8624, 2014.
[6] Thomas Naselaris, Kay Kendrick N., Shinji Nishimoto and Gallant Jack L.. “Encoding and decoding in fMRI,” in NeuroImage: Academic Press Inc., vol. 56, No. 2, pp. 400–410, 15 May 2011.
[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models,” in CoRR, 2021.
[8] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang and Humphrey Shi. “Versatile Diffusion: Text, Images and Variations All in One Diffusion Model,” 2024.
[9] Nishida Shinji and Nishida Satoshi. “Modeling and Decoding of Visual and Cognitive Brains,” in Journal of NICT: National Institute of Information and Communications Technology, vol. 64, No. 1, pp. 5–11, 2018.
[10] Miyawaki Yoichi and Kamitani Yukiyasu. “#8: Neural Decoding and Its Application,” in Journal of the Society of Instrument and Control Engineers: The Society of Instrument and Control Engineers, vol. 50, No. 10, pp. 888–894, October 2011.
[11] Ishii Masato. “Imadokino diffusion model Behind the magic behind the scenes that changed the world of image generation.,” in Computer Vision Saizensen 2023 Summer: Kyoritsu Shuppan Co., Ltd., pp. 9–41 2023.
[12] Gong Zhengxin, Zhou Ming, Dai Yuxuan, Wen Yushan, Liu Youyi and Zhen Zonglei. “A large-scale fMRI dataset for the visual processing of naturalistic scenes,” in Scientific Data, vol. 10, No. 1, p. 559, 2023.
[13] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai and Fei-Fei Li. “Imagenet: A large-scale hierarchical image database,” pp. 248–255, 2009, [2009 IEEE conference on computer vision and pattern recognition].
[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick and Piotr Dollár. “Microsoft COCO: Common Objects in Context,” 2015.
[15] Kingma Diederik P. and Welling Max. “Auto-Encoding Variational Bayes,” 2014 [2nd International Conference on Learning Representations. ICLR 2014, Banff, AB, Canada, Conference Track Proceedings, April 14-16, 2014].
[16] Patrick Esser, Robin Rombach and Björn Ommer. “Taming Transformers for High-Resolution Image Synthesis,” 2020.