Describing Semantic Representations of Brain Activity Evoked by Visual Stimuli

Eri Matsuo
Ichiro Kobayashi
Ochanomizu University
2-1-1 Ohtsuka, Bunkyo-ku,
Tokyo 112-8610, Japan.
g1220535@is.ocha.ac.jp
koba@is.ocha.ac.jp

Shinji Nishimoto
Satoshi Nishida
National Institute of Information
andCommunicationsTechnology
1-4, Yamadaoka, Suita-shi,
Osaka, 565-0871, Japan
nishimoto@nict.go.jp
s-nishida@nict.go.jp

Hideki Asoh
National Institute of Advanced
Industrial Science and Technology
2-3-26, Aomi, Koto-ku,
Tokyo, 135-0064, Japan
h.asoh@aist.go.jp

1 Introduction

Quantitative analysis of semantic activities in the human brain is an area of active study. With the development of machine learning methods and the application of such methods to natural language processing, many studies have attempted to interpret and represent brain activity with the semantics categories of words (Mitchell et al., 2008; Huth et al., 2012; Pereira et al., 2013; Stansbury et al., 2013; Horikawa et al., 2013; Nishida and Nishimoto, 2017). However, the previous studies examined word-level representation, and little is known about whether we could recover structured sentences from brain activity. In this paper, we propose a deep learning-based decoding method in which natural language descriptions of semantic contents in visual scenes are generated from scene-evoked human brain activity measured by functional magnetic resonance imaging (fMRI). In general, deep neural networks require a large amount of training data. However, assembling a large-scale brain activity dataset is difficult because observing brain activity data with fMRI is expensive and each human brain is different in its size and shape. To handle this problem, instead of training the whole model from scratch, we propose a model that associates the image features of the intermediate layer of a pre-trained caption-generation system with brain activity, which makes it possible to generate natural language descriptions of the semantic representation of brain activity. We train three regression models, i.e., ridge regression, three-layer neural networks, and five-layer deep neural networks (DNN), that learn the relationship between brain activity and deep-layer image features. The results demonstrate that the proposed model can decode brain activity and generate descriptions using natural language sentences.

2 Proposed Method

This study aims to generate natural language sentences that describe what a human being calls to mind using brain activity data observed by fMRI as input information. We combined a image $\rightarrow$ caption model and a brain activity data $\rightarrow$ image feature model (Sections 2.1 and 2.2). Figure 1 presents an overview of the proposed method.

2.1 Image $\rightarrow$ caption model (A)

We employed an image-captioning model (A) based on a DNN framework, i.e., the encoder-decoder (enc-dec) network (Cho et al., 2015; Vinyals et al., 2015), as the main component of the proposed model. In the enc-dec framework, by combining two DNN models functioning as an encoder and a decoder, the model encodes input information as an intermediate expression and decodes the information as an expression in a different modality. Generally, previous studies of image-captioning systems have proposed enc-dec models that combine two DNNs: one DNN extracts image features using a convolutional neural network and the other generates captions using a LSTM with the image features which correspond to an intermediate expression of the model Vinyals et al. (2015);

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

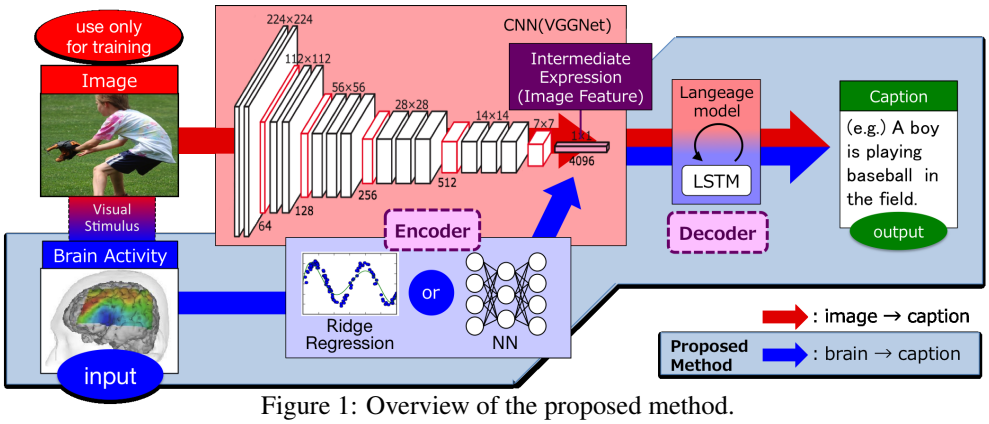

Figure 1: Overview of the proposed method.

Xu et al. (2015); Yao et al. (2016). Similar to such previous models, we constructed an image → image feature → caption model (A) employing VGGNet (Simonyan and Zisserman, 2015) as an encoder and a two-layer LSTM language model (LSTM-LM) (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014) as a decoder. We used pairs of image and caption data as training data.

2.2 Brain activity data → image feature model (B)

To apply the above image-captioning process to handle brain activity data rather than images, we constructed a model that predicts features extracted by VGGNet from images that evoke visual stimuli in the brain using fMRI brain activity data as input. In other words, the model encodes brain activity data into the intermediate expression in the image → caption model (A). We implemented and compared three models, i.e., a ridge regression model, a three-layer neural network model, and a five-layer DNN model, to determine which machine learning method is suitable for this model. The five-layer DNN model was pre-trained using stacking autoencoders (Bengio et al., 2006) to avoid delay and overfitting in training due to the lack of brain data. We used pairs of fMRI brain activity data and the images a subject observed as training data.

2.3 Process Flow

The process of the proposed method is as follows.

Step 1. *Encode brain activity as an intermediate expression.*
Model (B) predicts the feature of the image a subject watches from the input brain activity data evoked by the visual stimuli. In the followings step, the features are provided to model (A) and processed as the intermediate expression.

Step 2. *Word estimation by the LSTM-LM.*
The LSTM-LM decoder of model (A) predicts the next word from the image feature produced in Step 1 and the hidden states of LSTM at the previous time step.

Step 3. *Caption generation by iterative word estimation.*
A caption is generated by sequentially estimating words by repeating Step 2 until either the length of the sentence exceeds a predefined maximum or the terminal symbol of a sentence is output.

As mentioned above, we construct a brain activity data → image feature → caption model (C) by training the brain activity data → image feature model (B) and the image → image feature → caption model (A) individually and execute them sequentially in the prediction phase. Note that model (C) uses only fMRI brain activity data as input, i.e., without images.

3 Experiments

3.1 Experiment (A): image → caption model

3.1.1 Experimental settings

Microsoft COCO[1], which includes 414,113 pairs of images and their captions, was used as the dataset for the experiments. The hyper-parameters of the models used in the experiments were set based on previous studies: Vinyals et al. (2015); Mitchell et al. (2016); Cho et al. (2014).

[1] http://mscoco.org/

 A man is surfing in the ocean on his surfboard.  A black and white cat is sitting on the toilet.

Figure 2: Captions for randomly selected images

3.1.2 Results & Discussion

We confirmed the process of learning by the convergence of the perplexity of output sentences recorded for each epoch. Figure 2 shows the natural language descriptions for two images randomly selected from test images.

In the first example, a considerably reasonable natural language description was generated. In the second example, appropriate expressions for the subject of the generated sentence, i.e., a cat, and its color were selected. Reasonable captions were generated for the test images and the perplexity converged near 2.5; therefore, an appropriate model was built to generate natural language descriptions from the images.

3.2 Experiment (B): brain $\rightarrow$ image feature model

3.2.1 Experimental settings

As for the learning dataset for the corresponding relationships between brain activity and image features, we employed the brain activity data of a subject stimulated by natural movies (Nishimoto et al., 2011), i.e., the BOLD signal observed by fMRI, and still pictures taken from the movies provided as visual stimuli, which were synchronized with the brain activity data. In the natural movies, there are various kinds of movies about natural phenomenon, artifacts, humans, films, 3D animations, etc., whose length of time are a few tens of seconds. As the input data, we employed 65,665 voxels corresponding to the cerebral cortex part among $96 \times 96 \times 72$ voxels observed by fMRI,then learned the corresponding relationships between the brain data and the image features, whose dimensionality is 4,096, extracted from the image using VGGNet. We used 4,500 samples as training data (recorded every two seconds for 9,000 seconds), which is a small number for learning a DNN. We omit details of the learning settings.

3.2.2 Results & Discussion

We recorded the mean squared error (MSE) for each epoch and confirmed that the MSEs of the three models converged. For evaluation, we conducted an experiment to retrieve the images, which have similar image features to those estimated from brain activity data, from 82,783 images of the Microsoft COCO training dataset with MSE. Figure 3 shows the result of retrieving similar images.

As for ridge regression and three-layer neural network, those models retrieved proper images from most training data, so we confirmed that the models could extract proper image features from brain activity data. However, as for five-layer DNN, the same unrelated images were retrieved for all input brain activity data, and the results for the test samples were worse than those for training samples even when employing ridge regression or the three-layer neural network although the MSEs for the test dataset converged. The reason for this is that the input dimension, i.e., 65,665, was much larger than that of the parameters to be learned, and the number of training data samples, i.e., 4,500, was small. As a result, overfitting occurred due to the lack of adjustment of hyper-parameters.

3.3 Experiment (C): brain $\rightarrow$ caption model

3.3.1 Experimental settings

We built a model that generates a natural language description from brain activity data by combining the model in Experiment (A) and the three models in Experiment (B). We then generated descriptions based on the three methods, i.e., ridge regression, the three-layer neural network, and the five-layer DNN. In addition, we generated captions from the same images using model (A).

3.3.2 Results & Discussion

The natural language descriptions generated from the four brain activity data samples (i.e., two training data and two test data), and their images are shown in Figure 3.3.1. To compare the results, we also show the captions generated using the model in Experiment (A).

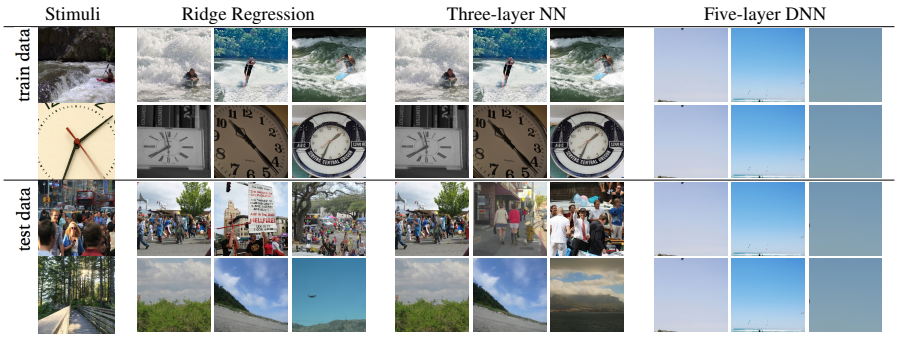

Figure 3: Exp. (B): Stimulation image and retrieved similar images (top-3)

	Stimuli	Ridge Regression	Three-layer NN	Five-layer DNN	Image → Caption Model
train data		A man is surfing in the ocean on his surf board.	A man is surfing in the ocean on his surf board.	A fire hydrant sitting on the side of an empty street.	A man is surfing in the ocean on his surf board.
		A pair of scissors sitting on the ground.	A close up of an orange and white clock.	A fire hydrant sitting on the side of an empty street.	A pair of scissors sitting on the ground.
test data		A group of people walking down the street.	A group of people standing next to each other.	A fire hydrant sitting on the side of an empty street.	A group of people standing next to each other.
		A bench sitting in the middle of an open field.	A man walking down the street with an umbrella.	A fire hydrant sitting on the side of an empty street.	A train traveling down tracks next to trees.

Figure 4: Exp. (C): Stimulation images and generated descriptions

Human understandable natural language descriptions were generated stably using only brain activity data. The descriptions generated using the brain activity data and those using the images were nearly the same for the models employing ridge regression and the three-layer neural network. Thus, we confirm that learning the corresponding relationships between brain activity data and image features was successful, and the proposed method functioned well. Taking into account of the results of Experiment (B), it was considered natural to find the same sentences were generated for all input information in the case of using the five-layer DNN model, and the test quality was low even when employing ridge regression or the three-layer neural network. Furthermore, as discussed relative to Experiment (A), the caption-generation model was learned properly, however the descriptions generated directly from the images were somewhat improper for the second test example. One of the reasons of this is the difference of quality of the images. Images in Microsoft COCO were prepared for image recognition, therefore, the content of the images was considerably understandable and describable using natural language. On the other hand, the natural movies were various types of pictures that including blurring, darkening, letters, animation, etc., i.e., images difficult to describe using natural language. Interestingly, a proper caption was generated using brain activity data with the three-layer neural network model compared to the image-captioning model for the second training example. It is unlikely that a human would confuse a clock with a pair of scissors. However, the image-captioning model made this mistake due to image processing errors in VGGNet. Thus, in this case, we assume that the image features obtained using brain activity data worked better than the features obtained directly from the images.

4 Conclusions

We have proposed a method to generate descriptions using brain activity data by employing a framework to generate captions for images using DNNs and by learning the corresponding relationships between brain activity data and the image features extracted using VGGNet. We constructed models based on three experimental settings for training methods, and we were successful in generating natural language descriptions from brain activity data evoked by visual stimuli. The quality of the descriptions was higher when using a three-layer neural network. In future, we plan to increase the amount of brain activity data, apply additional machine learning methods, i.e., Bayesian optimization, whitening, etc., and revise the hyper-parameters to increase prediction accuracy. Furthermore, we would like to investigate proper objective methods to evaluate the generated natural language descriptions.

References

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. 2006. Greedy layer-wise training of deep networks. *NIPS'06* 19:153–160.

K. Cho, A. Courville, and Y. Bengio. 2015. Describing multimedia content using attention based encoder decoder networks. *Multimedia, IEEE Transactions* 17(11):1875–1886.

K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP'14* .

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8).

T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani. 2013. Neural decoding of visual imagery during sleep. *Science* 340.

A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76(6):1210–1224.

M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daume III. 2016. Generating natural questions about an image. *ACL'16* .

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(1191).

S. Nishida and S. Nishimoto. 2017. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage* .

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* 21(19):1641–1646.

F. Pereira, M. Botvinicka, and G. Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial Intelligence* 194:240–252.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR'15* .

D. E. Stansbury, T. Naselaris, and J. L. Gallant. 2013. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* 79:1025–1034.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. *NIPS'14* .

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *CVPR'15* .

K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *ICML'15* .

L. Yao, N. Ballas, K. Cho, J. R. Smith, and Y. Bengio. 2016. Empirical upper bounds for image and video captioning. *ICLR'16 workshop* .