Analysis of Correspondence Relationship between Brain Activity and Semantic Representation

Kana Ozaki[1](ozaki.kana@is.ocha.ac.jp), Satoshi Nishida[23], (s-nishida@nict.go.jp)
Shinji Nishimoto[234] (nishimoto@nict.go.jp), Hideki Asoh[5](h.asoh@aist.go.jp), Ichiro Kobayashi[1] (koba@is.ocha.ac.jp)
[1]Graduate School of Humanities and Sciences, Ochanomizu University, Otsuka2-1-1, Bunkyo, Tokyo 112-8610, Japan
[2] Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Yamadaoka 1-4, Suita, Osaka 565-0871, Japan
[3]Graduate School of Frontier Biosciences, Osaka University, Yamadaoka 1-3, Suita 565-0871, Japan
[4]Graduate School of Medicine, Osaka University, Yamadaoka 2-2, Suita 565-0871, Japan
[5]National Institute of Advanced Industrial Science and Technology (AIST), Aomi 2-3-26, Koto, Tokyo 135-0064 Japan

Abstract

It is known that primary visual cortex uses a sparse code to efficiently represent natural scenes. Based on this fact, we built up a hypothesis that the same phenomenon happens at the higher cognitive function. Here we focus on semantic representation reflecting the meaning of words in the cerebral cortex. We applied sparse coding to the matrix consisting of paired data for both brain activity evoked by visual stimuli observed while a subject is watching a video, and distributed semantic representation made from the description of the video by means of a word2vec language model. Using this method, we obtained a dictionary matrix whose bases represent the corresponding relation between brain activity and the semantic representation. We then analyzed the characteristics of each base in the dictionary matrix. As a result, we confirmed that independent perceptual units were extracted with words representing their functional meaning.

Keywords: human perception; semantic representation; sparse coding; fMRI; dictionary learning; word embedding

Introduction

Recently, there are many studies that analyze what a person recalls while watching videos by observing his or her brain activity data via functional magnetic resonance imaging(fMRI). Huth et al. (A. G. Huth, Nishimoto, Vu, & Gallant, 2012) created a semantic map for semantic representation in the cerebral cortex by revealing the corresponding relation between brain activities with the words of WordNet (Miller, 1995) to represent the objects and motions in moving pictures. Furthermore, they systematically mapped semantic preference onto the human brain from the brain activity data obtained while subjects were listening stories(A. Huth, A. de Heer, L. Griffiths, Theunissen, & Gallant, 2016). Stansbury et al. (Stansbury, Naselaris, & Gallant, 2013) used latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) to assign semantic labels to still pictures using natural semantic descriptions synchronized with the pictures and discussed the resulting relationship between brain activity and the visual stimulus evoked by still pictures. On the base of these relationships, they created a model that classifiers brain activity into semantic categories, revealing the areas of the brain that deal with particular categories. Cukur et al. (Cukur, Nishimoto, Huth, & Gallant, 2013) estimated how a person semantically changes his or her recognition of objects from the brain activity data in cases where he or she pays attention to objects in a motion picture. Statistical models analyzing semantic representation in human brain activity have also attracted considerable attention as appropriate models for explaining higher order cognitive representations on the base of human sensory or contextual information. Furthermore, Nishida et al. (Nishida, Huth, Gallant, & Nishimoto, 2015) demonstrated that the skip-gram model, used in the framework of word2vec, as proposed by Mikolov et al.(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), is a more appropriate model than conventional statistical models used in quantitative analysis of semantic representation in human brain activity under the same experimental settings as that in the previous studies(Stansbury et al., 2013). In this paper, we refer to the representation by word2vec as semantic representation corresponding to the brain activity. First, we obtain dictionary bases by applying dictionary learning on a matrix which combines the human brain activity data obtained while subjects were watching videos and the semantic representation of the content of the video, and estimate new semantic representations corresponding to the test brain activity data using bases of the dictionary. Second, we analyzed some bases of the dictionary, and have confirmed that the independent brain recognition units were extracted with words representing their functional meaning.

Estimation of Semantic Representation from Brain Activity Data

Data

The data we used in the experiments are the brain activity data of subjects being stimulated by videos and the natural language sentences to describe their contents (Nishimoto et al., 2011). We used the brain activity data of 3 subjects (subject A, B and C) in total. Both data set of A and B contain 4500 samples for training data and 300 samples for test data, and the data set of C contains 9000 samples for training data and 600 samples for test data. We use the BOLD signals observed by means of fMRI as the brain activity data. The data of subjects A and B were sampled every two seconds, and the data of subject C were sampled every one second. When applying dictionary learning to the data, there is a con-

straint on the number of bases in a dictionary for the size of the target data: i.e., dimension of data $\leq$ number of bases $<$ number of samples. For instance, as for subject C, we set only 782 voxels which output more than 0.55 prediction accuracy in the previous study (Nishida et al., 2015) as the observation target among all 70,933 voxels of subject C's brain activity data. Here, prediction accuracy represents Pearson's correlation coefficient between the predicted brain activity values by means of a Ridge regression model from semantic representations of the sentences to describe the movies provided as visual stimuli and real observed brain activity data. The natural language sentences describing the contents of videos are the descriptions of the still pictures captured every one second from the videos being watched by the subject. These sentences of the still pictures are written by four annotators selected randomly from 40 annotators (see Figure 1). The brain activity data is associated with the data of the sentences describing the contents for one second each.

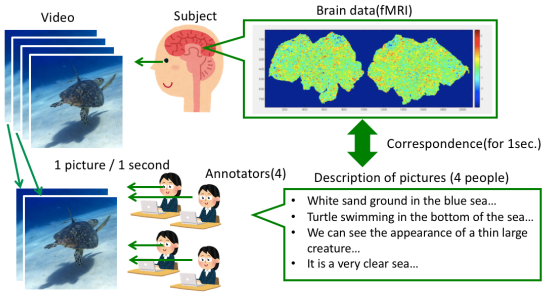


Figure 1: Brain activity and language description

Estimation Method

Our method is divided into two phases: learning and execution. In the learning phase, the brain activity data observed by using fMRI are converted into a matrix containing the values of BOLD signals whose rows and columns represent the voxels and samples, respectively. We call this matrix the brain activity matrix. Similarly, natural language sentences describing the videos are also converted into a matrix of semantic representation, which corresponds to distributed semantics made from the words whose part-of-speech are noun, verb, and adjective in the sentences. We call this matrix semantic representation matrix. We used 300 dimensional distributed representation vector trained with NINJAL Web Japanese Corpus (NWJC) in the skip-gram model. Those paired data, i.e., brain activity data and their annotated natural language sentences, are represented as a combined matrix of brain activity and semantic representation. Because there is approximately four to six second time lag in fMRI observation, we made the combined matrix taking account of the time lag. In the learning phase, the combined matrix consisting of 9000 samples is decomposed into a dictionary matrix and its coefficient matrix by dictionary learning for sparse coding. The dictionary matrix consists of bases combining the features of brain activity and those of semantic representation. The coefficient matrix is commonly used for the dictionary consisting of both bases of brain activity and semantic representation. In the execution phase, a semantic representation matrix is obtained by multiplying the dictionary matrix with a coefficient matrix derived from brain activity data via sparse coding.Here, we use Lasso-LARS as the algorithm for both dictionary learning and sparse coding. Table 1 shows the data characteristics we used in the experiment. As mentioned above, there is approximately four to six second time lag in the observation using fMRI, we considered the lag in combining both semantic representation matrix and brain activity data. Furthermore, because it takes time for dictionary learning with many samples, and subjects were watching the movies in which almost the same scenes last for several seconds, we thinned out the samples and reduced the size of the target matrix for dictionary learning. As for the data of subject A and B, we thinned out one in two, one in three, and one in four for 4500 samples, and as for subject C, we thinned out one in four and one in six for 9000 samples. As for evaluation, we employ cosine similarity between the distributed semantic vectors for both matrices; i.e., the one reconstructed from brain activity data with the dictionary and the one directly created from natural language sentences used as test data. Moreover, as the analysis for bases in the dictionary, we examined the bases whose cosine similarity are high in the experimental settings of Table 2, and investigated whether or not there is relationship between the words retrieved by the bases for the semantic representation and the pictures retrieved by the bases for brain activity data.

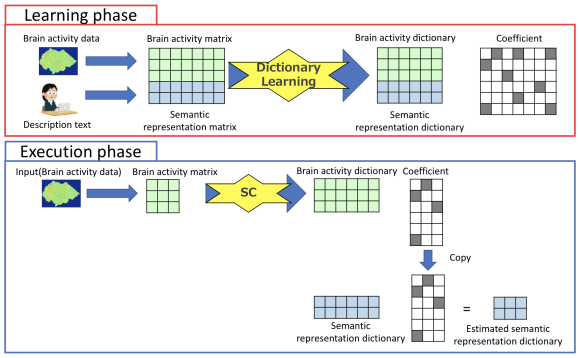


Figure 2: Overview of our method.

Results

Accuracy of estimated semantic representation

Table 2 shows the macro average of cosine similarity between the correct distributed semantic vectors and the estimated vectors with the settings for the numbers of bases and thinning, and observation time lag.

Analysis for bases

As shown in Table 2, in the case where subject is B, the number of bases is 900, and thinning out is one in two, cosine

Table 1: Dimension size of data and the number of bases

subject	original cortex	extracted voxels	semantic representation	combined matrix	number of bases
A	65665	481	300	781	800
B	68942	565	300	865	900
C	70933	782	300	1082	1100

Table 2: Cosine similarity between correct and estimated distribution semantic vectors of NL description

subject	number of bases	thinning	time lag	
			4sec	6sec
A	800	1/2sample	0.137	0.143
A	800	1/3sample	0.142	0.107
A	800	1/4sample	0.156	0.091
B	900	1/2sample	0.696	0.647
B	900	1/3sample	0.483	0.408
B	900	1/4sample	0.321	0.419
C	1100	1/4sample	0.187	0.210
C	1100	1/6sample	0.134	0.131

similarity is higher than any other settings for both cases of 4 and 6 seconds time lag. we examined the dictionary bases created in those two settings to investigate the relation between the bases for brain activity data and those for semantic representation.

Meaning of the semantic representation bases We analyze the meaning of the bases for semantic representation. Although there is not a direct word to express the meaning of a base for its semantic representation, it can be approximately expressed with close words. We measured the cosine similarity between each semantic representation base and all the word distributed representation vectors made by NINJAL Web Japanese Corpus, and retrieved the top five words with high cosine similarity to the semantic representation base. We regard those words as close words whose meaning is similar to the semantic representation base.

Samples reconstructed using the same base We confirmed whether there is common relation among the top five samples reconstructed by using a particular base with large coefficient. Because it is thought that a base for brain activity should represent some feature of human recognition when he or she watches a motion picture and a base for semantic representation should represent something to describe the characteristic of the recognition, it can be expected that the retrieved pictures are the ones which give us a common or a similar impression. We retrieved the pictures corresponding to the top five coefficients of a particular base. Each video sample is two seconds long, and we retrieved a still image at one second.

result As the method for selecting the samples to be analyzed, first of all, all the bases are numbered. We then randomly sampled 50 bases from each dictionary created in the two experimental settings, and adopted only the bases whose coefficient should not be zero for the top five pictures retrieved by the base. In the case of four second time lag setting, We could adopt 32 bases and 26 bases among 50 bases in the cases of four and six second time lag settings, respectively. To analyze the relation between the pictures retrieved by the bases of brain activity and the words retrieved by semantic representation, we conducted a questionnaire survey on nine people asking them to intuitively answer whether or not there is some relation between the top five pictures and the top five words retrieved from both 32 bases and 26 bases. Table 3 shows the result of the questionnaire in which the mean of ratio for the number of bases which people answered that there is some relation between them.

Table 3: Result of a questionnaire survey

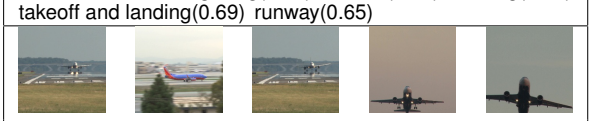
	4 seconds time lag	6 seconds time lag
Ratio of ”related”	63.19%	36.32%

We see from Table 3 that there should be some relation between the pictures and the words retrieved through both paired bases. Table 4 and Table 5 show the pictures and the words retrieved through a specific base in which all 9 people answered that there is some relation between them among the bases analyzed in the two experimental settings for observation time lag. In particular, we show only two samples of such specific bases for both observation time lag settings in the tables. In each table, for a specific base, the words retrieved from semantic representation, their cosine similarity to the semantic representation vector, and the pictures retrieved from the bases for brain activity with their coefficients are shown for the top five samples.

Discussion

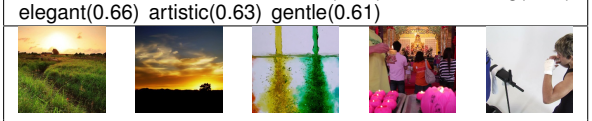
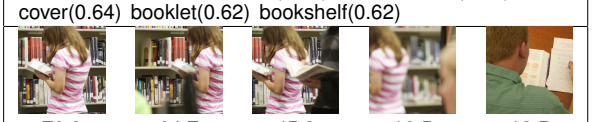
First, as for the result of the case with 4 seconds time lag settings, we refer to the two examples in which there is correlation between the retrieved words and the pictures. As for the base 422, an airplane is shown in all the pictures, and the similar words associated with an airplane such as ”gliding” and ”takeoff” are retrieved. As for the base 269, we can see that eyes are shown in four pictures among all five pictures, and the similar words associated with an eye such as ”eye” and ”pupil” were retrieved. In these pictures, it might be difficult to see that common appearance is the ”eye” at first glace, besides, we might pay attention to the features such as ”face”, ”animal”, etc., however, we see that the base for brain activ-

Table 4: Result of base analysis in 4-seconds time lag setting

Base number:422
Retrieved words : gliding(0.76) takeoff(0.76) landing(0.74) takeoff and landing(0.69) runway(0.65)

58.9 49.6 43.5 13.2 12.6
Base number:269
Retrieved words : eye(0.76) pupil(0.61) round and cute(0.60) look(0.60) stare(0.58)

42.3 17.6 12.7 12.7 12.2

Table 5: Result of base analysis in 6-seconds time lag setting

Base number:292
Retrieved words : beautiful(0.73) nice-looking(0.66) elegant(0.66) artistic(0.63) gentle(0.61)

26.5 15.0 11.6 11.2 10.4
Base number:57
Retrieved words : book(0.76) front cover(0.66) hard cover(0.64) booklet(0.62) bookshelf(0.62)

70.0 64.7 45.9 16.5 12.5

ity captured the feature of "eye" through the words retrieved by the paired base for semantic representation. Second, we discuss the result of the case with six-seconds time lag setting (see, Table 5). As for the base 292, colorful pictures are retrieved for the top four pictures with large coefficients, and similar adjectives whose meaning is "beautiful" are retrieved. This example shows us the fact that a base can express the concepts for adjective. As for the base 57, either books or bookshelves appear in all pictures, and the words related to "book" are also related. Other than those above examples, we confirmed that there are a lot of bases in which both retrieved words and pictures have correlation.

Conclusion

We applied dictionary learning and sparse coding to reconstruct semantic representation, which has expressed from brain activity data evoked by videos using distributed semantics provided by word2vec. Furthermore, by applying dictionary learning to a combined matrix of brain activity data and semantic representation, we succeeded in creating a dictionary whose bases reflect human brain recognition. By expressing the meaning of a base for the semantic representation obtained by dictionary learning with words, we have confirmed there is some relation in the samples obtained by using the same base with large coefficient. Hence it is considered that the bases for semantic representation obtained by sparse coding capture the features to efficiently reconstruct various semantic representations. In addition, among the bases obtained by dictionary learning, we found a base which hardly contributes to the reconstruction of samples. Therefore, as an assumption, it can be thought that there are not so many dimensions of features to express semantic representations in human brain. As future work, in this study, although we attempted to analyze brain activity with the constraint on the number of bases for sparse coding, we will analyze with a proper number of bases. Furthermore, based on the results of this study, we will investigate whether or not a specific brain activity labeled with words happens in a particular area of the brain in order to make a map of human recognition with words in the brain in a different way from (A. G. Huth et al., 2012; A. Huth et al., 2016).

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Cukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. In *Nature neuroscience.*

Huth, A., A. de Heer, W., L. Griffiths, T., Theunissen, F., & Gallant, J. (2016, 04). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*, 453-458. doi: 10.1038/nature17637

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc.

Miller, G. A. (1995). Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, *38*, 39–41.

Nishida, S., Huth, A. G., Gallant, J. L., & Nishimoto, S. (2015). Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. *Society Neuroscience Abstract*, *45*, 333.13.

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

Stansbury, D., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, *79*, 1025-1034.