Hierarchical Processing of Visual and Language Information in the Brain

Haruka Kawasaki[1], **Satoshi Nishida**[2], **and Ichiro Kobayashi**[1]
[1]Ochanomizu University, Japan
[2]Center for Information and Neural Networks,
National Institute of Information and Communications Technology, Japan
[1]{g1820509, koba}@is.ocha.ac.jp
[2]s-nishida@nict.go.jp

Abstract

In recent years, many studies using deep learning have been conducted to elucidate the mechanism of information representation in the brain under stimuli evoked by various modalities. On the other hand, it has not yet been clarified how we humans link information of different modalities in the brain. In this study, to elucidate the relationship between visual and language information in the brain, we constructed encoding models that predict brain activity based on features extracted from the hidden layers of VGG16 for visual information and BERT for language information. We investigated the hierarchical characteristics of cortical localization and representational content of visual and semantic information in the cortex based on the brain activity predicted by the encoding model. The results showed that the cortical localization modeled by VGG16 is getting close to that of BERT as VGG16 moves to higher layers, while the representational contents differ significantly between the two modalities.

1 Introduction

In recent years, many studies have been conducted to elucidate the information representation mechanisms of the human brain using deep learning. Studies using convolutional neural networks (CNNs) have confirmed the hierarchical processing of visual information in the brain (Yamins et al., 2014; Eickenberg et al., 2017). In addition, studies using deep learning models that deal with language have confirmed that it is possible to model the representation of semantic information in the brain (Nishida et al., 2021). However, most studies are conducted separately, and the similarities and differences in the brain information representation of both modalities have not been sufficiently discussed.

With this background, the objective of this study is to investigate on how the information localization and representation of both modalities are related to each other in the brain – we particularly aim to investigate the hierarchical characteristics of the cortical localization and representation contents of visual and language information in the cerebral cortex by using representational similarity analysis (RSA) (Kriegeskorte et al., 2008).

2 Related research

In pioneering work in modeling brain representations using deep learning, Yamins et al. (2014) showed that there is homology between hierarchical information representations in the human cortex under visual stimuli and those in CNNs, and Güçlü and van Gerven (2015) showed that complexity gradually increases with higher layers in hierarchical processing. In a study using functional magnetic resonance imaging (fMRI) and magnetoencephalography, Cichy et al. (2016) used deep learning model to show that spatio-temporal dynamics in the human brain cortex during visual object recognition is a hierarchical response. Eickenberg et al. (2017) have revealed the functional organization of the visual cortex of the human brain by analyzing brain activity with the aid of a deep learning model. Nonaka et al. (2021) introduced the brain hierarchy score, which indicates the degree of hierarchical response based on encoding and decoding to brain activity, and discussed what kind of deep learning models accurately represent the structure of the visual cortex of the human brain, showing that deep learning models with high accuracy in image identification do not necessarily represent the behavior of the visual cortex of the human brain.

On the other hand, in a study that models brain representations from semantic features of language, Huth et al. (2012) used fMRI to observe brain activity of subjects watching a two-hour natural video and labeled them using 1705 WordNet (Fellbaum, 1998)-based categories for objects and actions in the video, showing that these categories are not represented in specific brain regions but as locations

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 405–410
November 20–23, 2022. ©2022 Association for Computational Linguistics

in a continuous semantic space. Huth et al. (2016) constructed semantic maps in brain regions from brain activity induced by natural speech stimuli, and found that in most regions of the semantic system, there are specific semantic regions and groups of related concepts. Nishida et al. (2021) clarified that quantitative modeling of meaning using word2vec (Mikolov et al., 2013) and other methods is an effective means of estimating language activity in the brain through comparison with semantic structures evaluated from human behavior. Jain and Huth (2018) introduced LSTM (Hochreiter and Schmidhuber, 1997) to extract vectors for each word, used them in their encoding model, and achieved more accurate estimation than conventional models. In recent years, the construction of computational models that explain language processing properties in the brain using distributed semantic representations has played an important role. In this context, Sun et al. (2021) scrutinized the still unexplored relationship between the brain representation of sentences and distributed representations, and whether the linguistic features captured by distributed representations can better explain the correlation between brain activities in which sentences are given as linguistic stimuli, and showed the characteristics of distributed representations and their effectiveness.

Most of the above studies have explored the properties of visual and semantic brain processing separately. Therefore, the hierarchical processing from visual to semantic information in the brain is not well understood. In this study, we construct and compare encoding models based on these two different modalities, and investigate the characteristics of information localization and information representation content in the hierarchical processing of visual and semantic information.

3 Brain information analysis with RSA

3.1 Overview

Figure 1 illustrates an overview of our study. Firstly, we use fMRI to collect brain activity data while subjects are watching movies with either fixation or free viewing. We then extracted image features from the images cropped from the movies given to the subjects as stimuli using VGG16 (Simonyan and Zisserman, 2014) and linguistic features from the annotations assigned to the images using BERT (Devlin et al., 2019).

To predict the brain activity from the features extracted by those deep learning models, we construct encoding models using Ridge linear regression. Then, to investigate the hierarchical characteristics of cortical localization and representational contents of visual and linguistic information on the cerebral cortex, we apply RSA to analyzing the brain states predicted by the encoding models.

3.2 Encoding model

In this study, we employ the method by Naselaris et al. (2011) for the construction of encoding models. When constructing the encoding model, the target feature space and brain activity patterns are linearly regressed, and weights are learned so that the measured brain activity patterns and predicted brain activity patterns are close. The constructed encoding model is then applied to the evaluation data, and the prediction accuracy is evaluated. In general, Ridge liner regression is used as the regression method, and by observing the regression coefficients, it is possible to observe the behavior with respect to voxels.

3.3 Representational Similarity Analysis

RSA is a framework for characterizing representations of various modalities by representational dissimilarity matrices (RDMs) and comparing RDMs. An RDM is a matrix that allows us to retrieve the representational distance (or dissimilarity) of each modality. The dissimilarity in our study is calculated by correlation distance (1 - Pearson's correlation coefficient). Creating RDMs makes us possible to measure things that cannot be directly measured for similarity. In addition, RSA has the property that it does not require the definition of mappings, which is necessary when directly comparing activity patterns.

4 Experiments

We have conducted the following three experiments to investigate whether or not:(i) predictable brain regions are similar to both vision and language stimuli; (ii) cortical localization patterns are similar; (iii) representational content is similar. The numbers on the right side of Figure 1 correspond to the numbers of the experiments.

4.1 Experimental settings

fMRI data Brain activity data were obtained by fMRI at the Center for Information and Neural Networks, National Institute of Information and

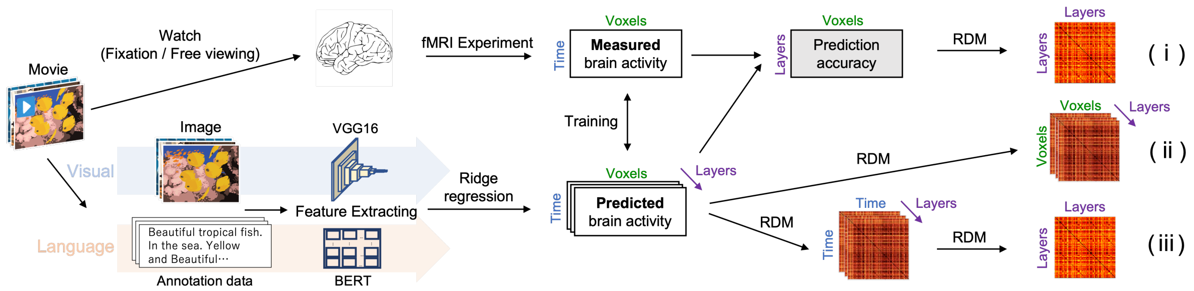


Figure 1: An overview of the experiments

Communications Technology (NICT). The brain activity data were collected by fMRI which is a 3T MRI (Siemens MAGNETOM Prisma), and the imaging parameters are TR 1 second and voxel size $2 \times 2 \times 2$mm. T1 structural images were also taken separately from the fMRI images, and were registered with the fMRI images using FreeSurfer (Dale et al., 1999). Only the voxels of the cerebral cortex extracted by this method were used in the analysis. Seven subjects (three males and four females, mean age 24.1 years) participated in the fMRI experiment. The experimental protocol was approved in advance by the Ethics Review Committee and Safety Review Committee of NICT, and written consent was obtained from all subjects before the experiment. Each subject watched a 2 hour 40 minutes compiled movie with sound in the fixation condition (gazing at a fixed point of view in the center of the screen) and the free viewing condition (moving the gaze freely). Of the 2 hours and 40 minutes of data obtained in each condition, 2 hours were used as training data for the model. The remaining 40 minutes of data consisted of four repetitions, which were averaged to 10 minutes and used as the evaluation data for the model.

Annotation data To extract linguistic features from the movies, we obtained written scene descriptions from five to six annotators for each one-second video scene. The annotators were native speakers of Japanese and did not participate in the fMRI experiment.

Encoding models for the experiments The same method was used to create encoding models based on image features and language features. A total of 40 encoding models were constructed using the features extracted from each of the VGG16 (using 8 layers) and BERT (12 layers in total) under fixation and free viewing conditions. A model that predicts the time series of brain activity using the time series of features as explanatory variables was trained by Ridge regression. In order to take into account the hemodynamic delay in the responses, we regressed the fMRI-observed brain activity data with the 3, 4, 5, and 6 seconds precedence features. In addition, 10-split cross-validation was conducted by shuffling the training data with 50 chunks, and the regularization term with the best average correlation coefficient was adopted. Using the learned encoding models, we evaluated the prediction accuracy of each voxel by obtaining Pearson's correlation coefficient between the predicted and measured fMRI signals to the same stimuli. In doing so, we rejected voxels with significant p-values ($p < 0.05$) corrected for false discovery rate.

4.2 Experimental results

In the following, we indicate total number of layers as n_layers, total number of voxels as n_voxels. We employ the rejected voxels with significant p-value in at least one of the 40 encoding models as the data used in all the following experiments.

(i) Predictable regions This analysis was performed to determine the similarity of brain regions that can be predicted by a total of 40 encoding models using features extracted from all targeted hidden layers of VGG16 and BERT as input. Prediction accuracies of all encoding models were used to create an RDM ($n\_layers \times n\_layers$) for each subject and averaged over all subjects. The upper figure of Figure 2 shows the RDM of ($n\_layers \times n\_layers$) and the lower figure shows it compressed into ($n\_layers \times 3$) using multi-dimensional scaling (MDS) and plotted on a 3-dimensional space. The closer the models are displayed to each other, the more similar brain regions they can predict. Both VGG16 and BERT are color-coded in the fixation and free viewing conditions, and visualized in a total of four colors according to the deep learning models and its conditions. The lighter colors indicate the lower layers and the darker colors the higher layers, and the numbers next to the dots indicate the number of the layer. From this result, it can be seen that brain regions where models are predictable become similar to that of BERT as the hierarchy of VGG16 increases from lower to higher layers.

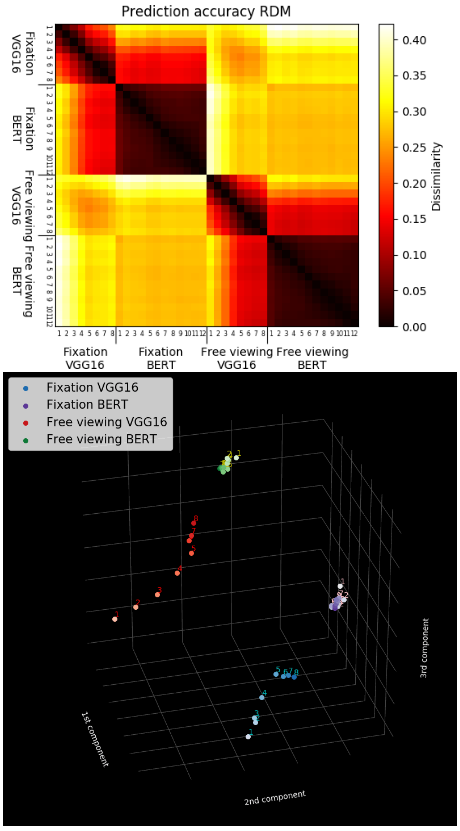
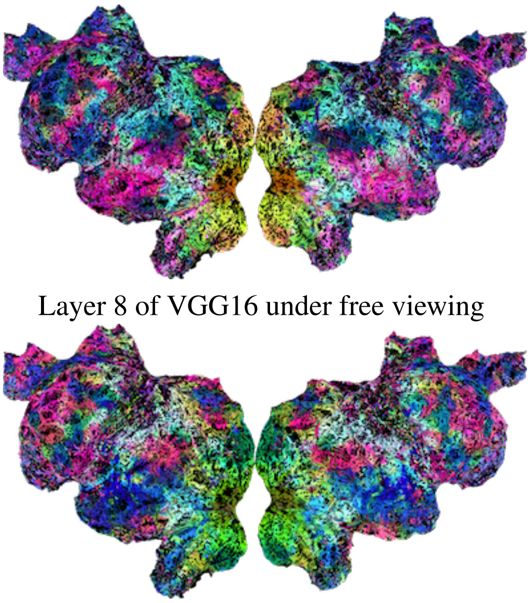


Figure 2: Similarity of predictable brain regions by the models

(ii) Cortical localization patterns We performed this analysis to see the similarity of cortical localization patterns for the VGG16 high layers and BERT. The brain activity predicted by each encoding model ($time \times n\_voxels$) was used to create an RDM ($n\_voxels \times n\_voxels$) of each layer for each subject. Figure 3 shows the visualization results of the predicted brain activity of one subject from the 8th layer, the highest layer among the targeted layers in VGG16 and the 12th layer of BERT under free viewing conditions. The RDM is reduced in dimensionality by uniform manifold approximation and projection (UMAP), and the colors are plotted on a flat map of the cortex created by means of Pycortex[1], with the colors of near objects being close to each other and distant objects being far apart. The results show that the pattern of similarity of the contents of voxel-wise information representation in the cortex is similar between the higher layer of VGG16 and that of BERT.


Layer 8 of VGG16 under free viewing

Layer 12 of BERT under free viewing

Figure 3: Similarity of cortical localization patterns

From (i) and (ii), we find that cortical localization becomes more similar to BERT as one moves from the lower to higher layers of VGG16. We now perform the experiment (iii) to see if the representational content also approaches BERT as one moves from the lower to higher levels of VGG16.

(iii) Representational content We performed this analysis to determine the similarity of representational content for 40 encoding models. Using the predictions of brain activity, we created RDMs of ($time \times time$) for each encoding model. All of these RDMs were then used to create an RDM of ($n\_layers \times n\_layers$) for each subject and averaged over all of them. Figure 4 shows the RDM among layers of VGG16 and BERT in terms of representational contents and the result of dimensionality reduction of the RDM with MDS and visualization on a 3-dimensional space. Figure coloring is the same as (i) of 4.2. From this result, it can be seen that there is a significant difference in the representation content between VGG16 and

[1] https://github.com/gallantlab/pycortex

BERT regardless of the fixation or free viewing conditions.

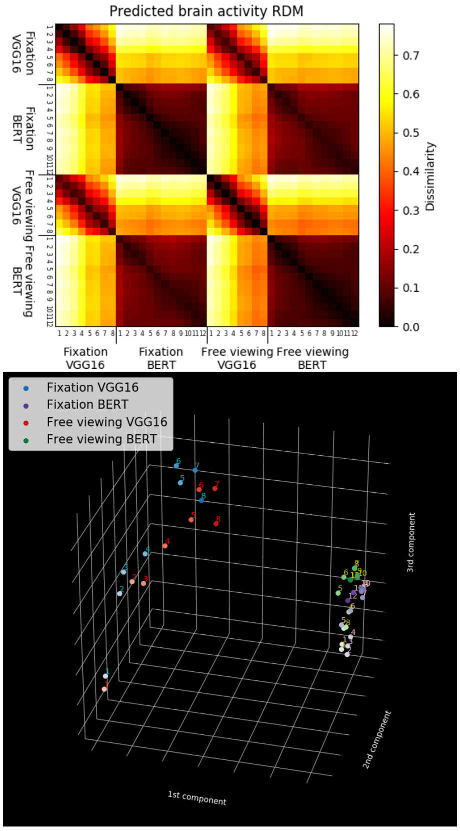


Figure 4: Similarity analysis of representation content

5 Discussion

In the experimental result of (i), we investigated the predictable regions of both models using RSA and found that the cortical localization becomes similar to that of BERT as VGG16 moves from lower to higher layers. In addition, from the result of the experiment (ii), it is observed that the pattern of similarity of information representation content is similar between the higher layer of VGG16 and that of BERT. We therefore estimated that the representational content in the two models is similar. However, from the result of experiment (iii), it was found that there is a significant difference in the representational content that can be modeled by VGG16 and BERT. In other words, our results suggest that VGG16 and BERT represent different brain information even in the same higher sensory cortex.

6 Conclusions

In this study, we have investigated the the hierarchical characteristics of cortical localization and representational content of visual and linguistic information on the cerebral cortex by means of RSA using prediction accuracy and contents. As a result, in the analysis of cortical localization using prediction accuracy, we found that VGG16, i.e., CNNs dealing with image features, was able to model the hierarchy in the cortical localization in the brain, and as it moved from lower to higher layers, it was able to predict brain regions closer to those predicted by BERT, i.e., DNNs dealing with linguistic features. Furthermore, in the analysis of information representation content with predicted brain activity, it was found that the higher layers of VGG16 can model complex cortical localization patterns in the cortex as well as BERT. However, we found a large gap between VGG16 and BERT in the comparison of the representational contents between the layers. These results suggest that visual information is represented in the same brain regions as linguistic information as it becomes more complex (e.g., category selection regions in the temporal cortex), but even within the same brain regions, there are significant differences between visual and linguistic information, and that modeling with VGG16 and BERT alone is not sufficient to fill in these differences. When cortical localization is similar between different modalities, we generally tend to conclude that the representational contents between them are also similar, but the results of this study suggest that the similarity relationship does not necessarily hold, which has an important message to encourage rethinking of the results of previous studies that tackle to elucidate brain information representation for different modalities based solely on the prediction accuracy of brain activity information. Based on this, we intend to continue to elucidate the characteristics between modalities.

References

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Torralba, Antonio an d Oliva. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual obj ect recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755.

Anders M Dale, Bruce Fischl, and Martin I Sereno. 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. 2017. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, 152:184–194.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Umut Güçlü and Marcel A J van Gerven. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across t he ventral stream. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(27):10005–10014.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, and Jack L Theunissen, Frédéric E andG̃allant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack 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.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositio nality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. 2011. Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410.

Satoshi Nishida, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. 2021. Behavioral correlates of cortical semantic representations modeled by word vectors. *PLoS Computational Biology*, 17(6):e1009138.

Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. 2021. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013.

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

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2021. Neural encoding and decoding with distributed sentence representations. *IEEE Trans. Neural Networks Learn. Syst.*, 32(2):589–603.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, and James J. Seibert, Darren an d DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.