Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6397–6404
Marseille, 20-25 June 2022

How Does the Experimental Setting Affect the Conclusions of Neural Encoding Models?

Xiaohan Zhang, [1,2] **Shaonan Wang,** [1,2] **Chengqing Zong** [1,2,3]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{xiaohan.zhang, shaonan.wang, cqzong}@nlpr.ia.ac.cn

Abstract

Recent years have witnessed the tendency of neural encoding models on exploring brain language processing using naturalistic stimuli. Neural encoding models are data-driven methods that require an encoding model to investigate the mystery of brain mechanisms hidden in the data. As a data-driven method, the performance of encoding models is very sensitive to the experimental setting. However, it is unknown how the experimental setting further affects the conclusions of neural encoding models. This paper systematically investigated this problem and evaluated the influence of three experimental settings, i.e., the data size, the cross-validation training method, and the statistical testing method. Results demonstrate that inappropriate cross-validation training and small data size can substantially decrease the performance of encoding models, especially in the temporal lobe and the frontal lobe. And different null hypotheses in significance testing lead to highly different significant brain regions. Based on these results, we suggest a block-wise cross-validation training method and an adequate data size for increasing the performance of linear encoding models. We also propose two strict null hypotheses to control false positive discovery rates.

Keywords: Neural encoding, fMRI, Naturalistic stimuli

1. Introduction

In recent years, neural encoding models have become increasingly important tools in studying the neural basis of language (Nunez-Elizalde et al., 2019). Neural encoding models are data-driven methods that explore the associative mapping between the linguistic stimuli and the evoked brain response in the data. In the fMRI literature, they have been used to investigate the cortical representations of a broad range of linguistic features (Mitchell et al., 2008; Wehbe et al., 2014; Huth et al., 2016; Wang et al., 2020; Sun et al., 2020; Jain et al., 2020; Zhang et al., 2020; Caucheteux et al., 2021). Many of these studies formalized the encoding models as linear regression models that trained with a cross-validation method to predict the brain responses to language stimuli. The conclusions reached in these studies highly depend on the prediction performance of encoding models. As the use the encoding models increases, it is critical to know whether and how various language-irrelevant factors affect prediction performance and conclusions of encoding models. One important but often neglected kind of such factors is the experimental settings.

As a data-driven method, the experimental settings would undoubtedly affect the performance of encoding models. However, it is unknown whether and how this will further affect the neural encoding conclusions. Existing work using neural encoding varied from the amount of neuroimaging data, the training method, and the significance testing methods used to test the statistical significance of encoding results. These experimental setting differences prevent a formal comparison between different works. More importantly, since the rate of false positive results tends to increase with high methodological flexibility (Carp, 2012), these differences may also impair the reproducibility and the reliability of encoding models. Therefore, a comprehensive investigation of how the variability of the experimental settings affect the encoding results is necessary.

This paper investigates three settings of encoding models, including the training method, the data size, and the way to perform significance testing. We ask two questions. First, how should we use neuroimaging data more efficiently when training and testing encoding models? Second, how should we conduct significance testing in order to control the false discovery results?

To address these issues, we collected the function magnetic resonance imaging (fMRI) data from 12 healthy subjects when they were listening to 60 Chinese stories. For each subject, we train voxel-wise encoding models with different training methods and different data size and analyze how these differences influence the prediction accuracy on different brain regions. Then, we test how different statistical testing methods affect the selection of language-sensitive voxels.

Our results demonstrate that inappropriate training method and insufficient data can substantially decrease the prediction performance on many brain regions, especially in the temporal lobe and the frontal lobe. Inappropriate significance testing may cause high false positive discoveries. Based on these results, we give the following practical suggestions on the using of encoding models:

- First, the cross-validation training of encoding

models using a block-wise form with block length longer than 32s can substantially increase the prediction accuracy.

- Second, for the linear encoding models, the size of neuroimaging data should be longer than 3 hours to avoid under-fitting. And, restricted by the expressive capacity of linear regression models, more data can hardly further improve the prediction accuracy.

- Finally, considering that encoding models can predict some voxels activation with permuted stimuli or predict their activation in resting state with higher-than-zero accuracy, the null hypothesis in significance testing should be designed carefully to control for these false discoveries.

2. Related Work

In this section, we briefly introduce the basics of neural encoding models and the three settings (i.e., the training method, the data size, and the significance test method) used in previous work.

Neural encoding models use a set of stimuli features to predict brain activities elicited by the stimuli. The parameters of neural encoding models are trained on a training set, usually with a cross-validation method. And then, these parameters are fixed to predict brain responses on a test set. The performance of encoding models can be assessed through Pearson correlation between predicted and actual brain responses or a classification task on a test set. After the testing process ends, a significance test is needed to ensure that the testing results are not coincidental.

An important step in cross-validation training is to split data into training, validation, and test sets. Previous work that collected fMRI with natural story stimuli often chose one story as a test set and conducted cross-validation on the remaining stories (Huth et al., 2016; Jain et al., 2020; Popham et al., 2021; Zhang et al., 2020). When splitting data into training and validation sets, previous work usually concatenated stories except the test story together and split data into k-folds and chose one fold as the validation set (Zhang et al., 2020; Schrimpf et al., 2020) or split data into blocks of a certain length and randomly choose some blocks as held-out sets (Huth et al., 2016). There are two problems with such a splitting strategy. First, there may be possible information leakage between the training and the validation sets. Because of the hemodynamic delay of BOLD signals, successive data samples are not independent. Inappropriate splitting may cause information leakage between the training and the validation sets. Second, choosing one story as the test story can avoid information leakage from the test set. But the prediction performance on the test story may be affected by the subject's condition when listening to this story (e.g., the subject may be absent-minded for a while) or the specific nature of the content of the story.

Apart from the training method, the collected amount of neuroimaging data also varied widely in previous work. Before discussing the data size used in previous work, we must first clarify one point: the stimuli modality and the temporal resolution of neuroimaging data collection are often different in different works. Existing work collected fMRI when subjects were reading language words (Wehbe et al., 2014) or sentences (Pereira et al., 2018) showed on the screen or listening to speech (Huth et al., 2016; Zhang et al., 2020; Caucheteux et al., 2021). The temporal resolution of fMRI collection also differed from 0.72s to 2s. Thus, the data size can be quantified by time length of recordings, the number of fMRI images, or the stimuli vocabulary. Here, we simply use total time length of collected fMRI time series to quantify the data size. The data size used in existing work varied from 45 minutes (Wehbe et al., 2014) to more than 5 hours (Zhang et al., 2020). It is unknown whether lower amount of data causes under-fitting in any brain regions. If so, the brain regions representing language information may be predicted with lower accuracy and thus lead to incorrect conclusions.

Statistical significance testing in fMRI research has long been a source of contention (Bhaumik et al., 2009; Button et al., 2013; Eklund et al., 2016). Inappropriate significance testing methods may fail to detect a true effect or result in a high false discovery rate. A vital preliminary to significance testing is to make clear what the null hypothesis is. Then, significance testing can be conducted to decide whether the null hypothesis should be accepted or rejected, using either a parametric or non-parametric test, depending on whether the distribution of the object is known. However, previous work often did not directly state their null hypothesis.

We sum up the significance testing methods used in existing work into three null hypotheses: $H_0 : r = 0$, $H_0 : r = r_v$, and $H_0 : r = r_{perm}$. r is the Pearson correlation between predicted and real fMRI signals of the test set. r_v is the correlation between two independent random vectors with the same length of test set fMRI signals. r_{perm} is the correlation between the actual fMRI signals and the predicted signal by encoding models trained with permuted stimuli. All three null hypotheses aim to test the Pearson correlation r between predicted and real BOLD signals, but they are based on distinct assumptions. The first null hypothesis assumes that a r greater than zero is sufficient to demonstrate brain response predictability. The second null hypothesis accounts for the bias of Pearson correlation by assuming that the r between predicted and real brain signals should be greater than that of two random vectors. The third hypothesis assumes that the linear correspondence learned by encoding models may be driven by factors unrelated to linguistic information, thus the r should be greater than the r_{perm} which is the correlation between the actual fMRI signals and the predicted signal by encoding models trained with

permuted stimuli. Conducting significance test under different null hypothesis may select different voxels.
Due to the problems identified above, we systematically investigate how the differences in all three settings influence the encoding results.

3. Our Methods

As shown shown in Figure 1, the overall framework contains three experimental settings, including training method (section 3.1), data size (section 3.2), and the statistical significance testing method (section 3.3). We investigate whether and how these different experimental settings affect the results of encoding models. The implementation details of data collection and language representation are described in section 3.4
To formalize the training and test process of encoding models, let X be the stimulus and Y be the brain activity elicited by X. In this paper, X is a sequence of words $w_1, w_2, ..., w_n$ and Y is a sequence of fMRI signals. The neural encoding models learn to map from stimulus X to the elicited brain response Y for every voxel. To align with the temporal delay of BOLD signal, the word vectors are convolved with a canonical hemodynamic response function (HRF)[1] and then downsampled to the sampling rate of fMRI. The model could be expressed as:

$$Y = downsample(conv(X, hrf)) \times W$$

By splitting data into training and test set, the regression weight matrix W is trained with Y_{train} and X_{train}. And then, W is used to predict Y_{test} based on X_{test}. The Pearson correlation is computed by $Pearson(Y_{test}, \hat{Y}_{test})$.

3.1. Training Method

The key question in training and evaluating ridge regression models with cross-validation is to split data into training, validation, and test set. Considering that randomly choosing one story as test set may have an effect of the session, we conduct nested cross-validation (Figure 2). That is, the training process contains two loops: the inner 5-fold cross-validation loop that chooses the best hyper-parameters for ridge regression and the outer loop that tests the encoding model on the test set.
Specifically, we choose one story as test set in each outer loop and run the outer loop for 60 times so that each story can be the test set in different loops. As shown in Figure 2, within each outer loop, we run a 5-fold cross validation by concatenating the remaining 59 stories together and then splitting them into training and validation set in the unit of different length of blocks. Note that the blocks in validation set are randomly selected among all blocks. We choose 7

[1]The canonical HRF is a mathematical model that describes what the BOLD signal would theoretically be in response to a neural impulse.

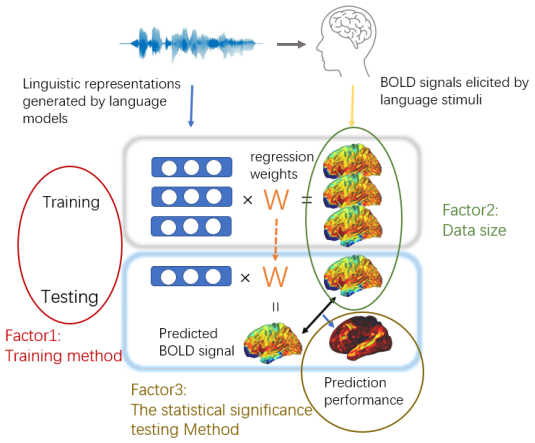


Figure 1: The overall framework

block lengths (labelled in Figure 2 as $c1$ to $c7$) plus an average-splitting (labeled as $c8$) that often used in previous work. The block lengths include two shorter than 32s and five as the integral multiple of 32s. The reason to choose 32s as a borderline is that 32s is usually taken as the length that an hemodynamic delay lasts. As mentioned above, this block-splitting strategy can not fully avoid the information leakage between the training and validation set. Hence, we also split data directly by stories (labeled as $c9$) to avoid the dependence between training and validation set and use results of this strategy as the ceiling, that is, the best results that the cross-validation can reach.
In the end, we have 9 data-splitting conditions and 60 Pearson correlation coefficients for each condition. For simplicity, we use r_{c_i} to indicate the correlation of the i_{th} condition averaged across all 60 stories and all brain voxels. By comparing the r_{c_i} between different conditions, we can know how the splitting method affect the encoding results and which is a better way to conduct cross-validation.

3.2. Data Size

Using the 9_{th} condition of nested cross-validation, we study how the data size affect the encoding results. We analyze whether small data size makes certain brain regions less predictable and whether there is sufficient data size such that adding more data could not increase the prediction accuracy. We use six different amounts of data, including $[10, 20, 30, 40, 50, 60]$ stories respectively. We compute the averaged Pearson correlation of each size and analyze which brain regions' prediction performance increases most when adding 10 more stories.

3.3. Statistical Significance Testing

The purpose of significance test is to test whether the prediction accuracy of encoding models is coincidental. Moreover, it should also test whether the prediction accuracy is driven by the mapping irrelevant to stimuli learned by encoding models. Based on this purpose,

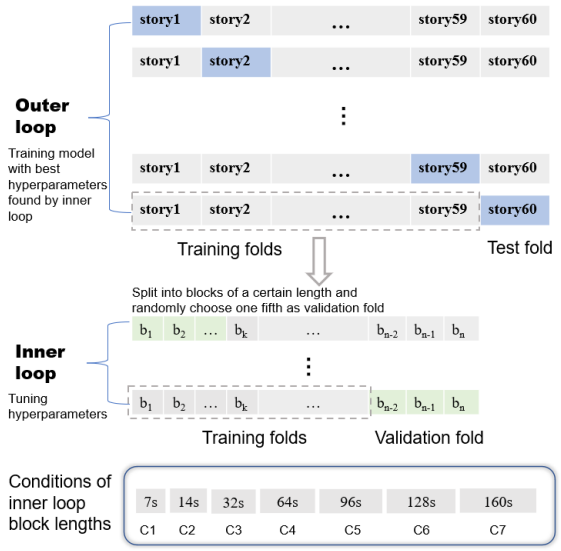


Figure 2: The nested cross-validation.

we adopt two null hypotheses used in previous work: $H_0 : \bar{r} = 0$, $H_0 : \bar{r} = \bar{r}_{perm}$. Besides, we also add one null hypothesis: $H_0 : \bar{r} = \bar{r}_{rest}$, which computes the Pearson correlation between the predicted fMRI signal $\hat{Y}$ and the resting state fMRI signal Y_{rest}. The assumption of both $H_0 : \bar{r} = \bar{r}_{perm}$ and $H_0 : \bar{r} = \bar{r}_{rest}$ is that the encoding models may predict the brain activities irrelevant to stimuli.

To test these null hypothesis within our nested cross-validation encoding framework, we compare the means of the Pearson correlations between those computed with encoding models and those computed under the null hypothesis. Hence, significance testing becomes the testing of mean difference between two populations. These three null hypothesis are tested both on subject-level and group-level. For subject-level, the correlations of all test stories of one subject are regarded as samples from one population. For group-level, the correlations of all subjects and all stories are regarded as samples from the same population. And then we test whether the population means of r are significantly larger than that of r_{perm} and r_{rest}.

3.4. Implementation Details

We collected the fMRI data from 12 healthy subjects when they were listening to 60 Chinese stories. All stories are collected from Renmin Daily Review website and are available at `https://www.ximalaya.com/toutiao/30917322/`. Each of these stories last from 4 to 7 minutes and the total time length of all 60 stories is about 5 hour.

The text of all stories are transcribed and then segmented into words. Then, we use a pre-trained BERT model (Devlin et al., 2018) to generate the representation of every word. The sub-word embeddings are averaged as the embedding of words.

For ridge regression, the regularization parameter is chosen amongst 40 values log-spaced between 10^{-3} and 10^{3}.

4. Results and Analysis

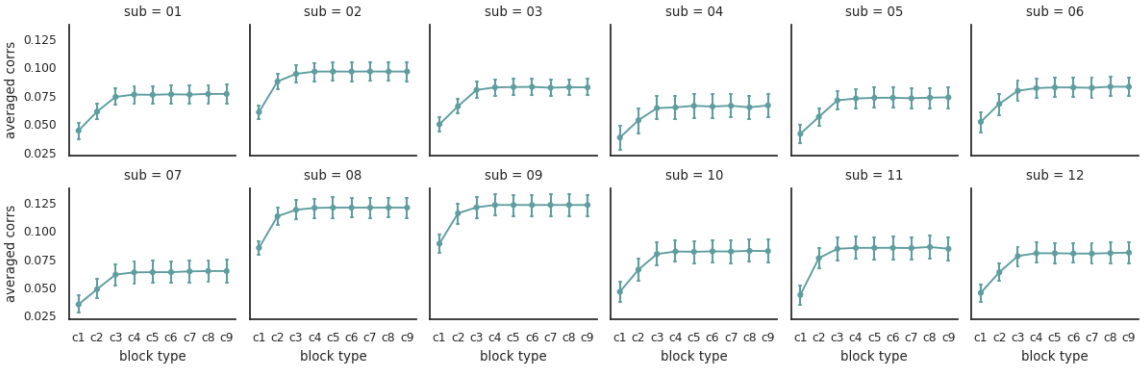
We trained encoding models separately for each subject and evaluated the performance of neural encoding models trained with different training methods and different data sizes for all 12 subjects. We also conducted the significance test under all three null hypotheses described in section 3.3. By comparing prediction performance under different experimental settings, we have several findings about how these experimental settings affect the performance of encoding models.

Splitting data with longer blocks can improve prediction performance As shown in Figure 3(a), in all 12 subjects, splitting data into training and validation sets with blocks less than 32s can substantially decrease the prediction accuracy. But as long as the block length exceeds 32s, increasing the block length can hardly improve or decrease the prediction results.

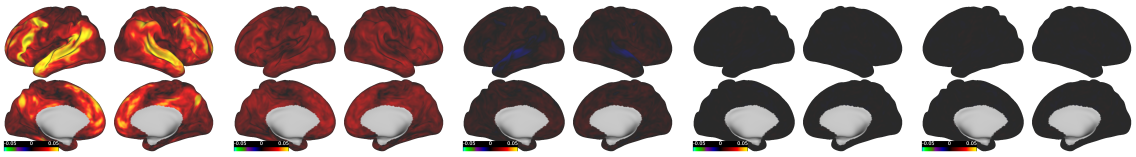
To intuitively show how the block length affects the prediction accuracy in the whole brain, we directly subtract the results of shorter blocks from that of longer blocks and average across all subjects. The results are shown in Figure 3(b)-(f). As illustrated, the prediction accuracy of the whole brain improves when block length increases from 7 to 14 seconds, with the temporal lobe and the frontal lobe showing the greatest improvement. When block duration increases from 14 to 32 seconds, there is still a whole brain rise, but it is more uniform across the brain. Continuing to increase the block length makes no difference in most brain regions and even a slightly decrease in the temporal and the frontal lobe. We also compare the results of the average-splitting ($c8$) and the story-level splitting ($c9$) in Figure 3(e) and Figure 3(f). As shown, the results of story-level splitting are nearly identical to the results of longer blocks, suggesting that blocks longer than 32 seconds is enough to avoid the influence of information leakage caused by temporal correlation in successive fMRI data.

Insufficient data makes the frontal and the temporal lobe substantially less predictable Figure 4(a) shows the prediction accuracy of encoding models with different data size. In general, the prediction accuracy increases as the amount of data increases for all subjects. And the increase of accuracy is more substantial when the data size is small, such as adding 10 stories to 20 stories.

To further analyze how the increase of data size improves the prediction accuracy, we subtract the results of each data size from the results after adding 10 stories. Figure 4(b) to 4(f) are subtraction results averaged across all subjects. Results shown that adding data mainly increases the prediction accuracy in the temporal lobe and the frontal lobe. Specifically, when adding the number of stories from 10 to 20, the predictability of almost all voxels increases. The increase of voxels in

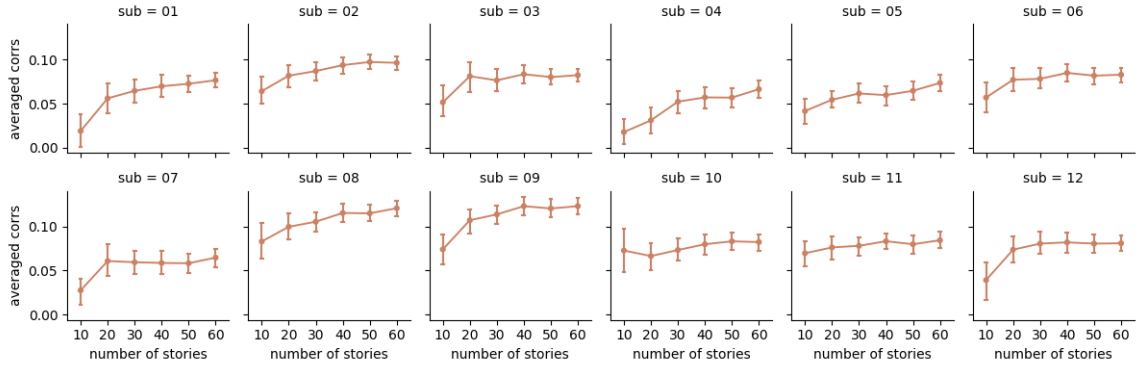


(a) Results of different block lengths.

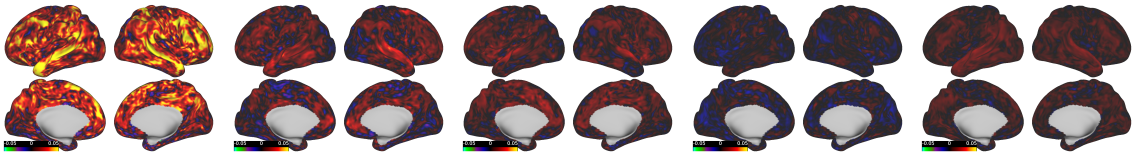


(b) $r_{c2} - r_{c1}$ (c) $r_{c3} - r_{c2}$ (d) $r_{c4} - r_{c3}$ (e) $r_{c8} - r_{c7}$ (f) $r_{c9} - r_{c8}$

Figure 3: Prediction results of different training-validation data-splitting conditions. (a) is the Pearson correlation averaged across all brain voxels. The error bar indicates the standard deviation of all 60 test stories. (b) to (f) are the subtraction results averaged across all subjects between different block conditions.



(a) Results of different data size



(b) $r_{20} - r_{10}$ (c) $r_{30} - r_{20}$ (d) $r_{40} - r_{30}$ (e) $r_{50} - r_{40}$ (f) $r_{60} - r_{50}$

Figure 4: Results of different data size. (a) is the Pearson correlations averaged across all brain voxels of different data size. The error bar indicates the standard deviation of all 60 test stories. (b) to (f) are the subtraction results averaged across all subjects between different data size.

frontal and temporal lobe is especially greater than in other parts of the brain. Adding the number of stories from 20 to 30 and from 30 to 40 increases the frontal and temporal lobe prediction results in a similar way, but with smaller increases. When the number of stories reaches 40, adding more can hardly improve prediction results.

The choice of null hypothesis has much greater impact on the significance test results than the test method Figure 5 shows the subject-level and group-level significance test results under three null hypotheses. As illustrated, significant voxels under different null hypotheses varied substantially, especially at subject-level. However, for the same null hypothesis, the results between parametric test and non-parametric

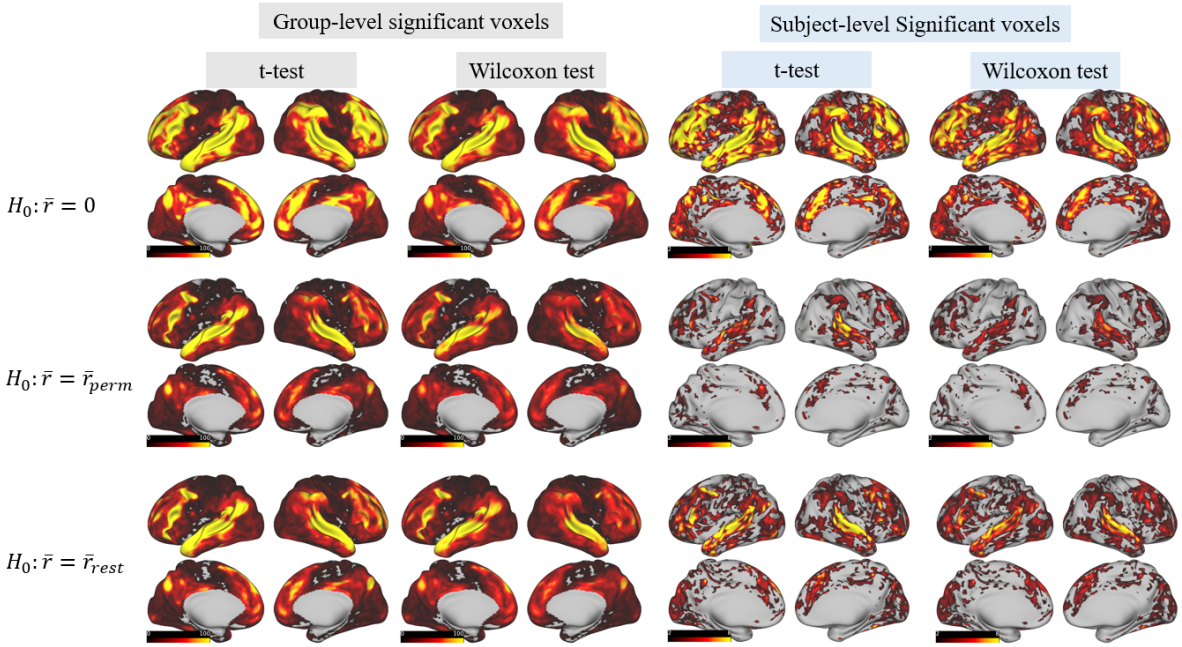


Figure 5: Results of significance test. Color-highlighted areas are significant voxels under different null hypotheses (FDR $q < 0.01$, the color bar represents values of $-\log q$)

test are pretty similar. This phenomenon suggests that choosing the appropriate null hypothesis is more important than choosing the statistical test method.

Among the results of all three hypotheses, the $H_0 : \bar{r} = 0$ selects the most voxels at both subject level and group level. However, among these significant voxels, many of them are not significant under the other two null hypothesis, suggesting that a $\bar{r}$ greater than zero may not enough to prove a voxel represents linguistic information. Because the encoding models can learn some mapping even with the permuted stimuli and with the rest fMRI signals. Furthermore, the results of subject-level test and group-level test are also different. Almost all voxels are statistically significant in group-level test. The reason is that with an increased number of samples, a small effect can also be significant.

5. Discussion

Exploring brain language processing with neural encoding models highly depends on the performance of encoding models in predicting brain responses. Our experiment shows that different experimental settings can lead to highly different encoding results. In this section, we discuss the results of our experiment.

Our first finding is about the cross-validation method used to train the encoding model. A main challenge for cross-validation is that the successive measurements of fMRI are temporally correlated. We find that splitting training-validation data with blocks longer than 32s can substantially increase the prediction performance. The reason may be that the influence of temporal correlation in fMRI data could be diminished with longer blocks. This is reasonable because the temporal correlation between fMRI images decreases as the distance increases. We also find a slight decrease on the temporal lobe and the frontal lobe when continuing to increase the block length. A possible explanation is that shorter blocks can better mix up the data from different stories and therefore choose a better hyper-parameter. From our experiment, a block length around 32s reaches the best performance of encoding models. This finding about the best block length is also very useful when using bootstrapping methods.

The second finding is about the data size used to train encoding models. As a data-driven method, there is no doubt that more data can lead to better results of encoding models. However, our concerns are whether small data would miss some brain regions and whether increasing the data can linearly increase the prediction performance. The experimental results show that when data size is small, adding data can substantially increase the prediction accuracy especially on the temporal lobe and frontal lobe. However, as the data size continues to increase, the improvement brought by new data becomes lower, and the prediction performance tends to be stable. This phenomenon may be explained by the limitation of the expressive capacity of linear models. As discussed by Ivanova et al. (2021), existing work focused excessively on the linear encoding model. It is quite reasonable to use linear models when the data size is small because complex models are prone to over-fitting on small data. However, as the amount of data available grows, more powerful models may be adopted in the future to better learn the mapping between deep language models and brain signals.

Finally, we discuss the statistical significance test-

ing. On one hand, in neuroscience studies, significance testing is an important step and various testing methods have been raised in the past decades (Nichols and Holmes, 2002; Winkler et al., 2014; Etzel, 2015; Lohmann et al., 2018). However, fewer attention has been paid on it in studies using encoding models. Our results show that choosing appropriate null hypothesis is very important in controlling false discoveries. But the three null hypotheses we considered are only attempts to control different bias that encoding models may learn in data. There may also be other bias that the encoding models can learn. More discussions are needed to further clarify this question. On the other hand, as mentioned by Hamilton and Huth (2020), many neuroscience studies paid too much attention to statistical significance and too little to effect size. As in our results, almost all voxels are statistically significant in the group-level testing. However, a small Pearson correlation may be significant but not important in explaining brain mechanisms. Future work can pay more attention to statistical power and effect size.

6. Conclusion

Neural encoding models predict brain responses based on the stimuli that elicited them. The present study systematically investigates how different experimental settings of encoding models affect the results. By exploring encoding models with different training methods, different data size, and different significance testing methods, we find that in cross-validation, splitting training-validation set with blocks longer than 32 seconds can reach substantially higher prediction accuracy than blocks shorter than 32 seconds. And for a linear encoding model, the recording duration of fMRI data longer than three hours may be sufficient to avoid the under-fitting training results. Finally, the careful choice of null hypothesis is very important to control for false discoveries. We considered three null hypotheses in this work, but more discussions are needed to clarify what is the best way to conduct significance testing.

As more researchers begin to use neural encoding models to study the neural basis of language, a thorough analysis of the settings of these models can aid in their proper application and replication. We hope this paper will encourage more discussions about the settings used in brain encoding analysis and improve the soundness of the research.

7. Bibliographical References

Bhaumik, D. K., Roy, A., Lazar, N. A., Kapur, K., Aryal, S., Sweeney, J. A., Patterson, D., and Gibbons, R. D. (2009). Hypothesis testing, power and sample size determination for between group comparisons in fmri experiments. *Statistical methodology*, 6(2):133–146.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, 63(1):289–300.

Caucheteux, C., Gramfort, A., and King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, pages 1336–1348. PMLR.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905.

Etzel, J. A. (2015). Mvpa permutation schemes: permutation testing for the group level. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 65–68. IEEE.

Hamilton, L. S. and Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience*, 35(5):573–582.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Ivanova, A. A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., and Isik, L. (2021). Is it that simple? linear mapping models in cognitive neuroscience. *bioRxiv*.

Jain, S., Vo, V. A., Mahto, S., LeBel, A., Turek, J. S., and Huth, A. (2020). Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. In *Advances in Neural Information Processing Systems*, volume 33.

Lohmann, G., Stelzer, J., Lacosse, E., Kumar, V. J., Mueller, K., Kuehn, E., Grodd, W., and Scheffler, K. (2018). Lisa improves statistical analysis for fmri. *Nature communications*, 9(1):1–9.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25.

Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman,

S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., and Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., and Fedorenko, E. (2020). Artificial neural networks accurately predict language processing in the brain. *bioRxiv*.

Sun, J., Wang, S., Zhang, J., and Zong, C. (2020). Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):589–603.

Wang, S., Zhang, J., Lin, N., and Zong, C. (2020). Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5):9201–9208.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, 92:381–397.

Zhang, Y., Han, K., Worth, R. M., and Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11(1):1877.