IST-01 Kamitani Lab Group Seminar Report

SUN YAQI

6930-35-7347

Intelligence Science and Technology Course

Introduction

The blood-oxygen-level-dependent (BOLD) response measured by functional MRI (fMRI) is a noisy and indirect measure of neural activity from which researchers try to infer neural function. The fMRI signal reflects the sum of the activity of around a million neurons within a cubic volume called a 'voxel'. With thousands of voxels sized 2mm * 2mm * 2mm, fMRI records brain activity at two-second intervals as a 4-dimensional brain data. The first three dimensions are spatial, and the 4th is temporal. Under these conditions, we can design experiments, obtain brain data from the subjects when they are performing the cognitive tasks.

Research purpose

Human can quickly convert the stimuli of seeing an image or hearing a sentence into a semantic understanding. In other words, we can easily judge whether a sound caption correctly describes a picture when the information from the image is clear enough. However, if we preprocess the picture in some way, such as gray scaling or blurring, our understanding of the image becomes hindered, and we may not be able to judge the consistency between a caption and the image.

In this experiment, a subject volunteered to perform this task. We hope to find some brain regions from brain data whose activity distinguishes the consistency between the caption and the image, which may point to the neural locations in the brain that handle semantic conflicts.

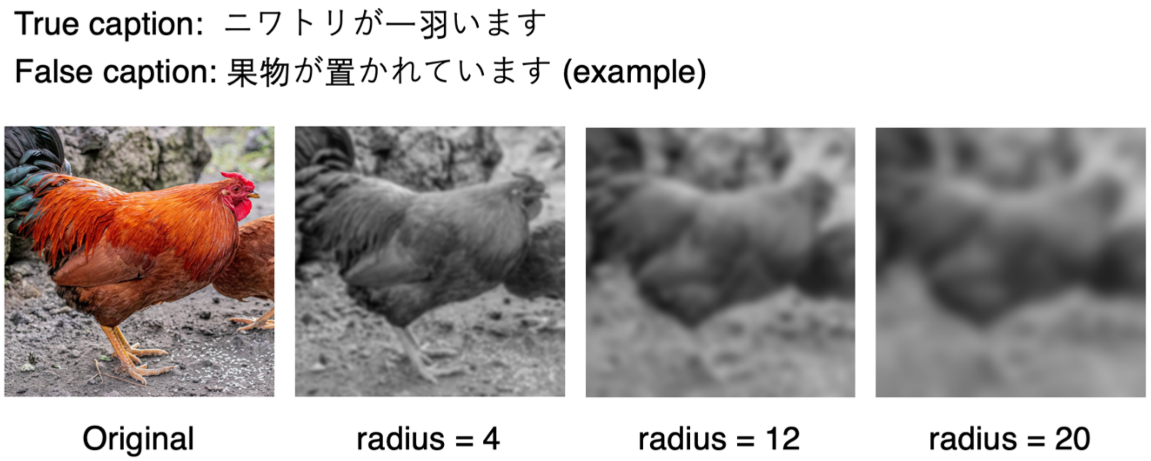
Furthermore, we want to compare the differences in brain activity when making confident judgments versus when unable to make judgments. We are curious to see if the brain data still provides directional information when we are subjectively unable to judge consistency.

Experiment

Dataset setup

First, we collected approximately 400 diverse images ranging from people, animals, food, to vehicles from the MS COCO (Microsoft Common Objects in Context) dataset and free image websites like Unsplash. Each image was manually annotated with a correct (True) caption.

Then, we processed the images, cropping all of them into squares and gray scaling them, as colored images can be very informative in some situations. We blurred the images using Gaussian filters with radius of 4, 12, and 20 at a 3:1:1 ratio. Different radiuses of an image will not be used at the same time. The effect of the image after processing with different radius is shown in Figure 1.



True caption: ニワトリが一羽います
False caption: 果物が置かれています (example)

Figure 1. An example of image processing and its True/False caption.

Finally, we assigned True/False captions at a 1:1 ratio. True captions correspond to the manually annotated descriptions that accurately depict the images, while False captions correspond to descriptions that are unrelated to the images. To create the False captions, we selected half of the images and randomly shuffled their corresponding True captions, ensuring that the shuffled captions did not accidentally match the new images. To generate sound for the assigned True/False captions, we used the 'say' command in macOS, which is a text-to-speech function that can read the text you entered. Each sound clip is less than 4 seconds in duration.

In summary, the prepared dataset for the experiment ultimately includes 360 images, with 180 images labeled with 'true' (corresponding to a True caption) and 180 images labeled with 'false' (corresponding to a False caption). 216 images are labeled with 'radius04', and 72 images each labeled with 'radius12' and 'radius20', distributed evenly among the True/False labels. The final label format for each pair of image and sound caption is *{serial number}_{radius04/ radius12/ radius20}_{true/false}*.

Process

In the real fMRI experiment, a subject answered the consistency questions for all 360 pairs of image and sound caption.

The entire experiment consisted of 12 runs, with each run having 30 trials. Each trial had three phases: t1, t2, and t3, with each phase lasting 4 seconds. During the t1 phase, only the image was visible; in the t2 phase, the image disappeared, and the sound caption was played; in the t3 phase, the participant responded to the question 'Does the image match the caption?'. The order of the trials was randomized and did not follow the serial number sequence. After each run, the participant could choose to take a break. Figure 2 illustrates the flow of a trial.

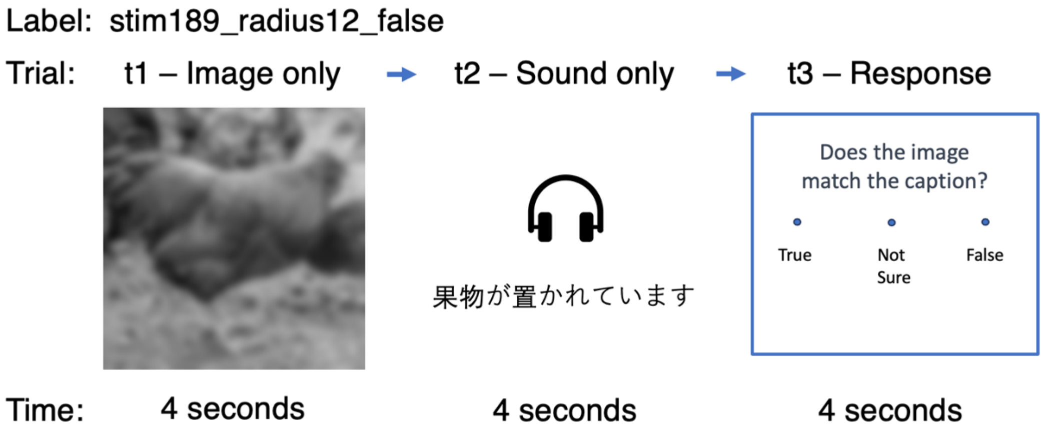


Figure 2. An example of a trail in fMRI experiment.

Result

Figure 3 summarizes the relationship between the consistency of the image and caption and the subject's response. Consistency has two possible settings (True/False), while the response has three options (1/2/3) representing the subject's judgment on consistency. 1 represents positive, 2 represents uncertainty, and 3 represents negative. Therefore, combining the 'True/False' label and response, there are six possible *Label_Response* combinations.

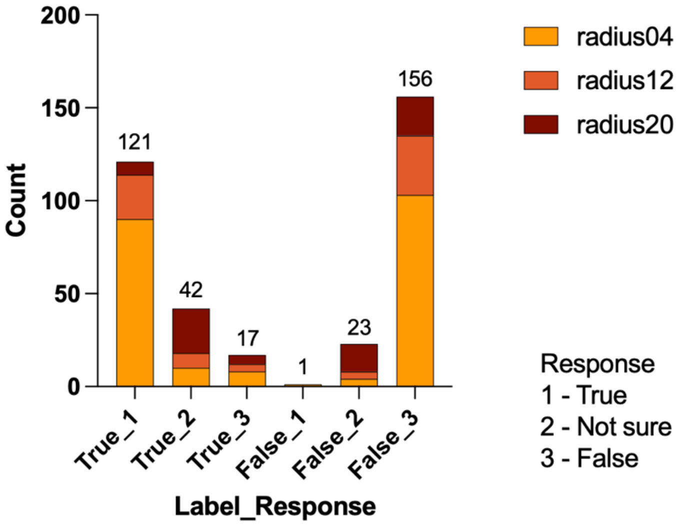


Figure 3. Experiment summary.

In agreement with the experimental setup, blurrier pictures (radius12, radius20) will make the subject more inclined to choose the uncertain response. Because each picture has a different difficulty of comprehension after a blur of the same radius according to the complexity of the picture. For example, human faces are still recognizable after radius20. Therefore, the above distribution is acceptable.

From this summary, we can additionally see that it is more difficult to judge the image and caption to be consistent than inconsistent. This is because, when we created the False captions, we did not manually annotate a plausible annotation; instead, we randomly assigned one. In most of the cases, the False caption has a very low correlation with the image.

Finally, the distribution of *Label_Response* for each run in Figure 4 shows that subject's performance is relatively stable, and there is not a single run with particularly low performance, so the experimental data does not need to be screened. The only False_1 was due to insufficient preparation of the data, which the False caption appeared to be close to the Ture caption of the image. The effect of False_1 will be discarded in the following analysis.

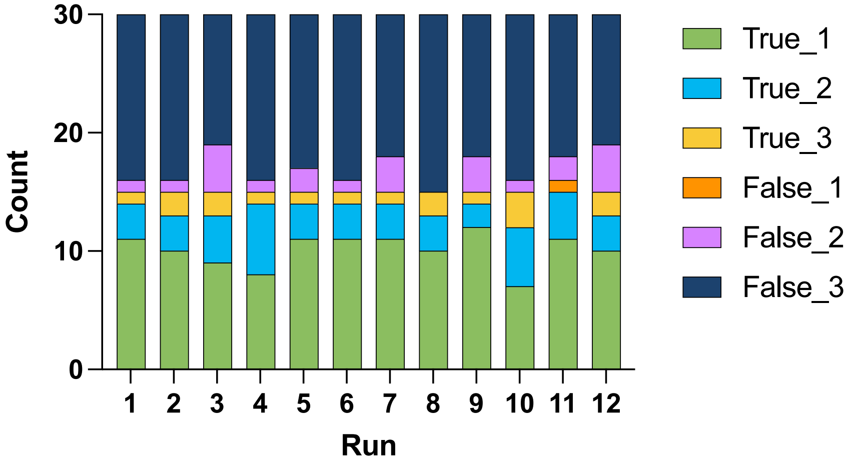


Figure 4. Experiment summary for each run.

Brain Data Analysis

Classifier model

The data distribution from this experiment is rather imbalanced. In some binary classifications, we had to use under-sampling, which involves randomly deleting samples from the majority class. However, employing under-sampling can result in the loss of information that might be invaluable to the model. Therefore, in binary classifications with a smaller sample size, we adopted a strategy of repeating random under-sampling several times and taking the average.

We utilized the Support Vector Machine (SVM) as the classifier because it can handle very high dimensions in the feature space and is very straightforward to apply. When employing SVM, since we deal with many voxels simultaneously, the feature selection (choice of voxels) plays a pivotal role in determining the classification accuracy. A common method, for instance, involves selecting voxels in a region of interest (ROI) based on known brain functional areas.

Model evaluation was carried out using 5-fold cross-validation, with Accuracy as the metric. Since we applied under-sampling, both the training and test sets are balanced. As a result, one can intuitively understand the model's performance compared to the chance level (1/2 for a binary classifier and 1/3 for a 3-class classifier). This is also the reason we didn't use over-sampling: with over-sampling, the test set is imbalanced, making it infeasible to use Accuracy and chance level as evaluation criteria.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Consistency classifier with confidence

In our experimental hypothesis, we hoped to detect differences in consistency after t2 (when the caption is heard). From the figure 5, although t2 achieved an accuracy of 0.61 in the inferior parietal cortex (IPC), and the IPC is known for its roles in basic attention, language, and multimodal sensory integration. Unfortunately, t2 did not show persuasive performance across all ROIs.

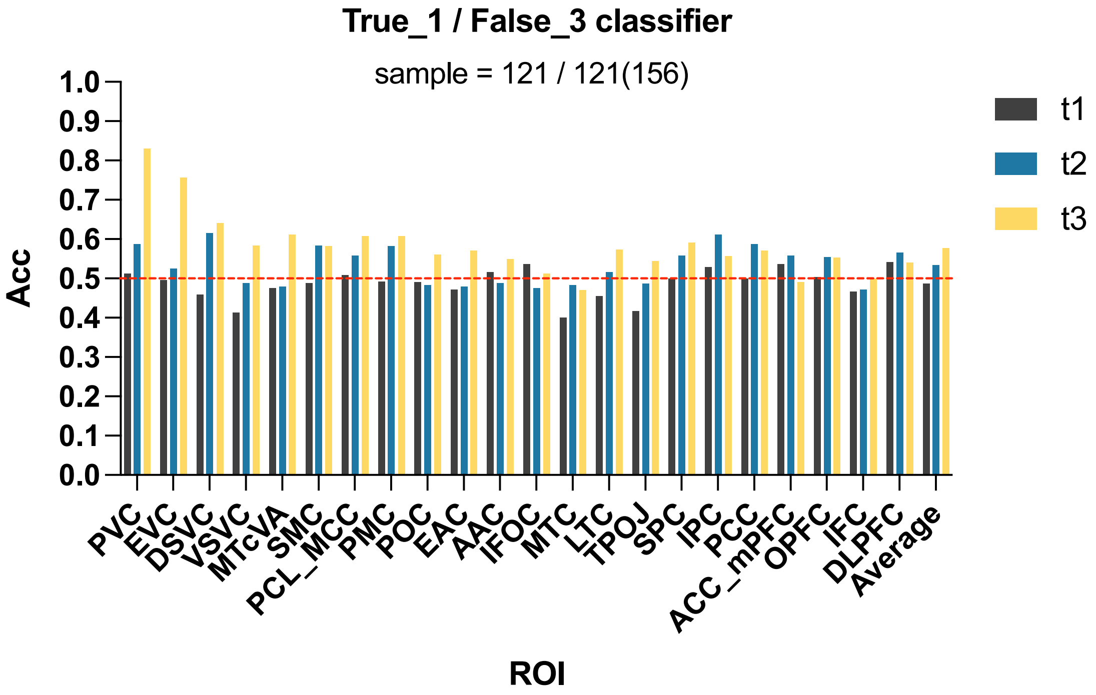


Figure 5. True_1 / False_3 classifier. *Sample = 121 / 121(156)* means False_3 is the majority class and was under-sampled to the same size of True_1. 156 is the actual size of False_3.

It is worth noting that we speculate the short duration of t2, being only 4 seconds with the average caption lasting around 3 seconds, means participants may not have heard the complete caption early in t2 and by the time they decide the judgment, t2 has already ended. In the following phase (t3), we observed a significant rise in accuracy in regions such as PVC, EVC, DSVC, and VSVC, which are involved in processing visual inputs. However, during t3, the subject would be looking at options on the left or right side of the screen, the observed increase might be influenced by the actions made during t3.

To verify this possibility, we can compare the True_1 / False_3 classifier with the True_1 / True_3 classifier in Figure 6, which similarly had options 1 and 3 on either side during the t3 phase. For this classifier, due to the particularly low sample size of True_3, we repeated it five times. We also observed a rise during t3, but this increase did not reach an accuracy greater than 0.8 as seen in figure 5. It is possible that during t3, some special features between consistency and inconsistency was detected. However, this change may overlap with the visual shifts caused by making a choice, and given the small sample size, it is challenging for us to determine the extent of this impact.

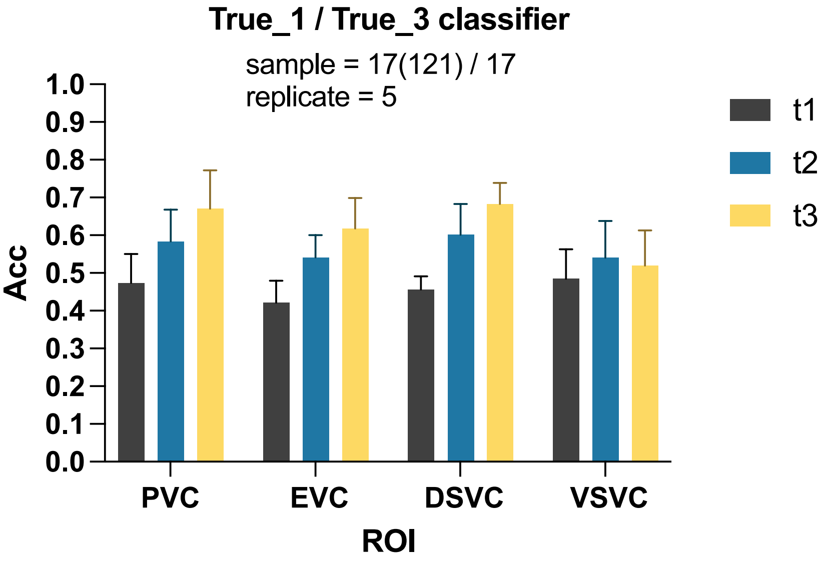


Figure 6. True_1 / True_3 classifier.

Consistency classifier with uncertainty

In situations where the Consistency classifier with confidence cannot make meaningful classifications, we can hardly expect the True_2 / False_2 classifier to perform well, as seen in Figure 7. Moreover, both True_2 and False_2 chose the central 'not sure' option during the t3 phase. Naturally, no increase was observed in the PVC and other ROIs during t3.

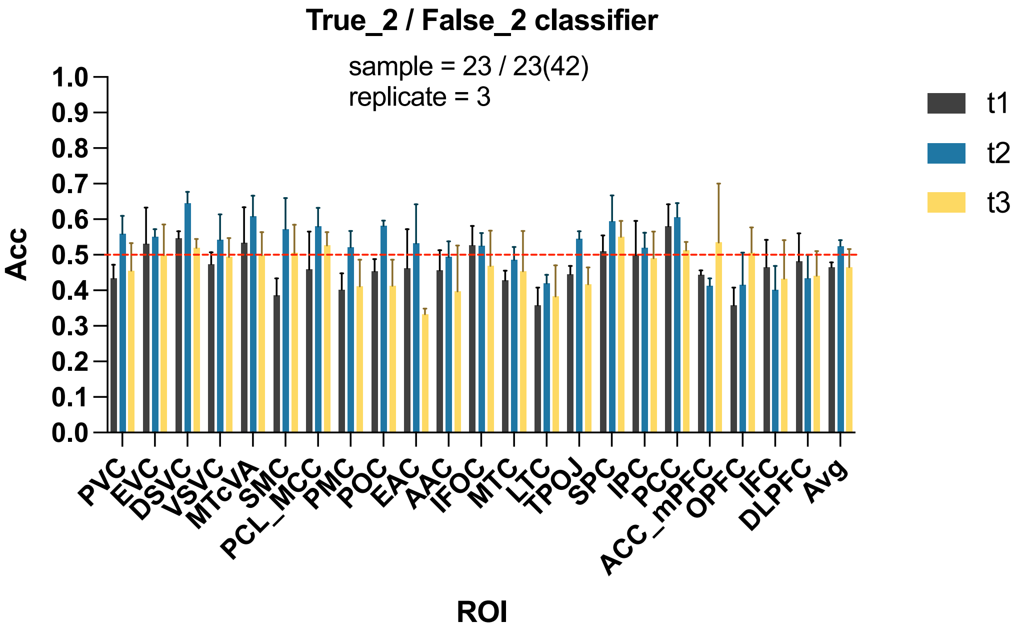


Figure 6. True_2 / False_2 classifier.

Radius classifier

We have images processed using three different radiuses of Gaussian filter. What difference in brain data can be observed when looking at these three types of images? Refer to Figure 7, the Radius classifier displayed performance greater than the chance level in PVC and EVC. Additionally, after t1, the image disappears, and the classification performance gradually decreases during t2 and t3, which aligns with the experimental facts.

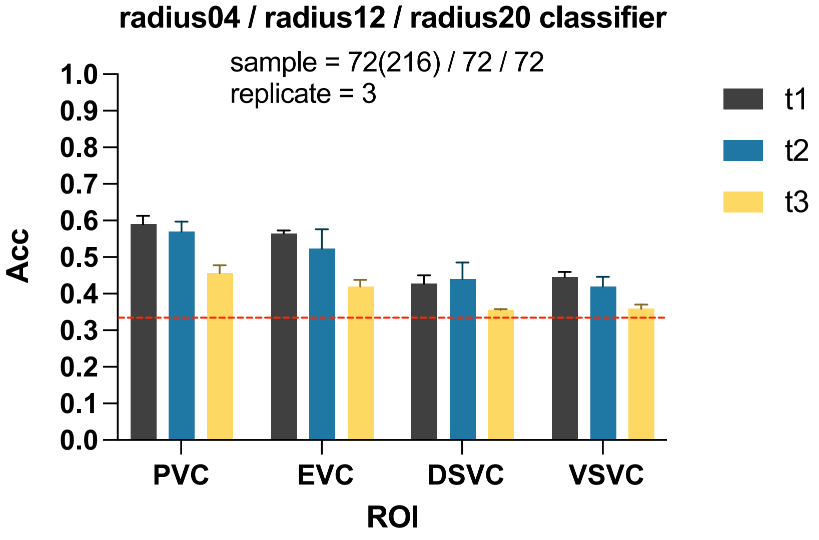


Figure 7. Radius classifier.

Confidence classifier

Is the brain's state different when the subject chose 'not sure' compared to confidently selecting 'True' or 'False'? To increase the sample size of small samples, considering that the classification performance of the True_2 / False_2 classifier is very close to the chance level, regardless of what the original label was, the subject chose 'not sure'. Therefore, in the confidence classifier, we combined True_2 and False_2 into one category. The results revealed that the classifier's accuracy exceeded 0.7, even 0.8, in PVC, EVC, and DSVC during t2 and t3.

However, this result is constrained by many factors, and we cannot derive a confident conclusion. For instance, in the 'not sure' class, the proportion of radius12 and radius20 is higher than that of True_1. This suggests that the performance exceeding the chance level in t1 and t2 might be contributed by the radius classifier. While the contribution of the radius is more prominent in t1 and t2, the high accuracy observed in t3 might also result from the subject choosing different options during the t3 phase.

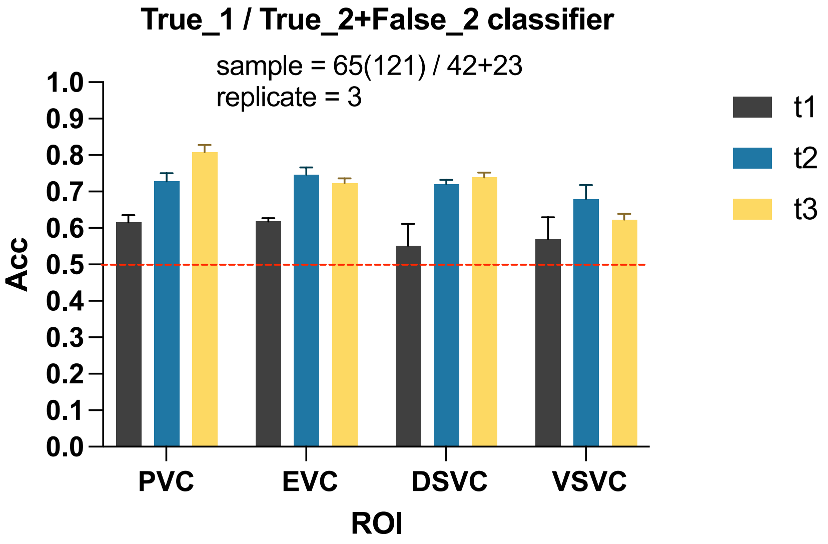


Figure 8. True_1 / True_2+False_2 classifier.

Discussion

The design of the experiment itself is relatively novel. Judging from the distribution of *Label_Response*, the grayscale and blur strategy effectively achieved the goal of quantitatively altering confidence. However, setting up an imbalanced dataset in a small experiment with a limited overall sample size was somewhat short-sighted. It happened because, in the beginning, we had high expectations for the True_1/False_3 classifier. When this classifier did not meet our expectations, it became challenging to analyze other minority classes adequately. For example, initially, changing the ratio of images with a radius of 4, 12, and 20 from 3:1:1 to 1:1:1 might have made the distribution more balanced.

Furthermore, since we did not account for the time it takes to listen to the caption in t2, adding a 2-second thinking interval would better exclude the effect of eye shifts in t3.

Lastly, it is possible that, for the brain, interpreting an image or a sentence are more complex and independent processes. Subsequent semantic comparison might be a relatively effortless task, not resulting in a significant impact. Thus, we might indeed be unable to observe this phenomenon at a resolution of 2 * 2 * 2 mm^3.

fMRI Experiment Methods

MRI acquisition

fMRI data were collected with a 3.0 Tesla MRI scanner (MAGNETOM Verio, Siemens) at Kokoro Research Center, Kyoto University. A multi-band interleaved T2*-weighted gradient-echo echo planar imaging (EPI) was performed to acquire functional images covering the whole brain (TR: 2000 ms; TE: 43 ms; flip angle: 80 degree; FOV: 192 × 192 mm; voxel size: 2 × 2 × 2 mm; slice gap: 0 mm; number of slices: 76; multiband factor 4). A T1-weighted magnetization-prepared rapid acquisition gradient-echo (MP-RAGE) image was acquired for each subject and used as an anatomical reference of the individual brain (TR: 2250 ms; TE: 3.06 ms; TI: 900 ms; flip angle: 9 deg; FOV: 256 × 256 mm; voxel size: 1.0 × 1.0 × 1.0 mm).

MRI data preprocessing

fMRI data were preprocessed with fmriprep (Esteban et al., 2019). Field map correction, head motion and slice timing correction were applied on fMRI images. fMRI images were then aligned to an anatomical reference of the individual brain.

fMRI signals were further preprocessed before the analysis. fMRI signals were temporally shifted 4s (2 volumes) to compensate for hemodynamic delays. Then, nuisance parameters (estimated six head motion parameters, linear trend, run-wise mean) were regressed out from the signal. Outliers in the signal of each voxel were clipped and replaced with the threshold value (above ±3 SD in each run). Finally, fMRI signals in the same stimulus presentation block (X s, N volumes) were temporally averaged.

Region of interest

For each subject, ROIs were defined based on anatomical alignment of HCP-180 parcellations (Glasser et al., 2016) to the subject’s brain.

Reference

Esteban et al. (2019) fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods, 16, 111-116. doi: 10.1038/s41592-018-0235-4

Glasser et al., (2016) A multi-modal parcellation of human cerebral cortex. Nature, 536, 171-178. doi: 10.1038/nature18933