Neuroscience Letters 783 (2022) 136709

Contents lists available at ScienceDirect

Neuroscience Letters

journal homepage: www.elsevier.com/locate/neulet

Research article

Testing the distributed representation hypothesis in object recognition in two open datasets

Shen Zhang [a,b,c], Zilu Liang [a,b,c], Chao Liu [a,b,c,*]

[a] State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China
[b] Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, 100875 Beijing, China
[c] Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, 100875 Beijing, China

ARTICLE INFO

Keywords:
Distributed representation
Multi-variate pattern analysis
Machine learning
Object recognition
Multi-variate connectivity

ABSTRACT

Neural representation has long been thought to follow the modularity hypothesis, which states that each type of information corresponds to a specific brain area. Though supported by many studies, this hypothesis surfers the pitfall of inefficiency for information encoding. To overcome difficulties the modularity representation hypothesis faced, researchers have proposed that information may be distributed represented in a specific brain area. The distributed representation hypothesis along with the multi-variate pattern approaches have made great success in detecting representation patterns in the previous decade. However, this hypothesis implicitly requires that the pattern should be transformed in a consistent way with respect to all of the represented information in the specific brain area. And the accuracy and validity of this prediction have never been thoroughly tested. Here in the present study, we tested this prediction in two open datasets compiling the object recognition. We validated the distributed representation patterns in the lateral occipital complex/ventral temporal gyrus where all six classifiers were capable of predicting the correct category represented. Furthermore, we correlated the classifiers' decision function values to the bold signals and found that the decision function value of the logistic regression classifier was exclusively correlated with activities of the same brain area in both datasets. These results support the distributed representation hypothesis and suggest that our neural system may be embedded within the algorithm of a specific classifier.

1. Introduction

Where and how our neural systems encode information are fundamental questions of neuroscience. One prevailing hypothesis of these questions is the modularity hypothesis which states that one specific brain area is responsible for the processing of a specific category of information. However, in the first trial to identify specific area for a memory trace, researchers found that the amount of memory retained for a specific event was only correlated with the size of the brain tissue removed but not the location of the brain area [25]. This instead support the distributed representation hypothesis which assumes information was encoded distributedly across all brain areas. The first convinced evidence for modularity hypothesis came from a lesion observation which found that impairment of a part of the inferior frontal gyrus disables the production of language. This area was later named Broca's area in memory of its discoverer. Recently, brain regions encoding a specific type of stimulus can be easily detected with functional magnetic resonance imaging (fMRI) using the general linear model (GLM) methods (i.e., the univariate method) [9]. In the GLM framework, neural responses to a given category of information are contrasted to some baseline condition, yielding brain regions where the activity is correlated with thereby processing this category of information. With this univariate method, in tasks require object recognition, researchers have revealed that the information of objects was encoded in the ventral temporal cortex [11,12], and within this area, the encoding of face information was located at a subarea called the fusiform face area [22]. Along with several other univariate studies, these findings largely support the modularity hypothesis.

However, the distributed representation hypothesis is yet abandoned. Though the modularity hypothesis and the univariate methods have made great success detecting brain areas responding to specific categories of stimuli, it is insufficient for encoding all the complex

* Corresponding author at: State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China.
E-mail address: liuchao@bnu.edu.cn (C. Liu).

https://doi.org/10.1016/j.neulet.2022.136709
Received 4 November 2021; Received in revised form 30 May 2022; Accepted 31 May 2022
Available online 3 June 2022

information in the world. As Haxby pointed out [14], the univariate methods tend to ignore submaximal responses and implicitly hypothesize that these weaker but significant brain responses play no roles in representation. This hypothesis seems improbable on two counts. First, this method is inefficient in information representation in the sense that it discards information in the submaximal responses thereby contradicting the method in analyzing population response representation. Second, it seems impossible that there exists a specific brain region for each stimulus category as any given face or object can be treated as a category [14]. In light of these defects, the distributed representation hypothesis was then refined, assuming that encoding of different stimuli can be in a shared brain region but is represented by different activation patterns [16]. To test the distributed representation hypothesis, multivariate approach (i.e. the Multivariate pattern analysis) was developed to detect the distributed representation of information [13,17,30].

This approach was formally expressed as machine learning methods. Usually, neural responses of one brain region of interest (ROI) from part of the data were used to train a classifier, and the prediction of the classifier was tested on the rest. Hig her than chance-level performance of the classifier reveals that the stimuli are distributed represented in this brain area. After it was proposed, the MVPA methods have been widely used in cognitive neuroscience [15,29] and found that a large amount of information such as object categories [16], emotion states [2] and decision values [21] are all distributedly represented in the brain. However, comprehensively testing of the distributed representation requires answering the question of how and where these representations are transferred to. That is, if information is indeed distributedly represented in some brain areas, the neural system must be able to read out the information in a consistent way for different stimuli. This means that, for a specific brain area, there should be a transformation method that can be applied to all distributed representations of information, and the transformed information should be transferred to the same brain area for all representations.

In the machine learning framework of MVPA, the neural response of a brain region can be treated as a point in a high-dimensional Euclidean space and the classifiers made their decision of what this neural response represent based on a scalar (i.e., the decision function value or decision value for short) representing the distance of this point to a high-dimensional plane (i.e., the decision plane).The decision function value of classifiers serves as a good candidate for what the distributed representation information transfer to. Indeed, researchers have used the decision function value of support vector machine (SVM) as a regressor in the GLM to explore where the distributed representation patterns are transformed [4]. However, this study didn't fully test the distributed representation hypothesis, they only used one classifier, (i.e., the linear SVM), and tested this multi-variate pattern transformation in only one dataset of two mental states (thus two categories of information). Fully testing the hypothesis requires finding a consistent way of information transformation for multiple categories. Furthermore, as the brain area may implement multiple information transformation methods, more than one classifier should be tested.

To fill this gap, in the present study, we fully tested the distributed representation hypothesis in two open datasets. All datasets tackled the object representation in the ventral pathway. The first dataset (the exploratory dataset) focused on a subset of animals, and the second (the confirmatory dataset) explored a subset of furniture. Specifically, in the first dataset, we identified several brain areas that respond to pictures of animals. Several classifiers were trained to detect what the activation patterns of these brain areas represent. And decision values of all these classifiers were submitted to a general linear model to find brain areas whose activity was correlated to the decision values. Then, we tried to replicate the results in the exploratory dataset, again we correlated brain activities with decision function values when the participants looked at the picture. We argue that it would be convincing evidence of the distributed representation if there exist one or more classifiers whose decision function value was correlated with the activities of the same brain areas across both datasets.

2. Methods

2.1. Exploratory dataset analyses.

2.1.1. Exploratory dataset

Detailed descriptions of the dataset are in the Openneuro reservoir https://openneuro.org/datasets/ds000241/versions/00002. Briefly, participants completed a recognition memory task with a slow event-related design in which pictures of six animal categories were presented (Fig. 1A). On each trial, six encoding events, each from one animal type, were followed by one probe event. In each encoding event, three exemplars of the animal were presented each for 500 ms without gaps. And in the probing event, another three exemplars (i.e., a probe) from one animal category were presented, participants were asked to determine whether the probe was identical to one of the encoding events. All events are followed by a fixation of 4500 ms. Twelve participants were included in the dataset.

2.1.2. Preprocessing

All fMRI analyses were performed using SPM12 (Wellcome Trust Center for Neuroimaging). Functional images were first slice-timing and then motion-corrected by first realigning the first image of each run to the first image of the first run and then all other images were realigned to the first image of each run. T1-weighted image was then coregistered to the mean functional image, and then segmented using SPM12′s new segmentation method. Deformation information images generated from the new segmentation procedure were then applied to each functional image to transform them to the standard MNI space, while resampling to a voxel size of 3 mm × 3 mm × 3 mm. All functional images were then smoothed with a Gaussian kernel of 8-mm full width at half maximum.

2.1.3. Localizer (GLM-1)

GLM-1 was aimed to find brain regions that respond to animals in the exploratory dataset. It was built on the smoothed images, we added the following regressors for each run: R1 was an indicator function for encoding events consisting of boxcar function for each event lasts for 1.5 s (same as the events' duration). R2 was an indicator function for the exemplars in probe events, which was modeled as a boxcar function of 1.5 s started from the onsets of the first exemplar. These two regressors were convolved by the SPM's canonical hemodynamic response function. Nuisance regressors included six head-motion parameters resulting from the realignment process, and run specific constant. A high-pass filter of 128 s was applied to remove low-frequency drifts and the GLM also included an AR(1) regressor to control for the fMRI's autocorrelations. SPM's default restricted maximum likelihood estimation was used to estimate all GLMs. We performed first-level contrast of encoding events for each participant (R1). This contrast compares the neural response at the time of encoding to the implicit baseline (i.e. neural response to the jitter). Then contrast images from all participants were submitted to the second-level analysis (SPM's standard summary statistics method) and one-sample test was used to search for the brain areas involved in the encoding of animals. We applied a peak level $p < 0.05$ family-wise error correction to control for the multiple comparisons.

2.1.4. Reliability Check

It's widely believed that the signal to noise ratio of the event related design is relatively low, hence, to check the reliability of our following decoding analysis, we estimated the temporal signal to noise ratio (tSNR) and contrast to noise ratio (CNR) of the data. For each voxel, the tSNR was calculated as its averaged bold signals divided by the standard deviation of the signal. The smoothed data was used to compute the tSNR. We computed the tSNR for each run of each subject, and used the median of the all runs as the tSNR estimates for each subject [38]. Then

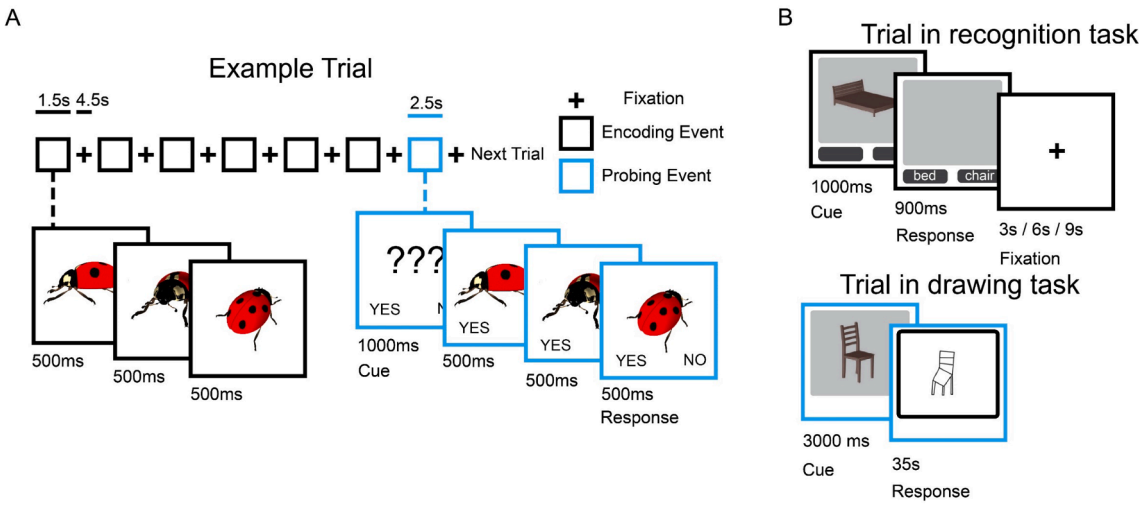

Fig. 1. Illustration of tasks in two datasets. A) Timeline of trials in the exploratory dataset. One trial consists of seven events of which six are encoding events, and one is the probing event. In each encoding event, participants saw three exemplars of one animal category each for 500 ms. In the probing event, a probing cue was presented for 1000 ms followed by three exemplars of an animal category each for 500 ms. Participants were asked to determine whether the three exemplars were identical to any of the encoding events in this trial. **B)** Timeline of trials in the confirmatory dataset. Top panel, trial in the recognition task. A picture of one piece of furniture was presented for 1000 ms, then participants were asked to choose the correct category of the furniture. Bottom panel, trials in the drawing task. Picture of one piece of furniture was presented for 300 ms, then participants were given 35 s to draw this furniture.

median across all subjects serves as a group tSNR estimate.

The tSNR measures how large the noise is regardless of the task design, we further used the contrast-to-noise ratio (CNR) to measure how reliable the contrast we interested in is. This CNR is defined as the mean signal difference between the condition we interested and the baseline condition divided by the standard deviation of the noise. In the current study we estimated the CNR by the contrast value of the encoding events divided by standard deviation of GLM1′s residuals. And as the scans of the bold signal is large, we thus estimated the CNR as the t-value of the contrast. We calculated CNR for each subject, and the Median across all subjects were again used as the group level CNR estimate.

2.1.5. *Trial-wise responses to object (GLM-2)*

GLM-2 estimated trial-wise activation for each object category, preparing data for the decoding analysis. Thus, GLM-2 was built on the un-smoothed data. It had the following regressors: R1s were a series of regressors each modeled an encoding event. Each regressor was a boxcar function with the same onset and duration (i.e., 1.5 s) of the event. R2 was an indicator function for the exemplars in all probe events, each lasted for 1.5 s from the onset of the first exemplar to the end of the probe. Nuisance regressors were also included for head-motion, run specific constants, AR(1) autocorrelations. The bold signal was high-pass filtered with a 128 s filter. Trial-wise neural responses to the objects were estimated using SPM's restricted maximum likelihood estimation.

2.1.6. *Decoding analyses*

All machine learning algorithms were implemented using the nilearn package and the scikit-learn package in Python. To explore the distributed representation in the lateral occipital complex and ventral temporal areas (LOC/VT area), we trained several classifiers and leveraged a leave one run out cross-validation strategy to test the sanity of classifiers. Specifically, we considered the following widely used classifiers: C1 which was a correlation classifier, C2 was a logistic regression, C3 was a logistic regression with L1 regularization, C4 was a logistic regression with l2 regularization, C5 was a support vector machine with linear kernel, and C6 was a support vector machine with non-linear kernel (the default rbf kernel).

In C1 (Correlation classifier), we averaged neural response across trials for each object category in the ROI and used this averaged neural activation as an activation template. Then, voxel-wise correlations between sample data and each activation template of an object category were defined as the decision function for this category. All other classifiers are natively binary, so scikit-learn's one vs. rest method was used to construct a corresponding multiple-class classifier for C2-C6. In this method, a two-class classifier was trained for each category, and the most confident category was used to predict the test sample data. Geometrically, the decision function of each category for C2-C6 is thus a signed distance between the sample data to a decision boundary for this category. For the linear classifier, the decision boundary is a plane while for the non-linear classifier it is a curved surface in a high dimensional Euclidean space. In the one vs. rest method used in scikit-learn, the category with the highest decision function value was chosen as the predicted category.

In the localizer analyses, two clusters survived the corrections of multiple comparisons (the bi-lateral lateral occipital complex/ ventral temporal cortex (LOC/VT) areas), then trial-wise BOLD responses from all voxels from these clusters were extracted as features for decoding. As the exploratory dataset was fully balanced in the sense that all object categories were presented the same times in each run and all runs had the same trials, we used decoding accuracy as our cross-validation measure. We adopted a leave-one-run-out approach to compute the decoding accuracy in which each run was chosen once as the testing data while data from all other runs were used as training data. Decoding accuracies were tested across participants using one-sample *t*-test against chance level. To further confirm the significance of the decoding, we used the following permutation methods to determine the p-value for each classifier. For each participant, we randomly assigned labels to data 100 times to generate 100 datasets, classifiers were then trained on these datasets to generate 100 null accuracies. A group-level null accuracy was generated for each classifier by averaging across one randomly sampled null accuracy from each participant. Then 10,000 group level null accuracy were generated to create an empirical null distribution for each classifier, and true group average accuracy was compared to this null distribution to determine the p-value. To further confirm the reliability of the decoding results, we compute the confusion matrix for each classifier. Specifically, for each object category revealed to the participants, the frequency of the categories predicted by each classifier were computed.

2.1.7. Information transformation and transfer

We then asked how the distributed representation is transformed and where this information is transferred. Critically, each classifier adopts a different information transformation method. All classifiers make their decisions based on the value of the decision function which is a transformed information of the distributed representation. Thus, the neural response to this decision-function value may serve as an indicator that which neural system implements the algorithm of classifiers. Notably, different classifiers reach their accuracies adjusting the decision boundaries by different loss functions. So even all linear classifiers apply a plane as their decision boundary, the decision value could differ a lot. This significant property of the classifiers allows us to dissociate information transformed by different classifiers.

Following these rationales, we fitted six GLMs (GLM-3 to GLM-9) each for a classifier. For each GLM, three regressors of interest were added. R1 was an indicator function for all encoding events which was modeled as boxcar function lasts from the onsets of the event and for 1.5 s. R2 was a parametric modulator of R1 with the decision function value of the classifier for the real category. R3 was a parametric modulator of R1 with a decision function value of the predicted category. Parametric modulators were not orthogonalized between each other. All other omitted details were the same as GLM-1. We were interested in two contrast that looked for areas whose BOLD response correlates with decision function values of the true category (i.e. the picture presented) or the predicted category (i.e., the category with the highest decision function value). We identified regions that passed whole-brain cluster correction at $p < 0.05$ combined with a voxel-wise threshold of $p < 0.005$. We applied this lenient voxel-wise threshold as the dataset only has 12 participants, a strict threshold may miss important effect.

2.2. Confirmation data analysis.

As our deduction, the distributed representation hypothesis requires all distributed representation in the same brain areas to be read out by the same transformation methods and to the same brain area. Thus, we repeated our analyses in a second dataset in which a different object subset (i.e. furniture) was used as stimuli. It has been reported that furniture like chairs or beds shared the same brain area with animals [16]. These analyses of the confirmatory dataset would provide a powerful test of the distributed representation hypothesis.

2.2.1. Confirmatory dataset (dataset 2)

The second dataset tackled the effect of drawing on object recognition (Fig. 1B, See https://openneuro.org/datasets/ds002241/versions/1.1.0 for details). Specifically, participants completed 6 runs of recognition task, and 2 runs of drawing task were performed after the fourth recognition run. Pictures of four furniture categories were presented to participants during the experiment. On each trial of the recognition run, a picture of the furniture was presented for 1000 ms followed by a 900 ms probe asking the participants to select the correct category of the furniture from two alternatives. In the drawing task, one picture of the furniture was presented for 3000 ms followed by a time window in which participants drew the picture for 35 s. Only two categories of the furniture were presented in the drawing runs, referred to as the trained categories. Thirty-one participants were included in the analyses of the original paper, in which data of 3 participants are incomplete in the reservoir, resulting in a data sample of 28 participants in our analyses. As for the exploratory dataset, we estimated the tSNR and CNR of the confirmatory dataset similarly as described in the exploratory dataset analysis.

2.2.2. Trial-wise responses to object (GLM-2′)

For all analyses ran on the confirmatory dataset, only recognition runs were included. As regressors of interests, GLM-2′ consisted of a series of indicator functions each was a boxcar function that lasts from trial onsets to the end (i.e. for 1.9 s). Other omitted details are the same as GLM-2. We again estimated trial-wise BOLD response to the objects' picture using SPM's restricted maximum likelihood estimation.

2.2.3. Decoding analysis

As a replication, we used brain regions identified in the localizer analysis for the exploratory dataset as the regions of interest for feature selection. All classifiers used in the exploring analysis were considered, and the same inference methods were used when comparing the accuracy of classifiers to the chance level. Similar to the exploratory dataset, confusion matrix was also calculated for each classifier.

2.2.4. Information transformation and transfer

Brain areas encoding decision function value for all six classifiers were detected in GLM-3′ to GLM-9′ (corresponding to GLM-3 to GLM-9). Similar to the exploratory dataset, we included the following regressors for each GLM: R1 was an indicator function for all encoding events which was modeled as boxcar function last from the onsets of the event and for 1.5 s. R2 was a parametric modulator of R1 with the decision function value of the classifier for the true category. R3 was a parametric modulator of R1 with a decision function value of the predicted category. Parametric modulators were not orthogonalized between each other. All other omitted details were the same as GLM-2. We were interested in two contrast that looked for areas whose BOLD response correlates with the decision function value of the true category or the predicted category.

We applied a more rigorous voxel-wise threshold of $p < 0.001$ for the confirmatory dataset, and a cluster-wise family-wise error correction of $p < 0.05$ was used to correct for multiple comparisons. To find a robust representation of the decision function value for a classifier, we tried to find brain regions that are both identified (i.e. overlapped) in two datasets. Furthermore, in order not to miss important effects, we used the brain regions encoding a specific decision function value of a classifier in the exploratory dataset as a searching volume to look for brain regions encoding this decision function value in the second dataset. Small volume correction was used to control for multiple comparisons when searching volumes were applied.

3. Results

3.1. Localizer and the signal to noise ratio

In the localizer analyses, only two clusters survived the multiple comparison correction, including the bilateral lateral occipital complex/ventral temporal gyrus (LOC/VT; Fig. 2A; Left part, peak voxel MNI coordinates: [-36, −58, −16], peak $t = 16.43$, $k = 180$; Right part, peak voxel MNI coordinates: [33, −79, 11], peak $t = 25.10$, $k = 407$). The tSNR of the exploratory dataset were shown in Fig. 2B. The tSNR of most voxels in the LOC/VT areas were above 100 (Fig. 2C), indicating a generally good signal to noise ratio. Moreover, the more relevant CNR of the LOC/VT areas were the largest of the whole brain (Fig. 2D and Fig. 2E). These results guarantee the reliability of our subsequent decoding analyses and thus the decision value representation analyses.

3.2. Distributed information representation.

As expected, all classifiers performed better in predicting the category of the stimuli than by chance (Fig. 3A, Table 1). Furthermore, confusion matrix for all classifiers showed that the true positive rate was the largest both in its role and column, indicating a good sensitivity and specificity (Fig. 3B).

3.3. Encoding of decision function value of different classifiers.

For the exploratory dataset, representation of the decision function value of different classifiers was provided in Table 2. However, no decision value representation was found for the true category. And we

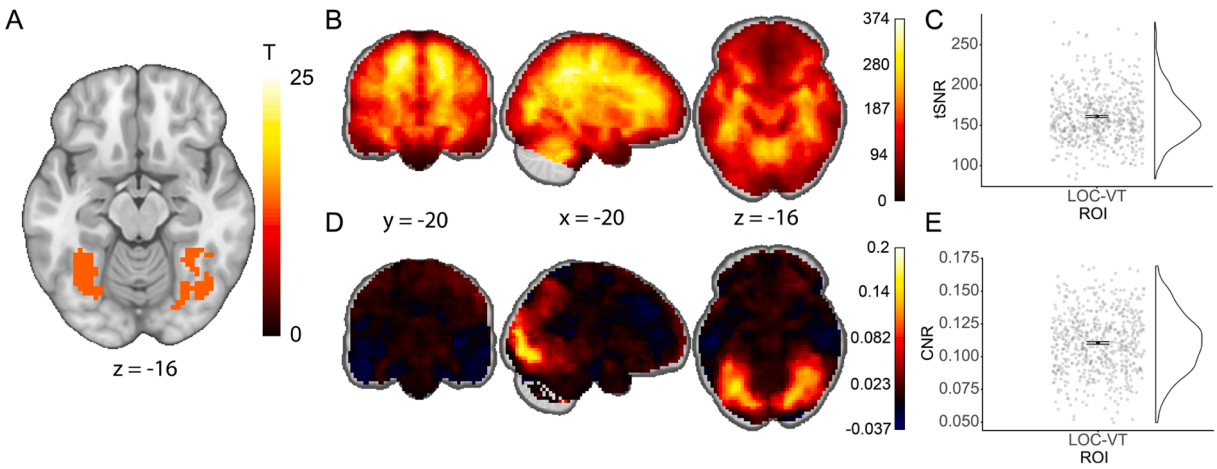

Fig. 2. The signal to noise ratio in the LOC/VT areas. A) Localizer analyses revealed significant brain response in the LOC/VT area to objects. **B)** Temporal signal to noise ratio of the whole brain and **C)** the LOC/VT areas. **D)** Contrast to noise ratio of the whole brain and **E)** the LOC/VT areas. Error bars in **C)** and **E)** indicate s.e.m. Black points indicate mean.

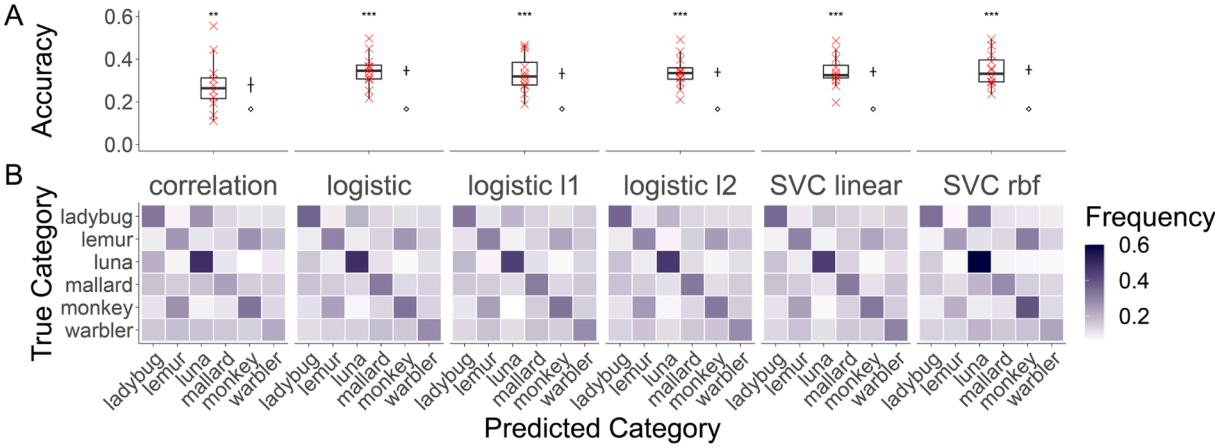

Fig. 3. Distributed representation of objects. A) All classifiers revealed higher than chance accuracy in the exploratory dataset. Boxplots show the accuracy of each classifier. Black crosses indicate mean and s.e.m. Diamonds indicate the mean of the non-distribution created from permutations. **p < 0.01, ***p < 0.001, all one-sample *t*-test. **B)** Confusion matrix showed good sensitivity and specificity for all classifiers.

Table 1
Classifiers' accuracy compared to chance level using *t*-test and permutations.

Classifier	T value	$p_{t\text{-test}}$	$p_{permutation}$
Correlation	3.16	0.009	< 0.0001
Logistic regression	8.08	< 0.001	< 0.0001
Logistic L1 regression	6.49	< 0.001	< 0.0001
Logistic L2 regression	7.87	< 0.001	< 0.0001
SVC linear	7.90	< 0.001	< 0.0001
SVC rbf	8.11	< 0.001	< 0.0001

found no decision value representation with respect to SVM classifiers.

3.4. Signal to noise ratio of the confirmatory dataset.

As the exploratory dataset, for the confirmatory dataset, the tSNR of most voxels in the LOC/VT areas were also above 100 (Fig. 4B), indicating a generally good signal to noise ratio. And the CNR of the LOC/VT areas were also among the largest of the whole brain (Fig. 4C and Fig. 4D).

3.5. Confirmation of distributed information representation.

We found that even in the confirmatory dataset, the LOC/VT areas identified in the exploratory dataset still represented the category information in a distributed manner which was evident that all classifiers reached higher than chance accuracy when predicting the correct category (Table 3, Fig. 5A). Also, confusion matrix results showed relatively good sensitivity and specificity for all classifiers (Fig. 5B).

3.6. Confirmation of encoding of decision function value

For the whole-brain analyses, brain areas representing each decision function value were reported in Table 4. The overlapped regions between two datasets were the superior temporal gyrus representing the decision function value of logistic regression (Fig. 6A, B, C). Furthermore, within the area almost all participants showed positive representation of this decision function value (Fig. 6D, E), revealing a high consistency between participants. We didn't find any brain areas after the multiple comparison correction in analyses of searching volumes.

4. Discussion

In the present study, we tested the distributed representation hy-

Table 2
Brain areas representing decision function value of each classifier in the exploratory dataset.

Classifier		Brain regions	Peak MNI coordinates x	y	z	T-value	Cluster size
Correlation							
	True category	None					
	Predicted category	Post central gyrus	3	−40	77	6.83	197
Logistic regression							
	True category	None					
	Predicted category	Left anterior insula	−42	11	−1	6.10	161
		Pre-central gyrus	15	2	62	5.84	246
		Dorsal anterior cingulate cortex	0	23	29	5.83	140
		Right fusiform gyrus	15	−67	−7	5.76	135
		Right superior temporal gyrus	60	−31	11	5.51	225
		Right anterior insula	39	8	−19	5.32	169
		Left inferior frontal gyrus	−63	−25	23	4.87	177
Logistic L1 regression							
	True category	None					
	Predicted category	Right superior temporal gyrus	57	−19	17	6.39	328
		Right anterior insula	33	11	14	6.20	291
		Left occipital gyrus	−48	−79	23	5.35	113
		Left inferior frontal gyrus	−57	−28	23	5.01	144
Logistic L2 regression							
	True category	None					
	Predicted category	Right Cerebelum	33	−52	−28	5.87	179
Linear SVM							
	True category	None					
	Predicted category	None					
Non-linear SVM (cbf kernal)							
	True category	None					
	Predicted category	None					

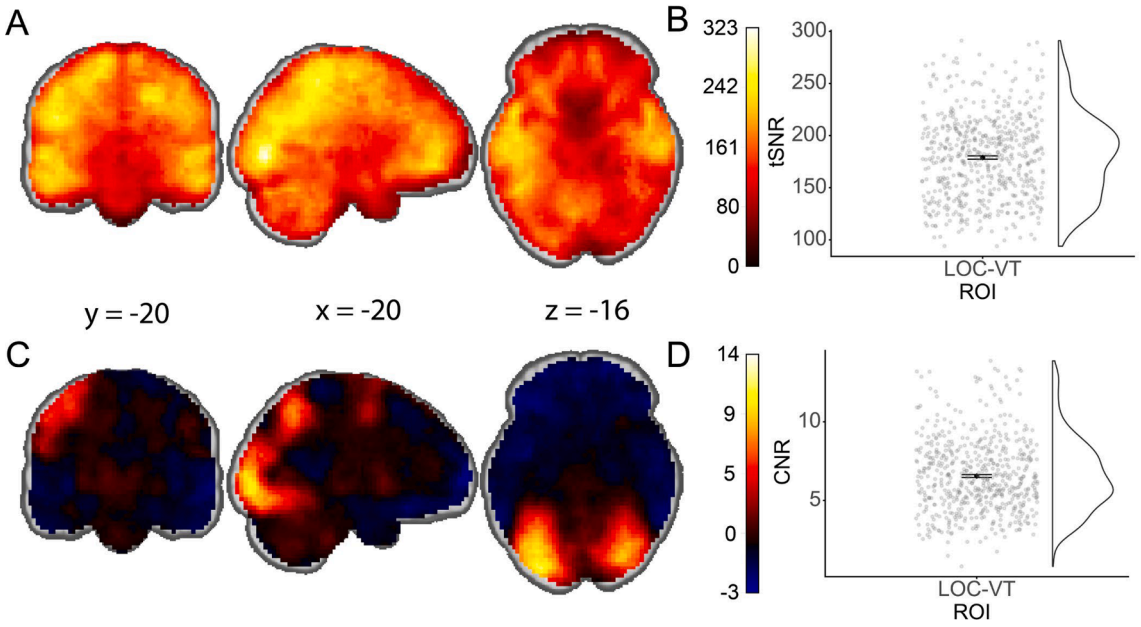

Fig. 4. Signal to noise ratio of the confirmatory dataset. A) Temporal signal to noise ratio of the whole brain and **B)** the LOC/VT areas. **C)** Contrast to noise ratio of the whole brain and **D)** the LOC/VT areas.

Table 3
Classifiers' accuracy compared to chance level using *t*-test and permutations.

Classifier	T value	$p_{t\text{-test}}$	$p_{permutation}$
Correlation	2.92	0.007	0.0019
Logistic regression	5.82	< 0.001	< 0.0001
Logistic L1 regression	5.22	< 0.001	< 0.0001
Logistic L2 regression	5.64	< 0.001	< 0.0001
SVC linear	5.04	< 0.001	< 0.0001
SVC rbf	8.58	< 0.001	< 0.0001

pothesis in two datasets. We first confirmed that the object categories were distributed represented in the LOC/VT areas, and we found that the decision function value of the logistic regression was correlated with the activities in the superior temporal gyrus in both datasets. Distributed representation hypothesis argued that the activation in one brain area may represent different stimuli in the way that activation of each voxel encodes an attribute of the stimulus [16]. These attributes should be combined to form a scalar signal to identify the stimulus. Thus, there should be a consistent way of transforming the distributed representation to a scalar for all stimuli represented in the same brain areas. To test

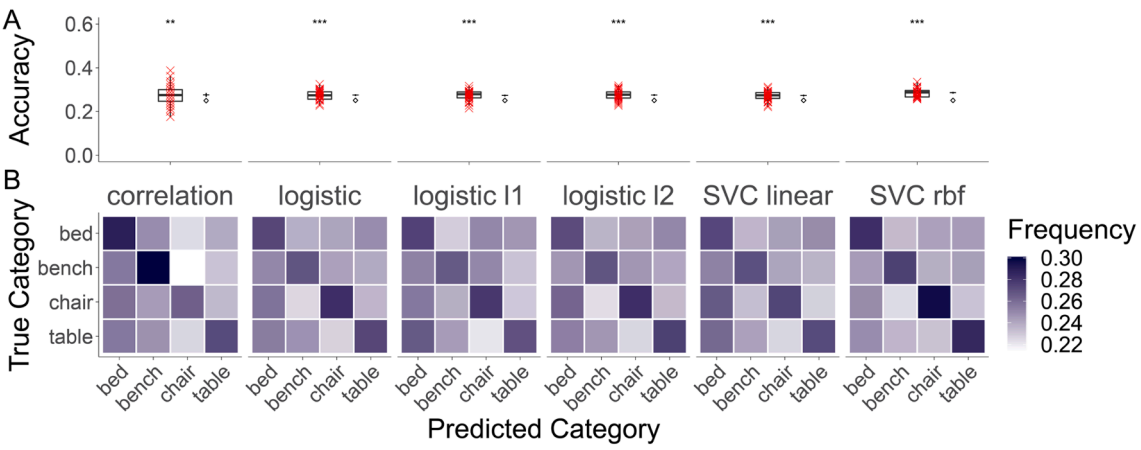

Fig. 5. Classification accuracy of all classifiers in the confirmatory dataset. A) All classifiers reached higher than chance accuracy. Black crosses indicate mean and s.e.m of the accuracy. Diamonds show the mean value of the null distribution created by permutations. **p < 0.01, ***p < 0.001, one-sample *t*-test. **B)** Confusion matrix showed good sensitivity and specificity for all classifiers.

Table 4
Brain areas representing the decision function value of each classifier in the confirmatory dataset.

Classifier		Brain regions	Peak MNI coordinates			T-value	Cluster size
			x	y	z		
Correlation							
	True category	None					
	Predicted category	None					
Logistic regression							
	True category	None					
	Predicted category	Right Superior temporal gyrus	63	−22	17	5.57	170
Logistic L1 regression							
	True category						
	Predicted category	None					
Logistic L2 regression							
	True category	Left dorsal lateral prefrontal gyrus	−27	50	26	4.68	90
	Predicted category	None					
Linear SVM							
	True category	Left dorsal lateral prefrontal gyrus	−27	53	23	4.91	165
	Predicted category	None					
Non-linear SVM (rbf kernal)							
	True category	None					
	Predicted category	None					

this hypothesis, we analysised two open datasets. The tSNR of the LOC/VT area in these two datasets were larger than 100 in these two datasets. It has been suggested that, for tSNR of 100, to detect an effect of 0.5% at $p = 5 \times 10^{-10}$ level requires about 1200 scans [10]. Scan numbers of these two datasets we considered are comparable to this amount, indicating a good power to detect small changes in BOLD signals. Moreover, we also calculated the CNR as a measure more related to the task. The CNR results showed that the LOC/VT areas we found in the localizer analysis had the largest CNRs among all voxels in the brain. Together with the tSNR results, this indicate that our decoding analysis and subsequent decision value representations are reliable.

Our results support the distributed representation hypothesis and we found that the potential transformation used when reading out the information in LOC/VT area is the logistic regression. Recently, several methods were proposed to investigate how multi-variate activation patterns were transformed to another multi-variate activation pattern [1,3,5,23,39]. These multi-variate connectivity methods provide another test tool for the distributed representation hypothesis. However, we choose the present method over these methods as the overlap between multi-variate patterns is harder to detect. To confirm the distributed representation in the LOC/VT areas, six classifiers were considered. The first five classifiers (C1-C5) are linear in the sense that the multi-voxel distributed information was transformed by a linear mapping (i.e., the decision surface is linear). The last classifier (C6) is an SVM of rbf kernel, which used a hyperplane to classify two classes, thus a non-linear classifier. A significant difference between each class of classifier is the way they adjust their decision boundaries. The classifiers we considered adjust the decision boundary by different loss functions (e.g., L1 or L2 regularization). So even classifiers with similar accuracy could have different decision values. This enable us to test different information transformation mechanism through the decision values of classifiers.

We found no brain area whose neural responses correlated with the decision function value of the SVM with rbf kernel even at a lenient voxel-wise threshold of 0.005. It has been proposed that, to recognize objects our neural system may apply multiple transformations to information received in which the last step of the transformation should be linear [6]. Our results support this idea by identifying the right superior temporal gyrus (rSTG) as the only region that responds to the decision value of a linear classifier (i.e. the logistic regression). To further explore the lateralization of this process, we also compare the decision value representation difference between the rSTG and the lSTG (the symmetrical part of the area rSTG). We found that the difference of decision value representation between these two areas was only revealed in the confirmatory dataset at 0.05 level which support neither strong symmetry or asymmetry.

The logistic regression use a logit function as its link function which is widely used in economic decision-making tasks. In these tasks, it has

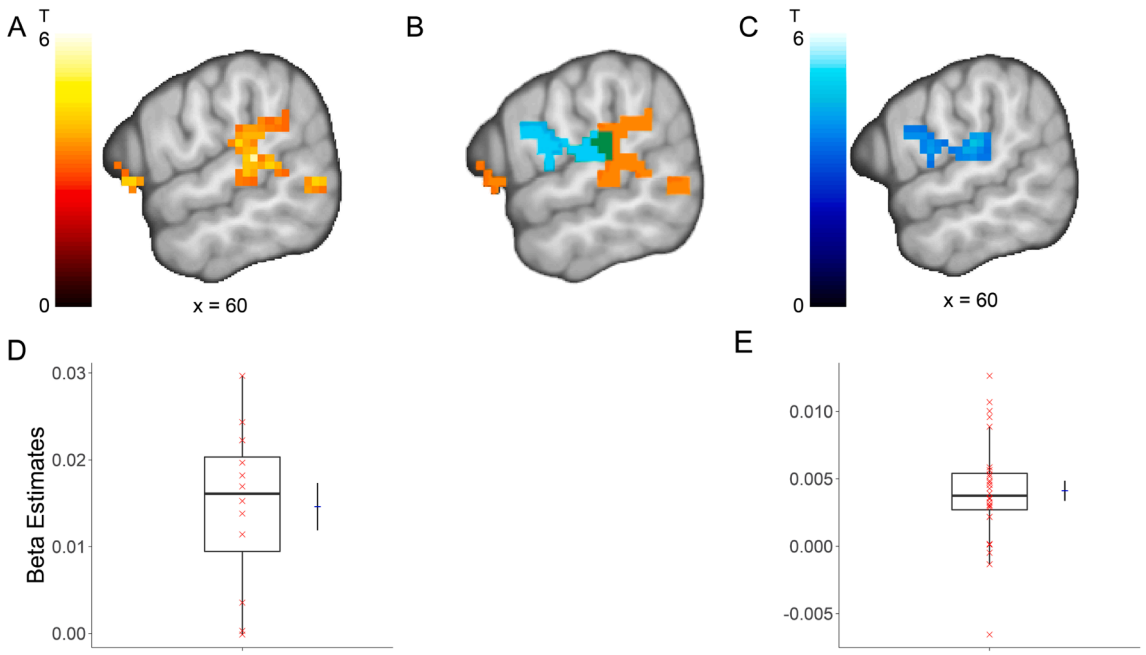

Fig. 6. Overlap of Distributed information transformation. The decision function value of logistic regression of predicted category was represented in the superior temporal gyrus both in the **A)** exploratory dataset and the **C)** confirmatory dataset. **B)** Overlap of areas in **A)** and **C)**. **D)**. Beta estimates extracted from **A)**. **E)** Beta estimates extracted from **C)**. Black crosses in **D)** and **E)** represent s.e.m. and mean.

been reported that the neural responses of the ventral medial prefrontal cortex (the vmPFC) at the time of decision were correlated with the decision value [31,33–34]. Our linear transformed neural responses mimic this decision value in the sense that the information is transformed from neural responses of other brain regions rather than environmental input. Another difference between classifiers is the regularization methods. The L1 regularization shrink less important feature weights to zero, serving as a method of feature selection [36], while the L2 regularization only lowers the weights of features of less importance [18,19]. We didn't find a reliable representation of the decision value of neither L1 nor L2 regularization logistic regression, this may indicate most features in the VT/LOC areas are important for the representation. However, the VT/LOC areas were identified in the localization analysis, which means we already performed data-specific feature selection in the analyses pipeline. This may be the cause of the failure to find brain responses related to the regularized logistic regression. Using the decision function value of a linear SVM, researchers have identified multi-variate pattern connectivity between the medial superior parietal lobule (mSPL) to the middle frontal gyrus (MFG), the superior frontal gyrus (SFG), the caudate, and the cuneus [4]. However, in the present study no regions' activities were correlated with the decision function value of SVM in all datasets. This seemingly contradictory result may be due to the different information transformations applied by different brain regions. Further studies would be needed to delineate the correspondence between classifiers and brain regions implementing them.

Our results support the distributed representation hypothesis, but don't fully falsify the modularity hypothesis. In fact, these two hypotheses need not contradict each other, such as, at the global level, the ventral pathway is processing what the object is while at the local level the LOC/VT areas represent the information of the object in a distributed way. It's still an open question of where the scaled boundary of these two types of representations lies. While some studies focused only on the distributed representation at a small local area [8,32,35], other studies found that representation of some information (e.g., emotional state) was distributed over the brain [2,24,37]. Factors influencing the scale of the distributed representation may include the modality or complexity of the information, which requires further exploring.

It has been reported recently that, neural network models trained to gain human-compatible performance can predict neural responses in particular brain areas [41]. And the fact that different neural networks can reach similar performance calls for a necessity for direct comparison between neural network algorithms [27]. Following these studies, instead of training the algorithms to recognize stimuli, we used several classifiers to decode the neural activities and correlated their decision function value to neural responses. Our methods are similar to fitting a single perceptron with different threshold functions. This pipeline compares the model transformed information directly to the neural activities searching for a brain feasible machine learning algorithm. As all classifiers reach similar performance to the logistic regression but only decision function value of logistic regression correlated robustly with some brain area, our results further emphasize directly correlating algorithm predicted information to real neural responses.

We showed that the distributed representation in the VT/LOC area may be transformed by a linear mapping, but how this linear mapping is implemented in the brain remains unclear. It has been proposed that linear mapping of information can be realized as multiple weighted neural connections [7,6]. These neural connection weights corresponded to the linear coefficients of the linear mapping. Thus, to implement linear mapping, neural connection weights should be correlated with linear mapping coefficients. Future study is needed to test this prediction. Furthermore, we only considered one possible non-linear transformation and found no evidence of its implementation. And most of the existing studies focus on the linear relationship between information and its transformation [40]. However, the non-linear transformation of information is common in the neural systems [7,20,26,28], The question of how non-linear transformation is implemented in the brain is still open.

5. Conclusion

Our results support the distributed representation hypothesis by showing that the distributed representation in the VT/LOC area can be read out by the same classifier (i.e., the logistic regression) and the decision function value of the classifier is represented in an overlapped area across two datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (31871094, 32130045), the Major Project of National Social Science Foundation (19ZDA363), the Beijing Municipal Science and Technology Commission (Z151100003915122), the National Program for Support of Top-notch Young Professionals. We thank Zhiqiang Sha for his valuable comments.

References

[1] S. Anzellotti, A. Caramazza, R. Saxe, Multivariate pattern dependence, PLoS Comput. Biol. 13 (11) (2017) 1–20, https://doi.org/10.1371/journal.pcbi.1005799.
[2] Y.K. Ashar, J.R. Andrews-Hanna, S. Dimidjian, T.D. Wager, Empathic Care and Distress: Predictive Brain Markers and Dissociable Brain Systems, Neuron 94 (6) (2017) 1263–1273.e4, https://doi.org/10.1016/j.neuron.2017.05.014.
[3] A. Basti, M. Mur, N. Kriegeskorte, V. Pizzella, L. Marzetti, O. Hauk, Analysing linear multivariate pattern transformations in neuroimaging data, PLoS ONE 14 (10) (2019) 1–23, https://doi.org/10.1371/journal.pone.0223660.
[4] Y.C. Chiu, M.S. Esterman, L. Gmeindl, S. Yantis, Tracking cognitive fluctuations with multivoxel pattern time course (MVPTC) analysis, Neuropsychologia 50 (4) (2012) 479–486, https://doi.org/10.1016/j.neuropsychologia.2011.07.007.
[5] M.N. Coutanche, S.L. Thompson-Schill, Informational connectivity: Identifying synchronized discriminability of multi-voxel patterns across the brain, Front. Hum. Neurosci. 7 (JAN) (2013) 1–14, https://doi.org/10.3389/fnhum.2013.00015.
[6] J.J. DiCarlo, D.D. Cox, Untangling invariant object recognition, Trend Cognit. Sci. 11 (8) (2007) 333–341, https://doi.org/10.1016/j.tics.2007.06.010.
[7] J.J. DiCarlo, D. Zoccolan, N.C. Rust, How does the brain solve visual object recognition? Neuron 73 (3) (2012) 415–434, https://doi.org/10.1016/j.neuron.2012.01.010.
[8] T. Ethofer, D. Van De Ville, K. Scherer, P. Vuilleumier, Decoding of Emotional Information in Voice-Sensitive Cortices, Curr. Biol. 19 (12) (2009) 1028–1033.
[9] K.J. Friston, A.P. Holmes, K.J. Worsley, J.-P. Poline, C.D. Frith, R.S.J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, Hum. Brain Mapp. 2 (4) (1994) 189–210, https://doi.org/10.1002/hbm.460020402.
[10] J. Gonzalez-Castillo, V. Roopchansingh, P.A. Bandettini, J. Bodurka, Physiological noise effects on the flip angle selection in BOLD fMRI, NeuroImage 54 (4) (2011) 2764–2778, https://doi.org/10.1016/j.neuroimage.2010.11.020.
[11] K. Grill-Spector, N. Knouf, N. Kanwisher, The fusiform face area subserves face perception, not generic within-category identification, Nat. Neurosci. 7 (5) (2004) 555–562, https://doi.org/10.1038/nn1224.
[12] K. Grill-Spector, Z. Kourtzi, N. Kanwisher, The lateral occipital complex and its role in object recognition, Vision Res. 41 (10–11) (2001) 1409–1422, https://doi.org/10.1016/S0042-6989(01)00073-6.
[13] M. Hanke, Y.O. Halchenko, P.B. Sederberg, S.J. Hanson, J.V. Haxby, S. Pollmann, PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data, Neuroinformatics 7 (1) (2009) 37–53, https://doi.org/10.1007/s12021-008-9041-y.
[14] J.V. Haxby, Multivariate pattern analysis of fMRI: The early beginnings, NeuroImage 62 (2) (2012) 852–855, https://doi.org/10.1016/j.neuroimage.2012.03.016.
[15] J.V. Haxby, A.C. Connolly, J.S. Guntupalli, Decoding Neural Representational Spaces Using Multivariate Pattern Analysis, Annu. Rev. Neurosci. 37 (1) (2014) 435–456, https://doi.org/10.1146/annurev-neuro-062012-170325.
[16] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, P. Pietrini, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex, Science 293 (5539) (2001) 2425–2430, https://doi.org/10.1126/science.1063736.
[17] M.N. Hebart, K. Görgen, J.-D.-D. Haynes, J. Dubois, The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data, Front. Neuroinf. 8 (January) (2015) 1–18, https://doi.org/10.3389/fninf.2014.00088.
[18] A.E. Hoerl, R.W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems, Technometrics 12 (1) (1970) 69–82, https://doi.org/10.1080/00401706.1970.10488635.
[19] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12 (1) (1970) 55–67, https://doi.org/10.1080/00401706.1970.10488634.
[20] H. Hong, D.L.K. Yamins, N.J. Majaj, J.J. Dicarlo, Explicit information for category-orthogonal object properties increases along the ventral stream, Nat. Neurosci. 19 (4) (2016) 613–622, https://doi.org/10.1038/nn.4247.
[21] Kahnt, T. (2018). A decade of decoding reward-related fMRI signals and where we go from here. *NeuroImage*, *180*(June 2017), 324–333. Doi: 10.1016/j.neuroimage.2017.03.067.
[22] N. Kanwisher, J. McDermott, M.M. Chun, The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception, J. Neurosci. 17 (11) (1997) 4302–4311, https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997.
[23] P.A. Kragel, M. Čeko, J. Theriault, D. Chen, A.B. Satpute, L.W. Wald, M. A. Lindquist, L. Feldman Barrett, T.D. Wager, A human colliculus-pulvinar-amygdala pathway encodes negative emotion, Neuron 109 (15) (2021) 2404–2412.e5, https://doi.org/10.1016/j.neuron.2021.06.001.
[24] P.A. Kragel, M. Kano, L. Van Oudenhove, H.G. Ly, P. Dupont, A. Rubio, C. Delon-Martin, B.L. Bonaz, S.B. Manuck, P.J. Gianaros, M. Ceko, E.A. Reynolds Losin, C. W. Woo, T.E. Nichols, T.D. Wager, Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex, Nat. Neurosci. 21 (2) (2018) 1–7, https://doi.org/10.1038/s41593-017-0051-7.
[25] K.S. Lashley, Mass Action in Cerebral Function, Science 73 (1888) (1931) 245–254, https://doi.org/10.1126/science.73.1888.245.
[26] N.J. Majaj, H. Hong, E.A. Solomon, J.J. DiCarlo, Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance, J. Neurosci. 35 (39) (2015) 13402–13418, https://doi.org/10.1523/JNEUROSCI.5181-14.2015.
[27] J. Mehrer, C.J. Spoerer, N. Kriegeskorte, T.C. Kietzmann, Individual differences among deep neural network models, Nat. Commun. 11 (1) (2020) 1–12, https://doi.org/10.1038/s41467-020-19632-w.
[28] E.H. Nieh, M. Schottdorf, N.W. Freeman, R.J. Low, S. Lewallen, S.A. Koay, L. Pinto, J.L. Gauthier, C.D. Brody, D.W. Tank, Geometry of abstract learned knowledge in the hippocampus, Nature 595 (7865) (2021) 80–84.
[29] K.A. Norman, S.M. Polyn, G.J. Detre, J.V. Haxby, Beyond mind-reading: multi-voxel pattern analysis of fMRI data, Trend. Cognit. Sci. 10 (9) (2006) 424–430, https://doi.org/10.1016/j.tics.2006.07.005.
[30] N.N. Oosterhof, A.C. Connolly, J.V. Haxby, CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave, Front. Neuroinf. 10 (July) (2016) 27, https://doi.org/10.3389/fninf.2016.00027.
[31] A. Rangel, C. Camerer, P.R. Montague, A framework for studying the neurobiology of value-based decision making, Nat. Rev. Neurosci. 9 (july) (2008) 545–556, https://doi.org/10.1038/nrn2357.
[32] N.W. Schuck, M.B. Cai, R.C. Wilson, Y. Niv, Human Orbitofrontal Cortex Represents a Cognitive Map of State Space, Neuron 91 (6) (2016) 1402–1412, https://doi.org/10.1016/j.neuron.2016.08.019.
[33] W. Schultz, Multiple reward signals in the brain, Nat. Rev. Neurosci. 1 (3) (2000) 199–207, https://doi.org/10.1038/35044563.
[34] W. Schultz, Dopamine reward prediction-error signalling: A two-component response, Nat. Rev. Neurosci. 17 (3) (2016) 183–195, https://doi.org/10.1038/nrn.2015.26.
[35] S. Suzuki, L. Cross, J.P. O'Doherty, Elucidating the underlying components of food valuation in the human orbitofrontal cortex, Nat. Neurosci. 20 (12) (2017) 1780–1786, https://doi.org/10.1038/s41593-017-0008-x.
[36] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1) (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[37] I. Vilares, M.J. Wesley, W.-Y. Ahn, R.J. Bonnie, M. Hoffman, O.D. Jones, S. J. Morse, G. Yaffe, T. Lohrenz, P.R. Montague, Predicting the knowledge–recklessness distinction in the human brain, Proc. Natl. Acad. Sci. 114 (12) (2017) 3222–3227, https://doi.org/10.1073/pnas.1619385114.
[38] V. di Oleggio, M. Castello, V. Chauhan, G. Jiahui, M.I. Gobbini, An fMRI dataset in response to "The Grand Budapest Hotel", a socially-rich, naturalistic movie, Sci. Data 7 (1) (2020) 1–9, https://doi.org/10.1038/s41597-020-00735-4.
[39] C.W. Woo, L. Koban, E. Kross, M.A. Lindquist, M.T. Banich, L. Ruzic, J.R. Andrews-Hanna, T.D. Wager, Separate neural representations for physical pain and social rejection, Nature Commun. 5 (May) (2014), https://doi.org/10.1038/ncomms6380.
[40] D.L.K. Yamins, J.J. DiCarlo, Eight open questions in the computational modeling of higher sensory cortex, Curr. Opin. Neurobiol. 37 (2016) 114–120, https://doi.org/10.1016/j.conb.2016.02.001.
[41] D.L.K. Yamins, H. Hong, C.F. Cadieu, E.A. Solomon, D. Seibert, J.J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex, PNAS 111 (23) (2014) 8619–8624, https://doi.org/10.1073/pnas.1403112111.