Information Sciences 507 (2020) 256–272

Contents lists available at ScienceDirect

Information Sciences

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

Fine-grained neural decoding with distributed word representations

Shaonan Wang [a,b,*], Jiajun Zhang [a,b], Haiyan Wang [a,b,c], Nan Lin [d,e], Chengqing Zong [a,b,f]

[a] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[d] CAS Key Laboratory of Behavioural Science, Institute of Psychology, Beijing, China
[e] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China
[f] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China

ARTICLE INFO

Article history:
Received 12 December 2018
Revised 14 August 2019
Accepted 18 August 2019
Available online 19 August 2019

Keywords:
Neural decoding
fMRI word decoding
Word class
Stimuli paradigm
Word embedding models
Informative voxels

ABSTRACT

fMRI word decoding refers to decode what the human brain is thinking by interpreting functional Magnetic Resonance Imaging (fMRI) scans from people watching or listening to words, representing a sort of mind-reading technology. Existing works decoding words from imaging data have been largely limited to concrete nouns from a relatively small number of semantic categories. Moreover, such studies use different word-stimulus presentation paradigms and different computational models, lacking a comprehensive understanding of the influence of different factors on fMRI word decoding. In this paper, we present a large-scale evaluation of eight word embedding models and their combinations for decoding fine-grained fMRI data associated with three classes of words recorded from three stimulus-presentation paradigms. Specifically, we investigate the following research questions: (1) How does the brain-image decoder perform on different classes of words? (2) How does the brain-image decoder perform in different stimulus-presentation paradigms? (3) How well does each word embedding model allow us to decode neural activation patterns in the human brain? Furthermore, we analyze the most informative voxels associated with different classes of words, stimulus-presentation paradigms and word embedding models to explore their neural basis. The results have shown the following: (1) Different word classes can be decoded most effectively with different word embedding models. Concrete nouns and verbs are more easily distinguished than abstract nouns and verbs. (2) Among the three stimulus-presentation paradigms (picture, sentence and word clouds), the picture paradigm achieves the highest decoding accuracy, followed by the sentence paradigm. (3) Among the eight word embedding models, the model that encodes visual information obtains the best performance, followed by models that encode textual and contextual information. (4) Compared to concrete nouns, which activate mostly vision-related brain regions, abstract nouns activate broader brain regions such as the visual, language and default-mode networks. Moreover, both the picture paradigm and the

* Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Beijing 100190, China.

E-mail addresses: shaonan.wang@nlpr.ia.ac.cn (S. Wang), jjzhang@nlpr.ia.ac.cn (J. Zhang), wanghaiyan2015@ia.ac.cn (H. Wang), linn@psych.ac.cn (N. Lin), cqzong@nlpr.ia.ac.cn (C. Zong).

https://doi.org/10.1016/j.ins.2019.08.043

model that encodes visual information have stronger associations with vision-related brain regions than other paradigms and word embedding models, respectively.

1. Introduction

Mind-reading technology, i.e., technology that interprets what the human brain is thinking, has long been a tantalizing topic for researchers. Recent studies have indicated that, given human brain activity as measured by functional magnetic resonance imaging (fMRI), it is possible to predict the corresponding stimuli. A pioneering study by Mitchell et al. [24] shows that it is feasible to build an entire linear forward model linking experimental stimuli to brain activation data by leveraging a representation of the stimulus and learning the mapping between that representation and its effect on activation across the brain. This paradigm can not only help reading out associated words from brain activation data but also allow us to study human language representation at the voxel level across the entire brain. Follow-up studies adopt the same paradigm and focus on analyzing different properties of words such as domain and concreteness [4,5]. Recent studies have shown that decoding is possible for larger-unit verbal stimuli, such as phrases [8], sentences [3,32,39,42], or text fragments [16,44]. There is also much work on decoding pictures or videos from brain activations [18,27,28,40,42,46]. This paper focuses on fMRI decoding of words, which are the basic units of language.

Most word decoding studies use nouns as stimuli, exploring how to decode brain activation of nouns and how nouns are represented in the brain [9,17]. However, other word classes have rarely been studied. Although word classes are fundamental to the grammatical systems of human languages, there are debates on the neurobiological basis of different word classes. Semantic model lines of research suggest that topographical differences in brain activation are driven by semantics and not by lexical class. These studies show that brain regions reveal clear category differences between concrete nouns and verbs but not abstract ones [25,35,41]. In contrast, lexical-grammatical lines of research support the notion that the human brain has dissociative neural correlates of semantic processing of nouns and verbs regardless of whether the words are abstract or concrete [21,48]. The difficulty with such research lies in the complex confound between lexical category and semantic meaning. In contrast to these works, we study the similarities and differences among the three classes of words (nouns, verbs, and adjectives) from the perspective of decoding performance. We further analyze their corresponding informative voxels to explore the neural basis of the representations of different word classes.

Aside from the categories of the word stimuli, input modalities and word embedding models also play important roles in neural word decoding. Previous fMRI word decoding experiments use different input modalities, such as pictures [24,29,31], verbal words [4,5], audio recordings [12] or combinations thereof [32]. Among these modalities, picture and verbal stimuli have been the most common input choices. Few studies have explored the effect of different input modalities in brain activation decoding tasks, especially when coupled with different word embedding models. However, both embodied and symbolic theories of conceptual organization hold that there is partial sharing and partial differentiation between the neural activation patterns observed for concepts activated via different stimulus modalities [2,10,13]. Therefore, we assume a difference between different input modalities in both the decoding performance and the neural basis. To test this hypothesis, we use three input modalities, i.e., a picture with a target word, a highlighted word in a sentence, and a highlighted word in a cloud of related words, and conduct decoding experiments using different kinds of word embedding models. Subsequent analysis of informative voxels from the different input modalities can reflect the differences and commonalities between their neural bases.

Furthermore, word embeddings, computer-manipulable representations of word meaning, bridge the gap between brain activation and word stimuli. Adding this intermediate layer of analysis allows the computer to probe the images in ways no human could, breaking them down into component parts associated with each aspect of a word's meaning. Recently, with the rapid development of neural networks, a range of new word embedding methods have been proposed and successfully used in a variety of natural language processing tasks [23,30,33]. At the same time, studies have shown that such word embedding models fit the neural activation data very well and significantly surpass human-elicited features [1]. The basic assumption here is that the better the performance of a model is, the more probable it is that the word embedding model reflects what happens in the human brain to understand the meaning of a word. Previous work has compared the effect of different word embedding models in fMRI word decoding tasks [1,47]. Our work differs from these efforts in that we focus on providing a more comprehensive understanding of fMRI word decoding by exploiting a broader range of word embedding models, including textual, visual, conceptual, multimodal and meta-word embedding models, and exploring their performance in decoding different kinds of words.

The principal motivation for this paper is to better understand the effect of fine-grained factors such as different word classes and stimulus-presentation paradigms in fMRI word decoding tasks by enforcing analogous comparison as much as possible. To that end, we present a large-scale evaluation of eight word embedding models and their combinations for decoding fine-grained fMRI data associated with three classes of words recorded from three stimulus-presentation paradigms. Specifically, we investigate the following research questions: (1) How does the brain-image decoder perform on different classes of words? (2) How does the brain-image decoder perform in different stimulus-presentation paradigms? (3) How

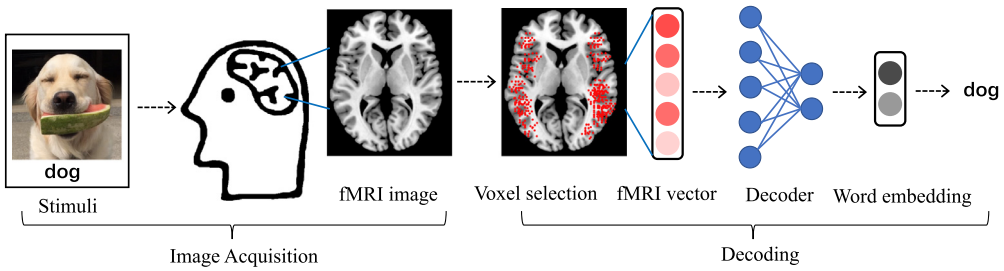

Fig. 1. Schematic of the fMRI word decoding task. After seeing a stimulus, the human brain encodes the properties of the stimulus, which activates certain regions of neurons and forms certain patterns of brain activation. Once we have brain activation data for a certain number of stimuli, we can train a mapping function between brain activations and word embeddings. Since each word embedding corresponds to a specific word, the trained decoder can be used to translate fMRI images into words.

well does each word embedding model allow us to decode neural activation patterns in the human brain? Furthermore, we analyze the most informative voxels associated with different classes of words, stimulus-presentation paradigms and word embedding models, to explore their neural mechanisms.

To summarize, our main contributions include the following:

(1) We conduct a comprehensive analysis of the influence of different factors on the fMRI word decoding task, drawing conclusions about the best way to decode fMRI images in different conditions. This is the first step toward the overarching goal of decoding the human mind which has great significance for people who cannot express thoughts in words.
(2) We provide experimental evidence about the neuroscientific question of the similarities and difference among human semantic representations when understanding different kinds of words with different stimulus-presentation paradigms. These results corroborate, generalize, and extend previous findings and highlight the value of introducing the state-of-the-art computational models to the study of language comprehension by the brain.

2. Background

This paper aims to provide a comprehensive understanding of different factors in fMRI word decoding tasks to find the best way to decode fMRI images under different conditions. Our work falls into the area of neural decoding, which is concerned with the hypothetical reconstruction of sensory and other stimuli from neural activations. Improving the performance of neural decoding not only constitutes an essential step for applications such as brain-machine interfaces but also allows us to better understand the related neural mechanisms.

Traditional decoding methods usually adopt hand-crafted or rule-based features to represent the experimental stimulus, such as extracting corner points, scale, orientation, etc., as image features [20] and collecting feature scores to represent word meaning [38]. These features cannot express fine graded differences between stimulus, especially at the semantic level. With the rapid advances in deep learning methods in recent years, researchers begun to use neural network models as feature extractors in neural decoding tasks for vision, auditory and language [6,16,19,46] and other neuroimaging tasks [34,45].

In this paper, we adopt distributed word representation models to represent word meanings, as is the currently dominant approach in the natural language processing (NLP) community. In the decoding experiments, we use eight state-of-the-art word embedding models and adopt the largest public fMRI dataset associated with word comprehension, which contains 180 concepts collected with three stimulus-presentation paradigms. Therefore, we performed comprehensive experiments to investigate the effects of different factors on the fMRI word decoding task. To our knowledge, this is the first work investigating the fine-grained factors of different word classes, experimental paradigms and word embeddings on fMRI word decoding tasks.

After representing stimuli in feature space, we must learn the mapping from brain activation space to the feature space. In general, the mapping between feature space and brain activation space is assumed to be linear because the features that are represented by a specific brain region should have the simplest possible relationship to its activity [26]. This assumption has supported by our previous experiments. We have adopted both linear and nonlinear mapping methods which achieve similar results on the neural decoding task [39]. Moreover, linear models have a simple interpretation and are relatively easy to estimate. Therefore, we adopt a linear mapping method in the experiments.

3. Experimental setup

fMRI word decoding is used to decode what the human brain is thinking and to interpret fMRI scans from people watching or listening to words. As shown in Fig. 1, the fMRI word decoding task contains several key parts.

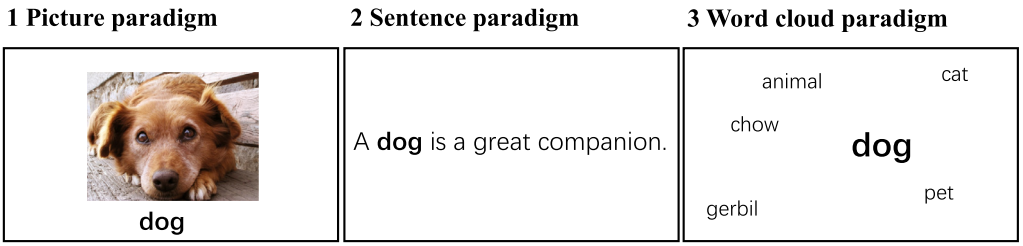

Fig. 2. Examples of a stimulus word in the three paradigms.

First, we collected fMRI images associated with human word understanding and preprocessed this dataset to obtain one fMRI image per word. Then, we selected a smaller set of voxel activations because the fMRI image contains approximately 200,000 brain voxel activations, while we have only a few hundred training examples. Next, we utilized word embedding models to learn representation vectors for stimuli words. After obtaining fMRI vectors and word vectors for a certain set of words, we were able to train the fMRI decoder. Finally, using the trained decoder, we predicted the word corresponding to a new fMRI image. We will introduce the details of each part in the following sections.

3.1. Image acquisition and preprocessing

Our experiments were conducted on the dataset from Pereira et al. [32] which is publicly available at https://osf.io/crwz7/. The dataset includes preprocessed functional activation data that were gathered from 15 participants while they were exposed to distinctive stimuli from three paradigms. Fig. 2 shows an example of a stimulus word in the three paradigms. In the sentence paradigm, participants were asked to read the sentences and think about the meaning of the target word (in bold) in the context in which it is used. In the picture paradigm, participants were asked to read the word and think about its meaning in the context of the accompanying image. In the word cloud paradigm, participants were asked to read the target word (bolded, in the center of the word cloud) and to think about its meaning in the context of the accompanying words.

The stimuli consisted of 180 words that were 131 nouns, 22 verbs, 21 adjectives and 6 adverbs. Within each scanning session, the 180 words were divided into two sets of 90 (performed separately for each participant and paradigm) and distributed across two runs. Each participant saw between 4 and 6 repetitions for each of the three paradigms. Across paradigms, each stimulus was presented for 3 s followed by a 2 s fixation period. The details of the experimental setup, materials and presentation scripts are available at https://osf.io/crwz7/wiki/home/.

The acquisition parameters and processing pipelines are described in detail in Pereira et al. [32]. Briefly, the fMRI scans were collected on a whole-body 3-Tesla Siemens Trio scanner with a 32-channel head coil, with an acceleration factor of 2, a repetition time (TR) of 2000 ms and an echo time (TE) of 30 ms. The functional data in each scanning session were corrected for slice timing, motion, and bias field inhomogeneity and high-pass filtered (at a 100 s cutoff). The images were then registered to the structural scan in their own session, then to the reference structural scan, and finally resampled into 2 mm isotropic voxels. The reference structural scan was registered to the Montreal Neurological Institute (MNI) template, and the resulting transformation was inverted to generate subject-specific versions of the various atlases and parcellations used. The responses to each stimulus were estimated using a general linear model (GLM) in which each stimulus presentation (picture, sentence, and word cloud) was modeled with a boxcar function convolved with the canonical hemodynamic response function (HRF). In this way, the time series data were converted to static vectors.

3.2. Voxel selection

We selected the most informative 5000 voxels following the voxel selection method from Pereira et al. [32]. In all experiments, we adopted a 10-fold cross-validation method (with 9 subsamples for training and the remaining subsample for testing) and performed voxel selection on the training set separately for each of the 10 subsamples.

3.3. Word embedding models

We utilized eight word embedding models that can be summarized into five categories: including textual (i.e., Word2Vec, fastText, and GloVe), linguistic (i.e., Dependency), conceptual (i.e., RWSGwn and ConceptNet), contextual (i.e., ELMo), and visual (i.e., VGG). We also computed combinations of these models using the best embedding model from each category (i.e., GloVe, Dependency, ConceptNet, ELMo, and VGG), resulting in four multimodal (i.e., Dep-VGG, Concept-VGG, GloVe-VGG, and ELMo-VGG) and six meta-word models (i.e., Dep-Concept, Dep-GloVe, Dep-ELMo, Concept-GloVe, Concept-ELMo, and GloVe-ELMo). Next, we will introduce these models separately.

- **Word2Vec**: Word2vec is a shallow neural network that reconstructs the context of a given word [23]. In our experiments, we use the pretrained skip-gram model from https://code.google.com/archive/p/word2vec/.

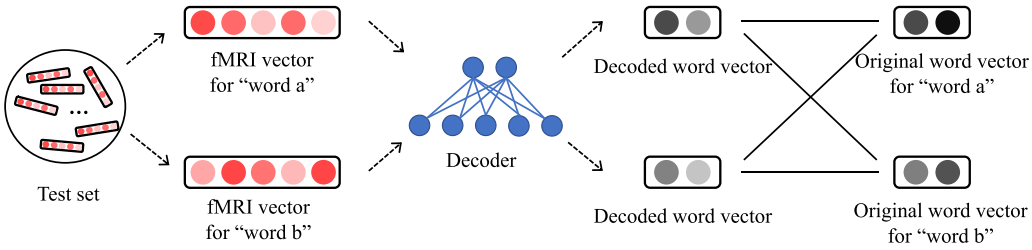

Fig. 3. Testing schematic. In the testing phase, two examples from test set were selected to compute the matching accuracy at each time. For word-class comparison experiment, we randomly chose 2 examples in each word class (or between each word class) as test set. For other experiments, we chose 18 examples (one of 10-fold) as test set, in which every two pair of examples was used to compute the matching accuracy.

- **fastText**: fastText is a modification of Word2Vec that takes morphological information into account [7]. Here, we use the pretrained model from https://fasttext.cc/docs/en/crawl-vectors.html.
- **Dependency-based Word2Vec**: The dependency-based Word2Vec introduced in [22] is a Word2Vec model in which the context of the words is computed based on the dependency relations. We use the pretrained dependency vectors from https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/.
- **GloVe**: GloVe is a count-based method that performs a dimensionality reduction on the cooccurrence matrix [30]. We use the pretrained GloVe vectors from https://nlp.stanford.edu/projects/glove/ on the Common Crawl corpus of 840B tokens.
- **RWSGwn**: RWSGwn embeddings are computed on WordNet with random walks and dimensionality reduction techniques [11]. We use the RWSGwn vectors from http://ixa2.si.ehu.es/ukb/.
- **ConceptNet**: ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled edges, representing the general knowledge involved in understanding language. The ConceptNet embeddings are computed on a ConceptNet graph using the PPMI method [37]. We used the pretrained ConceptNet vectors from https://github.com/commonsense/conceptnet-numberbatch.
- **ELMo**: ELMo is a deep contextualized word representation based on learned functions of the internal states of a deep bidirectional language model [33]. We use the pretrained model from https://allennlp.org/elmo.
- **VGG**: The VGG network is a convolutional neural network model trained for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [36]. In our experiments, we used the pretrained VGG-19 model from https://github.com/tensorflow/models/tree/master/research/slim#pre-trained-models and extracted features from the last hidden layer.
- **Multimodal**: Multimodal models utilize information from different modalities of information to learn word embeddings [43]. Here, we computed multimodal embeddings with VGG embeddings and four other types of embeddings: GloVe, Dependency, ELMo, and ConceptNet. We used two combination methods, i.e., averaging, which calculates the average of two vectors, and concatenation, which concatenates two vectors.
- **Meta-word**: Meta-word models utilize ensemble approaches of combining different word embeddings with the aim of learning word embeddings that contain more information. Here, we obtained meta-word embeddings by calculating averaged and concatenated combinations of GloVe, Dependency, ELMo, and ConceptNet vectors.

3.4. Decoding methodology

Following the work of Mitchell et al. [24], for every participant, the image decoder learns a mapping function between fMRI activation vectors and word embedding vectors using linear regression. The learned weights are used to make predictions about brain activation vectors associated with words that were not seen during training. We implemented ridge regression [15] as the mapping function. Suppose that $X \in \mathbb{R}^{m\times n}$ (training set) is the fMRI data matrix containing m examples, and $Y \in \mathbb{R}^{m\times p}$ is the corresponding word embedding matrix containing the same m examples, where n and p are dimensions of the imaging data vector and word embedding vectors respectively. We had the system learn the regression coefficients k (n-dimensional vector) and k_0 (a constant) that minimized

$$||Xk + k_0 - y||_2^2 + \lambda||k||_2^2, \tag{1}$$

for each column $y \in \mathbb{R}^{m\times 1}$ which is a single dimension of the Y matrix.

The regularization parameter λ is set separately for each dimension using generalized cross-validation within the training set. Each voxel is standardized across training stimuli in each imaging experiment, as is each word embedding vector dimension.

3.5. Training and evaluation setup

To avoid overfitting, we used the 10-fold cross-validation method, training on 162 examples and testing on the remaining 18 examples in each epoch. Voxel selection was conducted only with training examples. As shown in Fig. 3, in the testing

Table 1
Pairwise classification results for nouns, verbs and adjectives when different word embedding models are used. We show the average results from 15 subjects with standard deviations in parentheses. The bold scores in each row are the best results among the three word categories.

	Noun	Verb	Adjective
GloVe	**0.8768**(0.0792)	0.8544(0.0713)	0.8337(0.1081)
Word2Vec	**0.8386**(0.0942)	0.8309(0.0636)	0.8210(0.1028)
Fasttext	**0.8407**(0.0676)	0.8235(0.0766)	0.8077(0.0996)
RWSGwn	**0.8123**(0.0886)	0.7453(0.0771)	0.7425(0.1032)
ELMo	**0.9088**(0.0632)	0.8520(0.0797)	0.7993(0.1244)
ConceptNet	0.8646(0.0875)	**0.8702**(0.0695)	0.8249(0.0925)
Dependency	**0.8554**(0.0731)	0.8137(0.0755)	0.7891(0.0808)
VGG	0.9137(0.0618)	**0.9186**(0.0436)	0.8898(0.0504)
Average	**0.8639**(0.0769)	0.8386(0.0696)	0.8135(0.0952)

phase, we evaluated each of the semantic models by computing its accuracy in matching the two predicted word vectors with the two original ones. A match score was computed by analyzing the Pearson correlation between the predicted and original word vectors. If the sum of similarities for the correct pairing is higher than the one for the incorrect pairing, the matching accuracy is set to 1 for this round of cross-validation; otherwise, it is set to 0. Specifically, assume that the original vectors of words a and b is V_a and V_b and that the predicted vectors of words a and b are $V_{\hat{a}}$ and $V_{\hat{b}}$. The matching accuracy was defined as 1 if $Corr(V_a, V_{\hat{a}}) + Corr(V_b, V_{\hat{b}}) > Corr(V_a, V_{\hat{b}}) + Corr(V_b, V_{\hat{a}})$; otherwise, it was defined as 0. Thus, if the model chose the match at random, the expected accuracy would be 0.50. In the experiments, the reported results are the average accuracy across the 10-fold cross-validation.

Due to the relatively small number of examples in each word class (especially for verbs and adjectives), we had no way to successfully train an image-decoder for each word class. Therefore, we adopted another methodology to study the effect of different word classes. Specifically, we used 2 randomly selected examples in each word class (or between each word class) as the evaluation set and trained the model using the remaining 178 examples. This process was iterated 30 times, and the averaged accuracy was reported.

4. Experiments and discussion

To fully answer the questions raised in the introduction, we conducted a range of comparative experiments. In the following subsections, we present and discuss the results of the three research questions.

4.1. How does the brain-image decoder perform with different classes of words?

To compare the similarities and differences between the word classes, we decoded the neural activity patterns of two examples in the same word class at a time; their averaged pairwise classification results are presented Table 1.

As we can see, different classes of words behave differently when different word embedding models are used. Overall, the decoding performance for nouns is much higher than that for verbs or adjectives, partly because nouns take a large portion of all words (131/180 in the sample). Despite the fact, VGG and ConceptNet models were more effective in decoding verbs than nouns. This observation illustrates that different word classes have different properties that are best decoded with specific word embedding models. Moreover, verbs and adjectives were approximately equal in number (22/180 and 21/180 in the sample), but decoding accuracy was always higher on verbs than on adjectives with the different word embedding models. Two reasons may account for this phenomenon. First, word embedding models learn better representations of verbs than of adjectives. Second, the brain activation patterns associated with verbs are more distinguishable than those of adjectives. In addition, the standard deviation results of adjectives are much higher than those of nouns and verbs on average, indicating that a larger difference in understanding exists among subjects in understanding adjectives than in verbs or nouns.

To compare concrete and abstract words, we classified stimuli words according to their concreteness score. Specifically, we chose the 30 nouns and 10 verbs with the highest and lowest concreteness scores. As a result, we obtained 30 abstract nouns, 30 concrete nouns, 10 abstract verbs and 10 concrete verbs. For adjectives, we chose words with concreteness scores lower than 3 and obtained 20 abstract adjectives. The appendix shows all of the words used in our experiments, along with their concreteness scores (Table A.4).

To explore the detailed differences among concrete nouns, concrete verbs, abstract nouns, abstract verbs and abstract adjectives, we first calculated the pairwise classification accuracy in each word category. From Table 2, it is evident that different categories of words behaved differently when different word embedding models were used. VGG and ConceptNet models were particularly effective at decoding concrete verbs compared to other word categories. Dependency and RWSGwn models performed poorly at decoding abstract verbs. Overall, the five word categories achieved approximately the same standard deviation results with abstract nouns, whereas the results for abstract verbs were more uniform among the

Table 2
Pairwise classification results for both abstract and concrete nouns, verbs and adjectives when different word embedding models are used. We show the average results from 15 subjects with standard deviations in parentheses. Here, a- and c- are short for abstract and concrete. The bold scores in each row are the best results among the five word categories.

	a-noun	c-noun	a-verb	c-verb	a-adjective
GloVe	0.8433(0.0709)	0.8322(0.0954)	0.8074(0.0813)	**0.8474**(0.0979)	0.8356(0.0963)
Word2Vec	0.7822(0.1254)	0.8167(0.0915)	0.7985(0.0790)	**0.8430**(0.0995)	0.8411(0.0961)
Fasttext	0.7767(0.0955)	0.8078(0.0888)	0.8207(0.0918)	**0.8267**(0.1173)	0.7989(0.1072)
RWSGwn	0.7711(0.0790)	0.7289(0.1152)	0.6415(0.1312)	**0.7985**(0.1080)	0.7378(0.1107)
ELMo	0.8500(0.0926)	0.8289(0.1248)	0.8222(0.0935)	**0.8533**(0.0972)	0.8033(0.1106)
ConceptNet	0.8089(0.0742)	0.8400(0.0899)	0.8489(0.0878)	**0.8785**(0.0972)	0.8133(0.0956)
Dependency	0.8300(0.0759)	0.8200(0.1064)	0.7437(0.1032)	**0.8341**(0.1117)	0.7789(0.0665)
VGG	0.8778(0.0899)	0.8878(0.0768)	0.8815(0.0509)	**0.9304**(0.0605)	0.8733(0.0672)
Average	0.8175(0.0879)	0.8203(0.0986)	0.7956(0.0898)	**0.8515**(0.0987)	0.8103(0.0938)

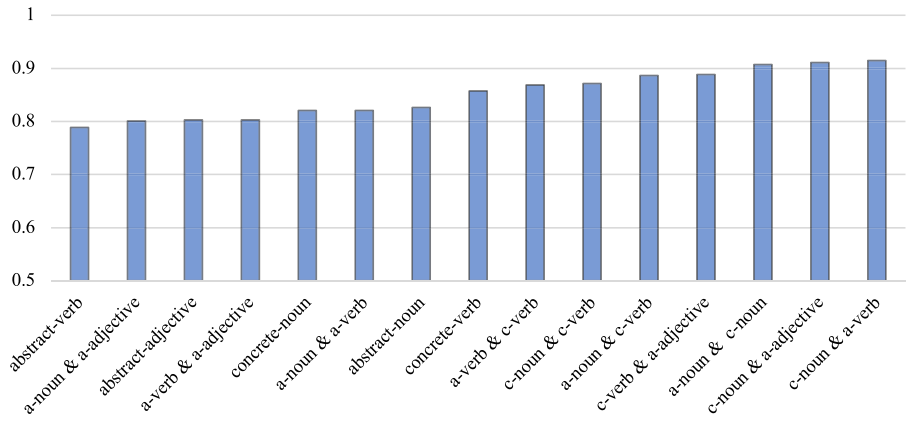

Fig. 4. Pairwise classification results for words within or between different word categories. The results are averaged among 15 subjects and eight embedding models. Here, a- and c- are short for abstract and concrete.

15 subjects. Moreover, among the five word categories of this dataset, concrete verbs could be decoded with the highest accuracy.

Then, we calculated the average pairwise classification accuracy within and between each word category. As shown in Fig. 4, pairwise classification results between concrete words and abstract words were much higher than those within concrete words or abstract words, revealing a clear gap between concrete words and abstract words. In addition, the difference between concrete nouns and concrete verbs was much higher than between abstract nouns and abstract verbs. This observation, from the perspective of decoding performance, supports the theory that brain activation is driven by semantics and not by lexical class.

To analyze the impact of different brain networks on different categories of words, we selected brain activations from each brain network. With the selected brain activation data, we learned their mapping functions with word embedding models and then evaluated the trained decoders with the pairwise classification method. Following Pereira et al. [32], we used 12 brain networks[1] that have been shown to be linked to higher-level cognition. Here, we analyzed two typical word embedding models with different characteristics, GloVe and VGG, the results of which are shown in Fig. 5.

The results of both the GloVe and VGG models show that abstract nouns depended on the language-related and DMN networks to a greater extent than concrete nouns did. This outcome is in accordance with dual-coding theory [14] which posits that concrete words are represented in the brain in terms of a visual and linguistic code, whereas abstract words are encoded only in the linguistic modality. In contrast, abstract verbs and concrete verbs had no obvious regular patterns and performed differently with the GloVe model and the VGG model. Overall, concrete verbs are decoded more effective than abstract verbs by most brain networks. Moreover, when comparing the results of concrete words, we found that concrete verbs were decoded much more effectively than concrete nouns by all brain networks. Among the three types of abstract words, the worst performance was obtained with abstract adjectives, and the GloVe and VGG models produced largely similar but slightly different results with abstract nouns and abstract verbs.

[1] The 12 brain networks are the languageParcelsConservative (languagePC), languageParcels (langaugeP), semantic, multiple demand (MD), default mode network (DMN), languageLH, langugeRH, visual_body, visual_face, visual_object, visual_scene, and visual.

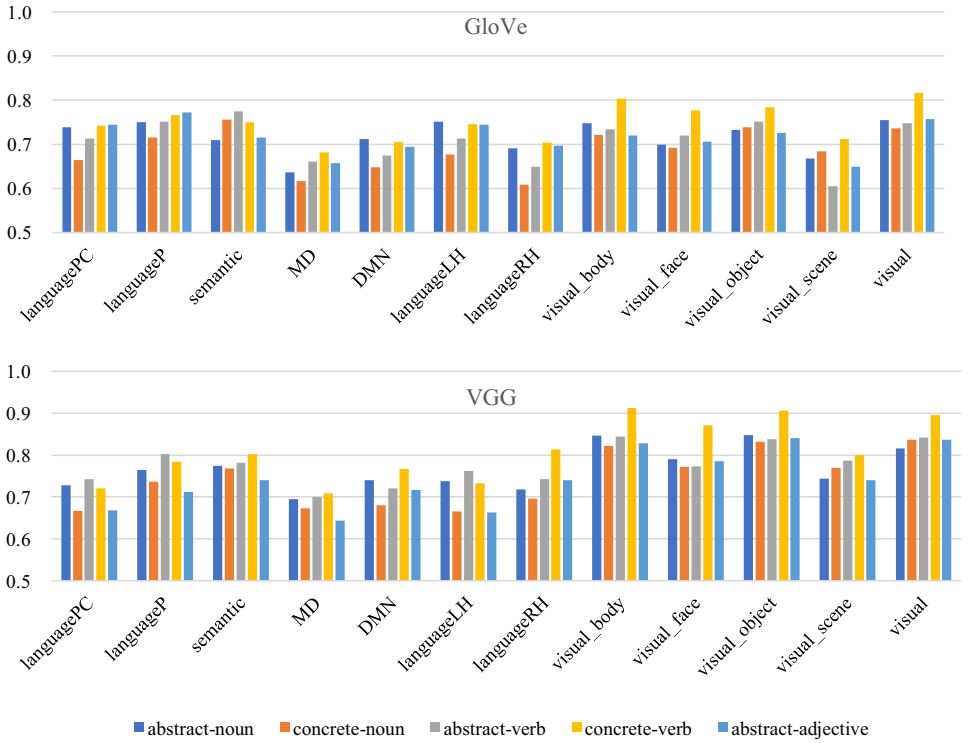

Fig. 5. Pairwise classification results with brain activation in voxels from 12 networks. The top panel shows the results with the GloVe model, and the bottom panel shows the results with the VGG model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

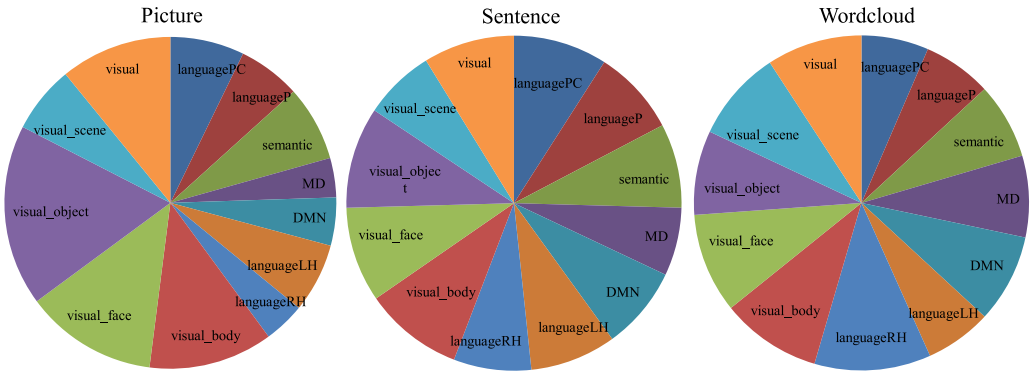

Fig. 6. The distribution of the average activation values in the 12 brain networks in which the brain activation data are collected from three different stimulus-presentation paradigms.

4.2. How does the brain-image decoder perform in different stimulus-presentation paradigms?

The question of how humans represent word meaning has been studied using different stimulus-presentation paradigms. In this paper, we investigate the commonalities and differences in decoding brain activities when viewing words in three different paradigms.

To form an intuitive impression of the relationship among brain activations collected from three stimulus-presentation paradigms, we show the distribution of their average activation values in the 12 brain networks. As shown in Fig. 6, there are clear differences among the three paradigms. Specifically, the picture paradigm activates more vision-related brain regions, while the sentence paradigm activates more language-related brain regions.

In Table 3, we present the results of the decoding methodology on fMRI data from the three paradigms. As shown clearly, the decoding accuracy of the picture paradigm was obviously the best, followed by the sentence paradigm, and the word cloud paradigm had the lowest accuracy. The picture paradigm achieved the lowest standard deviation, indicating that

Table 3
Pairwise classification results when three stimulus-presentation paradigms with the picture, sentence and word cloud paradigms coupled with different word embedding models are used. We show the average results from 15 subjects with standard deviations in parentheses. The bold scores in each row are the best results among the three paradigms.

	Picture	Sentence	Word cloud
GloVe	**0.7910**(0.0676)	0.6895(0.1331)	0.6271(0.0945)
Word2Vec	**0.7397**(0.0678)	0.6393(0.1031)	0.5925(0.0750)
Fasttext	**0.7852**(0.0510)	0.6791(0.1280)	0.6170(0.1014)
RWSGwn	**0.7087**(0.0557)	0.6363(0.0726)	0.5777(0.0628)
ELMo	**0.8059**(0.0503)	0.7307(0.1249)	0.6515(0.1042)
ConceptNet	**0.7980**(0.0793)	0.6935(0.1486)	0.6088(0.1230)
Dependency	**0.7739**(0.0524)	0.6793(0.1133)	0.6285(0.0877)
VGG	**0.8992**(0.0258)	0.6342(0.0597)	0.5854(0.0590)
Average	**0.7877**(0.0562)	0.6727(0.1104)	0.6111(0.0885)

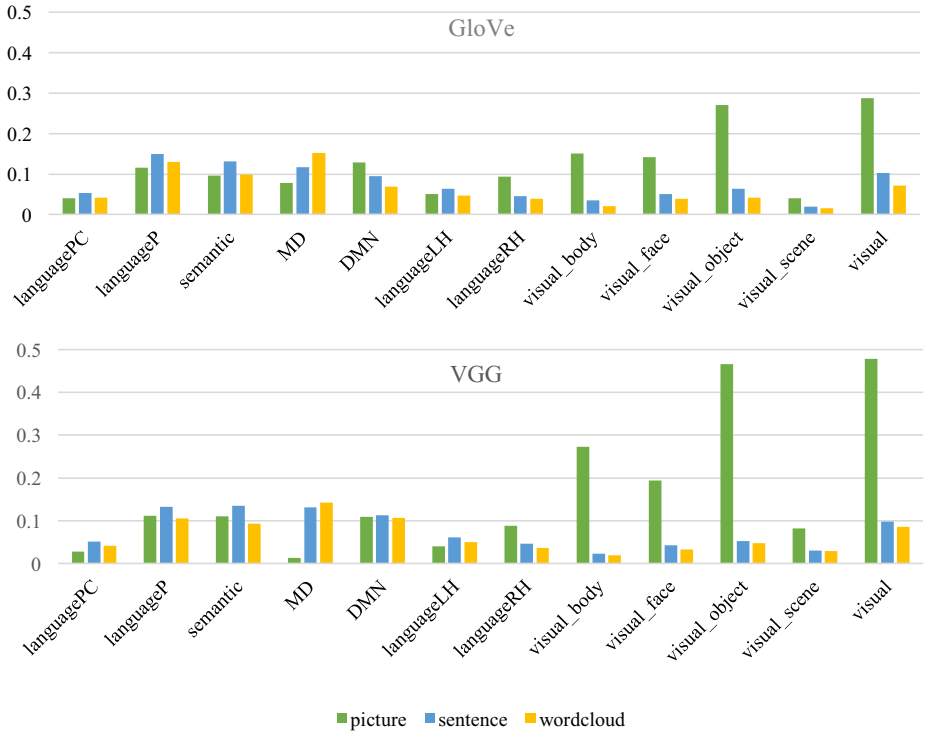

Fig. 7. Averaged distribution of informative voxels across the brain in 15 subjects with stimuli presented in picture, sentence and word cloud paradigms. The top panel shows results from the GloVe model, and the bottom panel shows results from the VGG model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this paradigm was the most stable of the three experimental paradigms among subjects in the fMRI word decoding task. Moreover, the distinction among the three paradigms was most obvious when the VGG model was used. The reason is that the VGG model encodes mostly the visual information of words and is thus most suitable for decoding brain activation data caused by visual stimulus inputs. Moreover, sentence and word cloud paradigms achieve the best results with the ELMo model, revealing that the ELMo model encodes word information that is more suitable to sentence and word cloud paradigms.

To analyze how the human brain responds to the different stimulus-presentation paradigms, we calculated the distribution of informative voxels in 12 different brain networks. As shown in Fig. 7, the picture paradigm had many more informative voxels in vision-related networks, while the sentence and word cloud paradigms had more informative voxels in the languageP, semantic and MD networks. This finding illustrates that different brain networks are responsible for processing stimuli presented in the different paradigms; specifically, the picture paradigm activates more vision-related brain networks, and the sentence or word cloud paradigms activate more language- and semantic-related brain networks.

In addition, we show the most informative voxels across the whole brain in Figs. 8 and 9. From these experiments, we found that the most informative voxels varied dramatically across subjects. Here, we show the results from sub-

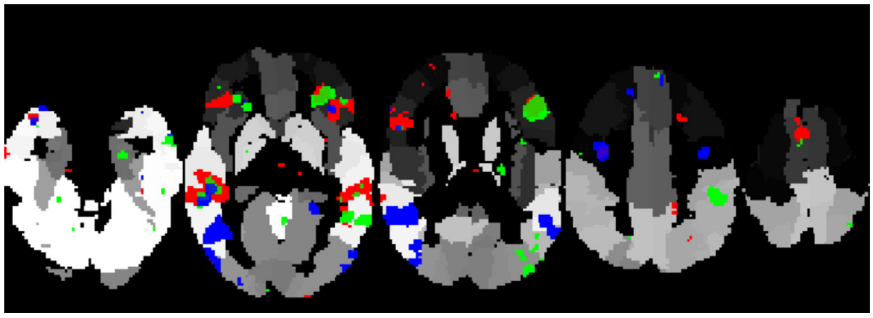

Fig. 8. The 5000 most informative voxels from the three stimulus-presentation paradigms with subject M15 when the GloVe model is used. The red color represents the picture paradigm, the green color represents the sentence paradigm and the blue color represents the word cloud paradigm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

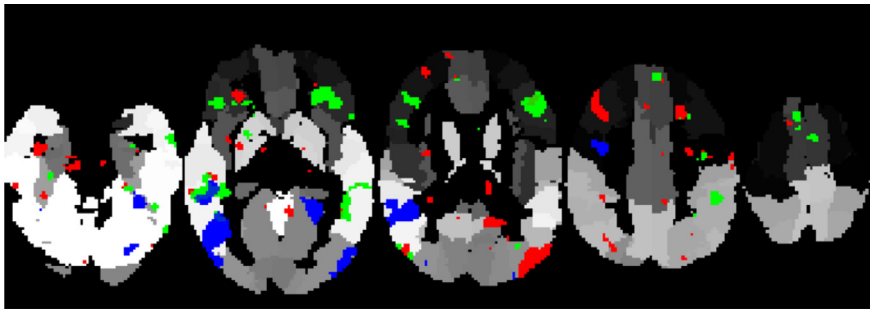

Fig. 9. The 5000 most informative voxels from the three stimulus-presentation paradigms with subject M15 when the VGG model is used. The red color represents the picture paradigm, the green color represents the sentence paradigm and the blue color represents the word cloud paradigm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ject M15, in whom the best average performance across all experiments is obtained. For the GloVe model with the picture/sentences/word cloud paradigms, there were 0/0/0 matching informative voxels among the 15 subjects and 504/303/262 matching informative voxels between the best two subjects, M15 and M02. For the VGG model with picture/sentences/word cloud paradigms, there were 0/0/0 matching informative voxels among the 15 subjects and 1237/393/253 matching informative voxels between M15 and M02.

Figs. 8 and 9 show that the most informative voxels from the three stimulus-presentation paradigms differed greatly, with average overlap ratios of 124/5000 and 128/5000, respectively, for the GloVe and VGG models. For both the GloVe and VGG models, these highest-accuracy voxels were meaningfully distributed across the brain. Specifically, the three brain regions with the largest numbers of informative voxels were the temporal lobe, occipital lobe, and fusiform gyrus for the picture paradigm and the temporal lobe, frontal lobe, and cerebellum for the sentence and word cloud paradigms. This outcome is consistent with our common-sense notion that the picture paradigm activates more brain regions that are associated with decoding visual inputs, while the sentence and word cloud paradigm activate more brain regions that are involved in higher cognitive activities. See Appendix Table A.5 for the exact number of informative voxels in each brain region.

4.3. How well does each word embedding model allow us to decode neural activation patterns in the human brain?

Different word embedding models encode different aspects of word information with different resources; thus, we expected to see some differences when decoding brain activations of words with these models. Furthermore, we used these embedding models to investigate the brain regions that are related to different word characteristics. Note that we use averaged brain activations from the three stimulus-presentation paradigms in this experiment.

First, we show the relationships among different types of word embedding models. Using the t-distributed stochastic neighbor embedding (t-SNE) method,[2] we compressed the high-dimensional word embeddings into the two-dimensional plane, in which words that cluster together have similar word embeddings. As shown in Fig. 10, the word embeddings trained with text corpora could clearly distinguish the word classes (nouns, verbs, adjectives, adverbs), especially with the GloVe and Dependency model, while the VGG model trained with images could not distinguish word classes well (see the t-SNE plots of other models in Appendix Fig. A.15).

[2] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

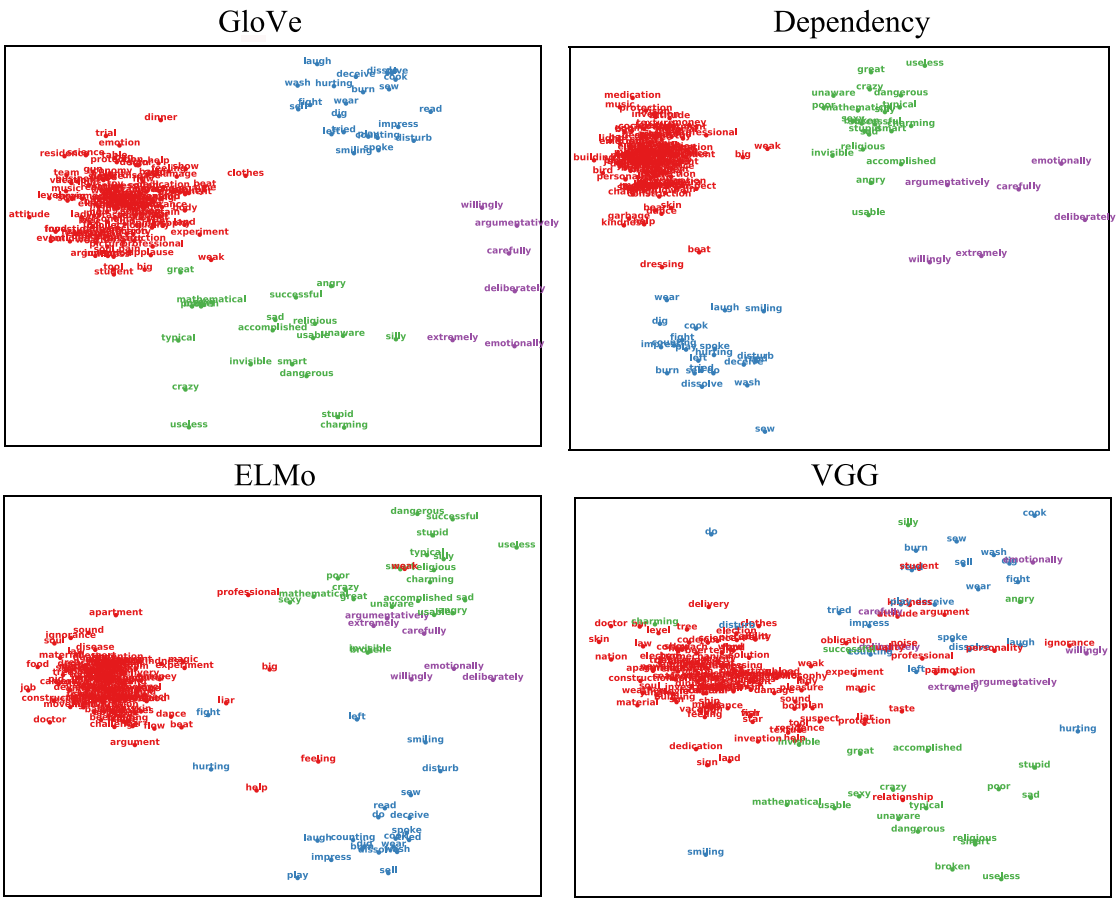

Fig. 10. The t-SNE plots of the 180 concepts with different word embeddings. The red dots are nouns, the blue dots are verbs, the green dots adjectives and the purple dots are adverbs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

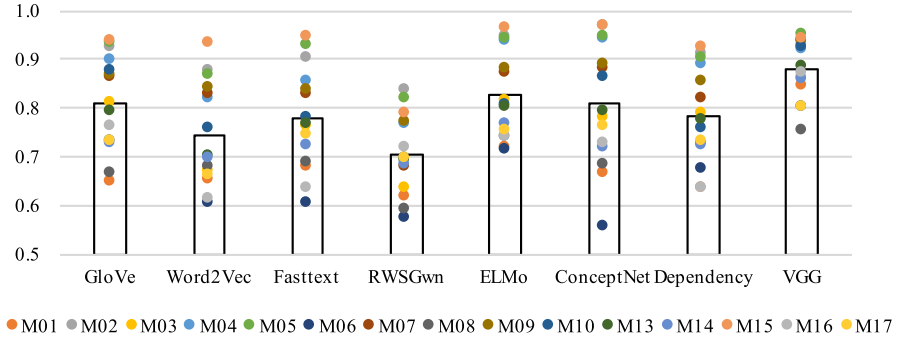

Fig. 11. Pairwise classification results when different word embedding models are used. The bars show the averaged results among 15 subjects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Subsequently, we characterized the pairwise classification results for each subject with different embedding models, as shown in Fig. 11. As the figure clearly shows, the VGG model achieves the best performance, followed by the ELMo model and the GloVe model, and the RWSGwn model obtains the worst results. This phenomenon can be partly explained by the particular characteristics of these models. The VGG model extracts word features from corresponding pictures and thus mostly encodes the visual information of the target word. The ELMo model extracts word features from sentences that contain the target word; thus, it mostly encodes the linguistic and pragmatic information of the target word. These two models are particularly useful in visual and language paradigms, achieving the best performance with averaged brain activation data. In contrast, the RWSGwn model learns word embeddings from a manually constructed knowledge base, mainly encoding word relation information. The manually constructed database is always incomplete, which diminishes the quality

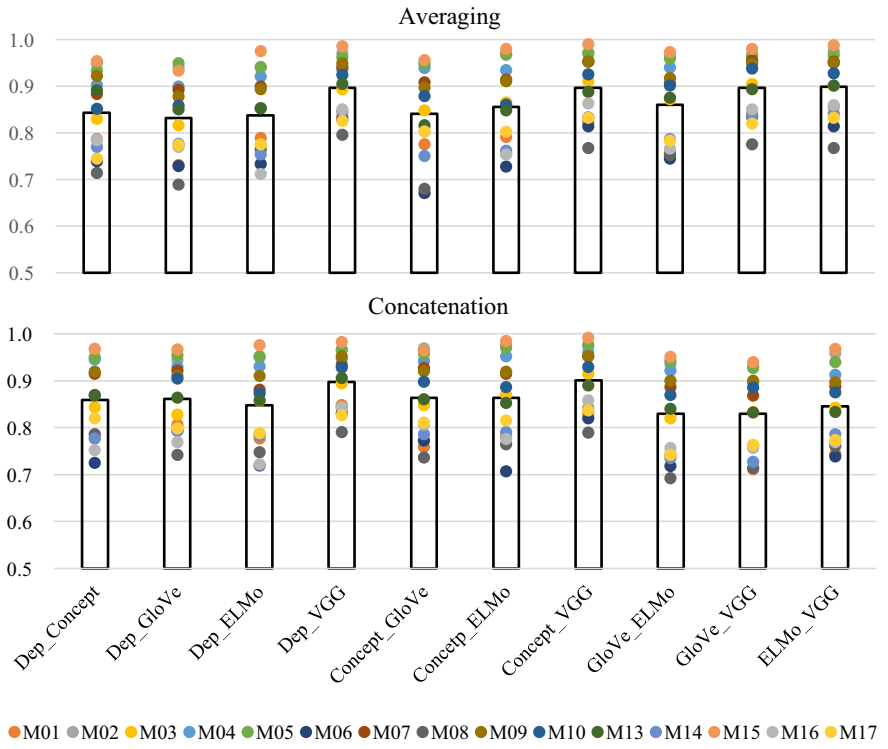

Fig. 12. Pairwise classification results with different multimodal and meta-word embeddings, which are averaged or concatenated combinations of two basic word embeddings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

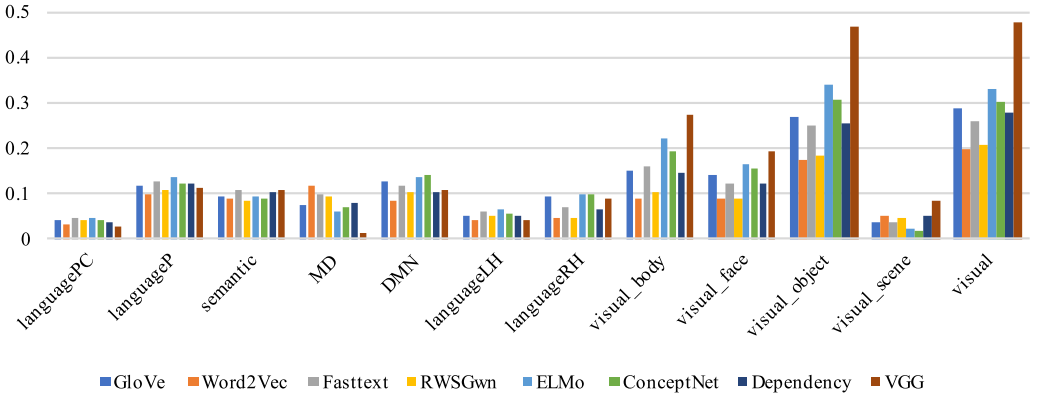

Fig. 13. Averaged distribution of informative voxels across the brain for 15 subjects using the different word embedding models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of word embeddings. In addition, there is a large gap between different subjects, which indicates that there may exist a distinctive understanding of the same word between different people, reflecting the complexity and difficulty of the fMRI word decoding task.

Next, we tested the multimodal and meta-word embedding models for decoding brain activation. From Fig. 12, we can see that multimodal and meta-word embeddings achieve better performance than the component embedding models, specifically for those with VGG embedding. Moreover, averaging and concatenation methods show similar results on most word embeddings. Exceptions include the Glove_VGG and ELMo_VGG models, which exhibit significantly better performance with the averaging method. A possible reason is that the GloVe and ELMo models encode overlapping information with the VGG model, in which case the averaging method can handle the data more effectively than the concatenation method can.

Then, to analyze the most informative voxels associated with the different word embedding models, we show the averaged fraction of the 5000 informative voxels in the 12 brain networks. As shown in Fig. 13, for all word embedding models,

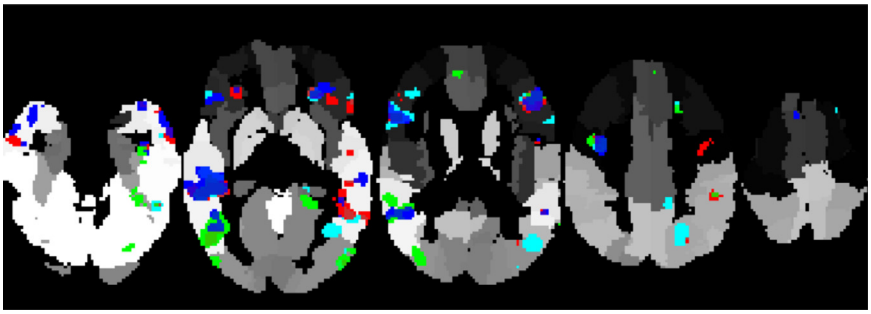

Fig. 14. The 5000 most informative voxels for subject M15 when different word embedding models are used. The red color represents the GloVe model, the green color represents the VGG model, blue color represents the ELMo model and the yellow color represents the Dependency model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a large portion of informative voxels are related to visual networks, illustrating that visual information is important for distinguishing words from each other. This phenomenon is most apparent with the VGG model, in which most of its informative voxels are in vision-related networks. This outcome is to be expected because the VGG model encodes mostly visual information of words.

Finally, Fig. 14 shows the most informative voxels across the whole brain when four different embedding models are used. From the experiments, we found that the most informative voxels varied dramatically across subjects. Specifically, for the GloVe/VGG/ELMo/Dependency models, there were 0/0/0/0 matching informative voxels across the 15 subjects, and 559/948/518/397 matching informative voxels between the best two subjects M15 and M02.

It is also apparent that the most informative voxels differed greatly among the four word embedding models; their average overlap ratio was 429/5000 among 15 subjects. In contrast to the VGG model, which relates mostly to vision-related brain regions, the GloVe, ELMo and Dependency models performed similarly and were strongly associated with vision-related and higher cognitive brain regions. See Appendix Table A.6 for the exact number of informative voxels in each brain region.

5. Conclusion and future directions

To comprehensively explain the influence of different factors on fMRI word decoding, we present a large-scale evaluation of eight word embedding models (five categories in total) and their combinations for decoding fMRI data recorded from three stimulus-presentation paradigms and three word classes. We investigated three research questions and analyzed the most informative voxels associated with the different classes of words, stimulus-presentation paradigms and word embedding models to explore their neural mechanisms.

The first question is how the brain-image decoder performs on different classes of words. We performed a quantitative analysis of our results to examine whether different word embedding models are most effective at decoding the brain activation for different categories of words. We found that the different word classes had different properties that could be decoded most effectively by different word embedding models. Concrete nouns and verbs were more easily distinguished than abstract nouns and verbs, which can be seen as computational support for semantic model research that has suggested that brain activation is driven by semantics and not by lexical class. Compared with concrete nouns, abstract nouns have an increased dependency on language-related and the DMN networks, which is in accord with dual-coding theory.

The second research question that we investigated is how the brain-image decoder performs in the different stimulus-presentation paradigms. To answer this question, we conducted a quantitative analysis on the decoding accuracy in the picture, sentence and word cloud paradigms. We found that the decoding accuracy was highest for the picture paradigm, followed by the sentence paradigm and the word cloud paradigm. Different brain networks are responsible for processing information from the different stimulus-presentation paradigms because the picture paradigm activates far more informative voxels in the vision-related networks, while the sentence and word cloud paradigms activate more informative voxels in the languageP, semantic and MD networks.

The third question we address is how well each word embedding model allows us to decode neural activation patterns in the human brain. To answer this question, we measure how well different word embedding models decoded brain imaging data. The results showed that the VGG model achieved the best performance, followed by the ELMo model and the GloVe model. The reason is that the VGG model encodes mostly the visual information of the target word, which is a key discriminative feature in our brain activation data. In addition, multimodal and meta-word embedding models achieve better performance than their component embedding models. Furthermore, visual information is important for distinguishing words from each other because a large portion of informative voxels are in the visual networks. In contrast to the VGG model, which is associated mainly with vision-related brain regions, the GloVe, ELMo and Dependency models are strongly associated with both vision-related and higher cognitive brain regions.

In conclusion, VGG embedding models are the best choice for decoding brain activation data collected from picture paradigms, especially for concrete verbs, while ELMo embedding models should be selected in the context of sentence and word cloud paradigms. Moreover, computational models can help investigate the neural basis of human language processing. Note that the conclusions drawn in this paper are based on a small set of datasets containing fMRI responses to 180 words. Further work will include collecting more brain activation data containing a larger number of words in each category. We also aim to distinguish more specific characteristics encoded in each word embedding model to learn their specific relations with brain activations. We believe that with the development of neuroscience and computational models, we will realize the long-term goal of decoding the human mind freely.

Declaration of Competing Interest

This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere. All authors have checked the manuscript and have agreed to the submission.

Acknowledgments

The research work has been funded by the Beijing Municipal Science and Technology Project No. Z181100008918017. We thank the anonymous reviewers for their efforts on this manuscript.

Appendix A

Table A.4
Total words and corresponding categories in the experiments. The bold one represents the chosen abstract words with lowest concreteness score, and italic one represents the chosen concrete words with highest concreteness score.

nouns	**ignorance (1.6), feeling (1.68), kindness (1.74), ability (1.81), sin (1.85), soul (1.86), attitude (1.97), obligation (2.04), pleasure (2.04), personality (2.08), challenge (2.11), philosophy (2.14), professional (2.14), dedication (2.16), elegance (2.18), quality (2.18), reaction (2.41), emotion (2.44), protection (2.5), marriage (2.51), investigation (2.52), help (2.56), economy (2.57), law (2.57), suspect (2.59), solution (2.61), charity (2.62), event (2.69), science (2.79), weak (2.79)**, relationship (2.8), argument (2.85), level (2.86), magic (2.9), electron (2.96), code (3), time (3.07), trial (3.07), shape (3.14), vacation (3.14), election (3.17), job (3.19), damage (3.2), business (3.28), experiment (3.38), plan (3.4), news (3.41), liar (3.43), disease (3.45), invention (3.48), pain (3.5), noise (3.52), illness (3.54), word (3.56), nation (3.62), movement (3.63), war (3.63), big (3.66), mechanism (3.67), delivery (3.69), sound (3.7), construction (3.72), flow (3.72), team (3.79), texture (3.83), weather (3.83), beat (3.97), device (3.97), show (3.97), dressing (4.03), collection (4.04), taste (4.07), king (4.1), art (4.17), light (4.21), material (4.26), applause (4.29), music (4.31), dance (4.32), lady (4.33), residence (4.34), dinner (4.5), picture (4.52), medication (4.53), money (4.54), land (4.57), movie (4.59), tool (4.6), sign (4.62), building (4.64), jungle (4.66), bar (4.67), prison (4.68), brain (4.69), doctor (4.69), garbage (4.69), star (4.69), driver (4.71), road (4.75), clothes (4.76), plant (4.76), *body (4.79), skin (4.79), food (4.8), gold (4.81), gun (4.83), seafood (4.83), dog (4.85), blood (4.86), engine (4.86), ship (4.87), sugar (4.87), bear (4.88), beer (4.88), apartment (4.9), bag (4.9), table (4.9), student (4.92), computer (4.93), dessert (4.93), toy (4.93), cockroach (4.96), mountain (4.96), hair (4.97), ball (5), bed (5), bird (5), camera (5), fish (5), pig (5), tree (5)*
verbs	**deceive (1.68), impress (2.07), tried (2.07), do (2.46), hurting (3.03), disturb (3.04), dissolve (3.15), play (3.24), sell (3.35), spoke (3.38)**, read (3.56), counting (3.57), *wear (3.61), left (3.7), smiling (3.88), sew (3.93), burn (4.11), fight (4.2), laugh (4.21), cook (4.32), dig (4.33), wash (4.35)*
adjectives	**typical (1.52), accomplished (1.7), smart (1.75), stupid (1.75), unaware (1.75), useless (1.78), great (1.81), usable (2), successful (2.07), dangerous (2.13), charming (2.14), silly (2.2), crazy (2.37), religious (2.5), mathematical (2.52), angry (2.53), sexy (2.64), poor (2.7), invisible (2.83), sad (3.07)**, broken (4.11)
adverbs	emotionally (1.72), extremely (1.84), willingly (2.04), argumentatively (2.07), carefully (2.15), deliberately (2.86)

Table A.5
The number of informative voxels in different brain regions for three stimuli paradigms combined with two embedding models. Results are averaged among the 15 subjects. The bold numbers are the three highest number in each paradigm.

	GloVe			VGG		
	Picture	Sentence	Word cloud	Picture	Sentence	Word cloud
Precentral	82.4	84.13	115.2	19.4	88.53	116.8
Frontal	520	**872.87**	**827.07**	91.93	**968.8**	**844.93**
Rolandic_Oper	15.93	66.2	43.73	1.4	45.07	47.87
Supp_Motor_Area	21.8	61.87	108.27	0.33	59.53	91.07
Olfactory	3.87	9	13.87	0.2	8	12
Rectus	8	35.13	33.87	0.2	36.27	45.8
Insula	29.73	64.93	83.27	2.47	65.47	48.53
Cingulum	48.07	113.6	110.13	1.27	139.93	143.33
Hippocampus	11	34.47	39.53	5.13	47.4	49.2
ParaHippocampal	48.07	58.6	60.33	108.6	101.8	64.73
Amygdala	2.6	9.07	9.27	0.47	6.8	5.8
Calcarine	128.33	110.2	97.73	159.33	61.73	64
Cuneus	23	34.73	30.2	24.07	34.4	40.4
Lingual	249.6	87.73	89.13	302	56.87	68.33
Occipital	**728.67**	266	205.2	**1159.93**	189.13	224.73
Fusiform	**543**	176	154.33	**904**	213.8	168.33
Postcentral	38.2	91.53	139.93	1.87	68.6	100.87
Parietal	110.27	162	218.67	21.33	197.2	205.13
SupraMarginal	42.47	60.73	56.73	23.47	84.67	93.6
Angular	70.07	63.13	68.47	35.13	56.07	99.73
Precuneus	64.47	139.4	118.67	41	166.07	132.67
Paracentral	11.8	44.8	32.13	0.4	14.8	16.07
Caudate	10.33	17.6	30.87	0.33	25.73	22.27
Putamen	14	16.6	27.87	0.47	24.33	23.6
Pallidum	2.47	11.87	6.93	0.4	6.93	7.6
Thalamus	12.73	21.67	25.6	1.93	30.4	28.4
Heschl	3.73	10	4.67	0.07	7.2	4.73
Temporal	**1212.87**	**938.6**	**753.33**	**1412.67**	**811.87**	**736.27**
Cerebelum	260.67	**366.8**	**373.47**	148	**356.27**	**464.53**
Vermis	12.33	32.87	44.67	1.6	29.6	36.47

Table A.6
The number of informative voxels in different brain regions for different word embedding models. Results are averaged among the 15 subjects. The bold numbers are the three highest number in each experimental paradigm.

	GloVe	VGG	ELMo	Dependency
Precentral	85.13	42.73	99.6	123.8
Frontal	**828**	465.73	**1051.6**	**920.53**
Rolandic_Oper	28.47	5.87	15.27	18.53
Supp_Motor_Area	62.47	5.6	61.27	45.4
Olfactory	4.4	1.33	3.53	4.93
Rectus	17.73	8.67	18.33	16.67
Insula	50.93	20.73	29.8	54.6
Cingulum	75.6	47.93	147.73	127.67
Hippocampus	39.6	20.73	21.53	66.27
ParaHippocampal	74.2	134.67	63.6	91
Amygdala	7.73	2.47	7.07	12.67
Calcarine	95.2	116.2	104.8	108.53
Cuneus	38.27	33.07	25.67	59.67
Lingual	133.13	178.67	85.87	94.2
Occipital	**394.6**	**741.93**	**334.87**	**343.47**
Fusiform	315.47	**615.73**	270.27	278.87
Postcentral	52	28.6	32.53	67.8
Parietal	146.67	105.67	116.27	204.2
SupraMarginal	74.73	42.47	84.8	84.67
Angular	87.4	54.47	72.27	69.27
Precuneus	137.8	129.87	140.27	137.2
Paracentral	27.93	1.6	3	14.13
Caudate	7.33	10.67	11.07	11.07
Putamen	13.33	9.8	18.67	8.67
Pallidum	7.2	2.07	7.73	1.73
Thalamus	10.07	5.4	9.73	20.07
Heschl	3.4	0.53	3.07	2
Temporal	**1268.93**	**1411.8**	**1295.53**	**1053.13**
Cerebelum	227.73	156	245.67	247.6
Vermis	10.6	6.07	12.8	12.27

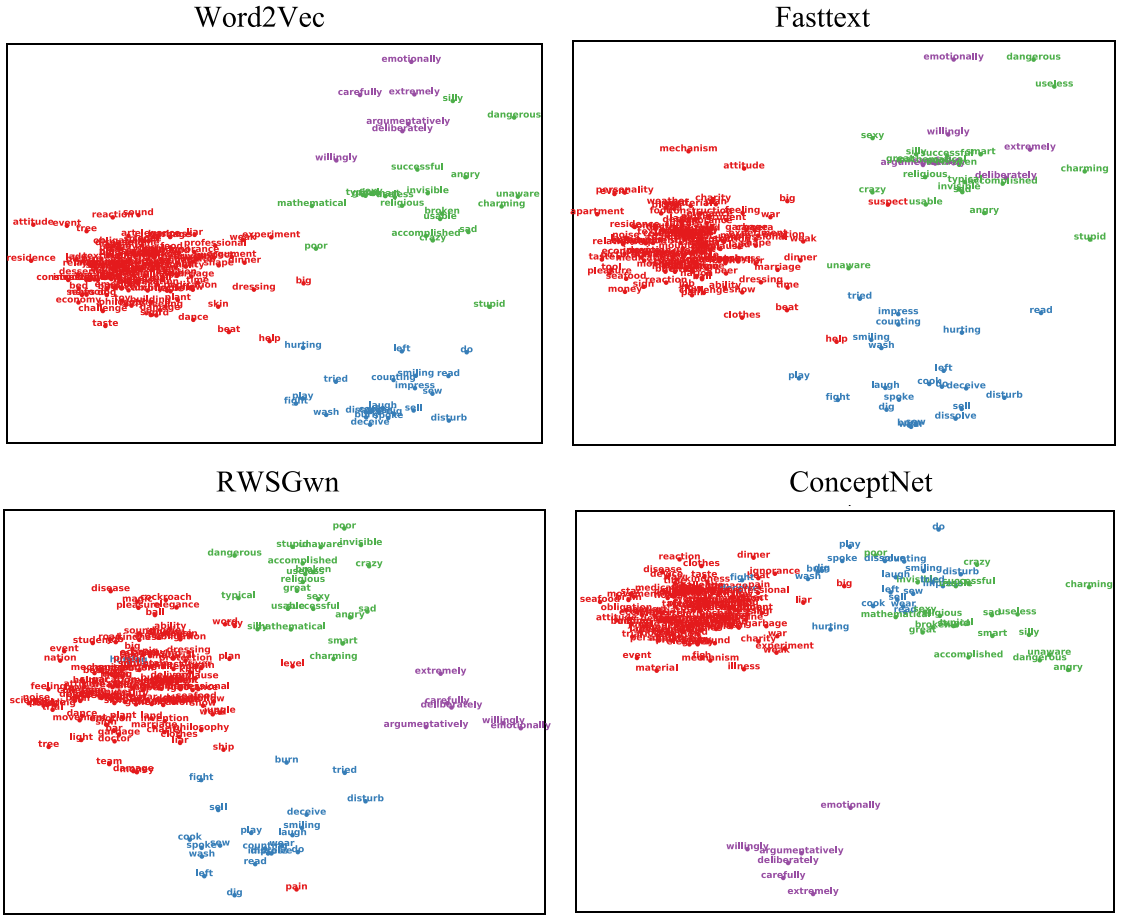

Fig. A.15. The t-SNE plots of the 180 concepts with different word embeddings. The red dots are nouns, the blue dots are verbs, the green dots adjectives and the purple dots are adverbs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

[1] S. Abnar, R. Ahmed, M. Mijnheer, W. Zuidema, Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity, in: Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), 2018, pp. 57–66.

[2] H. Akama, B. Murphy, L. Na, Y. Shimizu, M. Poesio, Decoding semantics across fMRI sessions with different stimulus modalities: a practical mvpa study, Front. Neuroinform. 6 (2012) 24.

[3] A.J. Anderson, J.R. Binder, L. Fernandino, C.J. Humphries, L.L. Conant, M. Aguilar, X. Wang, D. Doko, R.D. Raizada, Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation, Cerebral Cortex 27 (9) (2016) 4379–4395.

[4] A.J. Anderson, D. Kiela, S. Clark, M. Poesio, Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns, Trans. Assoc. Comput.Linguist. 5 (1) (2017) 17–30.

[5] A.J. Anderson, B. Murphy, M. Poesio, Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness, J. Cognit. Neurosci. 26 (3) (2014) 658–681.

[6] G.K. Anumanchipalli, J. Chartier, E.F. Chang, Intelligible speech synthesis from neural decoding of spoken sentences, Nature 568 (7753) (2019) 493–498.

[7] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.

[8] K.-m.K. Chang, Quantitative Modeling of the Neural Representation of Nouns and Phrases, 2011 Doctoral Dissertation.

[9] K.-m.K. Chang, T. Mitchell, M.A. Just, Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fmri activation, NeuroImage 56 (2) (2011) 716–727.

[10] J.A. Clithero, D.V. Smith, R.M. Carter, S.A. Huettel, Within-and cross-participant classifiers reveal different neural coding of information, Neuroimage 56 (2) (2011) 699–708.

[11] J. Goikoetxea, A. Soroa, E. Agirre, Random walks and neural network language models on knowledge bases, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1434–1439.

[12] G. Handjaras, E. Ricciardi, A. Leo, A. Lenci, L. Cecchetti, M. Cosottini, G. Marotta, P. Pietrini, How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge, Neuroimage 135 (2016) 232–242.

[13] J.V. Haxby, J.S. Guntupalli, A.C. Connolly, Y.O. Halchenko, B.R. Conroy, M.I. Gobbini, M. Hanke, P.J. Ramadge, A common, high-dimensional model of the representational space in human ventral temporal cortex, Neuron 72 (2) (2011) 404–416.

[14] M. Hiscock, Imagery and verbal processes, Psyccritiques 19 (6) (1974) 487.

[15] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[16] A.G. Huth, W.A. de Heer, T.L. Griffiths, F.E. Theunissen, J.L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex, Nature 532 (7600) (2016) 453.

[17] M.A. Just, V.L. Cherkassky, S. Aryal, T.M. Mitchell, A neurosemantic theory of concrete noun representation based on the underlying brain codes, PloS one 5 (1) (2010) e8622.
[18] K.N. Kay, T. Naselaris, R.J. Prenger, J.L. Gallant, Identifying natural images from human brain activity, Nature 452 (7185) (2008) 352.
[19] A.J. Kell, D.L. Yamins, E.N. Shook, S.V. Norman-Haignere, J.H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy, Neuron 98 (3) (2018) 630–644.
[20] S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain it cortical representation, PLoS Comput. Biol. 10 (11) (2014) e1003915.
[21] M. Laiacona, A. Caramazza, The noun/verb dissociation in language production: varieties of causes, Cognit. Neuropsychol. 21 (2–4) (2004) 103–123.
[22] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2, 2014, pp. 302–308.
[23] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[24] T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.-M. Chang, V.L. Malave, R.A. Mason, M.A. Just, Predicting human brain activity associated with the meanings of nouns, Science 320 (5880) (2008) 1191–1195.
[25] R.L. Moseley, F. Pulvermüller, Nouns, verbs, objects, actions, and abstractions: local fMRI activity indexes semantics, not lexical categories, Brain Lang. 132 (2014) 28–42.
[26] T. Naselaris, K.N. Kay, S. Nishimoto, J.L. Gallant, Encoding and decoding in FMRI, Neuroimage 56 (2) (2011) 400–410.
[27] T. Naselaris, R.J. Prenger, K.N. Kay, M. Oliver, J.L. Gallant, Bayesian reconstruction of natural images from human brain activity, Neuron 63 (6) (2009) 902–915.
[28] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, J.L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies, Curr. Biol. 21 (19) (2011) 1641–1646.
[29] M. Palatucci, D. Pomerleau, G.E. Hinton, T.M. Mitchell, Zero-shot learning with semantic output codes, in: Advances in Neural Information Processing Systems, 2009, pp. 1410–1418.
[30] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[31] F. Pereira, M. Botvinick, G. Detre, Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments, Artif. Intell. 194 (2013) 240–252.
[32] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S.J. Gershman, N. Kanwisher, M. Botvinick, E. Fedorenko, Toward a universal decoder of linguistic meaning from brain activation, Nat. Commun. 9 (1) (2018) 963.
[33] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2018, pp. 2227–2237.
[34] C.R. Ponce, W. Xiao, P.F. Schade, T.S. Hartmann, G. Kreiman, M.S. Livingstone, Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences, Cell 177 (4) (2019) 999–1009.
[35] F. Pulvermüller, F. Kherif, O. Hauk, B. Mohr, I. Nimmo-Smith, Distributed cell assemblies for general lexical and category-specific semantic processing as revealed by FMRI cluster analysis, Hum. Brain Mapp. 30 (12) (2009) 3837–3850.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014), arXiv:1901.06693.
[37] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: an open multilingual graph of general knowledge., in: AAAI Conference on Artificial Intelligence, 2017, pp. 4444–4451.
[38] G. Sudre, D. Pomerleau, M. Palatucci, L. Wehbe, A. Fyshe, R. Salmelin, T. Mitchell, Tracking neural coding of perceptual and semantic features of concrete nouns, NeuroImage 62 (1) (2012) 451–463.
[39] J. Sun, S. Wang, J. Zhang, C. Zong, Towards sentence-level brain decoding with distributed representations, in: AAAI Conference on Artificial Intelligence, 2019.
[40] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, S. Dehaene, Inverse retinotopy: inferring the visual content of images from brain activation patterns, Neuroimage 33 (4) (2006) 1104–1116.
[41] G. Vigliocco, D.P. Vinson, J. Druks, H. Barber, S.F. Cappa, Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies, Neurosci. Biobehav. Rev. 35 (3) (2011) 407–426.
[42] J. Wang, V.L. Cherkassky, M.A. Just, Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states, Hum. Brain Mapp. 38 (10) (2017) 4865–4881.
[43] S. Wang, J. Zhang, C. Zong, Learning multimodal word representation via dynamic fusion methods, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[44] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, T. Mitchell, Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses, PloS One 9 (11) (2014) e112575.
[45] D. Wen, Z. Wei, Y. Zhou, G. Li, X. Zhang, W. Han, Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: a brief overview and our opinion, Front. Neuroinformat. 12 (2018) 23.
[46] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, Z. Liu, Neural encoding and decoding with deep learning for dynamic natural vision, Cerebral Cortex (2017) 1–25.
[47] H. Xu, B. Murphy, A. Fyshe, Brainbench: a brain-image test suite for distributional semantic models, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 2017–2021.
[48] X. Yu, S.P. Law, Z. Han, C. Zhu, Y. Bi, Dissociative neural correlates of semantic processing of nouns and verbs in Chinese language with minimal inflectional morphology, NeuroImage 58 (3) (2011) 912–922.