Neural Encoding and Decoding With Distributed Sentence Representations

Jingyuan Sun, Shaonan Wang, *Member, IEEE*, Jiajun Zhang, *Member, IEEE*, and Chengqing Zong, *Senior Member, IEEE*

***Abstract*— Building computational models to account for the cortical representation of language plays an important role in understanding the human linguistic system. Recent progress in distributed semantic models (DSMs), especially transformer-based methods, has driven advances in many language understanding tasks, making DSM a promising methodology to probe brain language processing. DSMs have been shown to reliably explain cortical responses to word stimuli. However, characterizing the brain activities for sentence processing is much less exhaustively explored with DSMs, especially the deep neural network-based methods. What is the relationship between cortical sentence representations against DSMs? What linguistic features that a DSM catches better explain its correlation with the brain activities aroused by sentence stimuli? Could distributed sentence representations help to reveal the semantic selectivity of different brain areas? We address these questions through the lens of neural encoding and decoding, fueled by the latest developments in natural language representation learning. We begin by evaluating the ability of a wide range of 12 DSMs to predict and decipher the functional magnetic resonance imaging (fMRI) images from humans reading sentences. Most models deliver high accuracy in the left middle temporal gyrus (LMTG) and left occipital complex (LOC). Notably, encoders trained with transformer-based DSMs consistently outperform other unsupervised structured models and all the unstructured baselines. With probing and ablation tasks, we further find that differences in the performance of the DSMs in modeling brain activities can be at least partially explained by the granularity of their semantic representations. We also illustrate the DSM's selectivity for concept categories and show that the topics are represented by spatially overlapping and distributed cortical patterns. Our results corroborate and extend previous findings in understanding the relation between DSMs and neural activation patterns and contribute to building solid brain–machine interfaces with deep neural network representations.**

***Index Terms*— Brain–machine interfaces, distributed semantic representations, neural decoding, neural encoding.**

I. Introduction

UNDERSTANDING the mental representations of linguistic items that enable humans to communicate has always been one of the central goals in cognitive neuroscience. In recent years, functional neuroimaging approaches have provided remarkable insights concerning the neural representations of individual word concepts [1]–[4] and interconcept relations [5], [6]. However, characterizing the representations of more complex units, such as sentences, remains a considerable challenge [7]–[10]. How the neural circuits respond to sentence-level linguistic stimuli and whether brain activity can be directly decoded to infer what a subject is attending to are not fully resolved. These questions have, respectively, been studied as brain encoding and decoding.

Brain encoders predict the neural responses to the linguistic stimuli of interest, which sheds light on how the brain organizes and processes language through neural circuits. Brain decoders categorize and reconstruct the human perception by mining the neural activation patterns, potentially forming the basis of the noninvasive brain–machine interfaces [2], [4], [11]–[14]. Though distinct in motivation, the encoding and decoding models share a core purpose: the associative mapping between linguistic stimuli and the evoked brain responses. To computationally fit the mapping, the representations of functional magnetic resonance imaging (fMRI) signals and the linguistic stimuli are necessary.

To construct semantic representations of sentence stimuli, it is natural to draw upon related work in natural language processing (NLP). Distributed semantic models (DSMs) are currently the dominant text representation methods in the NLP community. Based on the idea that words similar in meaning occur in similar contexts, the original DSMs calculate semantic vectors of words by exploring the linguistic co-occurrence pattern of a given corpus. Then, in the constructed representation space, the similarity between two semantic vectors reflects the similarity of those words. The resulting vectors reliably predict human judgments in multiple tasks, from meaning similarity judgments to concept categorization [15], [16]. Neuroimaging studies have also demonstrated that the extracted word

Manuscript received September 11, 2019; revised April 9, 2020 and August 9, 2020; accepted September 16, 2020. This work was supported in part by the Natural Science Foundation of China under Grant 61906189, in part by the Beijing Municipal Science and Technology Project under Grant Z181100008918017, and in part by the Beijing Advanced Innovation Center for Language Resources and the Beijing Academy of Artificial Intelligence under Grant BAAI2019QN0504. *(Jingyuan Sun and Shaonan Wang contributed equally to this work.) (Corresponding author: Shaonan Wang.)*

Jingyuan Sun, Shaonan Wang, and Jiajun Zhang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shaonan.wang@nlpr.ia.ac.cn).

Chengqing Zong is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 200031, China.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3027595

vectors can be used to model semantic representations in the brain [10], [17], [18].

More recently, the family of DSMs has been extended beyond single words to express the meanings of sentences and discourses. Benefiting from the upsurge of deep neural networks, its methodology also evolved from solely co-occurrence counting to capturing much more diverse semantic, syntactic, and perceptive features of the given text. The sentence-level DSMs roughly fall into two classes: unstructured and structured models. Unstructured models, such as averaging among a word vector sequence, do not explicitly account for sentence structure, thus enjoying minimal parameters and fast training speed [19]. These models are also currently the most widely used DSMs to probe cortical sentence representations [18], [20]. In contrast, structured models can capture the sentence structure at the cost of a higher computational expense. Some of the structured models, such as BERT [21] and InferSent [22], have delivered excellent performance in predicting human similarity judgments and downstream NLP tasks, potentially better correlating with the human brain activation patterns. Nevertheless, compared with the wide adoption of unstructured approaches, most of them have scarcely been explored to probe the neural representations.

In this article, we bridge the latest advances of representation learning in NLP with brain encoding and decoding. As shown in Fig. 1(a), we build a bidirectional mapping between sentence stimuli and the corresponding brain activations, with 12 DSMs to encode sentence representations. Both unstructured and structured models are carefully selected, as outlined in Fig. 1(c), to conduct a comprehensive evaluation. In the encoding evaluation, the DSMs are first compared with respect to predicting cortical activities throughout 27 specific ROIs in language and visual networks. This approach not only allows us to confirm what kind of DSMs better account for brain activities but also revisits previous findings in splitting the cortical semantic system from the data-driven perspective. We find that encoders trained with transformer-based DSMs consistently outperform other unsupervised structured models and all of the unstructured baselines in predicting brain activities in both the language and the visual networks. Aggregating the results, the prediction accuracy for the left middle temporal gyrus (LMTG) and left lateral occipital cortex (LLOC) exceeds those of other regions. In the decoding evaluation, decoders trained with different DSMs are tested by identifying the stimuli from brain activation patterns. In addition to the reliable performance of the transformer-based models, we find that InferSent's decoding accuracy is comparable to those of the transformers in most subtasks.

Text representations produced by most neural network-based DSMs are dense and not easily interpretable. To gain deeper insights into the experimental results, we study two aspects. We first decompose what differences in features captured by DSMs best explain their different correspondence with brain activation patterns. Fueled by a set of probing tasks, the DSMs' abilities to account for the surface, syntactic, and semantic properties of a sentence are scored and correlated with the encoding and decoding accuracy. For the language networks, semantic probing scores of DSMs significantly correlate with encoding accuracy on more ROIs than syntactic and surface probing. We validate the findings with a set of ablation tasks. Second, we aggregate the encoding results of all DSMs on different topics and display them among the brain language and vision networks. We find that a topic is not exclusively selective to certain ROIs, but rather represented by spatially overlapping and distributed cortical patterns.

The findings of this article not only demonstrate the weaknesses and advantages of different DSMs in explaining cortical sentence processing. The results offer deeper insight into the connection between the two manifestations of mental meanings: the neural activation patterns and the extrinsic linguistic representations. We hope these findings could boost mutual promotion between understanding cortical linguistic representations and developing machine-learning semantic representation models.

The rest of this article is organized as follows. The experimental setup and the sentence stimuli for collecting the fMRI data sets are briefed in Section II-A. How to train and evaluate the brain encoders and decoders is detailed in Sections II-B–II-D. Section II-E introduces the 12 DSMs tested in this article, Sections II-F and II-G describe the probing and ablation tasks to explain the encoding and decoding performance of the DSMs. Section III presents the experimental results. In Section III-A, we report the neural encoding results. We first show the pairwise matching accuracy in language and visual networks, then explain the encoding performance with probing and ablation tasks, and finally discuss the topic distribution among different ROIs derived from the encoding performance. In Section III-B, we report the decoding results in a similar order to that in Section III-A. However, in Sections III-B3 and III-B4, we further show the distribution patterns of informative voxels and the decoding performance with voxels constrained in several functional networks.

II. Materials and Methods

A. Brain Imaging Data

We use the fMRI neural activation data published in [20], acquired on a whole-body 3-Tesla Siemens Trio scanner with a 32-channel head coil. Two experiments are carried out in the original work to collect imaging data. Experiment 1 involves the scanning of eight subjects, while experiment 2 involves the scanning of five subjects. The sentence stimuli are organized in the hierarchy of topic–passage–sentence in both experiments, as shown in Fig. 1(b).

In experiment 1, 96 passages are presented, each consisting of four sentences about a particular concept. The passages come from 24 broad topics (e.g., professions and opera), each providing basic information about the corresponding concept in the style of Wikipedia. In experiment 2, 72 passages are used, each consisting of three or four sentences about a particular concept. The passages also span a broad range of content areas that are unrelated to the topics in experiment 1 (such as skiing, dreams, opera, and bone fractures). The materials

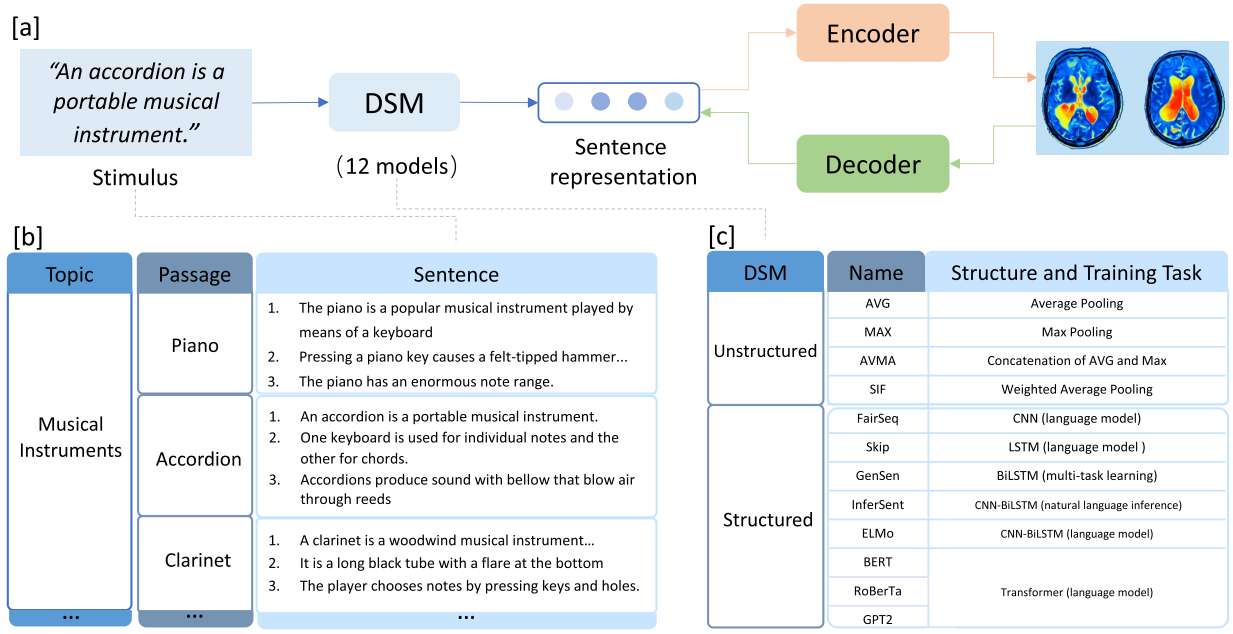

Fig. 1. (a) Neural encoding and decoding with DSM produced sentence representations. 12 DSMs are evaluated, respectively, transforming sentences to vector representations. Different encoders and decoders are trained based on each DSM. (b) Example of sentence stimuli organized in the hierarchy of topic–passage–sentence, taking the musical instrument topic as an instance. (c) Name, structure, and training tasks of every tested DSM.

include 48 Wikipedia-style passages and 24 first-/third-person narratives. The two experiments are comparable in their within and between-passage/topic semantic similarities.

All passages are presented sentence by sentence. Each sentence is presented for 4 s followed by a 4-s fixation gap. The entire set of sentences is presented three times. The participants are asked to read the sentences that they are presented to ensure attentive scanning. All subjects are scanned three times for every sentence stimuli. The scan is running consistently during the presence of sentence stimuli. Next, the acquired data series is corrected by slice timing and motion correction and then concatenated to align the sentence. The fixation gap is used to separate the sentences and distinguish brain activities of language processing from other (noisy) brain activities. Details of the experimental setup, materials, and presentation scripts are available online.[1]

To fully use the data set and comprehensively display the experimental findings, we discuss the encoding results on experiment 1's 384-sentence data and the decoding results on experiment 2's 243-sentence data.

B. Encoding Methodology

Regression-based encoding learns a mapping to predict the brain activities from the distributed sentence representations. Given the voxel matrix $X_e \in \mathbb{R}^{N_E \times N_V}$ and sentence representation matrix $Z_e \in \mathbb{R}^{N_E \times N_D}$ in the training set, where N_E denotes the number of examples, N_V denotes the number of voxels, and N_D denotes the number of dimensions for sentence representation, the regression coefficients of encoder W_e are estimated to minimize

$$||W_e Z_e - x_i||_2^2 + \lambda ||W_e||_1 \tag{1}$$

for each column z_i in Z, i.e., each dimension of the sentence vectors. λ is the regularization parameter that is 0.1 in our implementation.

The encoder is trained within the fivefold cross-validation procedure. In each fold, the regression parameters for each dimension are learned by predicting the brain images for 307 sentences from distributed representations and tested by predicting activation patterns for the 77 left-out sentences. The voxelwise normalization is carried out using a mean image derived from the training set, which is also subtracted from the test set. The cross-validation procedure is carried out on data from all five subjects with each of the DSMs, respectively. After fivefolds, this would result in 384 encoded brain activation vectors for each combination of subjects and DSMs. We evaluate the encoding results by running pairwise matching on the 384 encoded brain activation vectors. Pairwise matching is detailed in Section II-D.

C. Decoding Methodology

Opposite to encoding, decoding operates by estimating a semantic vector directly from the voxels, with each dimension predicted by a separate regression model. Formally, given the voxel matrix $X_d \in \mathbb{R}^{N_E \times N_V}$ and sentence representation matrix $Z_d \in \mathbb{R}^{N_E \times N_D}$ in the training set, where N_E denotes the number of examples, N_V denotes the number of voxels, and N_D denotes the number of dimensions for sentence representation, the regression coefficients of decoder W_d are estimated to minimize

$$||W_d X - z_i||_2^2 + \lambda ||W_d||_1 \tag{2}$$

[1]https://osf.io/crwz7/wiki/home/

for each column z_i in Z, i.e., each dimension of the sentence vectors. λ is the regularization parameter that is 0.1 in our implementation.

The decoder is also trained within fivefold cross validation. The regression parameters for each dimension are learned from mapping the brain images for 194 sentences to distributed representations and tested by predicting semantic vectors from the brain images for 49 left-out sentences in each fold. The voxelwise normalization is carried out using a mean image derived from the training set, which is also subtracted from the test set. The cross-validation procedure is carried out on data from all five subjects with each of the DSMs, respectively. After fivefolds, this would result in 243 decoded semantic vectors for each combination of a subject and a DSM. We evaluate the decoding results by running pairwise matching on the 243 decoded semantic vectors. Pairwise matching is detailed in Section II-D.

Each decoder is trained on a reduced image of 5000 voxels, approximately 10% of the number left after applying a cortical mask. Following the settings of [20], these voxels are selected by the degree to which they are informative about the distributed semantic vectors. For all subject and DSM combinations, the voxel selection is done separately in each cross-validation fold within the training set. Especially, before the training began, we learn regression models to predict each semantic dimension from the imaging data of each voxel and its 26 adjacent neighbors in 3-D. This yielded predicted values for each semantic dimension, which are then correlated with the values in the ground-truth sentence vectors. Finally, the informativeness score for each voxel is the average value of such correlation across dimensions.

D. Pairwise Matching

The encoding and decoding results are evaluated by the pairwise matching task, which is one of the most widely adopted metrics herein. We explain the pairwise matching task, taking brain decoding as an example. As shown in Fig. 2, we compute the correlation between the decoded vectors and the ground-truth sentence embeddings. If the decoded semantic vectors are more similar to their respective brain activation patterns than to the alternative, a successful matching is scored. Formally, for each possible pair of sentences S_i and S_j, let X_{S_i} and X_{S_j} denote the brain activation patterns aroused by S_i and S_j. Let Z_{S_i} and Z_{S_j} denote the sentence embeddings produced by a DSM, while D_{S_i} and D_{S_j} denote the decoded semantic vectors from brain images X_{S_i} and X_{S_j}. We score 1 for a matching if

$$\begin{aligned}\text{corr}\left(D_{S_i}, Z_{S_i}\right) + \text{corr}\left(D_{S_j}, Z_{S_j}\right) \\ > \text{corr}\left(D_{S_i}, Z_{S_j}\right) + \text{corr}\left(D_{S_j}, Z_{S_i}\right) \quad (3)\end{aligned}$$

else 0. corr denotes the correlation function where we adopt Pearson's correlation. Each possible pair of sentences will be explored. The final matching accuracy for each participant is the fraction of correctly matched pairs. If the model chose the match at random, the expected chance-level accuracy would be 0.50. Similar procedures are conducted for evaluating brain encoding results, while the difference lies in that we match encoded vectors to the ground-truth brain activities.

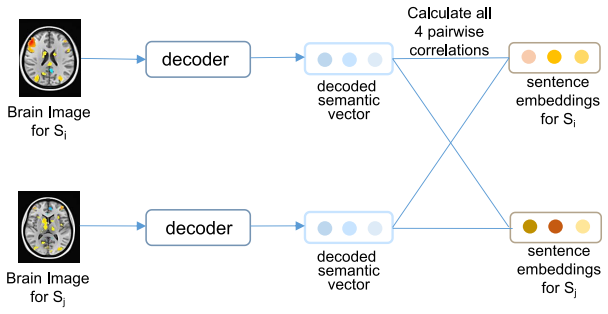

Fig. 2. Pairwise matching task for evaluating brain decoding, which is correct if vectors decoded from two images are more similar to the sentence embeddings for their respective stimuli than to the alternative.

Note that, for evaluating brain decoding, we adopt [20]'s setting where the pairwise matching task is further divided into three subtasks. These subtasks constrain S_i and S_j to be, respectively, from: 1) different topics (e.g., a sentence about a piano versus a butterfly); 2) different passages from the same topic (e.g., a sentence about a dragonfly versus a butterfly); and 3) different sentences within the same passage (e.g., two sentences about a piano).

E. Distributed Semantic Models

Moving beyond representations at the word level, multiple methods have been proposed to build distributed representations of higher level linguistic units, such as sentences. The sentence-level DSMs roughly fall into two classes: unstructured and structured models. Unstructured models, such as averaging among word embedding sequence, do not explicitly account for structural information of a sentence, thus enjoying minimal parameters and fast training speed [19]. They are also the most widely used methods to probe cortical sentence representations. In contrast, structured models can capture the sentence structure and model how words or phrases affect each other at the cost of higher computational expense. Though more powerful in expressibility, these models have not been fully explored in neural imaging studies. Representation models that deliver state-of-the-art performance in NLP tasks [23]–[25] are mostly structured, such as BERT [21] and GPT2 [26]. Structured models can be further classified into unsupervised and supervised methods.

1) Unstructured Models: A straightforward technique for representing a longer piece of text is to ignore the sequence structure and treat it as a bag of words. For a sentence, simply averaging the vectors for the individual words can produce representations that roughly match the human similarity judgments [19]. This approach is tantamount to an average pooling on the sequence of word vectors. Average pooling is also one of the most widely used methods to represent stimuli in sentence-level neural encoding and decoding [10], [18]. Averaging offers the advantage of aggregating the semantic information of every single word, but it also dilutes the most salient features of the sentence. Since only a small number of words in a sentence contribute most to its meaning,

max-pooling has also been adopted. It extracts the maximum value along each dimension of the word embeddings, selecting the most salient features of all dimensions in the sentence representation. Intuitively, features extracted by averaging and max-pooling capture complementary semantic information of a sentence as shown by Shen *et al.* [19]. The two extracted features can, thus, be concatenated as the third representation. Besides, Arora *et al.* [27] introduced SIF,[2] an improved average pooling weighted by word frequency that beats more complicated methods in standard natural language tasks.

2) Unsupervised Structured Models: Unsupervised structured models generally adopt the idea of language modeling, learning to simultaneously predict the next and previous sentences from the encoding of the current sentence. Recurrent neural networks (RNNs), convolutional networks (CNNs), and the transformer are currently the most mainstream building blocks of language modeling in the NLP community. Benefiting from their network architectures, the trained sentence encoders catch the syntactic information, such as word order and tree structure of a sentence.

Skip-thought[3] is a classical RNN-based language model. As proposed in [28], this approach is a natural extension to skip-gram, which is based on an encoder-decoder framework with a standard gated rectifier unit (GRU) encoder and a conditional GRU decoder. Given a sentence tuple $(\boldsymbol{s}_{i-1}, \boldsymbol{s}_i, \boldsymbol{s}_{i+1})$, let $w_i^1, \ldots, w_i^N$ be the words in sentence $\boldsymbol{s}_i$ and N denote the number of words in the sentence. At each time step, the encoder produces a hidden state $\mathbf{n}_i^t$, which can be interpreted as the representation of the sequence $w_i^1, \ldots, w_i^t$. The trained encoder maps a target sentence to maximize the probabilities of predicting the contexts. Thus, the objective optimized is the sum of the log-probabilities for the forward and backward sentences conditioned on the encoder representation

$$\sum_t \log P\left(w_{i+1}^t | w_{i+1}^{<t}, \mathbf{h}_i\right) + \sum_t \log P\left(w_{i-1}^t | w_{i-1}^{<t}, \mathbf{h}_i\right). \quad (4)$$

Sentences similar in distribution pattern can then be mapped to closer vectors in the produced representation space. For skip-thought, we choose the encoder's last layer and average its hidden states to produce sentence representation.

In addition to the RNN, both the CNN and transformer networks have been applied in language modeling. These networks generally share the idea of skip-thought as (3) and differ in the underlying structures. We choose FairSeq[4] to represent CNN-based models. The large-scale transformer-based language models have driven much recent progress in natural language understanding tasks. They currently dominate the family of unsupervised structured DSMs. We choose BERT [21], GPT2 [26], and RoBerTa [29] as typical transformer-based [30] models. It is worth noting that BERT is more frequently adopted than the other two transformer-based models in some recent work of probing brain activities with distributed representations. For example, Gauthier and Levy [13] discussed whether syntactic light BERT representations lead to better decoding performance, with the same brain imaging data as we use. Abnar *et al.* [16] proposed a variant of RSA to analyze the relationship between the layerwise representation of BERT with brain activation patterns.

3) Supervised Structured Models: InferSent[5] proposed in [22] is a supervised model trained on the Stanford Natural Language Inference (NLI) data sets. In this model, sentences are encoded by a bidirectional long short-term memory (LSTM) network and fed into a three-class classifier to conduct NLI. For a sequence of n words, a bidirectional LSTM computes a set of n vectors, and each vector is a concatenation of a forward LSTM and a backward LSTM that reads the sentence in opposite directions. Similar to previous methods, hidden states of the last encoder layer are transduced as the sentence representation. Two strategies are compared by the author: selecting the maximum value over each dimension of the hidden units (max-pooling) or by considering the average of each dimension (mean-pooling). In their experiments, the max-pooling strategy gives the best results in terms of accuracy on several downstream tasks, and such a setting is followed in this article's evaluation. InferSent is trained on a single supervised task, NLI, while sentence encoders trained with multitask strategy are also available, such as GenSen [31]. GenSen shares a single recurrent sentence encoder across multiple tasks, combining the inductive biases of diverse training objectives in a single model.

F. Probing Tasks

To have a clear understanding of what difference in the features captured by DSMs lead to their diverse performances in modeling the brain activation patterns, we probe the DSM's ability to encode surface, syntactic, and semantic features in sentence embeddings, and correlate them to the encoding or decoding performance. We first introduce the three probing tasks.

1) Surface Probing: We evaluate the extent to which the sentence length can be predicted from a DSM's representation. Length is considered as a surface feature of a sentence because it can be acquired without linguistic knowledge. We follow the task setting proposed by Conneau *et al.* [32]. 100k sentences as training sets and 10k sentences as validation and test sets are provided. These sentences are extracted from Toronto Book Corpus and in the 5-to-28 word range. They have been grouped into six equal-width bins by the number of words. In the probing task, a single hidden-layer MLP will be trained on the six-way classification of sentence length, taking only sentence embeddings as input. All the DSMs that we evaluate in this article will encode the sentences to finish this length classification task. The test accuracy is reported as the corresponding DSM's probing score for modeling sentence length. There might be other surface probing tasks available, but they have the risk of intrinsically correlating with the following syntactic and semantic probing that we implement. Thus, we only adopt sentence length probing.

2) Syntactic Probing: We probe a DSM's ability to encode syntactic information with the approach from [33], inspired by Gauthier and Levy [13]. This probing approach measures

[2]https://github.com/PrincetonML/SIF

[3]https://github.com/ryankiros/skip-thoughts

[4]https://github.com/pytorch/fairseq/tree/master/examples/language_model

[5]https://github.com/facebookresearch/InferSent

the degree to which syntactic analyses of a sentence can be reconstructed from its word embeddings. It operates on a parsed corpus, and we use sentences sampled from the Universal Dependencies English Web Treebank (UN-EWT) corpus. For a given DSM m and each sentence s_i in the corpus, let $w_1, \ldots, w_n$ denote words in s_i and $m(w_j)$ denote the DSM's embedding of word w_j. A transition matrix B is estimated so that

$$B\big(m(w_j) - m(w_k)\big)^T \big(B\big(m(w_j) - m(w_k)\big)\big)^2 \approx \big|m(w_j) \leftrightarrow m(w_k)\big| \quad (5)$$

where w_j and w_k denote the two words in a sentence. $|m(w_j) \leftrightarrow m(w_k)|$ denotes the number of edges separating the words in a dependence parsing tree of the sentence, which is actually a measure of syntactic distance. Once B is estimated on the train set, it can be applied to the test set to predict the distance between any two words in a sentence. This yields a $n \times n$ distance matrix for an n-word sentence, from which an undirected parsing tree can be derived with minimum spanning tree algorithm. The accuracy of such tree reconstruction is measured by calculating the unlabeled attachment score (UAS) to the ground-truth parsing of the sentence.

3) Semantic Probing: In this task, we probe a DSM's ability to account for the semantic information of sentences. We include the SICK data set [34] and the STS Benchmark [35] to evaluate the DSMs in predicting semantic relatedness between sentences. Sentence pairs in these data sets have been annotated with a relatedness score between 0 and 5. Similar to the surface probing task, a single hidden-layer MLP learns on top of DSM encoded embeddings to predict the relatedness score of two input sentences. The Pearson correlation between the predicted distribution and ground-truth score on the test set is reported. We average the performance on SICK and STS tasks as the semantic probing score of a DSM.

Once all the abovementioned probings are done, the scores are correlated with encoding and decoding accuracies of DSMs. Formally, taking the syntactic probing task and neural encoding as an example, we have the probing score $\mathrm{PS}_{\mathrm{syn}} : \{\mathrm{PS}^0_{\mathrm{syn}}, \mathrm{PS}^1_{\mathrm{syn}}, \ldots, \mathrm{PS}^i_{\mathrm{syn}}\}$ on every evaluated DSM i. Given every DSM's encoding accuracy $\mathrm{Acc}_{\mathrm{SB}} : \{\mathrm{Acc}^0_{\mathrm{SB}}, \mathrm{Acc}^1_{\mathrm{SB}}, \ldots, \mathrm{Acc}^i_{\mathrm{SB}}\}$ on a specific subject SB, we calculate the correlation as follows:

$$\mathrm{corr}\big(\mathrm{PS}_{\mathrm{syn}}, \mathrm{Acc}_{\mathrm{SB}}\big) = \mathrm{pearson}\Big(\Big\{\mathrm{PS}^0_{\mathrm{syn}}, \mathrm{PS}^1_{\mathrm{syn}}, \ldots, \mathrm{PS}^i_{\mathrm{syn}}\Big\}, \Big\{\mathrm{Acc}^0_{\mathrm{SB}}, \mathrm{Acc}^1_{\mathrm{SB}}, \ldots, \mathrm{Acc}_{\mathrm{SB}}\Big\}\Big). \quad (6)$$

There are three probing scores for each of the evaluated DSMs. They will be, respectively, correlated with encoding and decoding accuracy among all the subjects. Such a setting allows us to pin-point catching what difference in the linguistic features leads to the different encoding and decoding performance of the DSMs.

G. Ablation Tasks

We produce two ablation tasks to support the findings in the probing tasks, evaluating why a DSM might fail or succeed in modeling activities in a cause-effect manner. If a certain feature captured by the DSM did not matter in modeling the brain activities, the corresponding ablation would not decrease its encoding or decoding performance, and vice versa. Each ablation task is a modified form of the standard language modeling task. The training corpora of the task is manipulated to select against some particular features of linguistic representation, such as sentence meaning or grammatical structure. As a control, the model will also be tuned on original correct sentences in normal language modeling, denoted as control. All the sentences in the ablation data sets are extracted from the Toronto Book Corpus, which is widely used in pretraining language models, such as BERT and RoBerTa.

1) Language Modeling With Scrambled Word Order: An ablated language modeling task is designed to select against the fine-grained syntactic representation of models, as inspired by Gauthier and Levy [13]. This task is denoted as scramble, where the word order of every input sentence is randomly shuffled to remove all first-order cues to its syntactic structure. Note that only the word order is scrambled. No word is added, removed, replaced, or morphologically changed in the sentence. Tuning the language models on this data set means to predict a missing word from a bag of its content words, without structural hints.

2) Language Modeling With Content Words Modified: We next design a task select against fine-grained semantic representations of a DSM, compromising its ability in encoding the semantic relation between words in a sentence. This is achieved by feeding anomalous sentences that contained semantically implausible complements to the model. We make anomaly by interfering with the verbs and nouns that are usually the main content words of a sentence. Every single verb or noun in a sentence is replaced by another random verb or noun that is incompatible in meaning with its contexts. We make sure that such a replacement only causes semantic anomaly but not a syntactic error. We abbreviate this task as remword.

III. Experimental Results

A. Neural Encoding

In this section, we discuss the relationship between cortical sentence representations against DSM by using them as predictive models of the brain responses given sentence stimuli. In the encoding experiments, we first study how the tested DSMs predict the activities of regions of interest (ROIs), from several large-scale brain networks specifically linked to high-level cognition and/or semantic processing. The selected networks are as follows: 1) the fronto-temporal language selective network [36]; 2) the ensemble semantic system [37]; and 3) the visual network [38] (including scene, face, body, and object network). We compare the predicting accuracy delivered by different DSMs with a pairwise matching task, evaluating which kind of model consistently outperforms other baselines.

Knowing the encoding performance, we further study which of the linguistic features that DSMs captured better explained the accuracy of predicting brain activities. With a set of

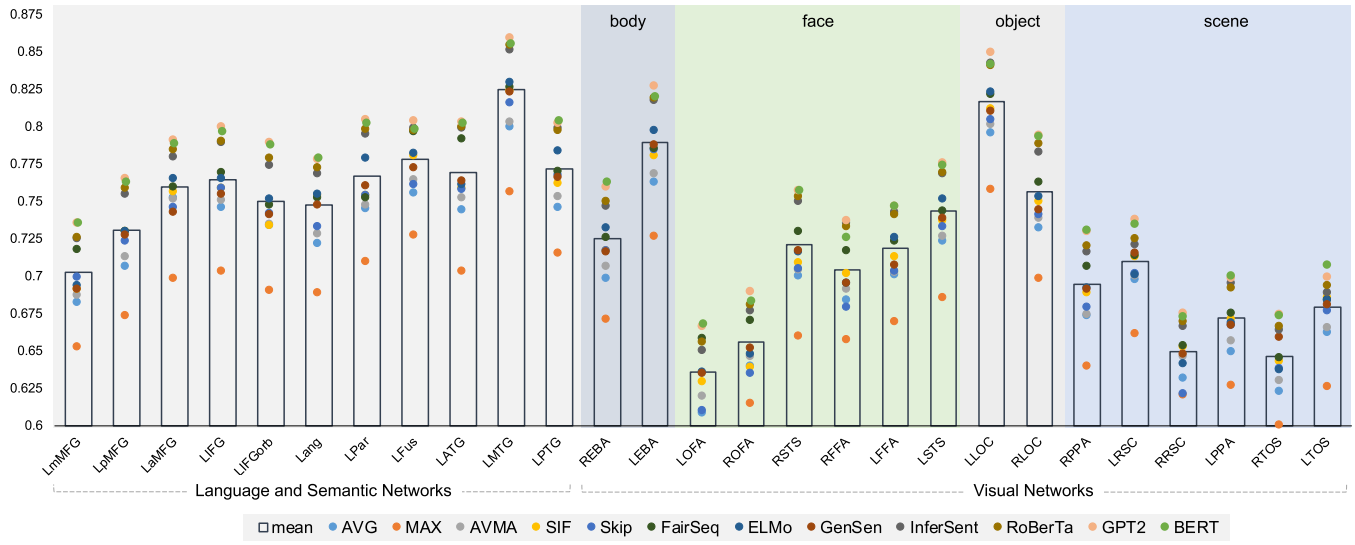

Fig. 3. Different DSMs' pairwise matching accuracy on ROIs in language and visual networks, averaging across encoders trained on different subjects (eight in total). The visual network is further decomposed to body, face, object, and scene area. Scatters with different colors denote the matching accuracy of encoder trained with different DSMs.

probing and ablation tasks, we offer a hint regarding how the selectivity of different linguistic features may influence the DSMs' encoding performance. In addition to catching linguistic features, DSMs can reflect the semantic category of a sentence in the embeddings. We show the topic encoding accuracy on different ROIs in language and visual networks by averaging across all the DSMs and subjects, which helps to reveal the topic selectivity of different brain areas. The findings from analyzing the encoding performance are validated by paired t-test with the Bonferroni correction. The results are reported throughout with a significance level of $\alpha = 0.01$.

1) Prediction Within Regions of Interest: We use encoders, respectively, trained with different DSMs to predict activation patterns of the brain language and visual networks. A DSM's encoding accuracies on eight subjects are averaged for each ROI, as shown in Fig. 3. We also estimate a null distribution of the encoding accuracy for each ROI, achieved by training encoders to predict the activities of those ROIs with random sentence vectors and evaluating by pairwise matching. The null encoding accuracy is consistent among the ROIs and falls within 0.502 ± 0.039 with 95% confidence intervals.

Each DSM has 27 encoding scores after averaging across all subjects, 11 for predicting the ROIs in the language and semantic networks and 16 for the visual networks. Pairwise comparison of the ROI scores of different DSMs reveals a DSM's overall correlation with the brain activation patterns in a functional network. In predicting the ROIs of the language and semantic networks, the three transformer-based models significantly outperform the other unsupervised structured models (ELMo, Skip-thought, and FairSeq) and all the unstructured baselines (all significant results have $p < 0.01$, and p-values are detailed in Table I(a) in the Supplementary Material). The structured models do not consistently outperform the unstructured models. For example, AVMA embedding is comparable with skip-thought in predicting the language networks. In predicting the ROIs of the visual networks, the three transformer-based models also significantly outperform the other unsupervised structured models and all of the unstructured baselines (all significant results have $p < 0.01$, and p-values are detailed in Table I(b) in the Supplementary Material). The exception is InferSent, a supervised structured model whose encoding performance is not significantly different from that of RoBerTa in predicting the visual networks ($p > 0.01$). Among all the tested models, max-pooling (MAX) delivers the lowest matching accuracy in predicting both the language and visual networks (all significant results have $p < 0.01$, where the p-values are detailed in Table I(c) and (d) in the Supplementary Material).

Every single ROI is predicted by 12 DSMs, so it is featured by a list of 12 matching scores. Pairwise comparison of the 12-DSM score of different ROIs can give a hint on an ROI's selectivity to sentence-level linguistic features. We find that LMTG is more accurately predicted than the other ROIs in the language networks (all significant results have $p < 0.01$, and p-values are detailed in Table II(a) in the Supplementary Material), including the left anterior and posterior temporal gyrus [LATG and left posterior temporal gyrus (LPTG)]. Matching accuracy on LmMFG (left mid-portion of the middle frontal gyrus) is significantly lower than other ROIs in the language networks (all $p < 0.01$, and p-values are detailed in Table II(b) in the Supplementary Material). In the visual networks, the LLOC is significantly more predictable by the DSMs than other ROIs, including left occipital complex (LOC) on the right hemisphere RLOC (all significant results have $p < 0.01$, and p-values are detailed in Table II(c) in the Supplementary Material).

2) Explanation With Probing and Ablation Task: To explain what difference in the feature they capture lead to their accuracy gap in predicting the brain, we conduct three probing tasks, correlate the probing score to the encoding performance, and depict the results in Fig. 4(a).

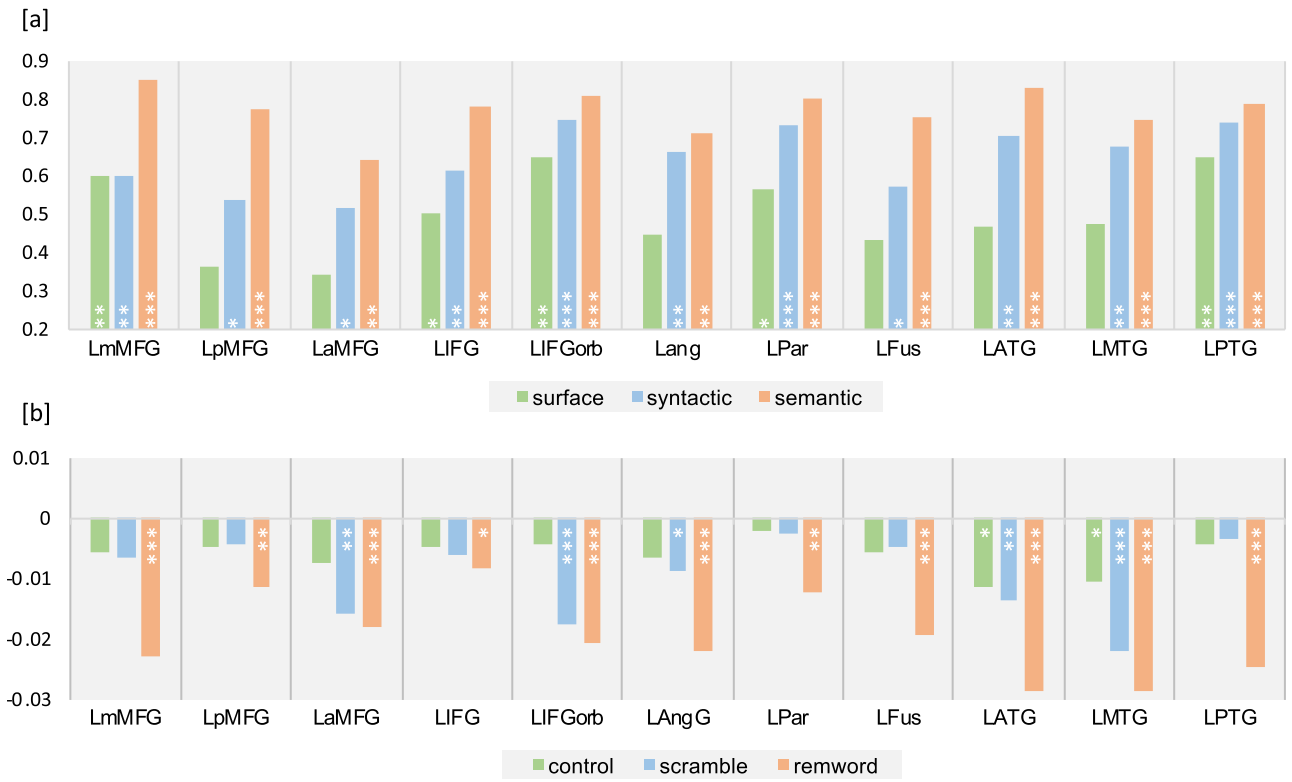

Fig. 4. (a) Correlation of probing task performance and encoding accuracy on the language network ROIs, reported in the Spearman correlation. (b) Pairwise matching accuracy of BERT after tuned on the ablation tasks, graphed relative to the encoding performance of its corresponding pretrained (untuned) model. The depicted result is averaged across eight subjects. $*$ at the bottom of bar denotes the significance of correlation. $*$ means $0.05 \leq p$-value < 0.1, $**$ means $0.01 \leq p$-value < 0.05, and $***$ means p-value < 0.01. No $*$ means p-value > 0.1.

As shown in Fig. 4(a), surface probing scores do not significantly correlate with the matching accuracy on all the ROIs in the language networks ($p > 0.01$ for each ROI). This means that difference in encoding sentence length in the embeddings does not reliably explain DSMs' different encoding performance in the language networks. Syntactic probing scores correlate with matching accuracies on lateral parietal regions (LPar), LPTG, and left inferior frontal gyrus-pars orbitalis (LIFGorb), with $p < 0.01$ for these three ROIs. This means that, in the three ROIs, the UAS of reconstructing dependence-parsing tree from different DSM's sentence embeddings accounts for a significant large portion of their differences in encoding accuracy. Semantic probing scores significantly correlate with matching accuracies on all ROIs ($p < 0.01$ for each ROI) except LaMFG ($p > 0.01$) in the language networks.

From the abovementioned experiments, we find that different DSM's semantic probing scores significantly correlate with the encoding accuracy on most of the ROIs in the language atlas. To further verify this finding, we pick BERT that reliably encodes the brain activities of the language network and conduct an ablation test. We systematically select against a model's syntactic or semantic representation by tuning them on the corresponding ablation data set and see what interference significantly influences DSM's encoding accuracy for ROIs in the language networks. As shown in Fig. 4(b), the interference of an ablation task leads to different results in different ROIs.

For BERT, tuning on the remword task yields decreased matching accuracy relative to its untuned baseline on 8 of the 11 ROIs (all $p < 0.01$), except left post-potion of middle frontal gyrus (LpMFG), left inferior frontal gyrus (LIFG), and LPar. On these eight ROIs, tuning on the remword task also leads to significantly lower matching accuracy than the control task (all $p < 0.01$). This corroborates with our findings that DSMs' different levels of semantic selectivity influence their performance in predicting the brain activities aroused by sentence stimulus. Tuning on the scramble task yields decreased matching accuracy on two ROIs: LIFGorb and LMTG (each $p < 0.01$). On the other ROIs, the differences in accuracy between tuning on the scramble task and the pretrained BERT baseline do not stand up to the statistical significance test.

3) Cortical Representation of Different Topics: The linguistic stimuli in the experiments are organized in the hierarchy of topic–passage–sentence, as shown in Fig. 1(b). Each time a sentence is correctly matched to its corresponding brain activation patterns, we record a value of 1 and otherwise 0. When all folds of cross validations are done, each sentence has a list of historical scores that are then averaged as this sentence's matching accuracy. The matching accuracies of all sentences from the same topic are further averaged to obtain the final matching score of that topic. All 384 sentences come from 24 topics. After pooling such metrics across all subjects and DSMs, each ROI receives a list of 24 scores that

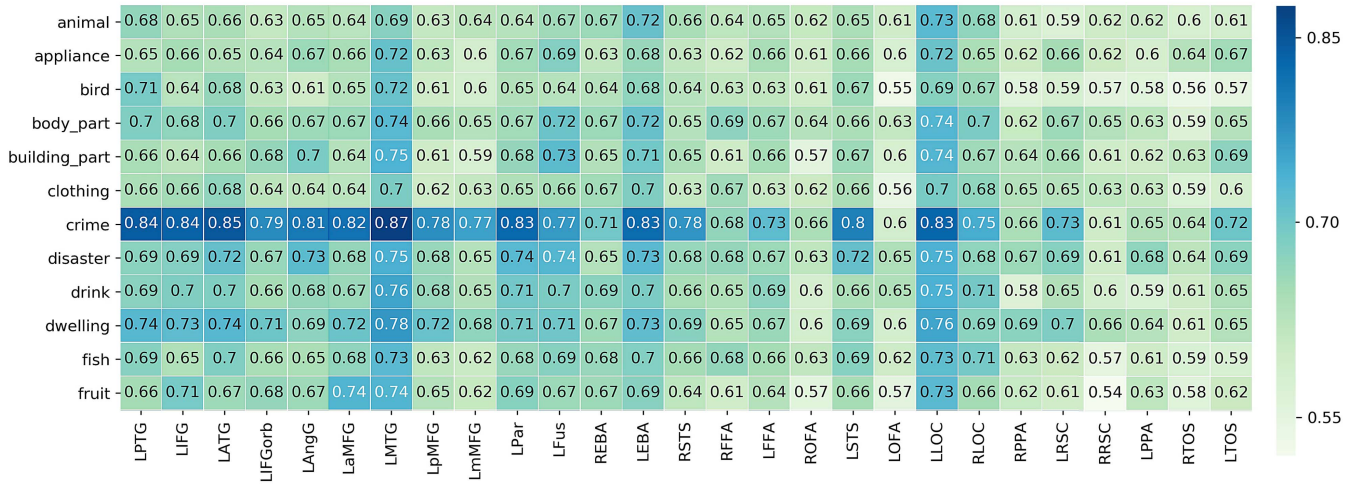

Fig. 5. Each topic's pairwise matching accuracy on ROIs in the language and visual networks, averaging across encoders trained on different subjects (eight in total) with different DSMs (12 in total). The horizontal axis represents different ROIs, while the vertical axis represents the topics. Only half of the 24 topics are depicted due to space limit, while the full result is shown in Fig. 3 in the Supplementary Material.

comprehensively reflect how well brain activities aroused by different topics can be predicted from that ROI. These results are depicted as a heat map in Fig. 5. Only half of the 24 topics are depicted due to space limitations; the full results are shown in the Supplementary Material. After pairwise comparison of different ROI's 24-topic scores, we find that LMTG is better predicted than other ROIs in the language and semantic networks among the topics (all significant results have $p < 0.01$, where the p-values are detailed in Table IV(a) in the Supplementary Material). LLOC is better predicted than other ROIs in the visual networks (all significant results have $p < 0.01$, where the p-values are detailed in Table IV(b) in the Supplementary Material) among all the topics. The difference between LLOC and LMTG is not significant ($p > 0.3$). These findings are consistent with the results in Fig. 3. Brain activities aroused by different topics tend to be more accurately predicted on LMTG and LLOC. However, this fact does not indicate that the topics are exclusively represented by a certain ROI. Oppositely, the heat map in Fig. 5 shows visibly that one topic can be predictable on multiple ROIs; also, in one ROI, there are multiple topics that can be predicted. Simply speaking, the results seek to indicate that many topics are represented by many ROIs. We found a topic is not exclusively selective to a certain ROIs but represented by spatially overlapping and distributed cortical patterns. For instance, the topic "crime" can be encoded on the posterior MTG (pMTG), fusiform gyrus (LFus), and inferior frontal gyrus (IFG).

B. Neural Decoding

In this section, we discuss the relationship between cortical sentence representations against DSMs by using them to decode sentence stimuli from brain responses. The regression model is trained and tested on different subsets of the 243 sentences in the fivefold cross validation for each participant. We first show the decoding performance with the informative voxels selected by the 12 DSMs. To determine which feature difference the DSMs capture may explain their different decoding performances, we probe the DSMs and correlate the probing score with the decoding accuracy. We also depict the DSM's decoding accuracy on different topics, revealing their possible semantic selectivity in Section III-B2. Since decoding is carried out with selected informative voxels, we illustrate the cortical distribution of the selected voxels among three functional networks. For a fair comparison, we conduct the pairwise matching test with voxels constrained in three functional networks in Section III-B3. Findings from analyzing the decoding performance are validated by paired t-test with the Bonferroni correction. The results are reported throughout with a significance level $\alpha = 0.01$.

1) Decoding With Informative Voxels: In this section, we conduct neural decoding with selected informative voxels and test with the pairwise matching task. Three subtasks in progressively finer granularity are included, matching sentences coming from: 1) different topics (e.g., a sentence about a piano versus a butterfly); 2) different passages from the same topic (e.g., a sentence about a dragonfly versus a butterfly); and 3) different sentences within the same passage (e.g., two sentences about a piano), for all possible pairs in every subtask. Matching sentences from the same passage is expected to be the most difficult subtask because sentences in one passage all describe the same object in the stimulus data set that we use. For two sentences of the same passage, it is possible for words to overlap and sentence meanings tend to be close.

To test the significance of results, we estimate null performance by training decoders with random sentence embeddings and evaluating the decoded semantic vectors with pairwise matching. We run the null experiment five times and obtain the average accuracy for each subject on the subtasks. Among the subjects, the null matching accuracy of the first subtask falls within 0.5006 ± 0.022, and the second falls within 0.5014 ± 0.035, while the third falls within 0.5061 ± 0.028, all with 95% confidence intervals. The null performance is consistent with the expected chance level accuracy (0.5) for pairwise matching.

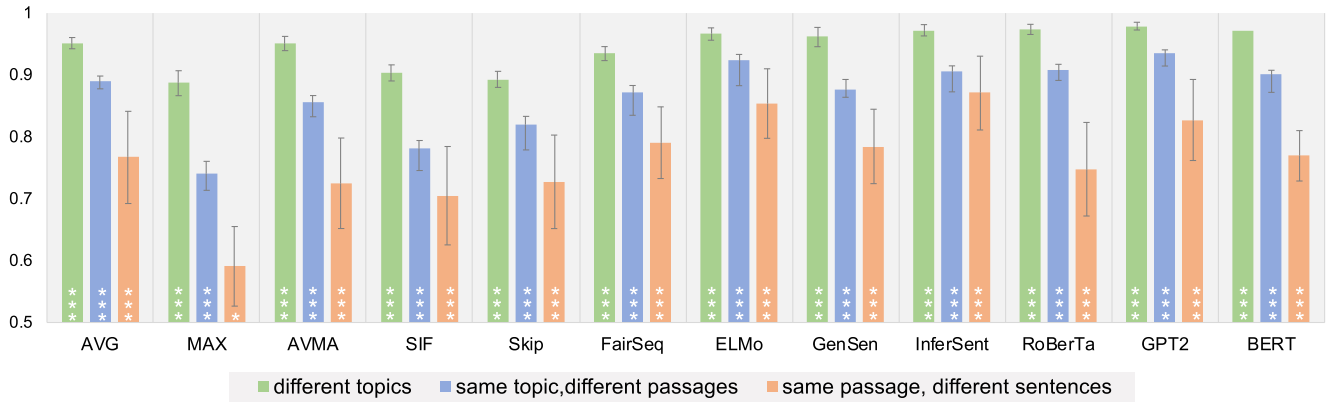

Fig. 6. Pairwise matching accuracy of brain decoding. The colored bar denotes averaged accuracy, while error bar denotes standard deviation of accuracy across different subjects. We report three measures: matching sentences from: 1) different topics (colored green); 2) different passages within the same topic (colored blue); and 3) the same passage (colored orange). $*$ denotes the significance level of comparing the corresponding DSM's decoding accuracy with the null decoding performance. $*$ means $0.05 \leq p$-value < 0.1, $**$ means $0.01 \leq p$-value < 0.05, and $***$ means p-value < 0.01. No $*$ means p-value > 0.1.

Fig. 6 shows the matching accuracy of different sentence representations. In decoding the sentence stimuli from different topics, all the tested representations perform significantly above the null level (all significant results have $p < 0.01$, and p-values are detailed in Table V(a) in the Supplementary Material). The three transformer-based models deliver matching accuracy higher than 0.9 (all significant results have $p < 0.01$, and the p-values are detailed in Table V(b) in the Supplementary Material). InferSent also yields an impressive performance that is not significantly different from that of RoBerTa ($p > 0.1$). Averaging achieves satisfactory performance but does not rank the top, even in the unstructured-based methods. We do not observe that structured models consistently outperform the unstructured models. In decoding the sentence stimuli from different topics, the performance pattern is also consistent across subjects. Except for SIF and skip-thought, the matching accuracy of each DSM among the subjects has a standard deviation lower than 0.015.

As the task becomes finer in granularity, the performance patterns change. In the third task of decoding sentences from the same passage, MAX does not significantly outperform the null level ($p > 0.01$). Skip-thought was inferior to AVMA in the first subtask ($p < 0.01$), but, in the third subtask, their performance difference is no longer significant ($p > 0.1$). In the third subtask, sentences from the same passage might use semantically related words to describe one single concept, as shown in Fig. 1(b). This means that, by merely pooling on the word embeddings, it is largely possible to produce similar sentence representations. We think that is where the structured and other auxiliary linguistic information becomes important. For sentences from different topics or different passages, words exhibit reduced overlap. The word embeddings alone may provide enough semantic information for a distinguishable sentence representation. Moreover, on the third subtask, the performance fluctuations caused by the subject gap are more obvious than for the first subtask when comparing the standard deviation of matching accuracy across the subjects.

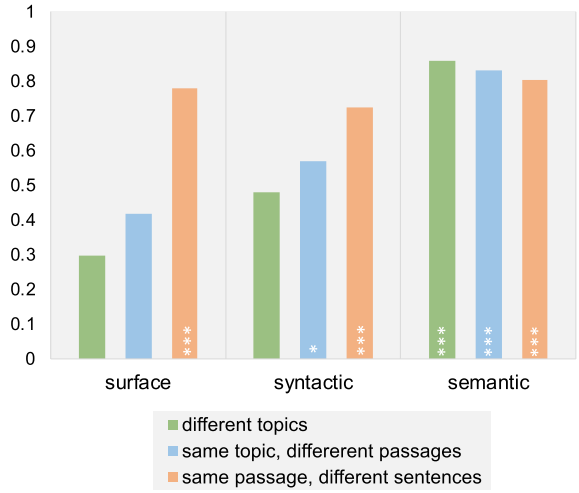

Fig. 7. Correlation of probing task performance and decoding accuracy on three subtasks, reported in the Spearman correlation. $*$ at the bottom of bar denotes the significance of correlation. $*$ means $0.05 \leq p$-value < 0.1, $**$ means $0.01 \leq p$-value < 0.05, and $***$ means p-value < 0.01. No $*$ means p-value > 0.1.

We validate the abovementioned results by correlating the decoding performance with the DSMs' probing score. As shown in Fig. 7, the scores of surface and syntactic probing do not correlate with decoding accuracy on the first and second subtasks, under the significance level of $\alpha = 0.01$. On the third subtask, the surface and syntactic probing scores significantly correlate with the decoding performance ($p < 0.01$ for both syntactic probing and surface probing). The semantic probing score correlates with the decoding performance in all the three subtasks ($p < 0.01$ for each subtask), but the correlation value tends to decrease as the subtask becomes finer in granularity. The results support our hypothesis that as the decoding becomes more fine-grained, syntactic, and surface features encoded by the DSMs can be helpful.

2) Decoding on Different Topics: Linguistic stimuli in the experiments are organized in the hierarchy of topic–passage–sentence. We are interested to know how the DSMs decode brain representation of sentences from different topics. So for,

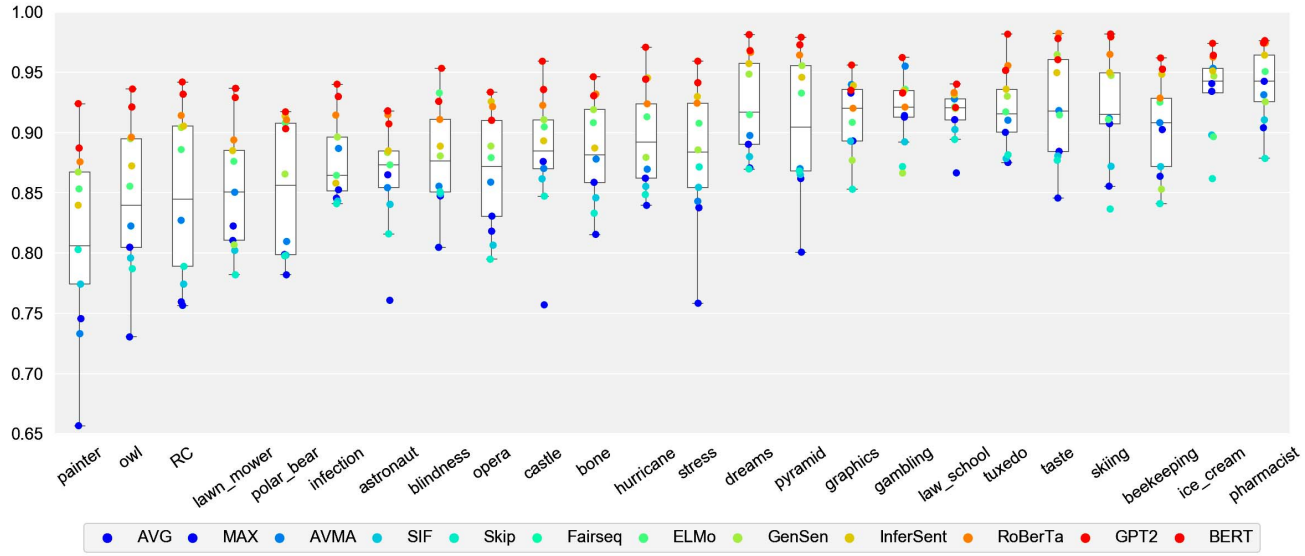

Fig. 8. Decoding accuracy on different topics. Scatters with different colors denote the matching accuracy of decoder trained with different DSMs. Every single illustrated accuracy is the average result among the five subjects. The distribution of the scores is further illustrated by the box plot.

in every tested DSM, we average its decoding score on all sentences from a specific topic. All five subjects are included in the calculation to qualify the results.

As depicted in Fig. 8, the DSMs show a tendency of topic selectivity; namely, a DSM does not perform uniformly when decoding different topics. For example, after pooling across different subjects, GPT2's highest mean matching accuracy is observed on the "dreams" topic while the lowest on the "painter" topic, with a gap of 0.094. The gap between the MAX's highest and lowest mean matching accuracy is even larger. MAX delivers its highest accuracy on the "ice-cream" and lowest on the "painter" topic, and the gap is 0.2839. If we compare the topicwise decoding accuracy on the five subjects but not the mean value among them, the difference between MAX's decoding accuracy on "ice-cream" and "painter" topic is significant ($p < 0.01$). In the 384-sentence data set, we find that brain responses to the crime-topic sentences are selectively more predictable than other topics. However, in the current 243-sentences data set, we do not find a topic that is consistently easier to decode than the other ones by the DSMs. We see from Fig. 9 that, on the "ice-cream" topic, DSMs perform more closely, and AVG, MAX, and AVMA deliver the highest matching accuracy on this topic. However, this does not mean that all the DSMs best decode the "ice-cream" topic. For example, RoBerTa's highest matching accuracy is observed on the "taste" but not the "ice-cream" topic.

3) Spatial Distribution of Informative Voxels: We decode with informative voxels that are selected by how well they predict the sentence representations. We map the full brain voxels to every single tested DSM in cross validation. The semantic vectors predicted from every voxel are correlated with the ground-truth sentence representations produced by a DSM. The average correlations on all 243 sentences are recorded as the informativeness score of a certain voxel. The 5000 most informative voxels are selected. Though selected by informativeness without any spatial constraints over the brain, voxels themselves belong to different brain areas with high-level cognitive functions. How sentences are neurally represented in the human brain remains an unsolved problem. We can gain some insights through studying the correspondence between sentence representations and functional brain networks. Following [20], we pick four brain networks: language networks [39], visual networks [36], default mode network (DMN) [38], and multiple demand (MD) network. We are particularly interested in the language networks that store the mappings between linguistic forms and meanings [39]. We show how the informative voxels distribute among these networks, as depicted in Fig. 9. We also estimate a null distribution by selecting voxels with random sentence embeddings. The null distribution is of all the voxels selected, $8.65\% \pm 0.50\%$ fall in the language networks, $6.58\% \pm 0.30\%$ fall in the visual networks, $8.88\% \pm 0.38\%$ fall in the DMN, and $15.32\% \pm 0.82\%$ fall in the MD (each with 95% confidence interval).

The informative voxels are not evenly distributed among the function networks. Language network voxels in the informative voxels selected by each DSM take percentage significantly higher than the null level (all significant results have $p < 0.01$, and p-values are detailed in Table VI(a) in the Supplementary Material). Visual network voxels significantly take percentage higher than the null level in voxels selected by DSMs except for AVG, ELMo, RoberTa, and GPT2 (all significant results have $p < 0.01$, and p-values are detailed in Table VI(b) in the Supplementary Material). DMN voxels do not consistently take percentage higher than the null level in the informative voxels selected by DSMs (p-values are detailed in Table VI(c) in the Supplementary Material). The distribution pattern of selected informative voxels is consistent among DSMs though the DSMs themselves can be largely different from each other in the underlying model structure. The rank of voxel percentage in the selected voxels is like language network $>$ visual network $>$ DMN $>$ MD.

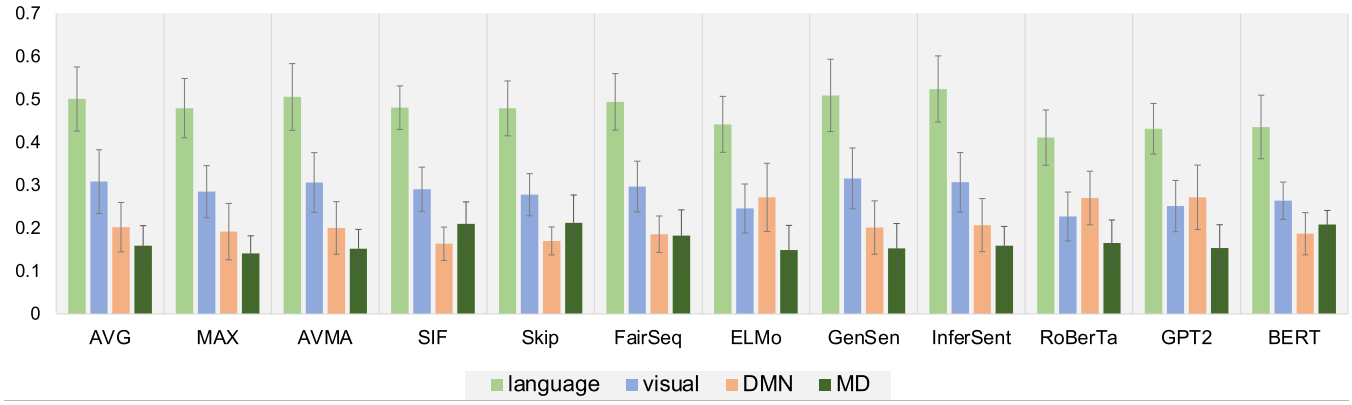

Fig. 9. Distribution of the 5000 most informative voxels in the language networks, visual networks, DMN, and MD selected by different DSMs. The reported result is averaged across different subjects (five in total), and the error bar denotes standard deviation.

4) Decoding Within Functional Networks: To demonstrate how well different DSMs decode a specific brain area, we further constrain the voxels to different brain networks and retrain the decoders. The decoding results on three subtasks are further averaged among subjects, and we show them in Fig. 10. Compared with decoding with selected informative voxels, constraining voxels in a specific network does not yield improvements in decoding accuracy. Such a trend becomes more clear in the third subtask (matching sentences from the same passage), where the decoding accuracy decreases by a maximum of 45% in maximum. Using voxels in the language atlas yields better decoding results than other the visual network and DMN. However, in the third subtask, we still observe an average accuracy decrease of 25%. BERT achieves 0.815 average accuracy with the selected voxels on the third subtask but only get 0.596 with the language network voxels. Though decreasing in scale, the performance rank of different DSMs is generally consistent with that of previous experiments in the language atlas. Unstructured models perform on par with most structured models on the first two subtasks. When decoding with the visual network and DMN voxels, we observe some patterns that differ from those of previous informative voxel-based decoding results. With visual voxels, all the unstructured models perform even below the chance level on the third matching task. With DMN voxels, InferSent tops in decoding accuracy on the first subtask, while BERT ranks the second.

IV. Discussion

A. Characterizing the Relationship Between Cortical Representations and DSMs

In this article, we characterize the relationship between cortical representations and distributed sentence embeddings through the lens of neural encoding and decoding. Though distinct in form and motivation, we obtain some common findings in encoding and decoding results.

We find that structured DSMs do not consistently outperform the unstructured models. For example, Skip-thought is not significantly different from AVMA in encoding the language and semantic networks. The two structured models tend to perform better in fine-grained decoding tasks, such as the third subtask of decoding, classifying sentences from the same passage. However, in the coarse-grained tasks, they perform closely to the unstructured ones. In the structured models, we further find that supervision is not a decisive factor for a DSM's encoding or decoding performance. The three unsupervised transformers significantly outperform GenSen, a supervised multitask model in encoding the language networks and in most subtasks of decoding. However, InferSent, another supervised model that we test, is comparable to RoBerTa in encoding the visual networks and the third subtask of decoding. Thus, it is worth studying the role of supervision in explaining a structured DSM's relationship with the brain.

The three transformer-based models surpass other unsupervised structured models and unstructured baselines in encoding the language and semantic networks. They also perform impressively in brain decoding with both selected voxels and voxels constrained in functional networks. These results lead us to highly recommend transformer-based models to be considered in sentence-level brain encoding and decoding. However, we should note that they do not fully resemble the neuron structure in the human brain. We observe that the differences in structure modeling and supervision do not seem to explain the DSM's relation with the brain. Therefore, it is desired to exhaustively study the underlying structure of DSMs (especially the transformer's attention-based architecture) and determine what leads to the high correlation between the distributed representations and the brain activation patterns aroused by the sentence stimuli. Given the excellent encoding and decoding performance, these models should create a solid baseline to be adjusted to bionically simulate and explain the cortical sentence processing. In this article, the transformer-based DSMs fine-tuned in downstream NLP tasks are not included in the encoding and decoding experiments except in the ablation test. It is promising to further explore how well these models trained in supervised tasks [23], [40] predict and decipher the human brain activities aroused by linguistic stimuli. Furthermore, the DSMs that we evaluate in this article are trained exclusively in linguistic modality. How the DSMs learned in a multimodal way [41], [42] probe the brain representations is also worth studying in the future work.

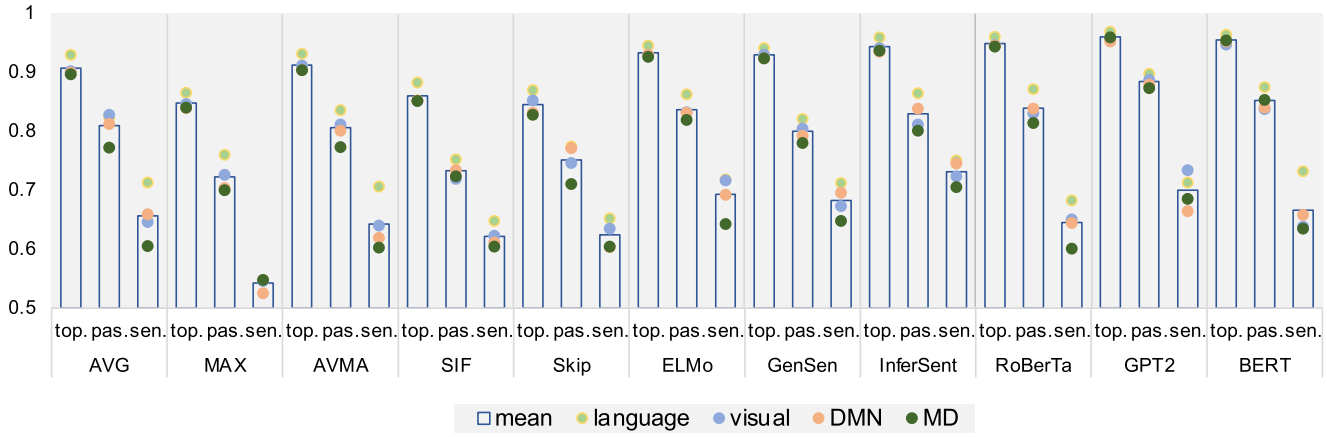

Fig. 10. Matching accuracy of sentence representations with voxels constrained in specific brain networks. The reported score is averaged among five subjects. The top., pas., and sen., respectively, mean the performance of matching sentences from: 1) different topics; 2) different passages within the same topic; and 3) the same passage.

B. Interpreting the Critical Linguistic Features for Neural Encoding and Decoding

The tested DSMs exhibit different performance patterns, and it is necessary to explain what difference in the feature captured contributes most to the encoding and decoding accuracy gap. Thus, we probe the DSMs' abilities to account for the surface, syntactic, and semantic features of a sentence. The probing scores are then correlated with the encoding and decoding accuracies to interpret by which feature they are most explained. This not only leads to a more clear empirical explanation of the experimental results but also it sheds light on the linguistic features critical in cortical language processing. If models better accounting for a certain linguistic feature also achieved higher performance in predicting and deciphering the brain activities among different subjects, such a feature should play a role in forming the cortical sentence representation.

Through the analysis, we find common patterns in the encoding and decoding results even though different data sets are used. DSM's performances on the semantic probing task best correlate with their encoding accuracy in the semantic and language networks and also highly correlate with the decoding accuracies in three subtasks. The semantic probing task evaluates how well the DSMs predict the semantic similarity and relatedness between sentences. Since the sentence-level DSMs exert different transformations on word vectors, the extent to which the semantic of original words can be conserved varies across the methods. For example, AVG treats every word equally, while the transformer-based models, such as BERT, explicitly weight the importance of every word to form a sentence representation. Information on important content words is more possible to take a higher proportion in the sentence representations produced by transformers. They do exceed other unsupervised models in encoding and most subtasks in decoding, indicating that its learned word importance weighting policy may share some patterns with the underlying sentence processing mechanism in the human brain.

Syntactic probing scores significantly correlate with the encoding accuracy of two ROIs of language networks. The ablation task on BERT also shows that tuning on word-order scrambled sentences only yields decreased matching accuracy in two ROIs of the language network. In decoding experiments, syntactic probing scores significantly correlate with the accuracy of the third subtask. A finding shared in encoding and decoding results is that the structured models do not consistently outperform the unstructured ones. Taking together, these results may indicate that modeling syntactic structure plays a role in forming cortical sentence representations but not the decisive role.

C. Suggesting the Semantic Selectivity of Different Cortical Areas

In both the encoding and decoding studies, the brain functional networks and ROIs have been probed by the distributed representations. Though the DSMs themselves might be built in very different ways, ROIs that can be predicted with high accuracy are virtually the same. The informative voxels selected in decoding also show consistent distribution patterns. We find, in the encoding results, that LMTG and LLOC are consistently better predicted among subjects and with different DSMs. Similarly, informative voxels in neural decoding also densely distribute in these areas. Without any *a priori* location constraints, nearly half of the voxels selected by the transformer-based models fall into the language networks. Based on the abovementioned results, we further discuss the semantic selectivity of different ROIs and DSMs. We find a topic is not exclusive to certain ROIs but represented by spatially overlapping and distributed cortical patterns. We also find that the topic selectivity pattern is not uniform across different DSMs. For the brain, one topic can be represented by multiple ROIs, and a single ROI can represent multiple topics. For DSMs, it is also difficult to find a topic that is exclusively decipherable by a certain model. Also, we note that the current pairwise matching-based encoding and decoding methodology emphasizes the distinction of a specific topic. High encoding and decoding accuracy of a specific topic in a way reflects that its evoked cortical activation patterns are more discriminant than other topics.

V. Conclusion

Humans beings have the unique capacity to communicate with language. Such ability is based on mental representations of meaning that can be mapped to linguistic items, but to which we have no direct access. Fortunately, the computational meaning representation of language is a well-developed field in the NLP community. Thus, we draw upon these representation models to probe and explain the human brain's language processing through the lens of neural encoding and decoding. In this article, encoders and decoders have been built and evaluated with a wide range of DSMs, including both classical unstructured models and state-of-the-art structured ones. Empirically, we show the cases where unstructured models can handle and where they fail to structured models in predicting and deciphering the brain activities. Based on the empirical results, we clarify what features contribute most to predicting and deciphering the cortical activities. We also confirm different ROIs and DSM's selectivity to topics. Our results not only corroborate and extend previous findings, highlight the value, and identify the potential of using DSMs to explain and decipher neural linguistic representations but also offer a deeper insight into the connection between the two manifestations of mental meanings: the neural activation patterns and the extrinsic linguistic representations. We hope that these could contribute to understanding cortical linguistic representations and developing machine-learning representation models in a mutual-promotion manner.

References

[1] M. Ghio, M. M. S. Vaghi, D. Perani, and M. Tettamanti, "Decoding the neural representation of fine-grained conceptual categories," *NeuroImage*, vol. 132, pp. 93–103, May 2016.

[2] M. A. Just, V. L. Cherkassky, S. Aryal, and T. M. Mitchell, "A neurosemantic theory of concrete noun representation based on the underlying brain codes," *PLoS ONE*, vol. 5, no. 1, p. e8622, Jan. 2010.

[3] R. A. Mason and M. A. Just, "Neural representations of physics concepts," *Psychol. Sci.*, vol. 27, no. 6, p. 904, 2013.

[4] T. M. Mitchell *et al.*, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, May 2008.

[5] S. M. Frankland and J. D. Greene, "An architecture for encoding sentence meaning in left mid-superior temporal cortex," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 37, pp. 11732–11737, Sep. 2015.

[6] S. G. Baron and D. Osherson, "Evidence for conceptual combination in the left anterior temporal lobe," *NeuroImage*, vol. 55, no. 4, pp. 1847–1852, Apr. 2011.

[7] A. J. Anderson *et al.*, "Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation," *Cerebral Cortex*, vol. 27, no. 9, pp. 4379–4395, Aug. 2016.

[8] J. Wang, V. L. Cherkassky, and 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 Mapping*, vol. 38, no. 10, pp. 4865–4881, Oct. 2017.

[9] E. Matsuo, I. Kobayashi, S. Nishimoto, S. Nishida, and H. Asoh, "Generating natural language descriptions for semantic representations of human brain activity," in *Proc. ACL Student Res. Workshop*, 2016, pp. 22–29.

[10] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, p. 453, 2016.

[11] F. Pereira, G. Detre, and M. Botvinick, "Generating text from functional brain images," *Frontiers Hum. Neurosci.*, vol. 5, p. 72, Aug. 2011.

[12] G. Handjaras *et al.*, "How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge," *NeuroImage*, vol. 135, pp. 232–242, Jul. 2016.

[13] J. Gauthier and R. Levy, "Linking artificial and human neural representations of language," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 529–539.

[14] S. Jat, H. Tang, P. Talukdar, and T. Mitchell, "Relating simple sentence representations in deep neural networks and the brain," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5137–5154.

[15] J. Sun, S. Wang, J. Zhang, and C. Zong, "Towards sentence-level brain decoding with distributed representations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7047–7054.

[16] S. Abnar, L. Beinborn, R. Choenni, and W. Zuidema, "Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains," 2019, *arXiv:1906.01539*. [Online]. Available: http://arxiv.org/abs/1906.01539

[17] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell, "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses," *PLoS ONE*, vol. 9, no. 11, Nov. 2014, Art. no. e112575.

[18] S. Nishida and S. Nishimoto, "Decoding naturalistic experiences from human brain activity via distributed representations of words," *NeuroImage*, vol. 180, pp. 232–242, Oct. 2018.

[19] D. Shen *et al.*, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 5964–5972.

[20] F. Pereira *et al.*, "Toward a universal decoder of linguistic meaning from brain activation," *Nature Commun.*, vol. 9, no. 1, p. 963, Dec. 2018.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2018, pp. 4171–4186.

[22] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 670–680.

[23] Z. Zhao, H. Lu, D. Cai, X. He, and Y. Zhuang, "Microblog sentiment classification via recurrent random walk network learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3532–3538.

[24] S. Wang and C. Zong, "Comparison study on critical components in composition model for phrase representation," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 3, p. 16, 2017.

[25] Z. Zhao *et al.*, "Long-form video question answering via dynamic hierarchical reinforced networks," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5939–5952, Dec. 2019.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[27] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[28] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 3294–3302.

[29] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pre-training approach," 2019, *arXiv:1907.11692*. [Online]. Available: http://arxiv.org/abs/1907.11692

[30] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[31] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal, "Learning general purpose distributed sentence representations via large scale multi-task learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[32] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2126–2136.

[33] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4129–4138.

[34] M. Marelli *et al.*, "A SICK cure for the evaluation of compositional distributional semantic models," in *Proc. LREC*, 2014, pp. 216–223.

[35] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity–multilingual and cross-lingual focused evaluation," 2017, *arXiv:1708.00055*. [Online]. Available: http://arxiv.org/abs/1708.00055

[36] E. Fedorenko, M. K. Behr, and N. Kanwisher, "Functional specificity for high-level linguistic processing in the human brain," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 39, pp. 16428–16433, Sep. 2011.

[37] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, "Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies," *Cerebral Cortex*, vol. 19, no. 12, pp. 2767–2796, Dec. 2009.
[38] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The brain's default network," *Ann. New York Acad. Sci.*, vol. 1124, no. 1, pp. 1–38, 2008.
[39] J. D. Power *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, pp. 665–678, Nov. 2011.
[40] Z. Zhao, H. Lu, V. W. Zheng, D. Cai, X. He, and Y. Zhuang, "Community-based question answering via asymmetric multi-faceted ranking network learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 3532–3539.
[41] D. Xu *et al.*, "Video question answering via gradually refined attention over appearance and motion," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1645–1653.
[42] S. Wang, J. Zhang, N. Lin, and C. Zong, "Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 5964–5972.

Jingyuan Sun is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, under the supervision of Prof. Chengqing Zong.

His current research interests include human and machine natural language understanding.

Shaonan Wang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2018.

She is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. Her current research interests include human and machine natural language understanding.

Jiajun Zhang (Member, IEEE) received the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include machine translation, natural language processing, and multilingual text analysis.

Chengqing Zong (Senior Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in March 1998.

He is currently a Professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. He has authored a book titled *Statistical Natural Language Processing* and coauthored a book titled *Text Data Mining*. His research interests include machine translation, dialog systems, and linguistic cognitive computing as well.

Dr. Zong is also a fellow of Chinese Association for Artificial Intelligence (CAAI), a member of the International Committee on Computational Linguistics (ICCL), and the President of the Asian Federation on Natural Language Processing (AFNLP). He has served many top-tier international conferences, such as Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP) 2021 as the Conference Chair, ACL-IJCNLP 2015 and International Conference on Computational Linguistics (COLING) 2020 as the PC Co-Chair, and Conference on Artificial Intelligence, the Association for the Advance of Artificial Intelligence (AAAI) 2019 and AAAI 2020 as the Area Chair. He also serves as an Associate Editor of *ACM Transactions on Asian and Low-Resource Language Information Processing* and a member of the Editorial Board of the *IEEE Intelligent Systems*.