ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230524

Tuning in to Neural Encoding: Linking Human Brain and Artificial Supervised Representations of Language

Jingyuan Sun[a;*], Xiaohan Zhang[b] and Marie-Francine Moens[a]

[a]Department of Computer Science, KU Leuven, Belgium
[b]Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, China

Abstract. To understand the algorithm that supports the human brain's language representation, previous research has attempted to predict neural responses to linguistic stimuli using embeddings generated by artificial neural networks (ANNs), a process known as neural encoding. However, most of these studies have focused on probing neural representations of Germanic languages, such as English, with unsupervised ANNs. In this paper, we propose to bridge the gap between human brain and supervised ANN representations of the Chinese language. Specifically, we investigate how task tuning influences a pretained Transformer for neural encoding and which tasks lead to the best encoding performances. We generate supervised representations on eight Natural Language Understanding (NLU) tasks using prompt-tuning, a technique that is seldom explored in neural encoding for language. We demonstrate that prompt-tuning yields representations that better predict neural responses to Chinese stimuli than traditional fine-tuning on four tasks. Furthermore, we discover that tasks that require a fine-grained processing of concepts and entities lead to representations that are most predictive of brain activation patterns. Additionally, we reveal that the proportion of tuned parameters highly influences the neural encoding performance of fine-tuned models. Overall, our experimental findings could help us better understand the relationship between supervised artificial and brain language representations.

1 Introduction

Neural encoding for language has long been a critical topic of interest in the intersection of natural language processing and cognitive neuroscience [1, 13]. The recent development of deep learning and pre-trained language representations [30, 32] has provided new opportunities for enhancing the precision and efficiency of neural encoding [31, 36]. Employing pre-trained language representations in neural encoding enables a more profound capture of the intricate relationships between linguistic stimuli and neural responses [3]. This, in turn, may foster the creation of superior language processing models [26]. Additionally, probing the relationship between artificial and neural representations of language can yield valuable insights into the fundamental mechanisms underlying language processing [24].

Despite the extensive research on unsupervised embeddings for English language in neural encoding [25, 33], there is a lack of studies exploring the use of supervised embeddings for neural encoding in other languages, such as Chinese [35]. Moreover, even the few studies that do adopt supervised embeddings for neural encoding in English still rely heavily on fine-tuning pre-trained models for task supervision [21]. However, fine-tuning has been shown to distort the knowledge learned from pre-training [16], which is inconsistent with the human brain's mechanism that does not require a major reformation of the brain's language network to learn a single new task. To address these gaps, this paper proposes the use of both fine-tuned and prompt-tuned supervised sentence embeddings to fit a neural encoding model for Chinese. Prompt-tuning, which protects pre-trained knowledge by freezing weights and learning additional embeddings to fit a task [18], has yet to be widely explored for neural encoding, and this paper is to address this gap.

In pursuit of this goal, we employ both partial and full fine-tuning as well as prompt-tuning to adapt the pre-trained language model to 8 different NLU tasks, individually. The aim is to discern the influence of task tuning on a Transformer model for neural encoding and identify which tasks result in the best encoding performance. We find that:

(i) Prompt-tuning on 5 of the 8 tasks yields supervised representations that significantly exceed fully fine-tuned peers in predicting brain activities in the language network[1]. However, on none of the 8 tasks do fine-tuned embeddings significantly outperform the prompt-tuned ones in neural encoding
(ii) Tuning on tasks that require a compositional understanding of entities and concepts yields supervised representations that are better at neural encoding than tuning on other tasks.
(iii) The proportion of tuned parameters highly influences the neural encoding performance of fine-tuned models.

In summary, this paper makes a triple-folded contribution. First, we propose a novel neural encoding framework with prompt-tuned supervised representations. We prove it to be a viable alternative to fine-tuning-based methods. Second, we demonstrate how different tuning methods influence a pre-trained Transformer in neural encoding with comprehensive experiments. Third, our findings indicate that the balance between protecting pre-trained knowledge and learning task-related features is important for optimal neural encoding performance. Overall, this work could help us better understand the relationship between task-tuned artificial and brain language representations.

[*] Corresponding Author. Email: jingyuan.sun@kuleuven.be

[1] A brain functional network is a collection of brain regions that consistently show coordinated activity for certain cognitive functions or during behavioral tasks.

2 Related Work

2.1 Neural Encoding with ANN representations

There has recently been a surge of discussions regarding the feasibility of ANNs as computational candidates for brain language representations due to their promising experimental results in NLU tasks [4, 26, 29]. In essence, previous studies have employed ANN representations to effectively predict brain responses to linguistic stimuli. Early research has proved their capability to predict brain activation patterns in response to simple verbal stimuli, such as individual words and phrases [11, 14]. Subsequent studies have further displayed their ability to predict and decipher the representations of more complex stimuli [1, 31, 27].

Previous studies have primarily focused on the use of unsupervised ANN representations optimized for language modeling or similar context-predicting goals, where supervised models have been paid relatively limited attention [28]. A related study by Oota et al., [21] utilized fine-tuning of a pre-trained BERT model on downstream tasks to compare neural encoding performance. However, full fine-tuning generally updates the entire set of pre-trained parameters and is distinct from the way the brain adapts to new tasks [16]. It raises the question as to whether fine-tuning is the most appropriate framework for obtaining supervised representations for probing the human brain. In light of this, we undertake a comprehensive comparison between fine-tuned and prompt-tuned representations for neural encoding.

Moreover, previous studies mainly targeted Germanic languages especially English, while Tibetan languages such as Chinese are under exploration. Partially limited by the scarcity of large-scale public brain imaging datasets for the Chinese language. Wang et al., ([35]) open-sourced a multi-modal neuroimaging dataset scanned from 12 subjects listening to 5 hours of naturalistic Chinese stories. We will base our experiments on this dataset. Wang et al., ([35]) also conducted basic neural encoding experiments mainly to evaluate the quality of the collected dataset but they did not employ any supervised representations.

2.2 Tuning Pre-trained Models

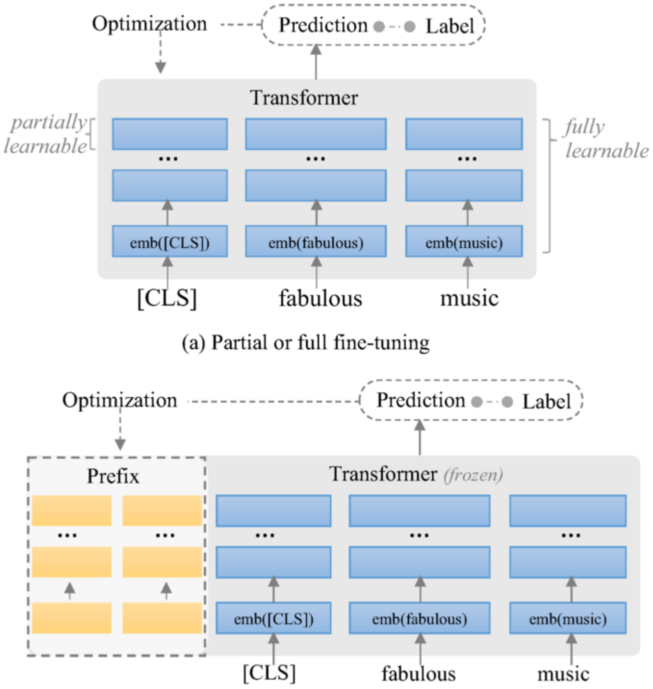
Fine-tuning. The recent decade has witnessed a significant advancement in the development of pre-trained language models, yielding remarkable improvements in performance across a wide spectrum of NLP tasks [23]. Fine-tuning is one of the most commonly employed techniques for task supervision which involves updating the pre-trained parameters to fit a specific task. Although fine-tuning typically results in favorable task performance, the computational time and space complexity is high due to the need of calculating gradients and saving optimizer states for all parameters [10]. Additionally, full fine-tuning has been proved to distort the general domain knowledge learned during pre-training, potentially undermining its cognitive plausibility [15].

Discrete prompting. Discrete prompting, as an alternative approach for fine-tuning, undergoes rapid development especially in the NLP field [19]. This technique leverages a pre-trained model's parameters to solve NLP tasks by querying the model with a discrete natural language prompt [2]. For example, to conduct emotion categorization, a sample such as "I love this lovely kitten" could be prepended with a prompt like "Categorize the emotion". The pre-trained language model would be asked to predict the subsequent tokens. Despite its advantage in not requiring training and only storing a single set of model parameters, discrete prompting's task performance may be highly impacted by the design of the prompt template and not be as optimal as fine-tuning in certain scenarios [17]. Therefore we do not choose discrete prompting as a baseline approach in this study.

Prompt-tuning. Prompt-tuning builds upon discrete prompting. It fixes the pre-trained parameters of a model and trains additional prompt embeddings to guide the model to learn a downstream task [22]. Prompt-tuning has been demonstrated in previous research to exceed discrete prompting on various tasks (as reported in [5]) and deliver task performance comparable to fine-tuning. Prefix-tuning [18] is one typical implementation of prompt-tuning. It prepends trainable embeddings called "prefixes" to the original input word embeddings and optimizes them. In this study, we employ the P-tuning V2 method [20], an optimized version of prefix-tuning that compares or even surpasses fine-tuning on GLUE [34].

3 Methods

In the following subsections, we will first introduce how fine-tuning and prompt-tuning work. We then demonstrate how the tuned representations are used to build a neural encoder.



(a) Partial or full fine-tuning

(b) Prompt-tuning

Figure 1: Tuning a Transformer on a downstream task, taking sentiment classification as an example: [a] with partial or full fine-tuning; [b] with prompt-tuning.

3.1 Tuning pre-trained Models

This subsection introduces fine-tuning and prompt-tuning, using the example of tuning a Transformer-based language model for a conditional generation task. The language model, denoted by $LM = p_\phi(y|x)$, takes an input context x and generates an output sequence of tokens y. The model is parameterized by ϕ, and $z = [x; y]$ represents the concatenation of x and y, with id_x and id_y denoting the corresponding sequences of indices. At each time step i, the activations of all hidden layers are represented as $h_i = [h_i^1, h_i^2, ...h_i^n]$, where h_i^j denotes the activation of the j-th Transformer layer at time step i, and h_i is the concatenation of all activation layers. The activations h_i are computed as a function of z_i and the past activations in their context, given by $h_i = \mathbf{LM}_\phi(z_i, h)$.

To compute the distribution for the next token, the model uses the hidden states $h_i^{(n)}$ from the last layer and applies a softmax function to the pre-trained matrix W_ϕ to generate the logits for the vocabulary. Specifically, the distribution for the next token is computed as $p_\phi\left(z_{i+1} \mid h_{\leq i}\right) = \operatorname{softmax}\left(W_\phi h_i^{(n)}\right)$, where $h_i^{(n)}$ denotes the last layer hidden states of h_i. This process enables the model to generate coherent and contextually-relevant sequences of tokens.

3.1.1 Fine-tuning

As shown in Figure 1, fine-tuning is based on a pre-trained set of parameters ϕ. During fine-tuning, a language model distribution p_ϕ is trained by optimizing the following log-likelihood objective:

$$\max_{\phi} \log p_\phi(y \mid x) = \sum_{i \in id_y} \log p_\phi\left(z_i \mid h_{<i}\right) \qquad (1)$$

The fine-tuning methods can be further divided into partial and full fine-tuning. Partial fine-tuning freezes part of pre-trained parameters ϕ and optimizes the remaining parameters such as these in the layers near the output layer. Full fine-tuning optimizes the full set of pre-trained parameters.

3.1.2 Prompt-tuning

We build upon the P-tuning method [20] to prompt-tune a pre-trained model on downstream tasks. In Figure 1, we illustrate the mechanism of P-tuning with a Transformer-based language model. P-tuning prepends a continuous trainable prefix to different layers of an ANN to guide its fitting on downstream tasks.

We now explain the mechanism of P-tuning following [18]. As shown in Figure 1, during P-tuning, a prefix, denoted as Pr, is prepended to the input sequence, resulting in the concatenation $z = [Pr; x; y]$. Pr is a continuous trainable embedding, its corresponding sequence of indices is denoted as Pr_{id_x} with length $|Pr_{id_x}|$. A trainable matrix M_θ^{Pr}, with dimensions $|Pr_{id_x}| \times dim(h_i)$, is initialized to store the prefix parameters. And

$$h_i = \begin{cases} M_\theta^{Pr}[i,:], & \text{if } i \in Pr_{\mathrm{id_x}} \\ \mathbf{LM}_\phi\left(z_i, h_{<i}\right), & \text{otherwise.} \end{cases} \qquad (2)$$

Training P-tuning is then to optimize the log-likelihood objective as in (1) with ϕ being fixed and θ being trainable.

3.2 Neural Encoder

To learn the mapping between computational models and brain activation, we train voxel-wise neural encoding models to predict fMRI signals from sentence stimuli for each of the 12 participants. The fMRI methodology measures the fluctuations in blood-oxygen-level-dependent (BOLD) signals, which undergo slow alterations post neuronal firing. To accommodate for the temporal delay between sentence stimuli and corresponding neural activity, the initial step involves the convolution of sentence embeddings with the canonical hemodynamic response function (HRF). The HRF delineates the transformation in the BOLD signal following neuronal firing. Then, the convolved features were used to predict fMRI signals with Ridge regression.

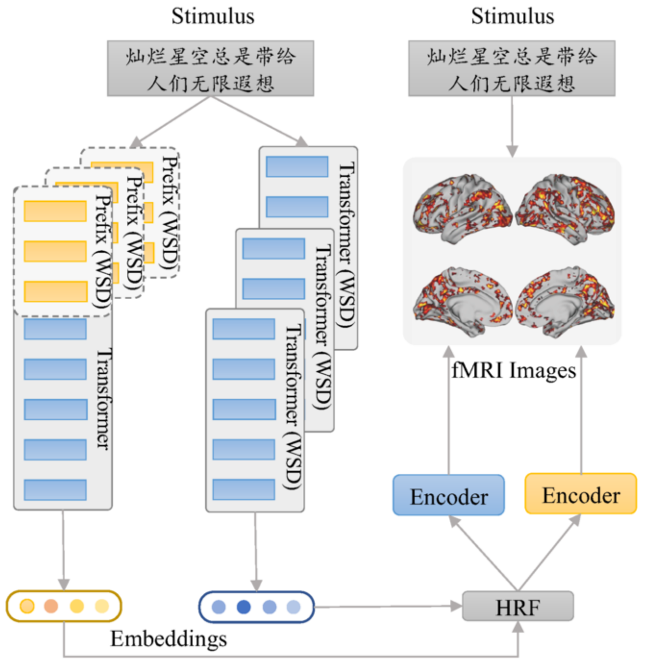


Figure 2: Building neural encoders with fine-tuned and prompt-tuned supervised language representations. The example stimulus is taken from the fMRI dataset that we use in this work and its English translation is "The brilliant starry sky always gives people endless reverie".

Specifically, within the training set, we have a voxel matrix $X_e \in \mathbb{R}^{N_E \times N_V}$ and a convolved sentence embedding matrix $Z_e \in \mathbb{R}^{N_E \times N_D}$. N_E signifies the number of examples, N_V represents the number of voxels, and N_D stands for the number of dimensions for sentence representation. The goal is to estimate the regression coefficients of the encoder W_e by minimizing the following objective function for each column x_i in X_e, which corresponds to the BOLD signal of each voxel:

$$||W_e Z_e - x_i||_2^2 + \lambda ||W_e||_1 \qquad (3)$$

Here, λ is the regularization parameter.

During the training process, the encoder is optimized using a 5-fold cross-validation procedure. This procedure is applied to data from each of all 12 subjects, with each kind of supervised sentence embedding as a separate input. To assess the performance of the encoder, we calculate the Pearson correlation between the predicted and measured fMRI signals.

4 Experimental Setup

In this section, we will first briefly introduce the brain imaging dataset we use. We then demonstrate the principle of task selection and some characteristics of the selected tasks. Code implementations are available online [2] .

4.1 Brain Imaging Data

The Chinese fMRI dataset used in our experiment is from [35]. It is one of largest existing fMRI datasets for naturalistic Chinese language narratives. In this study, 12 native Chinese speakers underwent scanning with a Siemens Prisma 3T scanner equipped with a

[2] https://github.com/soinx0629/sup_fmri_enc_chn/

Task	*Abbr.*	*Corpus*	*Description*
Short Text Classification	STC	TNEWS	This task is to categorize news titles from TouTiao news to one of 15 news categories (finance, technology, sports, etc.)
Long Text Classification	LTC	Iflytek	This task is to assigns app description into one of 119 topics (food, car rental, education, etc)
Coreference Resolution	CR	CLUEWSC	This task is to decide whether a pronoun and a noun (phrase) in a sentence co-refer. Sentences are hand-picked from 36 contemporary literary works in Chinese.
Natural Language Inference	NLI	OCNLI	This task determine whether the given “hypothesis” logically follows from the “premise”
Semantic Similarity Prediction	SSP	AFQMC	This task is to predict whether two sentences are semantically similar
Keyword Recognition	KR	CSL	Given an abstract and some keywords, the task is to tell whether the keywords are all original keywords of a paper.
Named Entity Recognition	NER	WikiANN	Locate and classify named entities mentioned in the text coming from Wikipedia articles
Question Answering	QA	CMRC	This task is to generate answers given questions and context. CMRC is a span-extraction QA task where the model is required to identify a span of text in the provided context that answers the question.

Table 1: 8 tasks for tuning and their abbreviations, corpus and descriptions.

64-channel receiver coil while they were engrossed in listening to a collection of 60 Chinese stories. These stories covered a broad range of topics, with each story extending from 4 to 7 minutes, culminating in approximately 5 hours of audio content. The text and audio of all stories were downloaded from Renmin Daily Review website[3], after which the text was manually checked to ensure its concordance with the audio content. There were 52,269 words in all stories, forming a vocabulary of 9,153 words. Subsequent to its collection, the fMRI data were preprocessed following the Human Connectome Project (HCP) pipeline [9]. The fMRI dataset is publicly available[4].

4.2 Tasks for Tuning

We choose tasks for this study based on two major principles following previous work [21]. First, tasks involving a diverse range of cognitive-linguistic skills are desired. Second, tasks that are commonly used in popular natural language processing (NLP) benchmarks, such as CLUE [37] are desired. Based on these principles, we choose the following 8 tasks for evaluation, including Long Text Classification (*LTC*), including Short Text Classification (*STC*), Natural Language Inferencing (*NLI*), Co-reference Resolution (*CR*), Keyword Recognition (*KR*), Semantic Similarity Prediction (*SSP*), Named Entity Recognition (*NER*) and Question Answering (*QA*). We present the details of these datasets shown in Table 1, including the task description and the corpus used.

4.3 Sentence Representations

In this study, we adopt the pre-trained RoBertTa-wwm-ext [6] model[5]. It is an improved version of naive RoBerTa model specially designed for Chinese. It achieved impressive performance on the CLUE benchmark as the base model. As shown in Figure 2, we adapt the RoBertTa with fine-tuning and prompt-tuning respectively on each of the 8 Chinese NLU tasks. Supervised representations for neural encoding are then generated by these tuned models. Specifically, we feed each of sentence stimuli presented for scanning the fMRI images to the tuned models and average the hidden states of the last layer to represent the sentence. The generated sentence embeddings have a dimension of 768. Both fine-tuned and prompt-tuned models are trained by ourselves on Nvidia 3080TI GPUs with 12GB VRAM. We conduct a hyper-parameter search to achieve optimal task performance.

5 Experimental Results

5.1 Prompt-tuned against Fine-tuned Representations

In this subsection, we compare the prediction performance with prompt-tuned and fully fine-tuned representations. We build neural encoders with task-tuned representations to predict brain activities within the language network, visual network, dorsal attention network, and default mode network. We conduct a group-level paired t-test with FDR correction to check the significance of comparison results. The results are shown in Figure 3.

As demonstrated in Figure 3a and 3b, prompt-tuning on 5 of the 8 tasks yields representations that exceed fine-tuned peers in predicting brain language network. The 4 tasks are LTC, STC, NLI, and CR. The most significant advantages of prompt-tuning representations over fine-tuned ones are achieved by tuning on the two classification tasks LTC and STC, where prompt-tuned embeddings significantly outperform in predicting the left superior temporal gyrus (LSTG), the dorsal area of left inferior frontal gyrus (LIFG) and middle frontal gyrus (LMFG). LSTG, LMFG, and LIFG contain acknowledged important components of the brain language network [7]. LMFG exhibits support for a dedicated "visual word form" system with orthographic specificity, which undergoes specialization to effectively represent orthographic knowledge [12]. LIFG is observed to facilitate verbal selection and possibly plays a domain-general role in the selection process for competing representations [8]. Tuning on the NLI task also yields prompt-tuned representations that exceed fine-tuned

[3] https://www.ximalaya.com/toutiao/30917322/
[4] https://openneuro.org/datasets/ds004078
[5] https://huggingface.co/hfl/chinese-roberta-wwm-ext

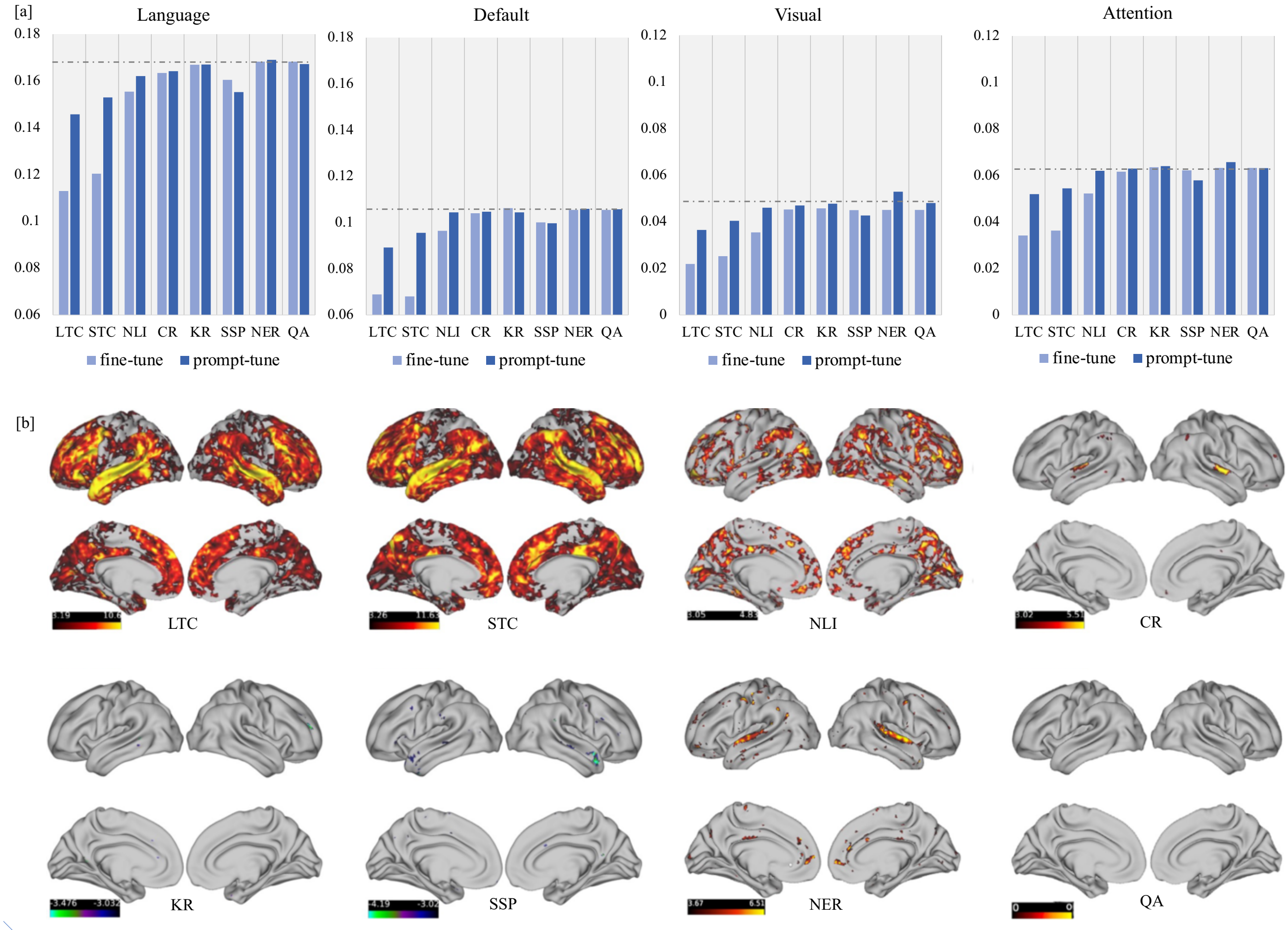


Figure 3: [a] Neural encoding performance for the brain language network, DMN, visual network and the dorsal attention network. Dashed lines denotes the performance of the un-tuned RoBerTa model. The Y-axis denotes the performance in Pearson's correlation. [b] Significance results for comparing the encoding performance of prompt-tuned against fine-tuned representations. Pairwise t-test with FDR correction is conducted and significance values are projected onto the inflated (top, lateral and medial views) cortical surface.

peers in predicting the left and inferior parietal lobule, especially the rostrodorsal and caudal area. Prompt-tuned CR representations surpass mainly in predicting the Superior Temporal Gyrus (STG). STG is an essential component of the brain's language network [38]. It is particularly important for extracting meaningful linguistic features from speech input. Except for these 4 tasks above, prompt-tuning on the NER leads to higher predicting correlation on some parts of the language and auditory network. These comparison result stands up to the significance test as depicted in Figure 3b.

Prompt-tuning on KR and QA tasks does not generate representations that significantly outperform fine-tuning. Specifically, the neural encoding performance of prompt-tuned and fine-tuned task representations do not differ significantly in predicting more than 95% of the cortical voxels. Only tuning 1 of 8 tasks, the SSP task, produce fine-tuned representations outperforming prompt-tuned ones on predicting some micro potions of the brain language network, such as small subregions within the left and right STG. The pattern of predicted performance is generally consistent across 4 functional networks. For example, not only within the language network but within the other 3 functional networks we also observe that prompt-tuned LTC, STC, and NLI representations significantly outperform fine-tuned ones in encoding correlation. Prompt-tuning on the other 4 tasks does not lead to significantly different performance from fine-tuning when predicting the major ROIs of the other 3 functional networks.

5.2 Tuned against Untuned Representations

In the previous subsection, we compare how prompt-tuned and fine-tuned representations are in neural decoding. It is a logical next step to ask how the tuned and untuned naive representations differ in predicting neural responses. We have depicted the neural encoding performance of the un-tuned model in 4 functional networks in Figure 3[a]. In this subsection, we go deeper to compare the prompt-tuned and fine-tuned representations respectively with the untuned model in predicting all cortical voxels. The significance of comparison results is checked by group t-test with FDR correction and depicted in Figure 4.

We find that only tuning on 2 of the 8 tasks significantly improves over the untuned model. As shown in Figure 4[a], tuning on NER and KR yields neural encoding performance that significantly exceeds untuned models on predicting ROIs of the language and auditory network. Prompt-tuning on the NER task surpasses the untuned model in encoding the left and right superior temporal gyrus (STG) especially the A4 and A5 area. It also surpasses in area 55b which

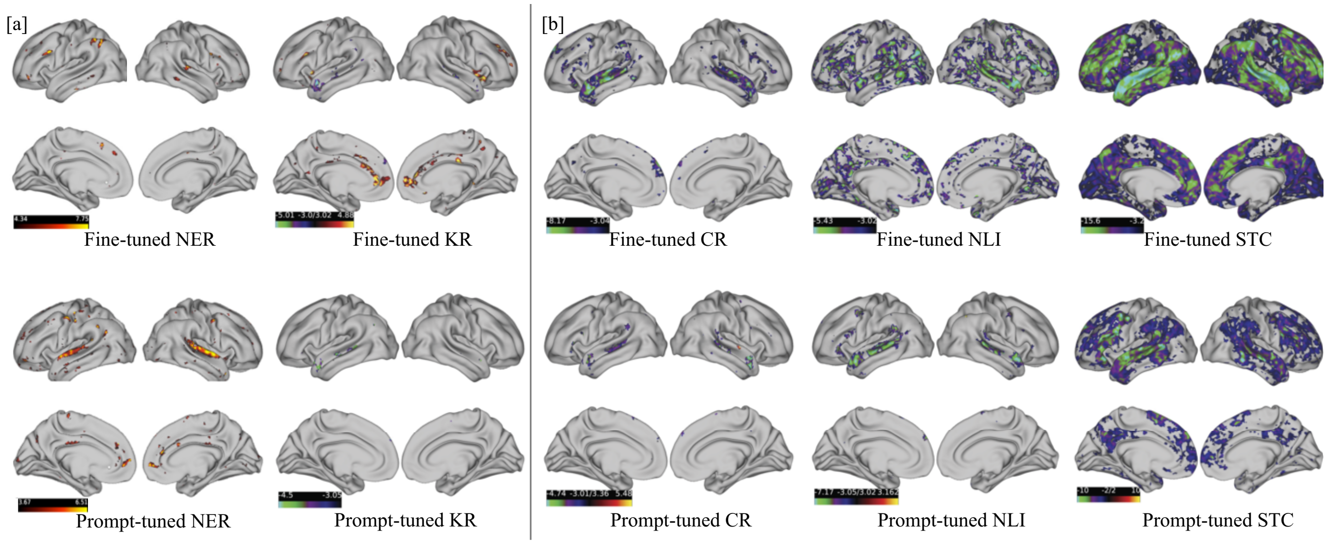


Figure 4: Significance results for comparing the encoding performance of prompt-tuned and fine-tuned representations against the untuned RoBerTa model. A pairwise t-test with FDR correction is conducted and significance values are projected onto the inflated cortical surface. Prompt-tuning on the two tasks in [a] significantly improves over the untuned model in some ROIs. Tuning on the three tasks in [b] underperforms the untuned model in encoding the major ROIs in brain language networks.

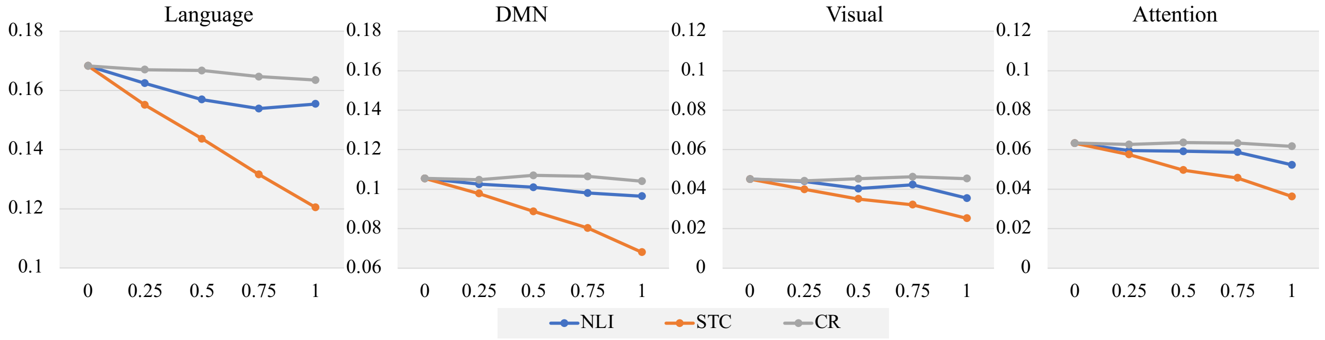


Figure 5: Influences of tuned parameter proportion on a fine-tuned model's neural encoding performance. The X-axis denotes the tuned proportion and the Y-axis denotes the encoding performance in Pearson's correlation.

is a part of dorsal auditory stream. Fine-tuning on the KR task improves over the untuned model in predicting the subgenual area of the cingulate gyrus (CG) and medial area of the right superior frontal gyrus (RSFG). Both KR and NER tasks require fine-grained compositional processing of entity and concept. So these empirical results might implicate that fine-grained representation of entities and concepts support the formation of brain language representation of sentences.

Fine-tuning and prompt-tuning on the NLI, STC, CR, SSP, and LTC task yield significantly worse neural encoding performance. Prompt-tuning exerts less influence on encoding performance than fine-tuning. We select 3 of the 5 tasks to display in Figure 4[b]. We find that tuning on these tasks decreases the neural encoding performance in major ROIs of the brain language network. Both fine-tuning and prompt-tuning on CR, NLI, and STC tasks significantly decrease neural encoding performances on STG. Tuning on STC further impairs the encoding performance on large potions of the middle frontal gyrus and Inferior parietal lobule, especially the rostrodorsal areas. Not only in the language network, NLI, and STC task tuning also significantly impairs encoding correlation in DMN and visual networks. The findings indicate that pre-trained knowledge distorted by fine-tuning contributes more to neural encoding than the task-related features learned for these 5 tasks.

5.3 *Influences of Tuned Parameter Proportion*

In previous subsections, we find that prompt-tuned representations of 5 tasks significantly outperform fine-tuned ones in encoding the brain language networks. We also demonstrate that even in the tasks where both fine-tuning and prompt-tuning decrease the encoding correlation compared to the untuned model, the influence of prompt-tuning is significantly lower than full fine-tuning. We analyze that one major difference between prompt-tuning and full fine-tuning is the proportion of tuned parameters. Prompt-tuning freezes all the pre-trained weights and learns additional embeddings to fit on a task, while full fine-tuning re-optimizes all the pre-trained weights.

In this subsection, we investigate whether the proportion of tuned parameters influences the encoding performance of fine-tuned representations. We partially fine-tune the pre-trained RoBerTa on the tasks found in the last subsection that lead to significantly worse encoding performance than the un-tuned model. We set the tuned parameter proportion to be 0.25, 0.5, and 0.75 respectively, and train neural encoders with these partially fine-tuned task representations. Altogether with the un-tuned and fully fine-tuned model, we depict the neural encoding performance in Figure 5.

As demonstrated in Figure 5, the tuned parameter proportion highly influences the encoding performances of fine-tuned representations, especially in the brain language network. Taking the classification task STC as example, we find a significantly positive correlation ($p < 0.01$) between the encoding performance with the tuned parameter proportion. With the increase of the tuned proportion, the encoding correlations of STC rapidly drop. A tuned parameter proportion of 1, which is fully fine-tuning, leads to the worst neural encoding performance. Not only in the language network but in the DMN and attention network, we also observe significant positive correlations between the tuned parameter proportion and encoding performance (all significant results have ($p < 0.01$)).

In summary, the findings of this subsection inform that the proportion of tuned parameters highly influences the encoding performance of fine-tuned representations. Especially for tasks on which tuning decreased the neural encoding performances, the more parameters tuned, the worse neural encoding performances go.

6 Discussion

In the previous section, we compare the encoding performance of prompt-tuned against fine-tuned representations, as well as the encoding performance of tuned representations against the untuned naive model. We also demonstrate the most predictive task representation on each cortical voxel. In this section, we provide further analysis and possible explanations for the experimental results.

First, we find that tuning on 5 of the 8 tasks yields prompt-tuned representations that exceed fine-tuned peers in predicting voxels within the brain language network. One major difference between fine-tuning and prompt-tuning is their treatment of the pre-trained weights. To learn a new task, fine-tuning generally updates the entire parametric space of a pre-trained model while prompt-tuning freezes the pre-trained weights and optimizes additional task-specific embeddings. Moreover, fine-tuning has been proven to distort the pre-trained features. So intuitively prompt-tuning can better protect the general domain knowledge acquired from pre-training than fine-tuning methods. Then the different encoding performances of fine-tuning and prompt-tuning can be at least partially explained by the importance of general domain knowledge in predicting the brain activations. If the general domain knowledge is more important than the task-related features for encoding the brain activities, then we can expect tuning methods that better protect pre-trained knowledge yields higher encoding correlation.

Second, we find that KR and NER representations more accurately predict brain activations in most cortical regions than other task-supervised representations, particularly in the regions of interest (ROIs) that compose the brain's language network. The KR task is to gauge a model's ability to determine whether the abstract can be summarized by the keywords. The NER task tests a model's ability to identify and classify named entities in text into predefined categories. The NLI and CR task representations rank second and third in prediction correlation. All these tasks require a compositional understanding of entities, concepts, and their semantic relations. This implies the possibility that capturing fine-grained concept and entity features supports the human brain in building representations for linguistic units.

Third, we investigate the impact of the proportion of tuned parameters on the encoding performance of fine-tuned representations. We suggest that the proportion of tuned parameters significantly affects the encoding performance of fine-tuned representations, particularly for tasks where tuning decreases the neural encoding performance. The more parameters tuned, the worse the neural encoding performances becomes. Notably, fully fine-tuning the model leads to the worst neural encoding performance. A larger tuned proportion means more pre-trained knowledge might be distorted by fine-tuning. So these findings indicate the importance of knowledge learned in Transformer pretraining for neural encoding. It is necessary for the supervised models to strike a balance between the protection of pre-trained knowledge and learning task-related features to achieve optimal neural encoding performances.

7 Conclusion

In this paper, we aim to link the representation of the Chinese language in supervised artificial neural networks (ANNs) and the human brain. Our investigation focuses on task tuning and its influence on the neural encoding of a pre-trained Transformer. To achieve this, we utilize prompt-tuning, a technique that is rarely explored in neural encoding for language, to generate supervised representations on 8 Natural Language Understanding (NLU) tasks. Our findings reveal that prompt-tuning leads to better representations for predicting neural responses to Chinese stimuli compared to traditional fine-tuning on 5 tasks. Interestingly, we observe that representations of tasks that requires fine-grained processing of concepts and entities are most predictive of brain activation patterns. We also discover that the proportion of tuned parameters significantly influences the neural encoding performance of fine-tuned models. In summary, our experimental findings provide insights into the relationship between supervised ANNs and brain language representations and could pave the way for future investigations into language representations in the brain.

Acknowledgements

This work is funded by the CALCULUS project (European Research Council Advanced Grant H2020-ERC-2017-ADG 788506).

References

[1] Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada, 'Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation', *Cerebral Cortex*, **27**(9), 4379–4395, (2016).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., 'Language models are few-shot learners', *Advances in neural information processing systems*, **33**, 1877–1901, (2020).

[3] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King, 'Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects', *arXiv preprint arXiv:2110.06078*, (2021).

[4] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King, 'Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects', in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3635–3644, Punta Cana, Dominican Republic, (November 2021). Association for Computational Linguistics.

[5] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen, 'Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction', in *Proceedings of the ACM Web Conference 2022*, pp. 2778–2788, (2022).

[6] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu, 'Revisiting pre-trained models for Chinese natural language processing', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, Online, (November 2020). Association for Computational Linguistics.

[7] Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher, 'Functional specificity for high-level linguistic processing in the human brain', *Proceedings of the National Academy of Sciences*, **108**(39), 16428–16433, (2011).

[8] Evelina Fedorenko and Sharon L Thompson-Schill, 'Reworking the language network', *Trends in cognitive sciences*, **18**(3), 120–126, (2014).

[9] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al., 'The minimal preprocessing pipelines for the human connectome project', *Neuroimage*, **80**, 105–124, (2013).

[10] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang, 'PPT: Pre-trained prompt tuning for few-shot learning', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8410–8423, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[11] Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini, 'How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge', *Neuroimage*, **135**, 232–242, (2016).

[12] Elizabeth A Hirshorn, Yuanning Li, Michael J Ward, R Mark Richardson, Julie A Fiez, and Avniel Singh Ghuman, 'Decoding and disrupting left midfusiform gyrus activity during word reading', *Proceedings of the National Academy of Sciences*, **113**(29), 8162–8167, (2016).

[13] Christopher R Holdgraf, Jochem W Rieger, Cristiano Micheli, Stephanie Martin, Robert T Knight, and Frederic E Theunissen, 'Encoding and decoding models in cognitive electrophysiology', *Frontiers in systems neuroscience*, **11**, 61, (2017).

[14] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant, 'Natural speech reveals the semantic maps that tile human cerebral cortex', *Nature*, **532**(7600), 453, (2016).

[15] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang, 'Fine-tuning can distort pretrained features and underperform out-of-distribution', *ArXiv*, **abs/2202.10054**, (2022).

[16] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang, 'Fine-tuning can distort pretrained features and underperform out-of-distribution', in *International Conference on Learning Representations*, (2022).

[17] Brian Lester, Rami Al-Rfou, and Noah Constant, 'The power of scale for parameter-efficient prompt tuning', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, (November 2021). Association for Computational Linguistics.

[18] Xiang Lisa Li and Percy Liang, 'Prefix-tuning: Optimizing continuous prompts for generation', in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, (August 2021). Association for Computational Linguistics.

[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig, 'Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing', *ACM Comput. Surv.*, (sep 2022). Just Accepted.

[20] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang, 'P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[21] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi, 'Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity?', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3220–3237, Seattle, United States, (July 2022). Association for Computational Linguistics.

[22] Guanghui Qin and Jason Eisner, 'Learning how to ask: Querying LMs with mixtures of soft prompts', in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5203–5212, Online, (June 2021). Association for Computational Linguistics.

[23] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang, 'Pre-trained models for natural language processing: A survey', *Science China Technological Sciences*, **63**(10), 1872–1897, (2020).

[24] Gideon Rosenthal, František Váša, Alessandra Griffa, Patric Hagmann, Enrico Amico, Joaquín Goñi, Galia Avidan, and Olaf Sporns, 'Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes', *Nature communications*, **9**(1), 2178, (2018).

[25] Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu, 'Exploring semantic representation in brain activity using word embeddings', *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 669–679, (2016).

[26] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko, 'The neural architecture of language: Integrative modeling converges on predictive processing', *Proceedings of the National Academy of Sciences*, **118**(45), e2105646118, (2021).

[27] Jingyuan Sun, Mingxiao Li, Zijiao Chen, Yunhao Zhang, Shaonan Wang, and Marie-Francine Moens, 'Contrast, attend and diffuse to decode high-resolution images from brain activities', *arXiv*, (2023).

[28] Jingyuan Sun and Marie-Francine Moens, 'Fine-tuned vs. prompt-tuned supervised representations: Which better account for brain language representations?', in *Proceedings of IJCAI*, Macau, China, (2023).

[29] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, 'Towards sentence-level brain decoding with distributed representations', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7047–7054, (2019).

[30] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, 'Distill and replay for continual language learning', in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3569–3579, Barcelona, Spain (Online), (December 2020). International Committee on Computational Linguistics.

[31] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, 'Neural encoding and decoding with distributed sentence representations', *IEEE Transactions on Neural Networks and Learning Systems*, **32**(2), 589–603, (2020).

[32] Jingyuan Sun, Shaonan Wang, and Chengqing Zong, 'Memory, show the way: Memory based few shot word representation learning', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1435–1444, Brussels, Belgium, (October-November 2018). Association for Computational Linguistics.

[33] Kiran Vodrahalli, Po-Hsuan Chen, Yingyu Liang, Christopher Baldassano, Janice Chen, Esther Yong, Christopher Honey, Uri Hasson, Peter Ramadge, Kenneth A Norman, et al., 'Mapping between fmri responses to movies and their natural language annotations', *NeuroImage*, (2017).

[34] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman, 'GLUE: A multi-task benchmark and analysis platform for natural language understanding', in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, (November 2018). Association for Computational Linguistics.

[35] Shaonan Wang, Xiaohan Zhang, Jiajun Zhang, and Chengqing Zong, 'A synchronized multimodal neuroimaging dataset for studying brain language processing', *Scientific Data*, **9**(1), 590, (2022).

[36] Shaonan Wang, Yunhao Zhang, Xiaohan Zhang, Jingyuan Sun, Nan Lin, Jiajun Zhang, and Chengqing Zong, 'An fmri dataset for concept representation with semantic feature annotations', *Scientific Data*, **9**(1), 721, (2022).

[37] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan, 'CLUE: A Chinese language understanding evaluation benchmark', in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772, Barcelona, Spain (Online), (December 2020). International Committee on Computational Linguistics.

[38] Han Gyol Yi, Matthew K Leonard, and Edward F Chang, 'The encoding of speech sounds in the superior temporal gyrus', *Neuron*, **102**(6), 1096–1110, (2019).