Cross-Modal Cloze Task: A New Task to Brain-to-Word Decoding

Shuxian Zou [1,2], **Shaonan Wang** [1,2], **Jiajun Zhang** [1,2*], **Chengqing Zong** [1,2,3]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{shuxian.zou, shaonan.wang, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Decoding language from non-invasive brain activity has attracted increasing attention from both researchers in neuroscience and natural language processing. Due to the noisy nature of brain recordings, existing work has simplified brain-to-word decoding as a binary classification task which is to discriminate a brain signal between its corresponding word and a wrong one. This pairwise classification task, however, cannot promote the development of practical neural decoders for two reasons. First, it has to enumerate all pairwise combinations in the test set, so it is inefficient to predict a word in a large vocabulary. Second, a perfect pairwise decoder cannot guarantee the performance on direct classification. To overcome these and go a step further to a realistic neural decoder, we propose a novel Cross-Modal Cloze (CMC) task which is to predict the target word encoded in the neural image with a context as prompt. Furthermore, to address this task, we propose a general approach that leverages the pre-trained language model to predict the target word. To validate our method, we perform experiments on more than 20 participants from two brain imaging datasets. Our method achieves 28.91% top-1 accuracy and 54.19% top-5 accuracy on average across all participants, significantly outperforming several baselines. This result indicates that our model can serve as a state-of-the-art baseline for the CMC task. More importantly, it demonstrates that it is feasible to decode a certain word within a large vocabulary from its neural brain activity.

1 Introduction

Neural decoding, i.e., using brain activity to make predictions of stimuli or mental states, is a challenging cross-discipline research area. It is crucial for developing brain-computer interfaces (BCIs) that allow people to communicate using brain signals instead of verbal or body language (Wolpaw et al., 2002; Haynes and Rees, 2006). With the development of brain imaging technology and computational models, two lines of work emerge. One is invasive decoding, which depends on invasive brain recording methods such as electrocorticography (ECoG). In recent years, several breakthroughs have been made in this field and demonstrated the feasibility to decode speech (Anumanchipalli et al., 2019; Makin et al., 2020; Moses et al., 2021) or handwriting (Willett et al., 2021) from neural activity at high accuracy and speed. Despite the impressive performance, invasive decoding is unlikely to be used except in rare medical situations since it needs invasive surgery on the brain.

In contrast, non-invasive decoding uses atraumatic neuroimaging technologies such as functional magnetic resonance imaging (fMRI) to collect brain signals, having wider applicable groups and applications prospects. However, progress in this field is relatively slow after the pioneering work of Mitchell et al. (2008) that shows the feasibility to discriminate an fMRI image between two words. For a decade, this pairwise classification task (shown in Table 1) has been used as default on non-invasive brain-to-word decoding (Palatucci et al., 2009; Pereira et al., 2013; Anderson et al., 2017; Pereira et al., 2018; Wang et al., 2020). Nevertheless, it is quite limited to developing practical neural decoders. On the one hand, to predict a word, it has to enumerate all pairwise combinations in the test set and thus is inefficient. On the other hand, a decoder with high pairwise accuracy can fail to capture the similarity structure of the gold semantic space (Minnema and Herbelot, 2019). And hence it may not perform well on classifying fMRI images into vocabulary words, which is the ultimate goal of brain-to-word decoding.

Recently, Affolter et al. (2020) argues that we should move on to a more difficult but direct classification task rather than staying on the simple pairwise classification. In their work, they experi-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2022, pages 648 - 657

Task	Input	Target	Input modalities	Target space
Pairwise (Mitchell et al., 2008)	An fMRI image for *dog*	*dog*	fMRI	Two words
Direct (Affolter et al., 2020)	An fMRI image for *dog*	*dog*	fMRI	Vocabulary from stimuli
CMC (Ours)	An fMRI image for *dog* Context: *a ____ is a great companion.*	*dog*	fMRI & text	Vocabulary from corpus

Table 1: Brain-to-word decoding tasks. Our CMC task takes an fMRI image and a context as input and outputs a word related to the fMRI image in a large vocabulary.

ment on direct classification (shown in Table 1) and demonstrate the feasibility of multi-class classification using fMRI data to a certain extent. However, their direct classifier cannot predict words that do not appear in the training set, so it cannot perform zero-shot learning (ZSL). ZSL is essential for a practical neural decoder because it is impossible to collect brain images for every word in the vocabulary used daily. In addition, they ignore the context of word stimuli, which is ready-to-use and can serve as a prompt for brain-to-word decoding.

To overcome these and facilitate the development of pragmatic neural decoders, we propose a new brain-to-word decoding task called **C**ross-**M**odal **C**loze (CMC) task. As illustrated in Table 1, the CMC task is to classify a brain image with a context into a word from a large vocabulary. Intuitively, the given context should provide extra information for predicting words by narrowing down possible candidates. In addition, introducing contexts into brain-to-word decoding may bring some inspirations for brain-to-text decoding word by word.

Furthermore, to address this task, we propose a general approach that leverages the pre-trained language model BERT (Devlin et al., 2019) to predict the target words. The challenge lies in how to extract useful features carried by brain signals that can be integrated into BERT to facilitate prediction. We handle this problem by combining regression and representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to transform fMRI data to feature vectors in a specific semantic space of BERT.

In this paper, we focus on non-invasive single-subject zero-shot brain-to-word decoding. Our main contributions can be summarized as follows:

- We propose a more challenging but practical **C**ross-**M**odal **C**loze (CMC) task for brain-to-word decoding, which is a departure from the naive pairwise classification task. Hopefully this task could serve as a bridge from decoding individual words to decoding continuous sentences, paving the way to build a practical neural language decoder.
- We propose a general approach to address the CMC task. In particular, we propose Representational Similarity Retrieval (RSR) method to extract feature vectors from fMRI images, which can also be used for direct classification.
- We perform extensive experiments on 24 participants from two fMRI datasets collected on English word stimuli. Experimental results show the effectiveness of our method, indicating that our method can serve as a strong baseline for the CMC task.

2 Related Work

2.1 Neural Decoding Tasks

In this paper, we focus on non-invasive decoding methods, especially fMRI, which provides the best spatial resolution among all non-invasive neuroimaging techniques. This line of research starts from Mitchell et al. (2008), who for the first time show that it is feasible to decode words from fMRI data by leveraging the semantic representations of words and learning a cross-modal mapping between fMRI images and word vectors. They adopt pairwise classification task to evaluate the learned neural decoders, which is a binary classification task that discriminates which one in two stimuli corresponds to the fMRI image. Since then, pairwise classification is widely used by researchers in non-invasive neural decoding to decode words (Palatucci et al., 2009; Pereira et al., 2011; Chang et al., 2011; Pereira et al., 2013; Anderson et al., 2017; Pereira et al., 2018; Wang et al., 2020) as well as sentences (Pereira et al., 2018; Sun et al.,

2019; Sun et al., 2020). Recently, there are some voices in the BCI community arguing that pairwise classification is quite limited and more challenging but direct tasks need to be set up to push current neural decoding to a higher level (Affolter et al., 2020; Zou et al., 2021). They directly classify an fMRI image into vocabulary words. In contrast to pairwise classification, direct classification is a much harder task since the decoder needs to predict the correct words in a much larger space. In this paper, we propose to address the CMC task, a multi-class classification task with brain image and context as input.

2.2 Neural Decoding Methods

Regression-based Decoding Regression-based decoding is a prevalent approach to address the pairwise classification task and it is designed with the goal to perform ZSL. It first leverages a word embedding model to represent the word stimuli and then learns a regression model from fMRI images to each semantic dimension of the word vectors (Palatucci et al., 2009; Pereira et al., 2018; Wang et al., 2020). The learned models can predict word vectors for new brain images, which are used for pairwise matching.

Similarity-based Decoding Based on RSA, Anderson et al. (2016) have proposed a similarity-based decoding method to address pairwise classification. The basic idea is to re-represent the neural activity in neural similarity space and the word vectors in semantic similarity space. Then the two similarity spaces are used for pairwise matching. It is a non-parametric method that does not require model training. However, how to construct similarity space for direct classification is non-trivial.

Deep Learning based Decoding To address direct classification, Affolter et al. (2020) train an end-to-end deep learning model, taking fMRI images directly as input without dimension reduction. Their model can output a predicted probability for each word in a small vocabulary. However, their model is designed specifically for the fMRI dataset from (Pereira et al., 2018). And extending it to other datasets is not easy. Besides, it needs more data for training compared to statistical models.

2.3 Pre-trained Language Model

The CMC task can be viewed as a combination of a direct classification task and a Cloze task. In natural language processing, Cloze task has been well addressed by BERT (Devlin et al., 2019), a pre-trained masked language model that randomly masks some of the words from the input and then predicts the masked word based on its context during pre-training. This pre-training strategy makes it especially appropriate for the Cloze task. We can use BERT to predict a word using only the context as input and this can serve as a weak baseline for the CMC task.

3 Task

First of all, we specify some notations and formalize the data set for our Cross-Modal Cloze (CMC) task for clarity. Let $\mathcal{D}^S_{train} = \{(\{x^S_i, c^t_i\}, y_i)|t = 1, \cdots, T_i, i = 1, \cdots, M\}$ be the training set for subject S, where x^S_i denotes a brain image evoked by word y_i from S, c^t_i denotes a context related to word y_i, T_i denotes the number of contexts for word y_i, and $y_1, \cdots, y_M$ denote M distinct words. For each word, we have only one brain image for each subject. Similarly, we define $\mathcal{D}^S_{test} = \{(\{x^S_i, c^t_i\}, y_i)|t = 1, \cdots, T_i, i = 1, \cdots, N\}$ be the test set, where y_i denotes a word that is not in the training set.

Now, we give the definition of our CMC task. Given a brain image x^S_i and a context c_i related to word y_i, the goal of CMC task is to predict y_i from a given vocabulary $\mathcal{V}$.

There are two major differences between our CMC task and the other two brain-to-word decoding tasks. The first one is that the CMC task takes a context as input in addition to an fMRI image. Notice that communication generally happens under a certain context. With the context as the background, it is relatively easier to guess what other people think since possible candidates usually fall in a much smaller space constrained by the context. Contexts are very useful and easily accessible information and it would be beneficial to use them in neural decoding. The second difference is the size of the target space. The decoding space in the CMC task is the vocabulary from a corpus instead of just the word stimuli as in the direct classification, let alone the pairwise classification.

Evaluation Metric For an input sample $\{x^S_i, c^t_i\}$ in $\mathcal{D}^S_{test}$, if the top-k predictions contain y_i or its synonyms, then the classification is deemed correct. Let n' denote the number of correct top-k classifications, top-k accuracy is computed using

the following equation:

$$\text{Top-k acc} = \frac{n'}{T_1 + \cdots + T_N} \quad (1)$$

Meanwhile, the predicted probability of the target word y_i (and its synonyms) can be used as an auxiliary metric. It is obtained through the following equation:

$$\text{Prob} = \frac{\sum_{i=1}^{i=N} \sum_{t=1}^{t=T_i} P(y_i | c_i^t, x_i^S)}{T_1 + \cdots + T_N} \quad (2)$$

4 Method

To address the CMC task, we propose a general two-step approach: 1) extract semantic features from fMRI images by cross-modal retrieval, and 2) fuse the semantic features into BERT to perform the Cloze task. The fMRI image in a masked sentence is like a "switched" code, and the intuition of our method is to decode the code by utilizing the codes in natural language.

4.1 Step 1: Feature Extraction

The goal of Step 1 is to extract semantic features from the fMRI images that can be directly fused into the hidden states in the embedding layer of BERT [1]. To this end, the intuitive way would be directly learning a cross-modal mapping from fMRI images to their word embedding extracted from the embedding matrix in the embedding layer of BERT (BERT embedding for short). However, by investigating the 5 nearest neighbours (NN) of each word in fMRI180 using BERT embedding, we find that BERT embedding does not capture semantics well compared to other widely used word embedding such as GloVe (Pennington et al., 2014) or the contextualized word embedding derived from deeper layers of BERT. And it suffers from a more severe "hubness" problem (Radovanovic et al., 2010), a problem that the same point tends to be nearest neighbors of many points in high-dimensional spaces. To overcome this, we introduce an intermediate word embedding and design a retrieval-based method. The basic idea is to use the intermediate word embedding to perform cross-modal mapping and then transform the predictions into the BERT embedding space by retrieval.

[1]We choose the embedding layer of BERT as a proof of concept and other layers are similar.

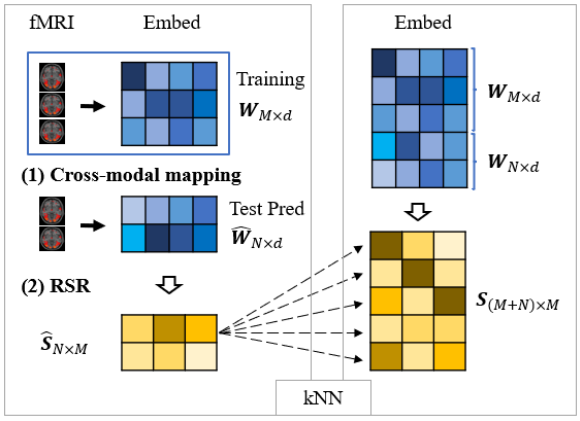

Figure 1: Feature extraction method, including two main steps: (1) Cross-modal mapping; and (2) Representational similarity retrieval (RSR). "Embed" represents the intermediate word embedding, such as GloVe, and it is different from BERT embedding.

Cross-Modal Mapping Let $\boldsymbol{W}'_{M\times d'}$ be the BERT embedding of words in $\mathcal{D}^S_{train}$ ($\mathcal{D}_{train}$ for simplicity) where d' denotes the dimension of the BERT embedding. Similarly, we have $\boldsymbol{W}'_{N\times d'}$ for words in $\mathcal{D}_{test}$. Let $\boldsymbol{W}$ denote the intermediate word embedding and d denote its dimension. As shown in Figure 1, we first use $\boldsymbol{W}$ in cross-modal mapping and train a linear regression model f_j to map x_i to $\boldsymbol{w}_{ij}$ $(i = 1, \cdots, M)$ for each semantic dimension j $(j = 1, \cdots, d)$ on the training set. Each regression model has $v + 1$ trainable parameters, where v denotes the number of selected voxels of fMRI images. And each model is trained independently. After the training, we use the mapping to obtain the predictions $\hat{\boldsymbol{W}}_{N\times d}$ for fMRI images in the test set $\mathcal{D}_{test}$.

Representational Similarity Retrieval Now the goal is to retrieve k NN words in the vocabulary for each predicted intermediate word vector of fMRI images. To this end, we construct a similarity space based on the M ground-truth intermediate word vectors in the training set. As shown in Figure 1, similarity[2] between the predicted embedding $\hat{\boldsymbol{w}}_i (i = 1, \cdots, N)$ and all M words in the training set are computed, resulting in an M-dimensional vector $\hat{\boldsymbol{s}}_i$ in the similarity space. Similarly, similarity between the ground-truth word embedding $\boldsymbol{w}_i (i = 1, \cdots, M + N)$ and all M words in the training set are computed, giving an M-dimensional vector $\boldsymbol{s}_i$ in the similarity space.

[2]We use Pearson correlation coefficient as the default similarity function in this work unless otherwise specified.

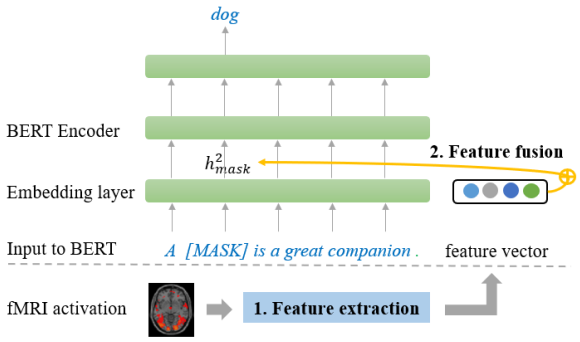

Figure 2: Feature fusion method. The pre-trained language model BERT is used to predict target words.

Based on these new representations, we retrieve k NN words in vocabulary $\mathcal{V}$ for each fMRI test sample x_i. For x_i, its k NN word indices are obtained through the following equation:

$$j_1, \cdots, j_k = \text{topk}(\{\text{sim}(\hat{\boldsymbol{s}}_i, \boldsymbol{s}_j) | j = 1, \cdots, |\mathcal{V}|\}) \tag{3}$$

where $\text{sim}(.,.)$ represents a similarity function, and $j_1, \cdots, j_k$ denote the indices of the top-k similar vectors in $\boldsymbol{S}_{(M+N)\times M}$. Then the feature vector $\boldsymbol{f}_i$ for fMRI test sample x_i is computed as follows:

$$\boldsymbol{f}_i = \frac{1}{k} \sum_{t=1}^{k} \boldsymbol{w}'_{j_t} \tag{4}$$

where $i = 1, \cdots, N$ and $\boldsymbol{w}'$ denotes BERT embedding.

4.2 Step 2: Feature Fusion

The goal of Step 2 is to integrate the feature vector $\boldsymbol{f}_i$ extracted from fMRI image x_i into BERT for better word prediction. Intuitively, if the feature vector $\boldsymbol{f}_i$ carries useful information of the target word y_i, then fusing it into the model should improve the performance in predicting y_i than merely using context c_i as input.

To be specific, let $\boldsymbol{h}^i_{mask}$ denote the hidden states of the `[MASK]` token in c_i, we directly update $\boldsymbol{h}^i_{mask}$ using the following equation:

$$\boldsymbol{h}^i_{mask} := (1 - \alpha)\boldsymbol{h}^i_{mask} + \alpha \boldsymbol{f}_i \tag{5}$$

where $\alpha \in [0, 1]$ is a tuning parameter that controls how much information to fuse in. The feature fusion method is shown in Figure 2. It only operates at the embedding layer of BERT and does not require fine-tuning the pre-trained model, which is quite straightforward.

In general, our two-step approach for the CMC task contains one trainable cross-modal transformation matrix (size $(v + 1) \times d$) and two main hyperparameters k and α. This pipeline approach is designed for small datasets considering that the sample size of a brain imaging dataset is often very small. And it is a general method that can be applied to any fMRI dataset.

5 Datasets

5.1 Brain Imaging Datasets

According to our knowledge, there are two open-source fMRI datasets collected from subjects exposed to English word stimuli and concentrated on thinking about the meaning of words. The first one is from (Mitchell et al., 2008), which contains 60 word stimuli. And the second one is from (Pereira et al., 2018), which consists of 180 word stimuli. For clarity, we denote them as fMRI60 and fMRI180 respectively.

fMRI60[3] fMRI60 contains the neural activity collected from 9 human participants while viewing 60 different concrete nouns. Some examples include: *carrot*, *dog*, *hammer*, *igloo*, *skirt*. During the brain recording process, each participant was shown a word and a small line drawing of the concrete object the word represents. The participants were asked to think about the properties of these objects. For each word, six fMRI scans with roughly 20,000 voxels are available. To reduce noise, we average the six scans to create a single fMRI image for each of the 60 words and each participant. The statistics of fMRI60 are shown in Table 2.

fMRI180[4] fMRI180 contains the neural activity observed from 15 human participants while viewing 180 content words. Some examples include: *ability*, *big*, *damage*, *experiment*, *seafood*. During the fMRI scanning process, each participant was shown a word presented in a sentence with itself in bold to highlight the relevant meaning. They were asked to think about the meaning of the target word in the context. There are two other paradigms as well, one uses pictures instead of sentences, and the other uses word clouds instead of sentences. For each word in each paradigm, 4-6 fMRI scans were taken with context varying and then were combined into a single fMRI image by using a general linear model. The data available online is one fMRI

[3] http://www.cs.cmu.edu/~tom/science2008/index.html
[4] https://osf.io/crwz7/

	Subject	Voxel	fMRI	Word	Sent	Syon
fMRI60	9	~20,000	60	60	360	0.28
fMRI180	15	~200,000	180	180	1080	0.29

Table 2: Statistics about the fMRI datasets. "Voxel" refers to the number of voxels (similar to pixels in a 2-dimensional image) containing in a 3-dimensional fMRI image. "Syon" denotes the average number of synonyms for the word stimuli.

fMRI60_CMC:	
carrot	the [MASK] is his favorite vegetable.
hammer	she puts the [MASK] down on the ground.

fMRI180_CMC:	
ability	he has the [MASK] to cultivate creativity.
damage	the accident left some serious [MASK].

Table 3: Context examples for word stimuli in the CMC datasets.

image per word per paradigm. To further reduce noise, we average the data across three paradigms to generate a single fMRI image for each of the 180 words, for each participant. The statistics of fMRI180 are shown in Table 2.

5.2 CMC Datasets

fMRI60_CMC For each of 60 words in fMRI60, we create 6 sentences, 4-13 words long (mean = 6.68, std = 1.57), and each containing the target word used in the intended sense. To create contexts for the CMC task, we remove the target word in its corresponding sentences by using a `[MASK]` token to replace it. Two context examples are shown at the top of Table 3. Furthermore, to create synonyms for the word stimuli, we first use WordNet (Miller, 1995) to find possible candidates and then manually proofread all the words to make sure they have the same meaning as the word stimuli. We obtain 0.28 synonyms per stimulus on average. Combining the brain imaging data, the contexts and the target words into the form we describe in Section 3, a dataset fMRI60_CMC is generated for the CMC task. It is publicly available at https://github.com/LittletreeZou/Cross-Modal-Cloze-Task.

fMRI180_CMC For words in fMRI180, we use the sentences in the presentation scripts in Pereira et al. (2018)'s experiment. These sentences are 4–11 words long (mean = 6.85, std = 1.22) and also contain the target words used in the intended meaning. Similarly, we create contexts and collect synonyms for each word stimulus (0.29 synonyms per stimulus on average). Then a dataset fMRI180_CMC for the CMC task is generated. Two context examples of fMRI180_CMC are shown at the bottom of Table 3.

6 Experiments

6.1 Experimental Settings

Voxel Selection As shown in Table 2, fMRI data is very high-dimensional, containing up to 200,000 voxels, while the sample size is very small. To avoid overfitting and reduce the computational complexity in cross-modal mapping, voxel selection is often performed to reduce the dimensions. Following the method proposed by (Pereira et al., 2018), we select the most informative 5,000 voxels for each subject in each fMRI dataset. Then we obtain a 5000-dimensional vector for each fMRI image. We use these fMRI vectors in the following experiments and still use the term "fMRI image" to refer to them.

Data Partition Since the CMC datasets are quite small, we split each dataset into 10 folds by word stimuli to allow cross-validation. For each fold, 8 folds are used for training, 1 fold is used for validation and 1 fold is used for test. The data partition is the same across subjects, with the same word stimuli in the same fold.

Models For the CMC task, we use BERT[5] without fine-tuning. In the cross-modal mapping, we use the most commonly used ridge regression in neural decoding literature as default. It is a linear regression model with L_2 regularization, which can regulate overfitting since we have 5,000 input features. The regularized hyperparameter is automatically optimized based on Pearson correlation coefficient of the predicted values and the true values on the validation data. And we experiment on three types of word embedding, including BERT embedding, GloVe[6] and contextualized embedding BERT LayerAvg[7]. Our best model uses BERT LayerAvg. The hyperparameter k is tuned to 5, and α is tuned to 0.7 based on the top-5 accuracy on the validation set.

[5]https://huggingface.co/bert-base-uncased

[6]https://nlp.stanford.edu/data/glove.840B.300d.zip

[7]For each word, 6 sentences containing that word are fed to BERT and the corresponding hidden states of the word in layer 7-12 are collected and further averaged into a word vector.

(%)	fMRI60_CMC			fMRI180_CMC		
	Top-1 acc	Top-5 acc	Prob	Top-1 acc	Top-5 acc	Prob
BERT	27.50	45.56	17.20	26.02	50.93	17.13
BERT-Direct Fusion	27.78	49.54	18.33	26.51	51.78	17.88
BERT-Retri Fusion (random)	24.81	44.01	16.04	25.91	50.83	17.60
BERT-Retri Fusion	**31.08**	**55.99**	**21.24**	**27.60**	**53.11**	**18.99**

Table 4: Performance on the CMC task. "BERT-Retri Fusion (random)" is a random version of "BERT-Retri Fusion" where feature vectors are randomly shuffled and do not match the contexts. The results are first averaged among 10-fold cross-validation data and then averaged across subjects. The best results are shown in bold.

(%)	Min	Max	Mean	SigFrac
fMRI60_CMC	48.61 +3.06	61.39 +15.83	55.99 +10.43	8/9
fMRI180_CMC	51.30 +0.37	55.19 +4.26	53.11 +2.19	11/15

Table 5: Statistical analysis of performance on subjects. Top-5 accuracy of "BERT-Retri Fusion" and the absolute improvement over BERT are reported in this table. "Min", "Max" and "Mean" denote the worst, best and average results on all subjects respectively. "SigFrac" denotes the fraction of participants with significant improvement over BERT.

6.2 Main Results

As shown in Table 4, the first three rows are three weak baselines for the CMC task. BERT achieves quite good results on both datasets, demonstrating the power of pre-trained models on Cloze task. Our method – BERT-Retri Fusion – achieves the best results on this task. On fMRI60_CMC, it achieves 31.08% top-1 accuracy and 55.99% top-5 accuracy on average across 9 subjects, outperforming BERT by absolute improvement of +3.58% and +10.43% respectively. On fMRI180_CMC, it achieves 27.60% top-1 accuracy and 53.11% top-5 accuracy on average across 15 subjects, outperforming BERT by absolute improvement of +1.59% and +2.19% respectively. Furthermore, the predicted probability of the target words in our method increases by 4.04% and 1.86% on the two datasets respectively, indicating our model is more confident about the correct answer. These results indicate three things: 1) The feature vectors derived from the fMRI data are informative and can be utilized by BERT to better predict the target word; 2) Our method is effective and can serve as a strong baseline for the CMC task; 3) It is feasible to decode an fMRI image with context as prompt into a word from a large vocabulary.

Moreover, when comparing the performance of BERT, BERT-Direct Fusion and BERT-Retri Fusion, the fusion of feature vectors does not necessarily result in significant improvement unless the feature vectors are good enough. When comparing the performance of BERT, BERT-Retri Fusion and its random version, fusing the mismatched feature vectors from other fMRI images into BERT decreases the performance a little bit while fusing the correct one increases the performance by a significant margin. This result indicates that the feature vectors derived from fMRI data by our method carry a certain amount of semantic information about the target words.

Finally, we investigate the performance of our method on different subjects. For each subject, we perform a significance test on the top-5 accuracy to see whether our method is significantly better than BERT. The data points are the 10-fold top-5 accuracy of our model and BERT. The statistical test we used is paired t-test with significance level 0.1 and FDR controlled for multiple comparisons (Benjamini and Hochberg, 1995). As shown in Table 5, on both datasets, all results of our method are better than BERT and most of them are statistically significant.

7 Analysis

Ablation Study The feature extraction step is a key step in our method. In this step, we use retrieval-based method. Hence the direct classification accuracy can also be used to evaluate the quality of feature vectors. We perform ablation experiments to understand the relative importance of each facet of our method. As shown in Table 6, using non-parametric RSR to match fMRI images and word vectors is better than using regression. And combining the two methods achieves the best

(%)	Direct Classification				CMC Task			
	fMRI60		fMRI180		fMRI60_CMC		fMRI180_CMC	
	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
Baseline	1.67	8.33	0.56	2.78	27.50	45.56	26.02	50.93
REG-NN	1.11	14.26	1.26	7.19	24.78	45.03	25.83	50.77
RSR	10.19	33.33	3.41	11.93	29.23	53.24	26.38	52.20
REG-RSR	**22.78**	**55.19**	**13.37**	**34.22**	**31.08**	**55.99**	**27.60**	**53.11**

Table 6: Ablation study on fMRI feature extraction. "Baseline" for direct classification refers to random baseline while for CMC task it refers to BERT. "REG" denotes regression. "REG-RSR" denotes our method. The results are first averaged among 10-fold cross-validation data and then averaged across subjects. The best results are shown in bold.

result, indicating that both regression and RSR are important for extracting fMRI semantic features.

Effect of Word Embedding In theory, our method can work with any type of word embedding. We perform experiments on two major types of word embedding, one is non-contextualized GloVe and the other is contextualized BERT Layeravg. As shown in Table 7, the performance on the CMC task using the two different types of word vectors are quite similar on both datasets. In contrast to previous work done on pairwise classification which focuses on finding better representations of words, the CMC task is not so sensitive to the types of word embedding.

(%)	Embedding	Top-1 acc	Top-5 acc
fMRI60_CMC	GloVe	29.85	54.88
	BERT Layeravg	31.08	55.99
fMRI180_CMC	GloVe	27.81	52.96
	BERT Layeravg	27.60	53.11

Table 7: Effect of word embedding used in our method for the CMC task. The results are first averaged among 10-fold cross-validation data and then averaged across subjects.

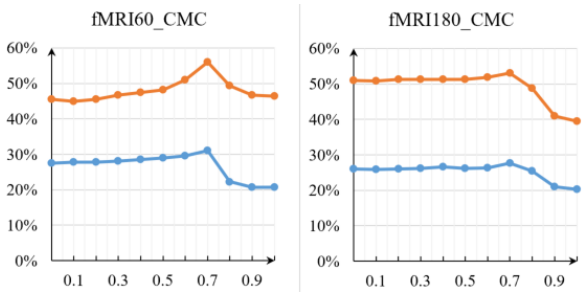

Figure 3: Effect of α. The blue line denotes top-1 accuracy while the orange one represents top-5 accuracy. k is set to 5 for all subjects.

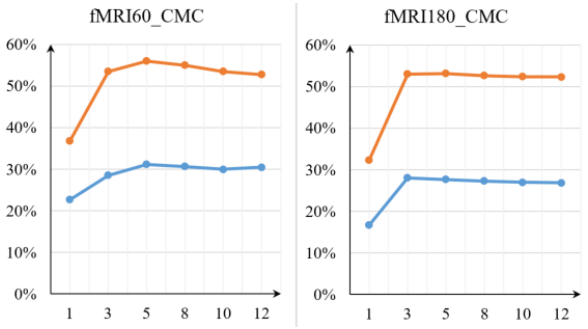

Figure 4: Effect of k. The blue line denotes top-1 accuracy while the orange one represents top-5 accuracy. α is set to 0.7 for all subjects.

Effect of α The hyperparameter α controls how much information from fMRI to fuse into BERT. As shown in Figure 3, as α increases from 0 to 0.7 gradually, the performance of the model steadily increases and reaches the maximum performance when $\alpha = 0.7$ on both datasets. This tendency demonstrates that the fusion of fMRI information is helpful for predicting words. However, when α becomes too large, the performance will drop quickly. We speculate that this is because the BERT embedding of the mask token is useful for predicting words, since that is how BERT was pre-trained.

Effect of k The hyperparameter k controls how many neighbors we want to engage with to derive the feature vector for an fMRI image. Intuitively, if we retrieve more words for an fMRI image, we have a larger probability to recall the target word. However, a larger k will diminish the utility of the feature vector since it is the average of word vectors corresponding to the k retrieved words. As shown in Figure 4, increasing k from 1 to 3 gives a large gain in decoding accuracy on both datasets. When $k > 5$, the decoding accuracy declines slowly. Generally, $k = 5$ is the best tradeoff between the prob-

ability of recalling the target word and the informativeness of the feature vector.

Limitations Our experiments are largely limited by three characteristics of fMRI signals, which are low temporal resolution with delayed hemodynamic response, noisy, 3D volume containing hundreds of thousands of voxels with small sample size. Correspondingly, there are three major limitations in our work. First and foremost, the context we use is not the actual context in which the fMRI images were collected. The reason that we use synthesised context is to avoid word-level alignment of fMRI data, which is currently too difficult when they are presented as continuous stimuli (Hollenstein et al., 2020). Second, to reduce noise, we use brain activity averaged across multiple trails rather than single-trail-based brain activity, which is different from the real scenarios of BCI applications. Third, we do not consider the spatial structure of brain images, but flatten them directly into vectors. While this is common practice in fMRI decoding, exploring spatial patterns may help deepen our understanding of the brain and improve neural decoding accuracy.

8 Conclusions and Future Work

In this paper, intending to build practical neural language decoders, we investigate the feasibility of large-vocabulary zero-shot brain-to-word decoding. Large-vocabulary classification is much harder than pairwise classification. By introducing context as a prompt, we formalize it as a cross-modal Cloze task, which alleviates the decoding difficulty while keeping the essence of neural decoding. Furthermore, if we assume the past and the future content of brain activity has been decoded, our CMC task can be viewed as a simplified version of brain-to-text decoding. Based on this task, we find that decoding brain activity into words from a large vocabulary is possible to a certain extent, which lays the foundation for decoding text word by word from the brain.

To move towards brain-to-text decoding, we can use a generative pre-trained language model to replace BERT. The biggest challenge lies in how to align fMRI signals to individual words when the stimuli are presented as a continuous time series of words. In the future, we are going to address this problem since it is fundamental for building powerful neural language decoders that translate brain activity into text.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 62036001, 62122088 and 61906189. We thank the anonymous reviewers for their insightful comments and suggestions.

References

Nicolas Affolter, Béni Egressy, Damián Pascual, and Roger Wattenhofer. 2020. Brain2word: Improving brain decoding methods and evaluation. In *Medical Imaging Meets Neurips Workshop-34th Conference on Neural Information Processing Systems (NeurIPS)*.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrew James Anderson, Benjamin D Zinszer, and Rajeev DS Raizada. 2016. Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Kai-min Kevin Chang, Tom Mitchell, and Marcel Adam Just. 2011. Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fmri activation. *NeuroImage*, 56(2):716–727.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John-Dylan Haynes and Geraint Rees. 2006. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on*

Linguistic and Neurocognitive Resources, pages 15–27, Marseille, France. European Language Resources Association.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Gosse Minnema and Aurélie Herbelot. 2019. From brain space to distributional space: The perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy. Association for Computational Linguistics.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, pages 1191–1195.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, volume 22, pages 1410–1418.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francisco Pereira, Matthew Botvinick, and Greg Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252.

Francisco Pereira, Greg Detre, and Matthew Botvinick. 2011. Generating text from functional brain images. *Frontiers in human neuroscience*, 5:72.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963–963.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7047–7054.

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

Shaonan Wang, Jiajun Zhang, Haiyan Wang, Nan Lin, and Chengqing Zong. 2020. Fine-grained neural decoding with distributed word representations. *Information Sciences*, 507:256–272.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. 2002. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2021. Towards brain-to-text generation: Neural decoding with pre-trained encoder-decoder models. In *NeurIPS 2021 AI for Science Workshop*.