Towards Brain-to-Text Generation: Neural Decoding with Pre-trained Encoder-Decoder Models

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

Abstract

Decoding language from non-invasive brain signals is crucial in building widely applicable brain-computer interfaces (BCIs). However, most of the existing studies have focused on discriminating which one in two stimuli corresponds to the given brain image, which is far from directly generating text from neural activities. To move towards this, we first propose two neural decoding tasks with incremental difficulty. The first and simpler task is to predict a word given a brain image and a context, which is the first step towards text generation. And the second and more difficult one is to directly generate text from a given brain image and a prefix. Furthermore, to address the two tasks, we propose a general approach that leverages the powerful pre-trained encoder-decoder model to predict a word or generate a text fragment. Our model achieves 18.20% and 7.95% top-1 accuracy in a vocabulary of more than 2,000 words on average across all participants on the two tasks respectively, significantly outperforming their strong baselines. These results demonstrate the feasibility to directly generate text from neural activities in a non-invasive way. Hopefully, our work can promote practical non-invasive neural language decoders a step further.

1 Introduction

Decoding language from human brain activities is crucial in building BCIs that translate brain signals into a coherent text. This technology is considered transformative for helping those who are unable to communicate due to some severe neuromuscular disorders [1]. Meanwhile, it also offers a tool for neuroscientists to study brain mechanisms. Two lines of neural decoding research have dominated this field: invasive decoding, based on invasive brain recording methods such as electrocorticography (ECoG); and non-invasive decoding, depending on atraumatic neuroimaging technologies such as functional magnetic resonance imaging (fMRI). In recent years, several breakthroughs have been made in invasive decoding and demonstrated the feasibility to decode speech [2–4] or handwriting [5] from neural activities at high accuracy and speed. Nevertheless, 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 is applicable to normal people without doing any harm. There have been some successful attempts in decoding words [6–8] and sentences [9–11] from fMRI data in the form of pairwise classification. The pairwise classification is a binary classification task that discriminates which one in two stimuli corresponds to the given fMRI image. The major limitations of this setting are: 1) to predict a word or a sentence, it has to enumerate all pairwise combinations in the test set and thus is inefficient; and 2) for sentence decoding, it can only select a fixed sentence from two options and thus can not produce flexible sentences.

NeurIPS 2021 AI for Science Workshop.

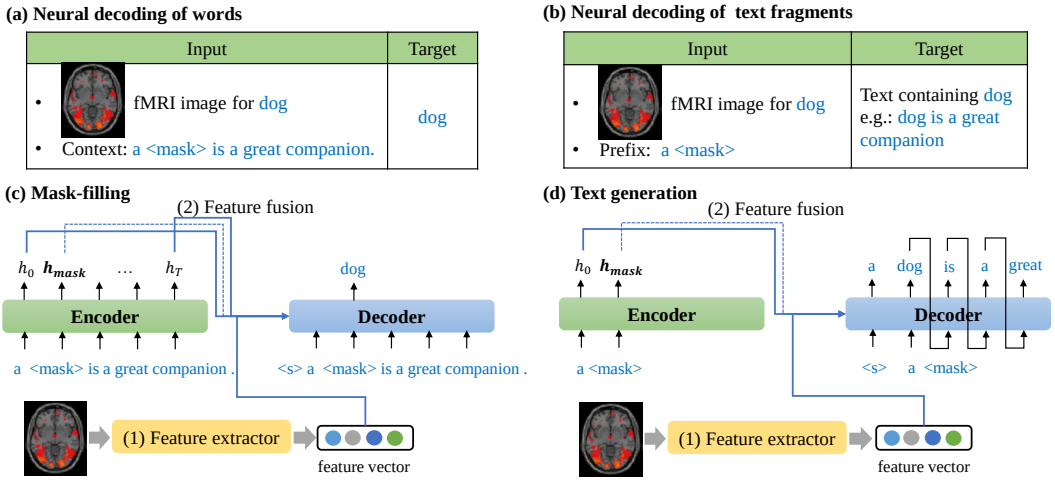

Figure 1: The two proposed neural decoding tasks are shown in (a) and (b) respectively. The mask token in the table is a placeholder indicating the lack of some words. The approach to address each task is shown in (c) and (d) respectively. The blue lines denote cross-attention.

Until now, little work has been done to directly generate text from fMRI images. Early work has attempted to generate natural language descriptions for brain activities evoked by motion pictures [12]. They formulate the neural decoding problem as an image-captioning task and employ a pre-trained image-captioning system to perform this task. Despite the novel method, they adopt perplexity as the evaluation metric, which can not reflect the relevance of the generated text to the given fMRI image. Recently, Affolter et al. [13] have tried to generate text conditioned on fMRI images using GPT-2 [14]. They first train a classifier that takes an fMRI image as input and outputs a probability vector over a vocabulary of 180 words. Then to generate the next word for the history, they directly adjust the output probability vector produced by GPT-2 by using the top-5 words predicted by the classifier. However, in their experiment, they use ground-truth information in prediction by limiting the top-5 words to always contain the target word.

To understand to what extent can we currently decode text from fMRI data, we take advantage of the useful and easily accessible context and design two neural decoding tasks with incremental difficulty. The first task, as shown in Figure 1(a), is to predict a word given an fMRI image and a masked sentence related to that word. The second task, as shown in Figure 1(b), is to generate a text fragment containing the target word given an fMRI image and a prefix. In general, the first task is an easier one while the second task is more closer to practical neural decoders.

To address the two tasks, we propose a general approach that leverages the powerful pre-trained transformer-based encoder-decoder model BART [15] by formulating them into a mask-filling task and a text generation task respectively. In summary, our contributions are: 1) we propose two neural decoding tasks to explore the feasibility of decoding text from fMRI data, paving the road to build practical neural decoders that translate neural activities into a coherent text; 2) we propose a general approach that leverages the powerful pre-trained encoder-decoder model to address the two neural decoding tasks; 3) we validate the effectiveness of our method and demonstrate the feasibility to generate text from fMRI data.

2 Neural decoding method

To address the two neural decoding tasks – the mask-filling task and the text generation task, we propose a general approach as illustrated in Figure 1(c)-(d). Our method contains two steps: 1) extracting semantic features from the fMRI image, and 2) fusing the extracted features into BART to predict a word or generate a text fragment. We choose BART as the backbone of our model for two reasons. One is that BART is a conditional language model. And the other is that BART is pre-trained using an in-filling scheme, where spans of text are replaced with a single mask token. Hence, BART is applicable and effective to both mask-filling and text generation tasks.

Feature Extraction Assume we have N fMRI images corresponding to N word stimuli, we first represent the word stimuli using word vectors derived from the last hidden layer of BART's encoder[1]. The word embedding matrix is denoted as $\boldsymbol{E} \in \mathbb{R}^{N*d}$, where d refers to the dimension of the word vectors. Following the conventions in the pairwise classification [6–10], we train ridge regression models to map fMRI samples to their corresponding word vectors under cross-validation setting. For each fMRI sample in the hold-out test data, its predicted word vector $\hat{\boldsymbol{e}}$ is used as a query to retrieve k nearest neighbor word vectors in the ground-truth semantic space $\boldsymbol{E}$ using the cross-domain similarity local scaling (CSLS) method [16]. The feature vector for each fMRI sample is obtained using the following equation:

$$\boldsymbol{f}_i = \frac{1}{k} \sum_{t=1}^{k} \boldsymbol{e}_{i_t} \tag{1}$$

where $i = 1, ..., N$ and $\boldsymbol{e}_{i_1}, ..., \boldsymbol{e}_{i_k}$ are the k retrieved word vectors. Intuitively, the feature vector can be viewed as a new representation of the fMRI sample in the semantic space where we perform feature fusion.

Feature Fusion As shown in Figure 1(c)-(d), feature fusion is performed at the last hidden layer of BART's encoder. The information from the neural activity and the information from the context are fused together in that layer and then propagated forward to the decoder. Operating at this particular layer allows the decoder to decide to what extent is the added information used in generating the next word since they interact by cross-attention. To be specific, assume we have an fMRI image corresponding to a word and a masked sentence / text fragment, the text is fed into BART as before. And then the hidden states of the mask token $\boldsymbol{h}_{mask}$ in the last hidden layer of the encoder is updated using the following equation:

$$\boldsymbol{h}_{mask} \coloneqq \boldsymbol{f}_i \tag{2}$$

where $\boldsymbol{f}_i$ is the feature vector corresponding to the fMRI image. Then the hidden states from the encoder are propagated forward to the decoder as it originally does.

In general, our decoding method is quite straightforward. For simplicity, we do not fine-tune BART in our experiment but use it to derive word vectors as well as to predict a word / words. The feature extraction step and the feature fusion step are decoupled and can be optimized independently.

3 Experiments

Datasets The brain imaging data is from [9], which contains 180 fMRI images corresponding to 180 content words collected from 15 human participants[2]. We select the most informative 5,000 voxels following the voxel selection method proposed by [9] to reduce the dimension of fMRI data. To prepare the context for our two decoding tasks, we use the sentences in the presentation scripts in the fMRI experiment and mask the target word in each sentence as illustrated in the upper part of Figure 2. Two samples of the dataset for each task are shown in the lower part of Figure 2. As a whole, we create a neural decoding dataset of 1,080 samples per task per participant.

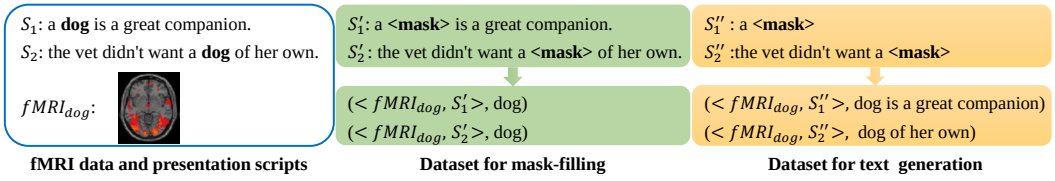

Figure 2: Illustrations of datasets. Two samples for each task are shown. The left side of the parentheses denotes Input while the right side represents Target.

[1]For each word, 6 sentences containing that word are fed into BART. Then we select the corresponding hidden states of the word in the last hidden layer of the encoder and average them into a word vector.

[2]For details about the dataset and the experiment paradigms, please refer to https://osf.io/crwz7/.

Models We adopt BART without any extra information fusion and fine-tuning as our baseline[3]. For neural decoding, we build our model on the same BART without fine-tuning. The hyperparameter k is tuned to 5. In the text generation task, we use greedy decoding, which is to select the most probable word on each step of generation. All the experiments are done under an 18-fold cross-validation setting, repeatedly using 16 folds for training, 1 fold for validation, and 1 fold for testing.

Evaluation metrics For the mask-filling task, we adopt top-1 accuracy as an evaluation metric following [13]. In the text generation task, if the generated text contains the target word, then it is deemed correct, so the evaluation metric is also top-1 accuracy.

Results of the mask-filling task For the mask-filling task, as shown in Table 1, the strong baseline achieves 17.13% accuracy, demonstrating the effectiveness of BART in predicting words. Based on BART, our method achieves 18.20% accuracy on average across 15 participants, outperforming the baseline by 1.07% absolute improvement. The minimum and maximum accuracy are 17.87% and 18.52% respectively, both surpassing the baseline. Furthermore, the improvements on 14 participants are statistically significant under paired t-test with p-value < 0.05 while the p-value for the worst subject is 0.051. These results demonstrate the feasibility to directly predict a word in a large vocabulary from an fMRI image, which is a departure from the traditional pairwise classification task.

Results of the text generation task As shown in Table 2, the strong baseline for the text generation task obtains 6.48% accuracy, which is much lower than the performance in the mask-filling task. This result shows that the text generation task is much harder than the mask-filling one. Despite the difficulty of the task, our model is able to achieve 7.95% accuracy on average across all participants, significantly outperforming the baseline by 1.47% absolute improvement. The minimum and maximum accuracy are 7.41% and 8.24% respectively, both surpassing the baseline. Moreover, the performances of our model on 14 participants are significantly better than the baseline. These results, for the first time, demonstrate that it is feasible to generate text related to the stimuli from fMRI images without using the ground-truth information. The experiment results show the effectiveness to use the pre-trained encoder-decoder model in neural decoding.

Table 1: Mask-filling results.

(%)	Baseline	Min	Max	Mean
Acc	17.13	17.87	18.52	18.20
ΔAcc		+0.74	+1.39	+1.07

Table 2: Text generation results.

(%)	Baseline	Min	Max	Mean
Acc	6.48	7.41	8.24	7.95
ΔAcc		+0.93	+1.76	+1.47

4 Conclusions and future work

In this paper, we have proposed two neural decoding tasks to understand to what extent can we currently decode text from brain images. From the perspective of engineering, these two tasks are departures from the traditional pairwise classification task. They can help to promote the development of practical non-invasive neural decoders that translate brain activities into a coherent text. From the perspective of neuroscience, a good neural decoder can serve as a useful adjunct to basic research in cognitive neuroscience. Future work will be move on to explore how voxels selected from different brain regions affect the accuracy of neural decoders. In this way, we can infer where in the brain semantics are represented.

In the field of natural language processing (NLP), pre-training has become a dominant paradigm. Language models pre-trained on text from a wide variety of sources have achieved remarkable results on language understanding and language generation tasks. In this paper, we have experimented with a pre-trained transformer-based conditional language model. In the future, we are going to experiment on more pre-trained models to investigate how different model architectures and pre-training objectives affect the performance of neural decoders. Intuitively, models that better correlate with brain activation data may better explain brain mechanisms. Through contrastive analysis of different pre-trained models in neural decoding, hopefully, we can provide some insights into how humans learn a language.

[3]https://huggingface.co/facebook/bart-base

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 61906189 and 62036001.

References

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

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

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

[4] 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. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

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

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

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

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

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

[10] 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, pages 7047–7054, 2019.

[11] 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.

[12] Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Generating natural language descriptions for semantic representations of human brain activity. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 22–29, 2016.

[13] Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: Decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[16] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.