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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decoding language from human brain signals is a crucial step to build communication brain-computer interfaces (BCIs). However, most of the existing studies in non-invasive decoding have focused on discriminating which one in two stimuli corresponding to the target word or sentence, while little work has been done to generate text from neural activities. To go a step further to text generation neural decoding, we propose to address two neural decoding tasks with different difficulties. The first task is to predict a word given a brain image and a context. Directly predicting words from neural activities is the first step towards text generation. Only if the first task is shown feasible can we further move on to the second task, which is to generate text given a brain image and a starting text. Furthermore, we propose a general approach that leverages the powerful pre-trained encoder-decoder model to address the two tasks at the same time. Experiment results demonstrate the feasibility to generate text from human brain activities and the effectiveness of our method. Hopefully, our work can promote practical non-invasive neural language decoders a step further.

1 Introduction

Decoding language from human brain activity is a crucial step to build a communication BCI, which are considered to be transformative for people who are unable to communicate due to some severe neuromuscular disorders [1]. 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 [2–4] or handwriting [5] from neural activities 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, which uses atraumatic neuroimaging technologies such as functional magnetic resonance imaging (fMRI) to collect brain signals, is likely to have wider applicability. Attempts have been successful in decoding words [6, 7] and sentences [8, 9] from fMRI data in the sense of pairwise classification. It is a binary classification task that discriminates which one in two stimuli corresponds to the given fMRI image, regardless of what kind of stimuli it is. However, simply discriminating two sentences is unable to build a practical neural decoder that can generate text.

Little work has been done to directly generate text from fMRI images. Early work has attempted to generate natural language descriptions for brain activity caused by motion pictures [10]. 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, their evaluation scheme is poorly

Submitted to NeurIPS 2021 AI for Science Workshop.

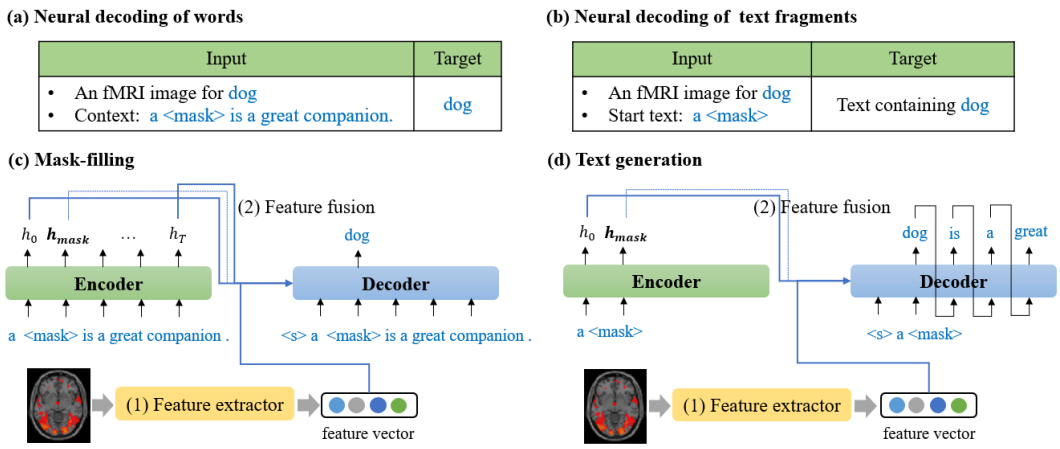

Figure 1: The two proposed neural decoding tasks are shown in (a) and (b) respectively. The approach to address each task is shown in (c) and (d) respectively. The blue lines in (c) and (d) denote cross-attention.

set up so that the actual performance of the model remains unknown. Recently, Affolter et al. [11] have tried to generate text conditioning on fMRI images using GPT-2 [12]. They train a classifier that takes an fMRI image as input and outputs a probability distribution on a vocabulary. Words corresponding to the top-5 probabilities are used to modify the next token prediction in text generation process. However, in their experiment, they only use those fMRI data where the classifier has a correct top-5 prediction, which is an unrealistic setting.

To understand to what extend we can 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 context related to that word.** Directly predicting words is the first step towards text generation. **The second one, as shown in Figure 1(b), is to generate text fragments containing the target word given an fMRI image and a start text.** Only if these two tasks are demonstrated feasible can we further move on to harder tasks such as generating text from a series of fMRI images.

Furthermore, we propose a general approach that leverages the powerful pre-trained sequence-to-sequence model BART [13] to perform both the two tasks by formulating them to a mask-filling task and a text generation task respectively. Our model can address the two tasks at the same time without any modification.

In summary, our contributions are: 1) we propose two neural decoding tasks to explore the feasibility of decoding text from fMRI data step by step, paving the road to build practical neural decoders that can translate neural activities into coherent text; 2) we propose a general approach to address the two neural decoding tasks at the same time; 3) we validate the effectiveness of our method and demonstrate the feasibility to generate text from fMRI data.

2 Background

In this work, we build our neural decoding method upon BART, a pre-trained model with a transformer-based encoder-decoder architecture [13]. The encoder is bidirectional like BERT [14] and takes corrupted text as input. The decoder is autoregressive like GPT [15], where each layer performs cross-attention over the final hidden layer of the encoder. The pre-training task involves a novel in-filling scheme, where spans of text are replaced with a single mask token. Hence, BART is applicable and effective to both text generation and mask-filling tasks.

3 Neural decoding method

To address the two neural decoding tasks, namely the mask-filling task and the text generation task, we propose a general approach as illustrated in Figure 1(c)-(d), which contains two steps: 1)

transforming the fMRI image to a feature vector in specific semantic space, and 2) fusing the feature vector into BART to predict word/words.

Feature Vectors 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 embeddings matrix is denoted as $\boldsymbol{E} \in \mathbb{R}^{N*d}$, where d denotes the dimension of the word vectors. Following the conventions in the pairwise classification [6–9, 16], we use ridge regression models under cross-validation setting to map each fMRI sample to its corresponding word vector. For each fMRI sample, we use its predicted word vector $\hat{\boldsymbol{e}}$ as a query to retrieve k nearest neighbor word vectors in the ground truth semantic space $\boldsymbol{E}$ using cross-domain similarity local scaling (CSLS) similarity measure [17]. The feature vector for each fMRI sample is obtained using 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 by adding information from the neural activity to the model besides the context. Operating at this particular layer allows the decoder to decide to what extend the added information is used in generating 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 input into BART as before and the hidden states of the mask token $\boldsymbol{h}_{mask}$ in the last hidden layer of encoder is updated using the following equation:

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

where $\boldsymbol{f}_i$ is the feature vector corresponding to the fMRI image.

In general, our decoding method is quite straightforward. We do not need to fine-tune BART but use it to derive word vectors as well as perform text generation. The two steps are decoupled and can be optimized independently.

4 Experiments

Dataset The brain imaging data is from [8], 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 [8] to reduce the dimension of fMRI data. To construct the context for our two tasks, we use the 6 sentences in the presentation scripts in the fMRI experiment for each word. For mask-filling task, we replace the target word in its corresponding sentences with a mask token (see Input in Figure 1 (a)). For text generation task, we replace the target word and all the following text with a single mask token (see Input in Figure 1 (b)). For each participant, we create two datasets by full joining the brain imaging dataset and the context dataset using stimuli as keys respectively. Therefore, we obtain two datasets, each with 1080 samples, for each participant. To choose a good model, we split each dataset into the same 18 folds, with 10 different words and its corresponding sentences and fMRI vectors in each fold to allow cross-validation.

Models We adopt BART without any extra information fusion and fine-tuning as our baseline[3]. In our neural decoding method, we build our model on the same BART as the baseline and we do not fine-tune the model. The hyperparameter k is tuned to 5. In the text generation task, we use greedy decoding. All the experiments are done under 18-fold cross-validation setting, repeatedly using 16 folds for training, 1 fold for validation and 1 fold for test.

[1] For each word, 6 sentences containing that word are fed to BART and the corresponding hidden states of the word in the last hidden layer of the encoder are collected and then averaged into a word vector.

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

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

Evaluation metrics For the mask-filling task, we adopt top-1 accuracy as an evaluation metric follows [11]. 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.

Mask-filling results 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 across participants are 17.81% and 18.52% respectively, all surpassing the baseline. Furthermore, for all participants, the improvements over the baseline are statistically significant under paired t-test with p-value < 0.1. These results demonstrate the feasibility to directly predict a word in a large vocabulary from a fMRI image, which is a departure from just discriminating two words. Moreover, they indicate that it may be possible to directly generate a sequence of words from fMRI data.

Text generation results 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 show that the text generation task is much harder than the mask-filling task. Despite the difficulty of the task, our model is able to achieve 7.95% accuracy on average across all participants, significantly surpassing the baseline by 1.47% absolute improvement. To be more specific, the performance of our model on each of the 15 participants is 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 any unrealistic assumptions. Essentially, the pre-trained encoder-decoder model is a conditional language model and can be viewed as a universal knowledge base. Intuitively, we are decoding text conditioning on personalized neural activities under the guidance of a universal knowledge base. This setting is in fact quite similar to the way we understand other's thoughts in daily communication. 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

5 Conclusions and future work

In this paper, we have proposed two neural decoding tasks, one is decoding a word and the other is decoding a sequence of words given an fMRI image and a specific context. On the one hand, these two tasks can help understanding to what extend we can currently decode language from neural activities. On the other hand, they can promote the development of practical non-invasive neural decoders that can translate brain signals into coherent text. Furthermore, we have proposed a general approach that can address the two tasks using one single model, which is to fuse the feature vector derived from the fMRI image to the powerful pre-trained encoder-decoder model to predict a word or generate a piece of text. Our method is simple and effective, demonstrating that it is feasible to generate text related to the stimuli from an fMRI image and a starting text. Last but not least, the pre-trained encoder-decoder model is essentially a conditional language model and can be viewed as a universal knowledge base while the fMRI data represents individual thoughts. Decoding personalized thoughts based on universal knowledge is in fact quite similar to the way people understand other's thoughts.

In the future, we are going to explore the feasibility to decode a sequence of words from a sequence of fMRI images by using state-of-the-art pre-trained language models. In this way, step by step, hopefully we could build a more powerful neural decoder to generate text, paving the way to create future practical non-invasive BCIs that may one day come into the public and help people with communication disorders to express their thoughts.

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] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[16] 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.

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