Decoding fMRI Data into Captions using Prefix Language Modeling

Vyacheslav Shen (shen9910@kaist.ac.kr)
School of Electrical Engineering, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, 34141 South Korea

Kassymzhomart Kunanbayev (kkassymzhomart@kaist.ac.kr)
School of Electrical Engineering, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, 34141 South Korea

Dae-Shik Kim (daeshik@kaist.ac.kr)
School of Electrical Engineering, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, 34141 South Korea

Abstract

With the advancements in Large Language and Latent Diffusion models, brain decoding has achieved remarkable results in recent years. The works on the NSD dataset, with stimuli images from the COCO dataset, leverage the embeddings from the CLIP model for image reconstruction and GIT for captioning. However, the current captioning approach introduces the challenge of potential data contamination given that the GIT model was trained on the COCO dataset. In this work, we present an alternative method for decoding brain signals into image captions by predicting a DINOv2 model's embedding of an image from the corresponding fMRI signal and then providing it as the prefix to the GPT-2 language model. Additionally, instead of commonly used Linear Regression, we explore 3D Convolutional Neural Network mapping of fMRI signals to image embedding space for better accounting positional information of voxels.

The code of our work is available by the following link: `https://github.com/slavaheroes/ccn24-brain-captioning`

Keywords: brain decoding; brain captioning; fMRI;

Introduction

It has been a common practice for brain decoding studies on the Natural Scenes Dataset (NSD) (Allen et al., 2021) to predict the embeddings of multimodal vision-language models like CLIP (Radford et al., 2021), GIT (Wang et al., 2022) from brain activations to use these predictions for further image and caption generation. Current SOTA brain captioning works like (Ferrante, Ozcelik, Boccato, VanRullen, & Toschi, 2023) (Scotti et al., 2024) leverage the GIT model to generate captions from fMRI signal.

However, this approach faces certain challenges that require attention. Primarily, the images in NSD come from COCO (Lin et al., 2014) which was used in training of GIT (Wang et al., 2022). According to the traditional train/test split of the NSD dataset (Takagi & Nishimoto, 2023) (Ozcelik & VanRullen, 2023), all images in the test set are present in the training split of the COCO, raising the concern of possible data contamination. It is important to ensure that the test set is completely independent to assess the ability to generalize unseen data fairly.

Additionally, in other brain decoding works including (Mai & Zhang, 2023) (Ozcelik & VanRullen, 2023)(Takagi & Nishimoto, 2023), fMRI voxels for a particular image are linearized using an ROI mask, followed by the application of Ridge Regression to map the fMRI voxels to the model embeddings. However, activations from regions not present in the ROI mask could be overlooked, and positional information of these voxels, which facilitate brain decoding, needs to be included.

To address the mentioned challenges, our study tests a new method utilizing DINOv2 (Oquab et al., 2023) embeddings and the GPT-2 language model (Radford et al., 2019). The main idea is to predict the DINOv2 embedding directly from its corresponding fMRI data using simple 3D-Convolutional ResNets (He, Zhang, Ren, & Sun, 2016), and feed them as a prefix to the language model to generate captions as proposed in (Mokady, Hertz, & Bermano, 2021).

Methodology

Dataset preprocessing

We follow the traditional NSD train/test split and data preprocessing of 4 subjects (sub1, sub2, sub5, sub7) obtained from GLM (*betas_fithrf_GLMdenoise_RR*) as it was utilized in Ozcelik and VanRullen (2023).

Given that, fMRI data represents a 4D array ($time \times W \times D \times H$), for Ridge Regression mapping from brain activity to the DINOv2 embedding space. We applied z-normalization for the linearized fMRI voxels[1] extracted from the NSDGeneral ROI mask, whereas for CNNs, a 3D input[2] at each time point was scaled between -1 and 1.

Brain Captioning

The scheme of our approach is presented in Figure 1. It has two parts: a brain and a captioning module. Both modules are trained separately.

The brain module is used to map fMRI activations into a DINOv2 embedding. DINOv2 was chosen due to its rich and robust visual features achieved by self-supervised learning compared to CLIP. It is trained using Mean Squared Error (MSE) loss where the ground truth label is the image embedding from the DINOv2-g model. We tried 3 mapping networks: Ridge Regression and two variations of 3D ResNet with 18 layers referred to as Shallow and Wide CNNs. Wide CNN has more feature planes than Shallow CNN.

The captioning module consists of a light transformer (Vaswani et al., 2017) and a language model. During training, the transformer converts the DINOv2 image embedding into prefix tokens that have the same dimensions as a word embedding. These prefix tokens are then used as inputs for the language model. The training objective is to predict caption tokens conditioned on the prefix autoregressively (Mokady et al., 2021).

At inference time, the brain module predicts the embedding of the seen image from the fMRI signal, which is then passed to the captioning module. While decoding from the language model, the beam search was employed to select the next token.

The training hyperparameters and design choices can be found in the provided GitHub link.

We compare the effectiveness of our approach with Ferrante et al. (2023), UniBrain (Mai & Zhang, 2023), MindEye-2 (Scotti et al., 2024).

Results

The best results obtained using Wide CNN and comparison with existing works are presented in Table 1. Using the eval-

[1] For sub1, input is a 1D vector of shape 15724

[2] For sub1, input has (81, 104, 83) shape

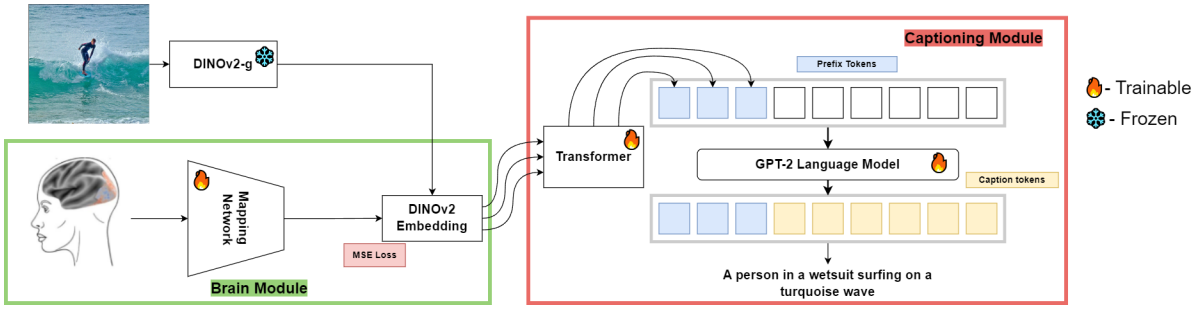

Figure 1: The scheme of our method. GPT-2 base model was used as the language model, while fMRI-DINOv2 embedding mapping

Table 1: Evaluation of captions from fMRI for sub1

	fMRI vs COCO			fMRI vs Image captions		
Metrics	UniBrain	MindEye-2	Ours	Ferrante et al. (2023)	MindEye-2	Ours
METEOR	0.170	0.248	**0.271**	0.305	0.344	**0.457**
ROUGE-1	0.247	**0.353**	0.346	-	0.455	**0.513**
ROUGE-L	0.225	**0.326**	0.316	-	0.427	**0.491**
Sentence	-	**47.9%**	39.7%	44.7%	**52.3%**	49.9%
CLIP-B	-	**73.7%**	68.7%	70.5%	**75.4%**	74.6%
CLIP-L	**86.1%**	63.8%	58.5%	-	67.1%	**67.4%**

uation metrics in Scotti et al. (2024), we assess our approach by comparing predicted captions from fMRI with the original COCO. Since Ferrante et al. (2023) and Scotti et al. (2024) evaluate the fMRI captions with the captioning model-generated captions, we also evaluate fMRI captions with the captions generated from the true image embeddings.

Compared with the original COCO captions, our approach outperformed previous works only in the METEOR metric and came close to the MindEye-2 in ROUGE metrics. However, the right half of Table 1 shows that our method is better in 4 metrics out of 6, meaning that our captions generated from fMRI are closer to the captions generated from the true image embeddings.

Table 2: Evaluation of Mapping Networks from fMRI to DINOv2 embedding space averaged for 4 subjects

	fMRI vs COCO		
Metrics	Ridge	Shallow CNN	Wide CNN
METEOR	0.263	0.267	**0.273**
ROUGE-1	0.331	0.340	**0.346**
ROUGE-L	0.300	0.312	**0.317**
Sentence	34.92%	36.71%	**38.91%**
CLIP-B	66.73%	67.22%	**67.79%**
CLIP-L	55.72%	56.65%	**57.59%**

Ablation on Mapping Network

Table 2 shows the results of different mapping networks used in the brain module. The advantage of using CNNs, Wide CNN in particular, is shown by outperforming Ridge Regression mapping in all captioning metrics.

Conclusion & Future Work

Using the abundant visual features extracted from the DINOv2 vision model, we could achieve competitive results compared to previous works on brain captioning. Our pipeline uses networks that haven't been trained on the COCO dataset, thus minimizing the issue of data contamination to zero. Additionally, the substitution of Ridge Regression with convolutional neural networks, that take whole fMRI data at one time-point as input, improves performance. This implies that information outside of the ROI mask and positional information can be helpful in brain decoding.

We plan to extend our work by introducing image-generation models to reconstruct seen images and conducting an analysis of the interpretability of the results. We believe that providing fMRI data as the prefix for the language models (Ye et al., 2023) can be promising in creating brain decoding frameworks designed for complex tasks like visual-question answering. Furthermore, we will continue to search for more efficient models to map fMRI voxels to the image-embedding space.

Acknowledgments

This work was supported by the Engineering Research Center of Excellence (ERC) Program supported by National Research Foundation (NRF), Korean Ministry of Science & ICT (MSIT) (Grant No. NRF-2017R1A5A101470823).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., ... Kay, K. N. (2021). A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*. Retrieved from https://api.semanticscholar.org/CorpusID:232059810

Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., & Toschi, N. (2023). Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Mai, W., & Zhang, Z. (2023). Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*.

Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... others (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, *13*(1), 15666.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., ... others (2024). Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*.

Takagi, Y., & Nishimoto, S. (2023, June). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 14453-14463).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., ... Wang, L. (2022). Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Ye, Z., Ai, Q., Liu, Y., Zhang, M., Lioma, C., & Ruotsalo, T. (2023). Language generation from human brain activities. *arXiv preprint arXiv:2311.09889*.