Reconstructing Natural Images from Brain Activity

Andrea Bosia
School of Computer Science Faculty of Engineering Supervisor Jinglei Lv

1. Introduction

Despite the tremendous advancement achieved, the brain remains the most mysterious organ of the human body. The advent of NN unlocked unprecedented new possibilities for understanding its functioning. Specifically, the research community has recently been focusing on the task of recreating visual stimuli form brain signal, particularly fMRI data. Such a challenge brings along the possibility of achieving a better understanding of how specific brain areas activate and thereafter can be mapped into the different features of a visual stimuli. This, in turns, allows for innumerable implementations that range from medical application to recreational ones.

A vast body of research has been produced in the last 5 years or so achieving gradual but notable improvement in term of the fidelity of the reconstructed input. [2][4][5] This work builds upon the successes of previous results and specifically leverages a two stages pipeline where the main module is a Diffusion Model conditioned via CLIP-Text. A specific focus is given to understanding how different latent representations condition the output fidelity.

2. Objective

The main focus is given to produce a pipeline capable of reconstructing the visual stimuli observed by a subject leveraging the brain signals captured by fMRI data. Various metrics such as CLIP Similarity, SSIM and PCC are used to evaluate the semantic and structural similarity between the ground truth image and the reconstructed one.

Alongside this primary objective great importance is also given to the neurobiological interpretability of the produced model. In other words, the goal here is to understand which voxels effect the most which features of the reconstructed output.

3. Dataset

The Natural Scene Dataset (NSD) represents, as of today, the biggest collection of (Image, fMRI signal) pairs available bridging cognitive neuroscience and artificial intelligence. Moreover, the roughly 10000 images shown to the 8 subjects are taken from the COCO image dataset meaning that annotations are available. [1] Tight ROIs are applied to the brain to only focus on areas with noticeable activation (which corresponds in this case to the visual cortex and specifically V1, V2, V3, hV4, VO, PHC, MT, MST, LO, IPS). Individual voxels are averaged across trials and represented as time series, describing the activation level over time.

4. Pipeline

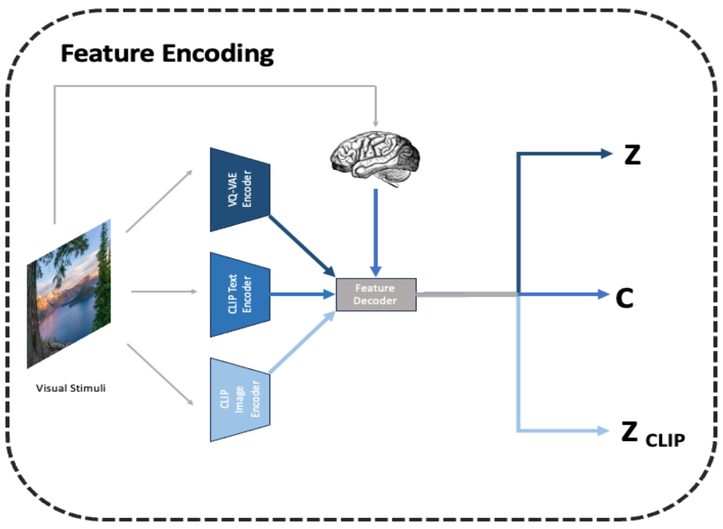

Figure 1. Three pre-trained encoder used to map the fMRI signal into three different feature spaces. The mapping is performed by a feature decoder that consists of three linear regression models.

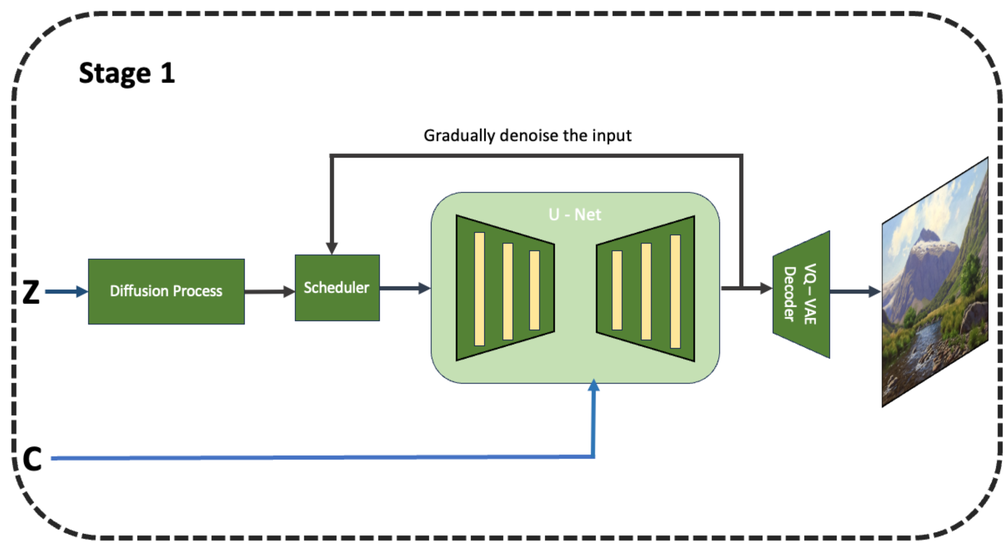

Figure 2. A Diffusion Model is used for inference. Noise is added to the latent Z based on a scheduler. The U-net, conditioned with the embedding C, predicts the noise. Z is denoised over several iterations and finally the output is decoded back to an image.

The first step in the pipeline (Figure 1), consists of mapping the fMRI data into three feature spaces. This is done using three L2-Regularised Linear Regression models. Such models map the fMRI data to three representations that specifically capture semantic (Z and C) and structural (Z_{CLIP}) details of the image stimuli.

Stage 1 (Figure 2) consist of the inference step of a Diffusion Model (which, depending on the computing resources available, can also be fine tuned). The input is represented by the mapping of the fMRI data into the VQ-VAE latent space (Z). Noise is added to such input throughout the diffusion process (controlled by the scheduler) and is then processed by a U-net which is in turn conditioned with the CLIP text representation (decoded form the fMRI) and the encoded noise level information. At each pass through the U-net noise is predicted and removed from the input Ultimately the Diffusion Model outputs the initial version of the reconstructed image.

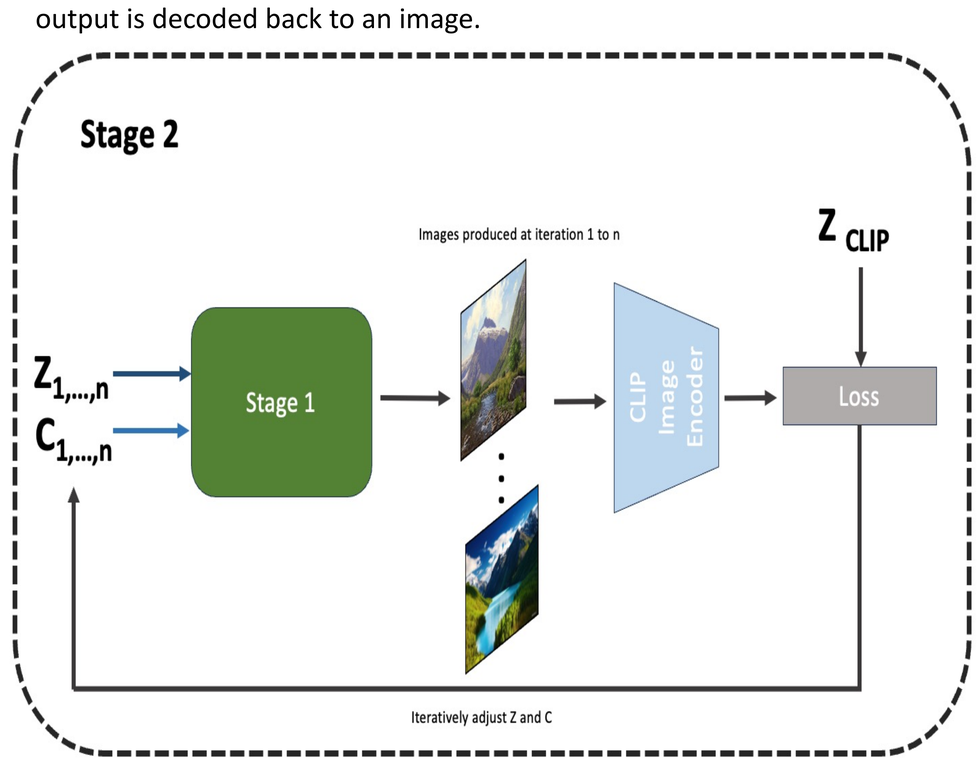

Figure 2. Z and C are fed into Stage 1, the output is encoded and compared with Z_{CLIP} from the first step of the pipeline. A loss is computed and used to iteratively adjust Z and C.

In Stage 2 (Figure 3), the modules from Stage 1 are plugged in into an iterative cycle. The image outputted is encoded using the CLIP Image Encoder. The obtained latent representation is compared with the original latent representation (Z_{CLIP}) derived from the fMRI during the very first step of the pipeline. The obtained Loss ($L_{structural}$) is backpropagated to adjust Z and C, whereas the weights of Stage 1 are kept frozen.

5. Results

The described pipeline successfully retains semantic and structural information in the reconstructed image. However, as it can be seen in Figure 3, there are instances in which the model struggles to reconstruct the GT visual stimuli. This is due to anatomical differences among subjects. Moreover, from Table 1, it can be noticed how both LVC and HVC are crucial to obtain the best possible reconstructive performances. Specifically, LVC appears to perceive more structural details, whereas HVC semantic ones, which is in line with neuroscience results. Lastly, a series of ablation studies highlighted how all feature spaces are crucial to the effectiveness of the pipeline. Specifically, *c* affects the most the semantic accuracy of the output, whilst *z* and Z_{CLIP} are more important for the structural fidelity.

Figure 3. On the left most column is depicted the ground truth visual stimuli shown to the subjects. The other columns show the reconstructed image for different subjects. [1]

Visual cortex		Semantic similarity ↑	Structual similarity ↑	
LVC	HVC	CLIP Similarity	SSIM	PCC
✓		0.552	0.338	0.051
	✓	0.554	0.219	0.030
✓	✓	**0.765**	**0.354**	**0.278**

Table 1. The table summarises the effect of different brain areas (i.e. LVC, HVC) on the semantic and structural similarity of the output. [3]

Model	Variants	Semantic similarity ↑	Structual similarity ↑	
		CLIP Similarity	SSIM	PCC
MindDiffuser	w/o *c*	0.549	0.346	0.218
MindDiffuser	w/o *z*	0.616	0.292	0.066
MindDiffuser	w/o Z_{CLIP}	0.597	0.253	0.183
MindDiffuser	completed	**0.765**	**0.354**	**0.278**

Table 2. The table reports a series of ablation studies where one by one a feature space (*c, z,* Z_{CLIP}) is discarded to evaluate its effect on the quality of the final output. [3]

[1] Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence.
[2] Chen, Z., Qing, J., Xiang, T., Yue, W. L., & Zhou, J. H. (2022, November). Seeing beyond the brain: Masked modeling conditioned diffusion model for human vision decoding. *arXiv*.
[3] Lu, Y., Du, C., Zhou, Q., Wang, D., & He, H. (2023). MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion.
[4] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv*.
[5] Takagi, Y., & Nishimoto, S. (2022). High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*.