High-resolution image reconstruction with latent diffusion models from human brain activity Summary

# Abstract

Reconstructing visual experiences from human brain activity offers a unique way to understand how the brain represents the world, and to interpret the connection between computer vision models and our visual system. While deep generative models have recently been employed for this task, reconstructing realistic images with high semantic fidelity is still a challenging problem. Here, we propose a new method based on a diffusion model (DM) to reconstruct images from human brain activity obtained via functional magnetic resonance imaging (fMRI). More specifically, we rely on a latent diffusion model (LDM) termed Stable Diffusion. This model reduces the computational cost of DMs, while preserving their high generative performance. We also characterize the inner mechanisms of the LDM by studying how its different components (such as the latent vector of image Z, conditioning inputs C, and different elements of the denoising U-Net) relate to distinct brain functions. We show that our proposed method can reconstruct high-resolution images with high fidelity in straight- \* Corresponding author forward fashion, without the need for any additional training and fine-tuning of complex deep-learning models. We also provide a quantitative interpretation of different LDM components from a neuroscientific perspective. Overall, our study proposes a promising method for reconstructing images from human brain activity and provides a new framework for understanding DMs. Please check out our webpage at https://sites.google.com/view/stablediffusion-withbrain/. 1

Notes:

* The use of additional semantic content from images to improve performance. This requires training new models from scratch with the given input system data
* Natural Scenes Dataset (NSD) for this project [1]. Please visit the NSD website for more details 1. Briefly, NSD provides data acquired from a 7-Tesla fMRI scanner over 30–40 sessions during which each subject viewed three repetitions of 10,000 images.
* Could we use pattern data between EEG and MRI results to translate these MRI scans into equivalent EEG patterns?
* Gaussian noise?
* Pearson correlation coefficient
* CLIP
* Ronneberger et al.

EEG Embedding Process

# Pre-Training

* masked signal modeling (MSM) -> using a large image data set to pre train the image generation model to re construct masked portions of images from the given image segment.
* high mask ratio for visual signals, low for natural language
* UNet with attention model (Whatttttttttt)
* Markov chain transitions

# Proposed Method

3 main components:

* masked signal pre-training for an effective and robust EEG encoder
  + Leverage masked signal modeling with noisy EEG data to train EEG encoder to extract contextual knowledge.
  + Given the high temporal resolution of EEG signals, we first divide them into tokens in the time domain, and randomly mask a certain percentage of tokens.
  + Subsequently, these tokens will be transformed into embeddings by using a one-dimensional convolutional layer
  + MAE [18] to predict the missing tokens based on contextual cues from the surrounding tokens
* fine-tuning with limited EEG-image pairs with pre-trained Stable Diffusion
* aligning the EEG, text, and image spaces using CLIP encoders
* Result provides EEG encoder that provides conditional features to stable diffusion via a cross attention mechanism.
* Want to close the distance between EEG and CLIP embeddings
* Stable diffusion fine tuning
  + Use off the shelf text to image model
  + Use CLIP to align the EEG, text, and image space

# Data Information

* 120,000 EEG samples from 400 subjects
* Channels range from 30 to 128.
* MOABB platform for pre-training made by BCI, have publicy available EEG data sets: https://moabb.neurotechx.com/docs/index.html
* These data contain a wide variety of EEG data, including tasks such as looking at an object, motor imagery, and watching videos.

A table with a check mark and numbers

Description automatically generated

* To facilitate pre-training, we have uniformly padded all the data to 128 channels by filling missing channels with replicated values.
* , every 4 adjacent time steps are grouped into a token and each token is transformed into a 1024-dimensional embedding through a projection layer for subsequent masked signal modeling.
* re reconstruction is performed on the entire set of 128 channels as a whole
* a collection of EEG recordings obtained from 6 subjects while they were shown 2000 images belonging to 40 different categories of objects from the ImageNet dataset
* Each category consisted of 50 images, and each image was presented for 0.5 seconds, followed by a 10-second pause for every 50 images
* 128-channel Brainvision EEG system, resulting in a total of 12000 128-channel EEG sequences
* version 1.5 of Stable Diffusion for image generation
* EEG signals are filtered within the frequency range of 5-95 Hz before pretraining
* signals are truncated to a common length of 512
* encoder is pre-trained for 500 epochs and finetuned with Stable Diffusion for another 300