2023 IEEE EMBS Special Topic Conference on
Data Science and Engineering in Healthcare, Medicine and Biology
December 7-9, 2023. Malta

2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology | 979-8-3503-8338-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF58974.2023.10404307

Linking Brain Signals to Visual Concepts: CLIP based knowledge transfer for EEG Decoding and visual stimuli reconstruction

Matteo Ferrante[1], Tommaso Boccato[1], Stefano Bargione[1] and Nicola Toschi[1,2]

***Abstract*— Decoding visual representations from human brain activity has emerged as a thriving research domain, particularly in the context of brain-computer interfaces. This study introduces a novel approach using a convolutional neural network (CNN) to classify images from the ImageNet dataset, leveraging electroencephalography (EEG) recordings. We collected EEG data from 6 subjects, each viewing 50 images across 40 distinct semantic categories. These EEG signals were transformed into spectrograms, serving as the input for training our CNN. A unique aspect of our model is the incorporation of knowledge distillation from a pre-trained image classification teacher network. This approach enabled our model to achieve a top-5 accuracy of 80%, notably surpassing a plain CNN baseline. Furthermore, we integrated an image reconstruction pipeline founded on pre-trained latent diffusion models. This innovative concatenation not only decodes images from brain activity but also provides a plausible reconstruction, facilitating rapid and subject-specific feedback experiments. Our work thus represents a significant advancement in the field, bridging the gap between neural signals and visual perception.**

I. Introduction

Electroencephalography (EEG) has emerged as a promising tool for decoding visual representations in the human brain. Recent advances have enabled decoding of complex visual stimuli from EEG signals, including categories from large image datasets like Imagenet [4], [1]. Convolutional and recurrent neural networks can classify EEG signals into image categories with promising accuracy. However, most studies focus on multisubject models, averaging EEG signals across individuals. This risks missing subject-specific neural representations. Single-subject models that tap into individual variability in visual processing may enable more detailed decoding, also providing a new layer of privacy for neural data, given that each model is specific for one subject and cannot be used in inference on other subjects' data. Reconstructing visual stimuli from EEG signals that have been elicited by them remains an active challenge. The low spatial resolution of EEG creates an ill-posed problem for reconstructing fine visual details, and current image reconstructions are limited to coarse features like shapes, colors, and textures. It also limits fine-grained decoding of visual features and image reconstructions. Rather than pixel-level recreations, semantic image reconstructions may be more feasible with EEG. Nevertheless, could be interesting to reconstruct in real-time images from EEG data and show this reconstruction to the subject during the experiment, enabling a biofeedback loop [2], so the subject can learn how to focus on images to improve classifier performances. Moreover, using EEG data will result in non invasive and portable measurements that facilitate the integration of these technologies into a brain-computer interface. Here, we propose a step forward in this direction with a pipeline (see Fig 1) that could train a single subject model during a short experiment (around 20 minutes of data collection, followed by a few minutes of training) and then perform quasi-realtime brain decoding (up to a couple of seconds to generate each image depending on specific hardware).

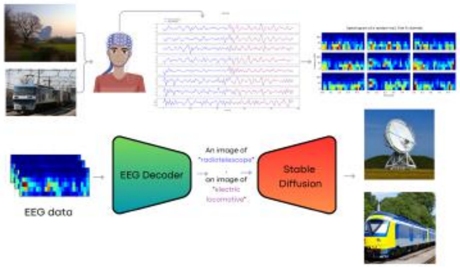
EEG data
EEG Decoder
An image of "radiotelescope"
an image of "electric locomotive"
Stable Diffusion

Fig. 1. Our pipeline can be described as follows: First, we record EEG data while the subject is viewing natural images. This data is then preprocessed and converted into spectrograms, which serve as the input for our neural network. Our EEG decoder is trained using a knowledge distillation method based on the CLIP model. The outputs from the EEG decoder, which are predictions of the image that elicited the EEG data, are then combined with an image generation pipeline. This end-to-end approach allows us to create images from the neural activity data captured by the EEG.

II. Material and Methods

We used EEG recordings from [4]. Data are acquired from 6 subjects as they viewed images from 40 ImageNet classes, with 50 images per class. Images were presented sequentially in 25-second batches, followed by 10-second pauses. This resulted in 2,000 total images over 1,400 seconds (23 minutes 20 seconds) of recording. Subjects participated in 4 recording sessions of 350 seconds each. 128-channel EEG data were acquired at 1KHz, yielding 11,466 sequences after excluding poor-quality recordings. This experimental design allowed us to examine EEG signals in response to a wide range of visual stimuli from ImageNet. By collecting multi-channel EEG during prolonged viewing, we obtained rich training data for decoding models. Raw EEG signals were preprocessed before using them to train our decoding models, using a notch filter at 49-51 Hz to remove power line noise and a Butterworth bandpass filter between 14 and 70 Hz to isolate frequency bands relevant to visual attention and object recognition and standardized the signals across channels. To generate inputs for our neural network, we

[1] University of Rome, Tor Vergata University of Twente, 7500 AE Enschede, The Netherlands matteo.ferrante@uniroma2.it

[2] Martinos Center For Biomedical Imaging, MGH and Harvard Medical School (USA)

split filtered EEG signals into 40 ms windows and computed spectrograms. This transformed each trial into a 128-channel image showing spectral power across time and frequency. In total, we obtained 2000 such 128-channel EEG spectrogram images for each subject to use for training and evaluating our convolutional neural network for EEG decoding. This multi-channel spectral representation captures both spatial and temporal dynamics in the EEG, enabling our model to learn features for classifying visual stimuli.

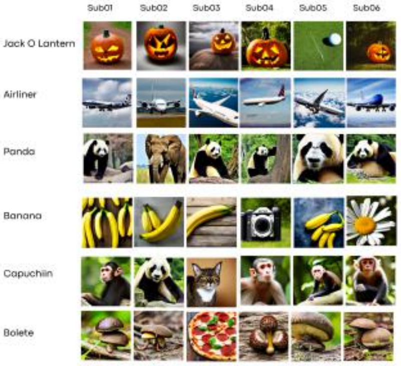

Fig. 2. Reconstructed images for a qualitative evaluation. On the left the target classes are presented and each column show result from a single subject

We implement a convolutional neural network (CNN) with residual connections to classify EEG spectrograms. The model has a front-end of convolutional layers with an increasing number of filters to extract spatial and temporal features. This is followed by global average pooling and fully-connected layers for classification. To train the CNN, we use a knowledge distillation approach [3]. First, we pretrain an image classifier using CLIP [5] features to predict the stimulus classes, achieving 99% accuracy. This provides ”soft targets” for teaching our EEG model. During training, we input EEG spectrograms to the CNN and CLIP image features to the teacher classifier. The CNN is trained to match the class probability distributions from the teacher. This distillation stabilizes training and improves model performance compared to training directly on class labels. At inference time, only the EEG-based CNN is used to predict classes from new spectrograms. By distilling knowledge from an image model, our CNN learns robust representations to decode visual stimuli from brain signals alone.

After training our EEG decoding model, we are able to predict ImageNet classes from new EEG spectrograms. To validate these predictions and reconstruct images that in principle could elicit the same neural activity, we leverage the Stable Diffusion generative model [6]. For each EEG prediction, we create a text prompt like ”an image of a {predicted class}”. We input this prompt along with random noise vectors to Stable Diffusion to generate new images matching the predicted class. This allows us to reconstruct visual stimuli solely from brain activity patterns. The EEG decoder predicts the class, while Stable Diffusion produces a representative image.

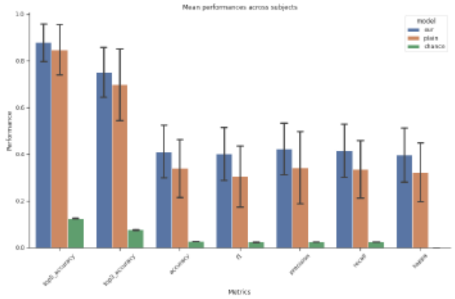

Fig. 3. Resuls for EEG decoder. **Our** is the CLIP based approach, **plain** is a vanilla CNN with the same architecture trained for classification and **chance** serves as comparison with chance level. Bars are average across subjects and error bars are standard deviations.

III. Results

We assess our model’s performance using a variety of metrics, including top-5, top-3, top-1 accuracy, F1 score, and kappa score (normalized for chance). As illustrated in Figure 3, our knowledge distillation CNN consistently surpasses both a plain CNN baseline and a random classifier. In particular, our model attains a top-3 accuracy of 77%, signifying that it can correctly identify the image from a mere 0.5-second recording of EEG activity in over three out of four instances when proposing three classes. On a qualitative level (refer to Fig. 2), we visualize the model’s predictions across six classes from the Brain2Image paper, with one column dedicated to each subject. Our model demonstrates a reliable ability to predict the correct class from EEG signals for the majority of subjects and categories. However, some errors do occur in the initial predictions; for instance, ”bolete” is occasionally confused with ”pizza,” or ”banana” is mistaken for ”margherita.”

IV. Conclusions

We proposed a novel CLIP-based knowledge distillation CNN can effectively decode semantic categories from EEG spectrograms and subsequent image synthesis from brain activity leveraging latent diffusion models.

References

[1] Y. Bai, X. Wang, Y. pei Cao, Y. Ge, C. Yuan, and Y. Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals, 2023.

[2] S. Enriquez-Geppert, R. J. Huster, and C. S. Herrmann. Eeg-neurofeedback as a tool to modulate cognition and behavior: A review tutorial. *Frontiers in Human Neuroscience*, 11, 2017.

[3] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015.

[4] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM ’17, page 1809–1817, New York, NY, USA, 2017. Association for Computing Machinery.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022.

This work was supported by NEXTGENERATIONEU (NGEU) and funded by the Italian Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006)– A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); by the MUR-PNRR M4C2I1.3 PE6 project PE00000019 Heal Italia; by the NATIONAL CENTRE FOR HPC, BIG DATA AND QUANTUM COMPUTING, within the spoke ”Multiscale Modeling and Engineering Applications”; the EXPERIENCE project (European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 101017727); the CROSSBRAIN project (European Union’s European Innovation Council under grant agreement No. 101070908).