Single Subject Human EEG decoding into natural images with CLIP-based knowledge distillation and latent diffusion models

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

***Abstract*— Decoding visual representations from human brain activity is an active area of research with applications for brain-computer interfaces. In this work, we developed a convolutional neural network to classify images from the Imagenet dataset based on electroencephalography (EEG) recordings. EEG data were recorded from 6 subjects viewing 50 images from 40 semantic categories. Spectrograms were generated from the EEG signals and used to train a CNN model with knowledge distillation from an image classification teacher network. Our model achieved top-5 accuracy of 80% in decoding EEG signals to Imagenet categories, outperforming a plain CNN baseline. On top of that, we concatenated an image reconstruction pipeline based on pre-trained latent diffusion models that can enable fast and subject-specific feedback experiments, decoding images from brain activity and then showing a plausible reconstruction to the subject.**

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 [8], [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 images from EEG signals remains an active challenge. The low spatial resolution of EEG creates an ill-posed problem for reconstructing fine visual details. Current image reconstructions are limited to coarse features like shapes, colors, and textures. Rather than pixel-level recreations, semantic image reconstructions may be more feasible with EEG. Approaches like generative adversarial networks (GANs) [5] show promise for creating semantically meaningful reconstructions from EEG. Overall, the poor spatial resolution of EEG limits fine-grained decoding of visual features and image reconstructions. EEG provides a useful macro-level window into visual processing in the brain. Multimodal approaches that combine EEG with imaging modalities like fMRI could help overcome the limitations of EEG alone. While on fMRI, given the increased spatial resolution, is possible to reconstruct images with low-level agreement [3], [6], [7], [11]. Nevertheless, could be interesting reconstructing in real-time images from EEG data and show this reconstruction to the subject during the experiment, enabling a feedback loop [2], so the subject can learn how to focus on images to improve classifier performances. 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).

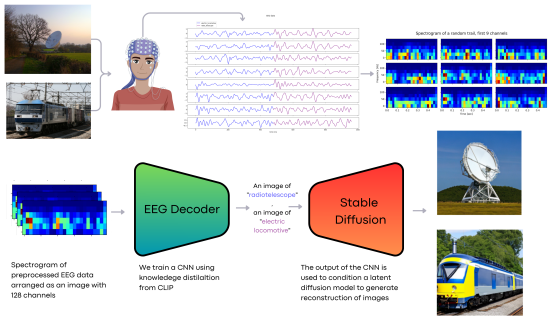

Fig. 1. Our pipeline can be described as follows: First, we record EEG data while the subject is viewing natural images. This data is then preprocessed and converted into spectrograms. These spectrograms serve as the input for our neural network. Our EEG decoder is trained using a knowledge distillation method based on the CLIP model. The outputs from the EEG decoder, which are predictions of the image that induced 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 [5]. 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, 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.

We preprocessed the raw EEG signals before using them to train our decoding models. First, we applied a notch

[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)

filter at 49-51 Hz to remove power line noise. Next, we used a bandpass filter between 14 and 70 Hz to isolate frequency bands relevant for visual attention and object recognition. We then standardized the signals across channels to normalize the scale. To generate inputs for our neural network, we divided the 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 captured both spatial and temporal dynamics in the EEG, enabling our model to learn robust features for classifying visual stimuli.

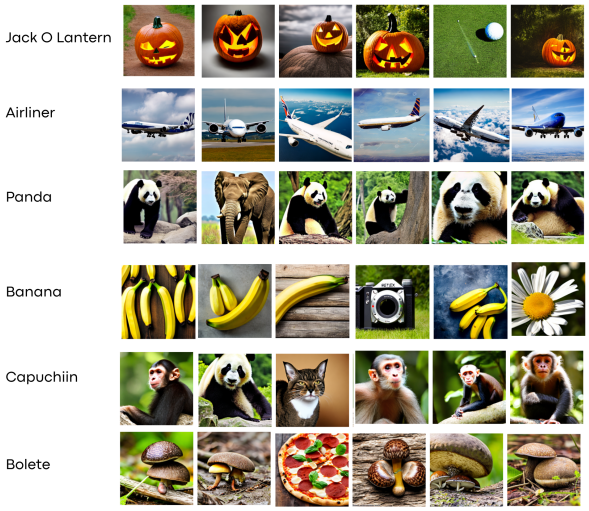

Fig. 2. Reconstructed images for a qualitative evaluation. On the left the target classes are presented and each column show result from a single subject

We implemented a convolutional neural network (CNN) with residual connections to classify EEG spectrograms. The model had a front-end of convolutional layers with increasing number of filters to extract spatial and temporal features. This was followed by global average pooling and fully-connected layers for classification. To train the CNN, we used a knowledge distillation approach [4]. First, we pretrained an image classifier using CLIP [9] features to predict the stimulus classes, achieving 99% accuracy. This provided "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 was trained to match the class probability distributions from the teacher. This distillation stabilized training and improved 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 learned robust representations to decode visual stimuli from brain signals alone.

After training our EEG decoding model, we were able

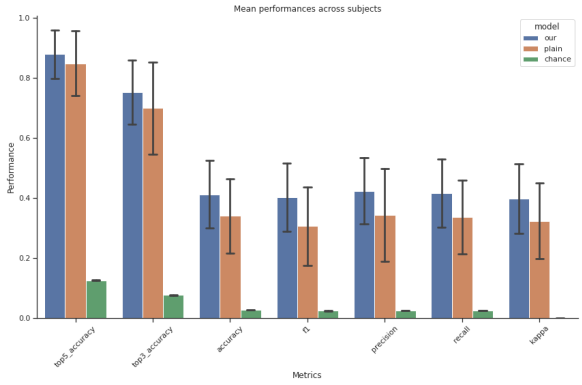

Fig. 3. Resuls for EEG decoder. **Our** is the CLIP based approach, **plain** is a vanilla CNN with the same architecture trained for classification and **chance** serves as comparison with chance level. Bars are average across subjects and erorr bars are standard deviations.

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 leveraged the Stable Diffusion generative model [10] For each EEG prediction, we created 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 allowed us to reconstruct visual stimuli solely from brain activity patterns. The EEG decoder predicted the class, while Stable Diffusion produced a representative image. By generating images from EEG, we could visualize and validate what our model learned to decode from neural signals.

III. Results

We evaluated our model using metrics like top-5, top-3, top-1 accuracy, F1 score, and kappa score (normalized for chance). As shown in Figure 3, our knowledge distillation CNN systematically outperformed both a plain CNN baseline and a random classifier.

Specifically, our model achieved top-3 accuracy of 77% which means that from 0.5 seconds of recording of EEG activity, can guess in more than 3 out of 4 cases the images proposing 3 classes.

Qualitatively (see Fig. 2) we visualized model predictions on 6 classes from the Brain2Image paper with one column per subject. Our model was able to reliably predict the correct class from EEG signals for most subjects and categories, although, sometimes there are some error in the first predictions, for example, "bolete" is confused with "pizza" or "banana" is confused with "margherita".

IV. Conclusions

In summary, our CLIP-based knowledge distillation CNN could effectively decode semantic categories from EEG spectrograms. Further work is needed to improve generalization across subjects and classes. But these initial results demonstrate the feasibility of EEG-based decoding of complex visual concepts.

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] M. Ferrante, F. Ozcelik, T. Boccato, R. VanRullen, and N. Toschi. Brain captioning: Decoding human brain activity into images and text, 2023.

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

[5] 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.

[6] F. Ozcelik, B. Choksi, M. Mozafari, L. Reddy, and R. VanRullen. Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs, Feb. 2022. arXiv:2202.12692 [cs, eess, q-bio].

[7] F. Ozcelik and R. VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion, 2023.

[8] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah. Generative adversarial networks conditioned by brain signals. pages 3430–3438, 10 2017.

[9] 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.

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

[11] Y. Takagi and S. Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, 2023.