Computers in Biology and Medicine 178 (2024) 108701

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models

Matteo Ferrante [a,*], Tommaso Boccato [a], Stefano Bargione [a], Nicola Toschi [a,b]

[a] Department of Biomedicine and Prevention, University of Rome Tor Vergata (IT), Italy
[b] Athinoula A. Martinos Center for Biomedical Imaging, MGH and Harvard Medical School (US), United States of America

ARTICLE INFO

Keywords:
EEG decoding
Brain decoding
Image reconstruction
BCI vision

ABSTRACT

Decoding visual representations from human brain activity has emerged as a thriving research domain, particularly in the context of brain–computer interfaces. Our study presents an innovative method that employs knowledge distillation to train an EEG classifier and reconstruct images from the ImageNet and THINGS-EEG 2 datasets using only electroencephalography (EEG) data from participants who have viewed the images themselves (i.e. "brain decoding"). We analyzed EEG recordings from 6 participants for the ImageNet dataset and 10 for the THINGS-EEG 2 dataset, exposed to images spanning unique semantic categories. These EEG readings were converted into spectrograms, which were then used to train a convolutional neural network (CNN), integrated with a knowledge distillation procedure based on a pre-trained Contrastive Language-Image Pre-Training (CLIP)-based image classification teacher network. This strategy allowed our model to attain a top-5 accuracy of 87%, significantly outperforming a standard CNN and various RNN-based benchmarks. Additionally, we incorporated an image reconstruction mechanism based on pre-trained latent diffusion models, which allowed us to generate an estimate of the images that had elicited EEG activity. Therefore, our architecture not only decodes images from neural activity but also offers a credible image reconstruction from EEG only, paving the way for, e.g., swift, individualized feedback experiments.

1. Introduction

Electroencephalography (EEG) is increasingly recognized as a valuable instrument for decoding visual representations within the human brain. The primary advantage of EEG lies in its non-invasive nature and its ability to provide real-time insights into human brain function via electrical activity recordings from the scalp. Despite its spatial resolution constraints, its unparalleled temporal resolution renders it ideal for real-time applications.

Recent technological advancements have facilitated the decoding of intricate visual stimuli from EEG signals, notably from expansive datasets such as ImageNet [1,2]. Both convolutional (CNN) and recurrent neural networks (RNN) have demonstrated efficacy in classifying EEG signals into distinct image categories with appreciable accuracy. The successful decoding of complex visual stimuli from EEG signals can pave the way for innovative neural prosthetics and biofeedback systems. Translating brain activity patterns into decoded image categories or reconstructions could potentially offer visually impaired individuals a semblance of artificial vision. Additionally, EEG decoding can revolutionize brain-centric image searches, communication platforms, and augmented reality interfaces. Real-time visualizations of decoded brain activity can also usher in novel neurofeedback paradigms, facilitating self-regulation of brain states through integrated EEG decoding and external visual feedback mechanisms [3].

However, a predominant focus in current research is on multi-subject models, which involve averaging EEG signals across multiple participants. This methodology may overlook the nuances of individual-specific neural representations. Models tailored to individual participants could offer a more granular decoding and introduce an added dimension of data privacy, as each model is uniquely calibrated for a specific individual, precluding its application to others. Also, in spite of recent progress, the task of reconstructing visual stimuli based on the EEG signals they elicit remains a formidable challenge. The inherent low spatial resolution of EEG poses difficulties in reconstructing detailed visual nuances. Presently, image reconstructions predominantly capture broader features, such as shapes, colors, and textures, thereby constraining the depth of visual feature decoding and image reconstructions. To overcome this obstacle, instead of attempting pixel-precise reproductions, a more pragmatic approach might be semantic image reconstructions. For example, approaches like generative adversarial

* Corresponding author.
E-mail address: matteo.ferrante@uniroma2.it (M. Ferrante).

https://doi.org/10.1016/j.compbiomed.2024.108701
Received 26 January 2024; Received in revised form 31 May 2024; Accepted 1 June 2024
Available online 7 June 2024

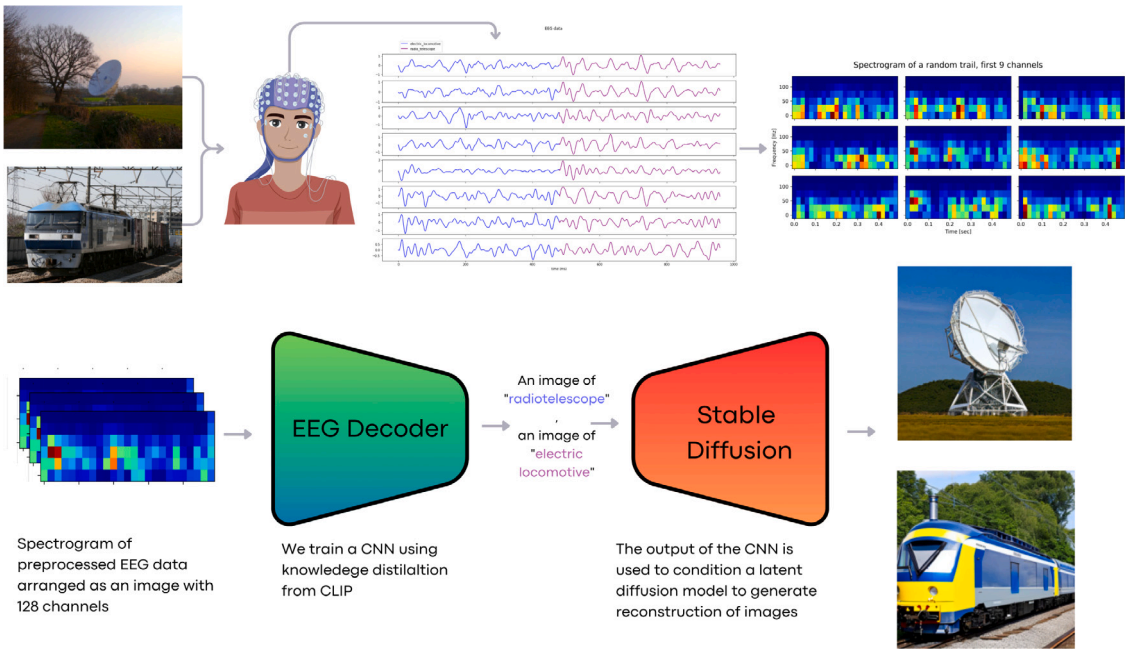

Fig. 1. Our pipeline can be described as follows: First, we record EEG data while the subject is viewing natural images. This data is then preprocessed and converted into spectrograms, which serve as the input for our neural network. Our EEG decoder is trained using a knowledge distillation method based 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 reconstruct images from the neural activity data captured by the EEG.

networks (GANs) [4] show promise for creating semantically meaningful reconstructions from EEG. 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. Using fMRI, the higher spatial resolution, is possible to reconstruct images with low-level agreement [5–8]. 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 [3], so the subject can learn how to focus on images to improve classifier performances. This research aims to improve existing methodologies for translating perceptual experiences from EEG patterns, with a focus on real-time applications. We present a methodology that advances this field, outlining a pipeline (as shown in Fig. 1) that facilitates the training of a single-subject model within a limited experimental timeframe, leading to near-real-time brain decoding.

The central innovation of this work lies in the proposed methodology for addressing the challenge at hand. Our goal is to achieve semantic decoding of visual content from electroencephalogram (EEG) activity, which is commonly very noisy and comes with reduced spatial resolution, limiting the chance of achieving fine-grained decoding of image details. To this end, we initially train an EEG classifier using an asymmetric student–teacher knowledge distillation approach [9] In this context, the 'teacher' model is the pre-trained Contrastive Language-Image Pre-training (CLIP) model [10], which generates class probabilities from images. Unlike traditional frameworks, where the "student" model learns to replicate the "teacher's" outputs with either reduced capacity or on a corrupted version of the same stimuli, our approach involves feeding EEG activity into the "student" model. This compels the student model to learn how to predict class probabilities based on neural signals. Following the training phase, we retain only the EEG decoder, which we then integrate with a generative model based on latent diffusion. This combination is employed to produce novel images that possess semantic content derived from EEG signals

2. Related works

EEG are widely processed in the context of brain–computer interfaces (BCI) to perform brain decoding for a wide variety of tasks [11–15]. A number of prior works have explored decoding visual representations from EEG signals using deep learning models. Kavasidis et al. [16] were among the first to propose generating images from EEG data. They recorded EEG while participants viewed ImageNet images, and used an Long Short Term Memory (LSTM) model combined with variational autoencoders or GANs to reconstruct images. The key difference is they aimed for class-level image generation rather than detailed reconstruction and focuses on processing data in the time domain. Spampinato et al. [17] also analyzed EEG responses to ImageNet stimuli. They trained an LSTM encoder to classify EEG signals into image categories. For reconstruction, they trained a separate CNN regressor to predict EEG features from images and replaced the EEG signal with this encoder model. Palazzo et al. [18] extended [17] using contrastive learning to align EEG and visual image features. However, their goal was improving image classification rather than reconstruction, and various challenges emerged [19]. Singh et al. [20] proposed an EEG-to-image GAN framework but focused on smaller (i.e. with fewer images) datasets of characters and shapes. In this work, we propose a modularized pipeline for reconstructing detailed photorealistic visual stimuli (i.e. images) directly from EEG brain signals, using a CLIP based knowledge distillation of a convolutional neural network trained on time–frequency decomposition (TFD) and generative diffusion synthesis, generating semantically plausible and visually similar images reconstruction to the original stimuli.

3. Material and methods

This section delineates the methodology adopted and the dataset utilized. The datasets, sourced from ImageNet EEG [4] and THINGS-EEG2 [21], are publicly accessible. All computational experiments and model training were conducted on a server outfitted with four NVIDIA A100 GPU cards (each with 80 GB RAM connected via NVLINK) and 2 TB of system RAM. The codebase was developed using Python 3.9, leveraging libraries such as Pytorch, Pytorch Lightning, and scikit-learn for model implementation. Code is freely accessible at https://github.com/matteoferrante/EEG_decoding.

3.1. Data

The ImageNet-EEG recordings employed in this study were sourced from [22]. These recordings were obtained from six participants who were exposed to images from 40 distinct ImageNet [23] classes, with each class comprising 50 images. The sampling rate for these recordings was 1000 Hz. The image presentation protocol involved sequential display in 25-second intervals, succeeded by a 10-second intermission. In each display interval, images are shown sequentially for 0.5 s each. This protocol yielded a total of 2000 images spanning 1400 s (or 23 min and 20 s) of recording time. Each subject underwent four recording sessions, each lasting 350 s. The experiments utilized a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch, Brainproducts) for EEG data collection. Brainvision amplifiers and data acquisition systems were used to record the EEG signals at a sampling rate of 1000 Hz with 16-bit resolution. The EEG data resulted in 11,466 sequences post the exclusion of recordings of suboptimal quality. The comprehensive nature of this experimental design facilitated the examination of EEG responses to a diverse array of visual stimuli from ImageNet. The multi-channel EEG recordings, captured during the viewing of thousands of stimuli, furnish a rich dataset conducive for training decoding models. For further detail about acquisition protocol please see the original article [22]. To show the generalization ability of our method, we also included another dataset, from the THINGS initiative collection, named THINGS-EEG2 [21] This dataset comprises a collection of EEG readings taken at high temporal precision, recording reactions to pictures of objects against a natural backdrop. It encompasses data from 10 participants, covering 82,160 instances across 16,740 different image scenarios. Image stimuli belong to the THINGS Image dataset, spanning across 1854 different classes. In this work, the EEG activity is recorded while 1654 categories were shown as part of the training set and the other 200 categories were shown as test set. Since the very fine granularity of concepts of this dataset makes the problem more complex, we obtained pseudo-labels for the entire dataset using a K-Means over the CLIP embeddings of all images. We used a k-Elbow approach to identify the optimal number of clusters (that turned out to be 8 in this case) and trained a K-Means clustering to predict cluster labels to re-label this dataset. EEG data was processed as described before. Since the cluster labels cannot be used as conditioning for the generative part because they are obtained via an unsupervised method and hence would require a human re-labeling, we adopted a simpler approach for the second dataset, restricting the analysis to the classification part.

3.2. Preprocessing

Prior to utilizing the EEG signals for training our decoding models, a series of preprocessing steps were executed. Initially, a notch filter in the 49–51 Hz range was applied to mitigate power line interference. Subsequently, a second-order band-pass Butterworth filter, ranging between 14 and 70 Hz, was employed to focus on frequency bands pertinent to visual attention and object recognition. The signals were then standardized across channels. For the purpose of neural network input generation, the filtered EEG signals were segmented into 40 ms windows moving each time 20 ms. Time–frequency decompositions (TFD) were computed for these segments using the short-time Fourier transform (STFT), converting each trial into a 128-channel image that depicted the spectrum across both time and frequency dimensions. For the ImageNet-EEG this process yielded 2000 EEG spectrogram images, each with 128 channels, for every subject. For the THINGS-EEG2 dataset, we applied the same preprocessing steps, resulting in 16,540 spectrograms for and 200 spectrograms for testing. Given the dataset's highly detailed categorization into 1854 distinct classes, with no overlap between training and testing categories, traditional classification methods were deemed inappropriate for handling the data. To address this challenge, we re-run the classification using pseudo-labels generated by a clustering algorithm (K-means). This algorithm was applied to the image embeddings derived from the pretrained CLIP model, leading to the formation of 8 classes. The decision to use 8 clusters was based on the K-Elbow method, which searched within a range of 2 to 20, ensuring these classes were consistently represented across both training and testing datasets. Spectrograms were then used for training and evaluation of our convolutional neural network tailored for EEG decoding. This multi-channel spectral representation encapsulates the spatial and temporal intricacies of the EEG, allowing our model to extract features essential for visual stimuli classification. It is worth noting that the preprocessing described herein is specific to the architecture proposed in this study. Alternative baselines adopted slightly varied preprocessing techniques, such as direct time domain data analysis, starting from the same filtered data in the time domain. These variant preprocessing methodologies are elaborated upon in 3.6.

3.3. Decoding pipeline

Our approach employs a CNN with integrated residual connections to classify EEG TFDs. The architecture begins with a series of convolutional layers, progressively increasing the number of filters to effectively extract both spatial and temporal features. Subsequent to this, global average pooling and fully-connected layers are utilized for classification tasks. For the training of the CNN, we adopt a knowledge distillation methodology [9]. Initially, an image classifier is pretrained using CLIP (Contrastive Language-Image Pre-Training) [10] features to anticipate the stimulus classes, achieving a commendable accuracy of 99%. This pretrained classifier furnishes "soft targets" to guide our EEG model. During the training phase, EEG spectrograms are fed into the CNN, while CLIP image features are directed to the teacher classifier. The objective is to train the CNN such that it aligns with the class probability distributions produced by the teacher. This distillation approach not only stabilizes the training process but also enhances the model's performance in comparison to direct training on class labels. For inference, only the EEG-based CNN is deployed to predict classes from novel time–frequency decompositions. Through the distillation of knowledge from the image model, our CNN is equipped to derive robust representations, enabling the decoding of visual stimuli solely from EEG signals.

Post the training of our EEG decoding model, it becomes capable of predicting ImageNet classes from fresh EEG TFDs. To validate these predictions and reconstruct images that could potentially induce analogous neural responses, we employ the Stable Diffusion generative model [24]. For every EEG prediction, a text prompt such as "an image of a predicted class" is formulated. This prompt, in conjunction with random noise vectors, is input into Stable Diffusion to generate images congruent with the predicted class. This methodology facilitates the reconstruction of visual stimuli exclusively from neural activity patterns. The EEG decoder identifies the class, while Stable Diffusion fabricates a semantically coherent image. A comprehensive diagram of the decoding pipeline is depicted in Fig. 1, and the knowledge distillation procedure is illustrated in Fig. 2.

3.4. Reconstruction pipeline

Diffusion models are generative frameworks trained to invert a noise diffusion process, facilitating image synthesis. Stable Diffusion operates as a latent diffusion model, proficient in generating lifelike images from random noise vectors, conditioned by textual descriptions. The model's strategy involves the iterative addition of noise to genuine images, followed by the learning of a parametric denoising function to eradicate the noise over multiple timesteps. By repetitively applying the denoising function, the model can produce lifelike images, conditioned on textual descriptions. This iterative denoising offers tight control over image generation, guided by text at every iteration. In the sampling phase, Stable Diffusion accepts a text prompt and progressively diffuses noise vectors until they converge into an image that aligns semantically

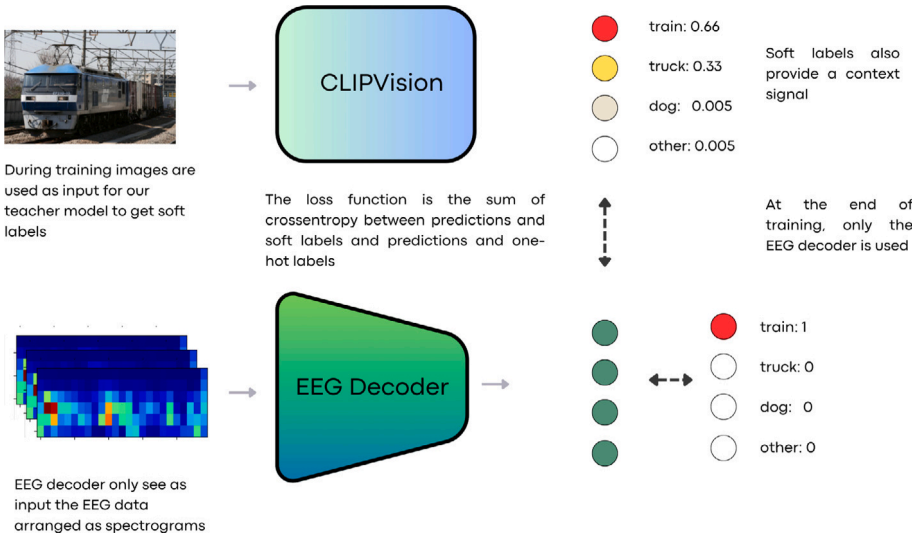

Fig. 2. Illustration of the training procedure. Knowledge distillation facilitates the training of a compact "student" model to emulate the outputs of a more extensive "teacher" model. This enables the student to achieve performance levels akin to larger models, even when initiated from distinct yet related inputs.

with the provided description. For the task of reconstructing images from EEG signals, Stable Diffusion's text conditioning capability proves invaluable. The EEG decoder outputs a label indicative of the visual stimulus class. This discrete label is then employed to generate corresponding images via Stable Diffusion, bypassing the need for direct pixel reconstruction. This approach facilitates the synthesis of plausible image reconstructions based on the decoded semantic category from neural activity patterns. This model-centric strategy also addresses the inherent resolution constraints of EEG for high-fidelity decoding. The guided diffusion modeling ensures the generation of visualizations that are both realistic and interpretable to human observers.

3.5. Knowledge distillation

Knowledge distillation facilitates the transfer of insights from a comprehensive, pretrained teacher model to a more compact student model [9]. This process empowers the student model to attain performance metrics that are typically associated with larger models. Consider $f_t(x)$ as the output vector of class probabilities produced by the teacher model for a given input x, representing the stimulus image. Similarly, let $f_s(e;\theta)$ denote the student model, characterized by parameters θ, where e represents the EEG recordings obtained during the presentation of stimulus x. The student model is trained through knowledge distillation by minimizing:

$$\mathcal{L}(\theta) = \alpha \mathcal{L}_{CE}(fs(e;\theta), y) + (1-\alpha)\mathcal{L}_{KD}(fs(x;\theta), f_t(x)) \tag{1}$$

Here, $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predictions of the student model and the actual ground truth labels y. In contrast, $\mathcal{L}_{KD}$ denotes the distillation loss, capturing the difference between the outputs of the student and teacher models. The temperature parameter T is employed to modulate the probability distribution of the teacher:

$$\mathcal{L}_{KD}(fs, ft) = -\sum_c \frac{\exp(f_{t,c}/T)}{\sum_{c'}\exp(f_{t,c'}/T)} \log \frac{\exp(f_{s,c}/T)}{\sum_{c'}\exp(f_{s,c'}/T)} \tag{2}$$

Training the student model to replicate the comprehensive probability distribution of the teacher facilitates the transfer of insights regarding inter-class relationships, offering a richer supervisory signal than mere ground truth labels. In our implementation, we set $\alpha = 0.5$ and $T = 1$ after initial empirical experimentation. For EEG decoding, a linear classifier was trained atop the CLIP [10] CLS tokens. CLIP, an acronym for Contrastive Language-Image Pre-Training, is a neural architecture trained to correlate images and text through contrastive learning. Comprising an image encoder and a text encoder, CLIP is trained to discern whether an image-text pairing is congruent or not. The image encoder in CLIP, a vision transformer (VIT), embeds images into latent representations. Throughout its training, CLIP cultivates an embedding space where semantically congruent images and texts are proximate. A pivotal element of the image encoder is the CLS token, an auxiliary token introduced to the network's input, enabling the encoder to generate a holistic representation of the entire image. A linear classifier was trained atop this CLS token for every image in the training dataset to predict the appropriate class. This amalgamation of CLIP and the classifier served as the teacher model, functioning as a bridge between EEG spectrograms and image classes. The student CNN, when exposed solely to EEG data, derives insights from both the teacher's distributions and the true labels. This distillation process accentuates the student's focus on neural patterns pertinent to visual recognition, enhancing convergence, accuracy, and generalization. By assimilating insights from a domain expert in image processing, the streamlined student decoder becomes adept at extracting visual representations from EEG signals.

3.6. Baselines

In order to underscore the efficacy of employing computer vision techniques for EEG signal decoding, we assessed a spectrum of baseline methodologies, spanning from conventional machine learning paradigms to contemporary neural network architectures.

Recently, several studies with remarkable results have been published on this dataset [4,22,25]. However, a subsequent analysis [26] revealed that, despite the methodological advancements being valid and innovative, the reported performance metrics are significantly inflated. This inflation is attributed to erroneous data preprocessing. Specifically, some preprocessing filters can induce temporal correlations between data points before splitting them into training and test sets, leading to information leakage. In response to these findings, follow-up counter-analyses [27] have demonstrated that, when eliminating this effect, the results remain valuable, albeit lower than initially reported. Therefore, in situating our work within the broader context of existing literature, in order to maintain best practices and avoid leakage, we have opted for the most conservative approach as outlined in the above-mentioned papers [26,27]. For similar reasons, in this paper we also provide an extensive set of baselines for performance comparison.

Initially, we employed a basic baseline wherein the raw EEG signals were standardized, squared, and subsequently averaged across channels. Following this, a Logistic Regression classifier was trained on the

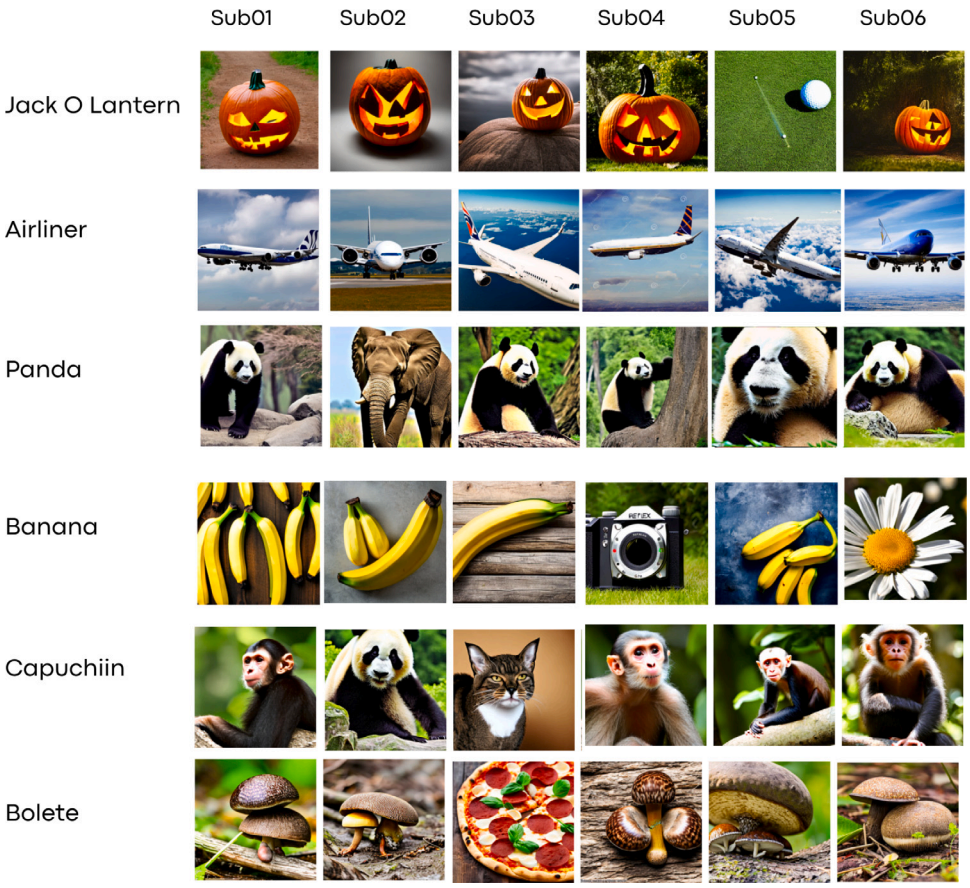

Fig. 3. Reconstructed images. Left column: target classes; subsequent columns: results from individual participants.

resultant data. An extension of this approach involved applying the Logistic Regression classifier to EEG signals that were averaged over an 80-point sliding window. In another variant we executed PCA on the windowed average EEG, preserving 29 components that accounted for 95% of the variance, prior to classifier training. Notably, these methodologies overlook the inherent spatial and temporal intricacies of the EEG signal. The main advantage of using the PCA is providing orthogonal features to the model that already integrate relevant spatiotemporal relationships. In this context, a recent proposition by CEBRA [28] demonstrated a deep learning technique that employs contrastive learning to project neural data onto lower-dimensional manifolds conducive for decoding. In alignment with this, we projected our EEG data onto a 32-dimensional manifold, utilizing CLIP features as a guiding mechanism. The value was chosen to be close to the number of features used in the PCA, picking the closest power of 2. This offers a robust nonlinear neural baseline that effectively harnesses both spatial and temporal patterns.

In terms of neural network architectures that directly process EEG time series data, we examined both a LSTM model and a 1D convolutional network (CNN) equipped with temporal convolutions. Both architectures incorporated 4 layers and were regularized using dropout, ensuring a consistent parameter count across models.

Further, we explored CNNs that operate on 2D representations of the EEG, thereby leveraging computer vision methodologies. One such model treated the raw EEG traces as a 2D image. Another model employed a wavelet decomposition utilizing the Daubechies db4 wavelet from PyWavelets [2] [29], which has been recognized as an efficient time–frequency representation for EEG [30]. Our final CNN baseline ingested the short-time Fourier transform (STFT) of the EEG, processed with a 40 ms window.

This ensemble of baselines, ranging from classical signal processing to avant-garde deep learning, offers a holistic comparative framework and accentuates the significance of spatiotemporal neural network modeling in the realm of EEG decoding. The computer vision-oriented strategies adeptly harness the structural nuances present in the multi-channel EEG.

For consistency, all neural networks were evaluated with a similar parameter count range (1.1–1.2 M). Each was trained using the Adam optimizer at a learning rate of $3e-4$. Additional training specifications included an early stopping callback with a 10-epoch patience based on validation loss variations, a batch size of 64, gradient clipping at a magnitude of 1.0, and a maximum epoch count set to 50.

4. Results

In this section, we present the outcomes for both datasets. For the initial dataset, ImageNet-EEG, we provide a comprehensive overview of the entire process, including classification outcomes and qualitative image reconstruction. This approach is feasible due to the relatively small (40) number of categories, allowing us to condition the generative model directly using the class labels. For the THINGS-EEG2 dataset, we focus exclusively on the classification results derived from the pseudo-labels assigned by the clustering algorithm, with the main objective of demonstrating the decoding of semantic information from EEG data".

4.1. Performance evaluation

Several metrics are available to evaluate the performances of a classification model [31–34]. In our case, the efficacy of our model is evaluated using a comprehensive set of metrics: top-5, top-3, top-1 accuracy, F1 score, and the normalized kappa score to evaluate

Table 1
Performance comparison of decoding baselines. The table presents the mean values accompanied by the standard deviation (enclosed in parentheses) for each evaluation metric across all participants. Results from Palazzo et al. [27] are reported from the original paper in the same setting used here. The first part of the table reports results for ImageNet-EEG dataset, while the second part report comparison between our method and plain CNN on the THINGS-EEG2 dataset.

Method	Metrics [Mean (Std)]				
	Accuracy	Top3 Accuracy	Top5 Accuracy	F1	Kappa
LR on average square signal	0.3600 (0.1313)	0.6619 (0.1758)	0.8156 (0.1619)	0.3493 (0.1375)	0.3435 (0.1345)
LR on windowed signal	0.0205 (0.0058)	0.0636 (0.0083)	0.1092 (0.0110)	0.0156 (0.0054)	0.0009 (0.0061)
LR on PCA windowed signal	0.0175 (0.0040)	0.0536 (0.0084)	0.0961 (0.0063)	0.0097 (0.0047)	0.0020 (0.0039)
CEBRA + kNN	0.0240 (0.0050)	0.0831 (0.0116)	0.1402 (0.0136)	0.0223 (0.0061)	−0.0012 (0.0056)
LSTM	0.3605 (0.0938)	0.7376 (0.1226)	0.8868 (0.1030)	0.3392 (0.0894)	0.3437 (0.0960)
Conv1d	0.2623 (0.0511)	0.6013 (0.0826)	0.7971 (0.0851)	0.2582 (0.0520)	0.2432 (0.0524)
Knowledge distillation on eeg (img)	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Knowledge distillation on wavelet	0.4060 (0.1154)	0.7490 (0.1282)	0.8787 (0.1007)	0.3889 (0.1148)	0.3905 (0.1183)
plain CNN on spectrograms	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Knowledge distillation on STFT	**0.4120 (0.1131)**	**0.7530 (0.1068)**	**0.8782 (0.0806)**	**0.4027 (0.1133)**	**0.3966 (0.1160)**
Palazzo et al. [27]	0.3350 (0.089)	–	–	–	–
Knowledge distillation (THINGS-EEG2)	**0.58 (0.04)**	–	–	**0.52 (0.036)**	–
Plain CNN (THINGS-EEG2)	0.52 (0.03)	–	–	0.48 (0.032)	–

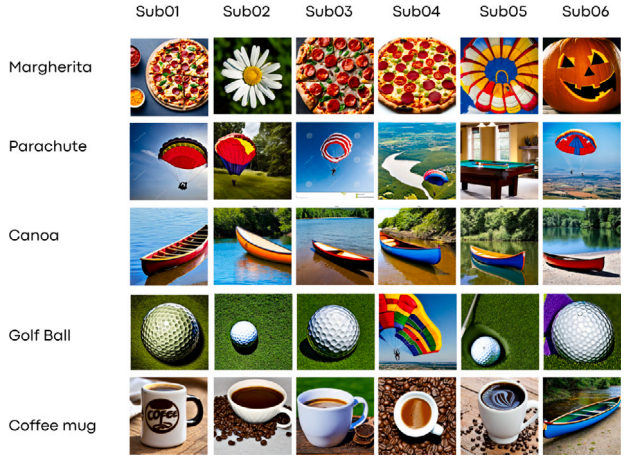

Fig. 4. On the left, the target classes are presented and each column show result from a single subject.

performances. Fig. 5 demonstrates that our knowledge distillation CNN consistently outperforms both the standard CNN baseline and a random classifier. Notably, the proposed approach – employing a CNN on TFD with CLIP-based knowledge distillation – exhibits superior performance compared to the same network without the distillation technique. This superiority is further evident when juxtaposed with other baselines detailed in Table 1.

Table 1 provides a summarized view of the decoding performance across various methods applied to EEG data. Clear trends in accuracy emerge across model types. Classical machine learning baselines, which utilize averaged or PCA-reduced EEG, yield near chance-level accuracy, underscoring the inadequacy of hand-engineered features for decoding intricate visual stimuli. An exception is the Logistic Regression model trained on squared data averages.

Conversely, deep learning models that harness spatiotemporal EEG TFDs patterns consistently achieve superior accuracy. Both convolutional and recurrent neural networks processing raw EEG time series deliver satisfactory results. Yet, the best performance is reached by models using 2D representations of multi-channel EEG. Specifically, CNNs fed with TFD computed using wavelet-transformed or spectrogram images both surpass 85% in top-5 accuracy, underscoring the benefits of computer vision techniques that learn directly from 2D structures in signal processing. Both wavelet and spectrogram decompositions seem to encapsulate pertinent time–frequency domain information for decoding. A closer examination of the top-3 and top-5 accuracy metrics reveals a consistent trend: deep learning models outclass classical baselines. The elite CNNs achieve over 75% in top-3 accuracy, implying that in approximately 3 out of 4 trials, the true label ranks within the top three predictions. The performance gap relative to the LSTM network is also noteworthy. This accentuates the efficacy of 2D convolutions in discerning the pertinent semantic categories from EEG patterns. The consistency of the top-5 accuracy across deep learning models suggests potential inherent challenges in precisely mapping EEG to granular image labels. However, the models adeptly identify the overarching category within their top predictions, underscoring the viability of EEG-based visual concept decoding. Finally, using our 8 clustering-derived pseudo-labels, we also verified that our approach outperforms a plain CNN baseline on the THINGS-EEG2 dataset. Our final model trained with knowledge distillation was able to achieve a top-1 prediction of 58%, hence discovering semantic content of the seen image from the neural data and confirming previously results".

From a qualitative perspective, Figs. 3 and 4 showcase examples of predicted and reconstructed images. While the model predominantly identifies the correct visual concept from EEG patterns, minor category confusions do arise. For instance, "bolete" might be misinterpreted as "pizza", or "banana" as "Margherita". Nevertheless, the model's ability to accurately discern the overarching semantic category and produce corresponding reconstructions is noteworthy.

In conclusion, our findings underscore the pivotal role of neural networks and image-centric representations in harnessing the rich multidimensional EEG representation. Directly classifying TFD inputs using a computer vision approach emerges as the potent strategy for EEG-based decoding.

5. Discussion

The primary objective of this study was to decode and reconstruct visual representations from EEG-recorded human brain activity. By employing deep convolutional neural networks trained on EEG TFD and guided by the CLIP-based knowledge distillation technique, we managed to predict image classes from the ImageNet dataset with an accuracy of 87% in the top-5 category. This knowledge distillation approach yielded a marked improvement in performance when compared to a baseline model and other data processing methodologies. While the model's predictions were generally reliable for the majority of participants, it did exhibit some confusion between closely related classes. The capability to extract the semantic content of image stimuli from non-invasive EEG recordings presents significant implications for the future of brain–computer interfaces. The methodology we developed for image reconstruction could potentially pave the way for a form of artificial vision, where decoded contents from a user's neural activity are visualized in real-time. Furthermore, our model introduces the possibility of innovative neurofeedback experiments, wherein participants could receive instantaneous visual feedback of decoded EEG patterns,

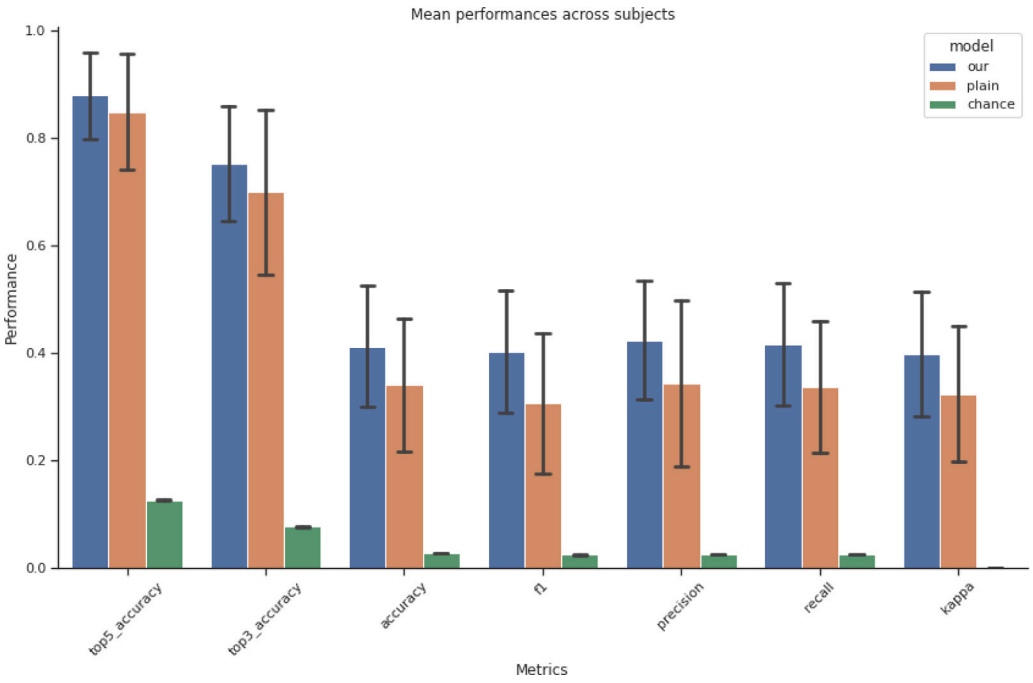

Fig. 5. Results for EEG decoder. **Ours** is the CLIP-based approach, **plain** is a vanilla CNN with the same architecture trained for classification and **chance** serves as a comparison with chance level. Bars are average across participants and error bars are standard deviations.

facilitating the voluntary self-regulation of brain states [3]. However, the study is not without limitations. EEG serves as a macroscopic lens into the brain's visual processing mechanisms. To address the limitations of EEG's spatial resolution, integrating it with other imaging techniques, such as fMRI, which boasts superior spatial resolution, is a promising avenue. Such multimodal strategies have shown potential in reconstructing images with a higher degree of detail [5–8]. Also, the model in its current configuration has not been optimized for decoding images outside the 40 categories or the 8 clusters used in the experiments, suggesting a need for further refinement. The variability in EEG decoding abilities across different participants or sessions, influenced by cognitive and neural factors, remains a topic that warrants deeper exploration. One of the significant concerns in EEG decoding revolves around the inadvertent extraction of personal perceptual data, which must be rigorously addressed. Our methodology places a strong emphasis on the creation of subject-specific models. This ensures that the decoding process is both consensual and uniquely tailored to the individual, mitigating potential ethical concerns. This approach not only necessitates voluntary participation but also minimizes the risk of misinterpretations due to the model's specificity to individual neural patterns. The rapid training methodology we have introduced also holds promise for real-time feedback paradigms using models tailored to individual participants, with a couple of seconds in inference time needed to predict class and generate the image on an A100 GPU. As the field of deep learning and generative models continues to evolve, we anticipate parallel advancements in EEG decoding and reconstruction capabilities.

6. Conclusions

In conclusion, our research demonstrates the capability of an integrated EEG decoding system using a novel knowledge distillation technique paired with latent diffusion models. This approach not only advances theoretical understanding but also holds significant promise for practical applications. The potential real-world applications of this technology are vast, including the development of assistive technologies for individuals with disabilities, enhancing communication for those unable to speak or use traditional input devices, and improving neurorehabilitation methods. One immediate application could be in the realm of augmented and virtual reality, where users could manipulate environments directly through neural inputs, creating a more immersive and intuitive user experience. Moreover, the integration of our decoding approach with existing technologies could lead to more responsive and adaptive systems, tailored to individual neurological profiles for personalized user interfaces. Future work will focus on refining the decoding accuracy and efficiency of the system, exploring the integration with other modalities like fMRI for improved spatial resolution and incorporating real-time feedback mechanisms to enhance learning and adaptation in the brain–computer interface. Additionally, further research into the ethical implications and the security of neural data in such applications is paramount to ensure privacy and consent in the use of this technology. The methodologies and findings from this study could significantly influence the development of next-generation brain–computer interfaces by providing a framework that employs advanced machine learning techniques to interpret and translate complex neural signals into actionable outputs. This could eventually lead to breakthroughs where brain–computer interfaces may offer seamless integration between human cognitive states and machine operations, heralding a new era of interaction between humans and technology In this study, we demonstrated the potential of deep neural networks, coupled with generative diffusion models, to reconstruct visual experiences directly from non-invasive EEG recordings from two independent datasets achieving a top-1 accuracy in prediction of 40 classes of 45% (and a top-5 accuracy of 87%) on the ImageNet-EEG dataset and a top-1 accuracy of 58% in prediction of 8 semantic clusters on the THINGS-EEG2 dataset. The application of knowledge distillation from language-image pretraining enabled our convolutional decoder to effectively extract semantic information from brain activity patterns. This capability significantly surpassed the performance of classical signal processing baselines. By generating images based on the predicted labels, we were able to produce visualizations that closely align with the decoded neural activity. Our emphasis on creating subject-specific models not only ensures a certain degree of privacy but also underscores the unique capabilities of EEG data in decoding individual mental representations. These techniques, which focus on translating neural signals into their corresponding images, can kickstart significant

Table 2
CNN Classifier Network Structure.

Layer	Type of Operation	Details
cnn_model	Classifier	–
net	Sequential	–
layer_0	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(17, 64, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(64), Dropout(p=0.2), GELU
- residual	Conv2d	Conv2d(17, 64, kernel_size=(3,), stride=(2,), padding=(1,))
layer_1	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(64, 64, kernel_size=(3,), stride=(1,), padding=(1,))
– adn	ADN	BatchNorm2d(64), Dropout(p=0.2), GELU
- residual	Identity	–
layer_2	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(64, 128, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Conv2d	Conv2d(64, 128, kernel_size=(3,), stride=(2,), padding=(1,))
layer_3	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(1,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Identity	–
layer_4	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
– adn	ADN	BatchNorm2d(128), Dropout(p=0.2), GELU
- residual	Conv1d	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
layer_5	ResidualUnit	–
- conv	Sequential	–
– unit0	Convolution	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
- residual	Conv2d	Conv2d(128, 128, kernel_size=(3,), stride=(2,), padding=(1,))
reshape	Reshape	–
final	Sequential	–
- 0	Flatten	Flatten(start_dim=1, end_dim=-1)
- 1	Linear	in_features=896, out_features=*num classes*, bias=True

advancements in the domains of brain–computer interfaces and neural prosthetics, as well as human–computer interaction research. Overall, our findings highlight the potential of non-invasive brain imaging as a tool to provide insights into the human cognitive experience.

CRediT authorship contribution statement

Matteo Ferrante: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tommaso Boccato:** Writing – review & editing, Writing – original draft, Conceptualization. **Stefano Bargione:** Validation, Investigation. **Nicola Toschi:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported and funded by: NEXTGENERATIONEU (NGEU); the Ministry of University and Research (MUR); the National Recovery and Resilience Plan (NRRP); project MNESYS (PE0000006, to NT) - A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); the MUR-PNRR M4C2I1.3 PE6 project PE00000019 Heal Italia (to NT); the NATIONAL CENTRE FOR HPC, BIG DATA AND QUANTUM COMPUTING, within the spoke "Multiscale Modeling and Engineering Applications" (to NT); the European Innovation Council (Project CROSSBRAIN - Grant Agreement 101070908, Project BRAINSTORM - Grant Agreement 101099355); the Horizon 2020 research and innovation Programme (Project EXPERIENCE - Grant Agreement 101017727). Matteo Ferrante is a Ph.D. student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma.

Appendix

Architecture details

Here we report some additional details, such as the network's architecture structure. Each CNN model has the following structure and was trained using the Adam optimizer at a learning rate of $3e-4$. Additional training specifications included an early stopping callback with a 10-epoch patience based on validation loss variations, a batch size of 64, gradient clipping at a magnitude of 1.0, and a maximum epoch count set to 50 (see Table 2).

Detailed performances

Table 3 shows details of classification of ImageNet-EEG dataset for all classes. In examining performance metrics across various classes, we note significant variances in precision, recall, and F1-scores, indicating the model's strengths and weaknesses in classifying diverse items. High-performing classes like "parachute" and "sorrel" demonstrate the model's efficacy with high precision and recall, suggesting these classes have unique, easily distinguishable features. In contrast, lower-performing classes such as "anemone fish" and "revolver" exhibit challenges in accurate identification, likely due to feature overlaps with other classes or insufficient training data. The impact of semantic similarity is evident, where high-performance classes typically show little resemblance to others, aiding their classification. For instance,

Table 3
Detail of average classification performances for ImageNet-EEG for each class.

Class	Precision	Recall	F1-Score
sorrel	0.71	1.00	0.83
parachute	.90	0.90	0.95
iron	0.50	0.67	0.57
anemone_fish	0.20	0.33	0.25
espresso_maker	0.33	0.33	0.33
coffee_mug	0.14	0.40	0.21
mountain_bike	0.50	0.33	0.40
revolver	0.20	0.14	0.17
giant_panda	0.67	0.20	0.31
daisy	0.36	0.44	0.40
canoe	0.50	0.56	0.53
lycaenid	0.43	0.38	0.40
German_shepherd	1.00	0.71	0.83
running_shoe	0.17	0.17	0.17
jack-o'-lantern	0.44	0.36	0.40
cellular_telephone	0.33	0.25	0.29
golf_ball	1.00	0.67	0.80
desktop_computer	0.50	0.46	0.48
broom	0.18	0.40	0.25
pizza	0.22	0.33	0.27
missile	0.50	0.25	0.33
capuchin	0.50	0.33	0.40
pool_table	0.50	0.71	0.59
mailbag	0.09	1.00	0.17
convertible	0.22	0.20	0.21
folding_chair	0.38	0.60	0.46
pajama	0.56	0.62	0.59
mitten	0.55	0.50	0.52
electric_guitar	0.44	0.33	0.38
reflex_camera	0.25	0.25	0.25
grand_piano	0.40	0.50	0.44
mountain_tent	0.88	0.78	0.82
banana	0.43	0.50	0.46
bolete	0.62	0.45	0.53
digital_watch	0.12	0.12	0.12
African_elephant	0.60	0.50	0.55
airliner	0.36	0.44	0.40
electric_locomotive	0.50	0.33	0.40
radio_telescope	0.75	0.60	0.67
Egyptian_cat	0.50	0.80	0.62

"parachute" has distinct features unlike any other class. However, low-performing classes like "espresso maker" and "coffee mug" may share commonalities, leading to confusion and incorrect class prediction even if this could be considered a good semantic approximation of the context if derived from neural activity. For instance, consider a scenario where the model incorrectly identifies a "coffee mug" as an "espresso maker". While this constitutes an error, it is noteworthy that the misclassification still falls within the same semantic realm of concepts related to coffee. The employment of a knowledge distillation approach is deliberately designed to nurture such similarities, ensuring that when errors occur, they remain semantically closer to the original concept. This strategy aims to mitigate the impact of mistakes by aligning them more closely with the underlying theme or category of the target object.

Fig. 6 shows examples of the clusters pseudo-labels obtained using K-Means on the CLIP CLS embeddings of training images. Since it is based on pseudo-labels, there is not an exact match with a specific class, however we can qualitatively infer the following groupings:

- Outdoor and Tactical Equipment: The first row (pseudo-label 0) seems to contain items related to outdoor activities or tactical equipment
- Sports Equipment: The second row (pseudo-label 1) appears to be sports equipment, including balls and a bicycle.
- Clothing and Accessories: The third row (pseudo-label 2) includes various items of clothing and personal accessories.
- Mixed food and small objects: The fourth row (pseudo-label 3) includes small objects, cakes and ice.

Table 4
Detail of average classification performances for THINGS-EEG2 for each classs.

Class	Precision	Recall	F1-Score
0	0.375	0.387	0.381
1	0.421	0.571	0.485
2	0.778	0.438	0.560
3	0.300	0.273	0.286
4	0.500	0.190	0.276
5	0.167	0.231	0.194
6	0.689	0.861	0.765
7	0.625	0.568	0.595

- Home and Living: The fifth row (pseudo-label 4) has items commonly found in a home or associated with living spaces, like furniture and appliances.
- Plants and Nature: The sixth row (pseudo-label 5) seems to focus on plants or elements commonly found in gardens.
- Animals: The sixth row shows animals, both wild and domestic.
- Food and Kitchen: The seventh row (pseudo-label 7) has images of food and items related to the kitchen or food preparation.

Table 4 summarizes details of performances on pseudolabels for THINGS-EEG2 dataset.

Based on the Table 4 and the image clusters (Fig. 6, we can comment on the performance of each pseudo-label cluster as follows:

Outdoor and Tactical Equipment (Class 0): This class has moderate precision and recall, suggesting the model has a reasonable ability to identify items within this cluster, but there is still room for improvement in recognizing and distinguishing these objects with greater accuracy.

Sports Equipment (Class 1): With a precision slightly above average and a relatively high recall, this cluster is better identified by the model, indicating that the distinctive features of sports equipment are more easily recognized.

Clothing and Accessories (Class 2): This class has the highest precision but lower recall, which could mean that while the items classified as clothing and accessories are often correct, the model is missing quite a few actual instances of this class.

Mixed Food and Small Objects (Class 3): The low precision and recall in this cluster imply that the model struggles significantly with this category. The heterogeneity of the group may contribute to this difficulty, as it combines various unrelated items.

Home and Living (Class 4): This class also has low precision and recall scores. Similar to the mixed food and small objects class, the diversity of items in home and living could be leading to challenges in accurate classification.

Plants and Nature (Class 5): This class has the lowest performance metrics across all clusters, with both precision and recall below 0.2. It suggests that the model has substantial difficulty in recognizing and categorizing these images accurately.

Animals (Class 6): The model performs best in this class, showing high precision and recall. This indicates that the model can effectively identify and categorize animal images, which might be due to more distinctive and recognizable features in these images compared to other classes.

Food and Kitchen (Class 7): The performance here is quite good, with both precision and recall above 0.5. The model is reasonably competent at identifying items related to food and kitchen, which might be due to their specific shapes and contexts that are easier to learn.

The variations in performance across the clusters may be influenced by the intrinsic properties of the items within them. Clusters with more visually distinct and less varied items (like Animals and Food and Kitchen) are classified more accurately. In contrast, clusters containing a wide range of heterogeneous items (like Mixed Food and Small Objects and Home and Living) tend to have lower performance

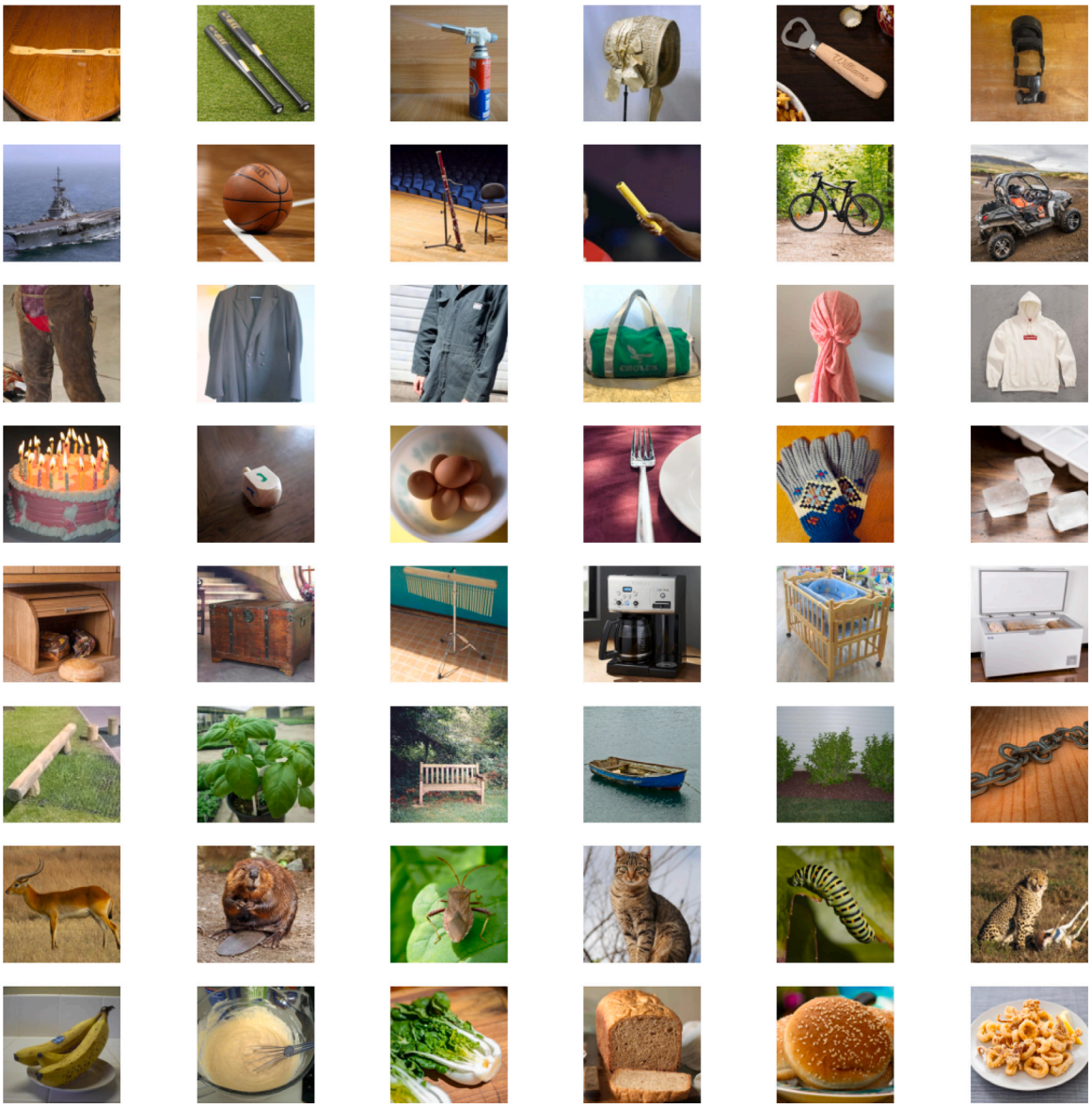

Fig. 6. Example of clusters for THINGS-EEG2. Each row is a different cluster with some examples to have a qualitative idea of the semantic content.

metrics, indicating a need for model improvements in these areas, perhaps through better feature extraction methods or more representative training data.

References

[1] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, D. Giordano, M. Shah, Generative adversarial networks conditioned by brain signals, 2017, pp. 3430–3438.

[2] Yunpeng Bai, Xintao Wang, Yan pei Cao, Yixiao Ge, Chun Yuan, Ying Shan, DreamDiffusion: Generating high-quality images from brain EEG signals, 2023.

[3] Stefanie Enriquez-Geppert, René J. Huster, Christoph S. Herrmann, EEG-neurofeedback as a tool to modulate cognition and behavior: A review tutorial, Front. Hum. Neurosci. 11 (2017).

[4] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, Mubarak Shah, Brain2Image: Converting brain signals into images, in: Proceedings of the 25th ACM International Conference on Multimedia, MM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1809–1817.

[5] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, Nicola Toschi, Brain captioning: Decoding human brain activity into images and text, 2023.

[6] Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, Rufin VanRullen, Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned GANs, 2022.

[7] Furkan Ozcelik, Rufin VanRullen, Brain-diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion, 2023.

[8] Yu Takagi, Shinji Nishimoto, High-resolution image reconstruction with latent diffusion models from human brain activity, 2023, bioRxiv.

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, Distilling the knowledge in a neural network, 2015.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning transferable visual models from natural language supervision, 2021.

[11] Raheel Zafar, Aamir Saeed Malik, Nidal Kamel, Sarat C. Dass, Jafri M. Abdullah, Faruque Reza, Ahmad Helmy Abdul Karim, Decoding of visual information from human brain activity: A review of fMRI and EEG studies, J. Integr. Neurosci. 14 (2) (2015-06) 155–168.

[12] Tianwei Shi, Ke Chen, Ling Ren, Wenhua Cui, Brain computer interface based on motor imagery for mechanical arm grasp control, Inf. Technol. Control. 52 (2023) 358–366.

[13] Kola Venu, P. Natesan, Optimized deep learning model using modified whale's optimization algorithm for EEG signal classification, Inf. Technol. Control 52 (2023) 744–760.

[14] Eglė Butkevičiūtė, Liepa Bikulčienė, Tatjana Sidekerskienė, Tomas Blažauskas, Rytis Maskeliūnas, Robertas Damaševičius, Wei Wei, Removal of movement artefact for mobile EEG analysis in sports exercises, IEEE Access 7 (2019) 7206–7217.

[15] N. Murali Krishna, Kaushik Sekaran, Annepu Venkata Naga Vamsi, G.S. Pradeep Ghantasala, P. Chandana, Seifedine Kadry, Tomas Blažauskas, Robertas Damaševičius, An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using EEG signals, IEEE Access 7 (2019) 77905–77914.

[16] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, Mubarak Shah, *Brain2Image*: Converting brain signals into images, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017-10-23, pp. 1809–1817.

[17] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2017-07, pp. 4503–4511.

[18] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, M. Shah, Generative adversarial networks conditioned by brain signals, in: 2017 IEEE International Conference on Computer Vision, ICCV, IEEE, 2017-10, pp. 3430–3438.

[19] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, Jeffrey Mark Siskind, Training on the test set? An analysis of spampinato et al. [31], 2018.

[20] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, Shanmuganathan Raman, EEG2IMAGE: Image reconstruction from EEG brain signals, 2023.

[21] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, Radoslaw M. Cichy, A large and rich EEG dataset for modeling human visual object recognition, NeuroImage 264 (2022) 119754.

[22] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Mubarak Shah, Nasim Souly, Deep learning human mind for automated visual classification, 2019.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, Hierarchical text-conditional image generation with CLIP latents, 2022.

[25] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, Ying Shan, DreamDiffusion: Generating high-quality images from brain EEG signals, 2023.

[26] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B Wilbur, Hari M. Bharadwaj, Jeffrey Mark Siskind, Training on the test set? An analysis of spampinato et al. [31], 2018.

[27] Simone Palazzo, Concetto Spampinato, Joseph Schmidt, Isaak Kavasidis, Daniela Giordano, Mubarak Shah, Correct block-design experiments mitigate temporal correlation bias in EEG classification, 2020.

[28] Steffen Schneider, Jin Hwa Lee, Mackenzie Weygandt Mathis, Learnable latent embeddings for joint behavioural and neural analysis, Nature 617 (7960) (2023) 360–368.

[29] Gregory R. Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, Aaron O. Leary, Pywavelets: A python package for wavelet analysis, J. Open Source Softw. 4 (36) (2019) 1237.

[30] Abdulhamit Subasi, EEG signal classification using wavelet feature extraction and a mixture of expert model, Expert Syst. Appl. 32 (4) (2007) 1084–1093.

[31] Minakshi Boruah, Ranjita Das, CaDenseNet: a novel deep learning approach using capsule network with attention for the identification of HIV-1 integration site, Neural Comput. Appl. 35 (23) (2023) 17113–17128.

[32] Minakshi Boruah, Ranjita Das, Identification of DNA motif using likelihood and attention based pooling method in the GRU framework, in: 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering, ICRAIE, 6, 2021, pp. 1–5.

[33] André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr, Douglas G. Manuel, A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 4.

[34] Andreas Holzinger, André Carrington, Heimo Müller, Measuring the quality of explanations: The system causability scale (SCS): Comparing human and machine explanations, KI - Künstliche Intell. 34 (2) (2020) 193–198.