Exploring EEG-Image Classification and EEG-Based Caption Retrieval

Diego Bermúdez Sierra
Carnegie Mellon University
Pittsburgh, PA 15213
dabermud@andrew.cmu.edu

Santiago Bolaños Vega
Carnegie Mellon University
Pittsburgh, PA 15213
sbolaosv@andrew.cmu.edu

Abstract

This project investigates whether visual perception can be decoded directly from electroencephalography (EEG) signals and aligned with multimodal representations such as text. For the midterm milestone, we implemented the baseline EEG classifier provided in the dataset framework and reproduced its reported performance, achieving comparable accuracy across subjects. We also completed image–caption retrieval experiments using a pretrained CLIP model to establish a multimodal baseline and verify our data and evaluation pipelines. In the next phase, we will extend this foundation by developing a unified architecture that leverages convolutional and transformer blocks to capture temporal and spatial dependencies in EEG data. This model will align EEG embeddings with CLIP's semantic space through contrastive and knowledge-distillation losses, and explore parameter-efficient fine-tuning methods such as LoRA to improve cross-subject generalization and stability.

1 Overview and Context

1.1 Motivation

Understanding how the human brain encodes visual information remains one of the central challenges in computational neuroscience and artificial intelligence. Decoding visual perception from EEG signals offers a unique opportunity to explore the neural correlates of vision using a non-invasive, affordable, and high–temporal–resolution recording technique. Beyond its neuroscientific relevance, this line of research has practical implications for brain–computer interfaces (BCIs), enabling communication or control for individuals with limited mobility, as well as cognitive monitoring in assistive and adaptive systems. Recent advances in multimodal learning, particularly through models like CLIP, have demonstrated that joint visual–language embedding spaces can capture rich semantic relationships across modalities. Bridging EEG with such representations could reveal whether neural signals carry semantic information similar to that encoded by deep vision–language models. This connection between brain activity and high–level machine representations may open new pathways for interpretable AI and cross-modal brain decoding.

1.2 Objective

The primary objective of this project is to evaluate whether EEG signals recorded while participants view images can be effectively mapped into a multimodal embedding space shared with textual representations. Specifically, our goals are:

- To design and train an EEG encoder capable of classifying image categories from brain activity across multiple subjects and sessions.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

- To develop a projection mechanism that aligns EEG embeddings with CLIP's text embeddings, enabling EEG–caption retrieval.
- To analyze the model's ability to generalize across subjects and evaluate retrieval performance using metrics such as Recall@K, BERTScore, and CLIPScore.

2 Related Work and Background

2.1 Literature Review

Recent progress in brain decoding and multimodal learning provides the foundation for this work. Large-scale EEG datasets such as Wang et al. [5] have enabled systematic modeling of visual perception from neural signals, revealing that EEG data encode category-level information that can be learned through deep architectures. At the same time, advances in multimodal representation learning, most notably contrastive models like Radford et al. [4], have shown that shared image–text embedding spaces can capture rich semantic structure across modalities. Prior multimodal EEG–vision studies [3] further demonstrate that neural activity can be aligned with visual features extracted by deep networks, suggesting that cross-modal learning can extend beyond vision to language-aligned spaces.

Building on these insights, our project explores whether EEG representations can be projected into a semantic space learned by large vision–language models. To improve training efficiency and stability, we intend to incorporate ideas from knowledge distillation [1] and low-rank adaptation [2], allowing the EEG encoder to learn from pretrained model outputs while adapting multimodal parameters with minimal computational overhead. Together, these studies motivate a unified framework for decoding perceptual representations from EEG through modern multimodal and transfer-learning techniques.

2.2 Background

Electroencephalography (EEG) is a non-invasive method for measuring brain activity through scalp-recorded electrical signals. It offers millisecond-level temporal resolution, making it well suited for studying perceptual and cognitive processes, though it is limited by noise, inter-subject variability, and relatively low spatial resolution. In visual recognition studies, EEG captures rapid neural responses following stimulus presentation, which reflect the brain's processing of visual features and object categories.

The dataset used in this project contains multi-subject and multi-session EEG recordings collected during image-viewing tasks, with 122 channels sampled at 1000 Hz and preprocessed using standard filtering techniques. Each trial corresponds to a short segment of brain activity aligned with image onset. Traditional EEG classification methods based on shallow or fully connected networks capture low-level temporal patterns but often fail to generalize across subjects. Recent architectures using transformers better model global temporal–spatial dependencies, offering a more powerful representation for decoding visual semantics. Our work builds on these advances, aiming to align EEG representations with multimodal embedding spaces to explore how visual information is encoded in brain activity.

3 Methodology

3.1 Model Description

At this stage of the project, we have implemented the baseline EEG classification model described in Section 4. These models establish the foundation for our framework and were used to validate the preprocessing pipeline and confirm dataset integrity. Further details on their structure and evaluation are provided in the next section.

3.2 Dataset

The dataset used in this project consists of electroencephalography (EEG) recordings collected from 13 human subjects while they viewed natural images from 40 object categories. The visual stimuli were drawn from a curated subset of ImageNet, covering diverse classes such as animals, vehicles, tools, furniture, and food. Each image was presented on a gray background for one second, followed

by a one-second blank screen. Subjects were instructed to fixate on the center of the screen and minimize movement to reduce noise in the EEG recordings.

This study focuses on the low-speed visual presentation condition, which provides a higher signal-to-noise ratio and more stable visual evoked responses. Each subject completed five sessions, with four runs of 50 trials per session, resulting in approximately 13,000 EEG samples per participant. EEG signals were recorded from 128 channels at a sampling rate of 1000 Hz using a BioSemi ActiveTwo system. Each trial corresponds to a one-second segment aligned with stimulus onset, resulting in an EEG sample of shape $(128, 1000)$. All signals were preprocessed using band-pass and notch filters for artifact removal and normalized per channel.

EEG Signal Characteristics and Model Rationale. Figure 1 illustrates the temporal and spatial variability of EEG signals, where different channels exhibit distinct oscillatory patterns and asynchronous fluctuations over time. This complexity motivated the use of an architecture capable of capturing both localized temporal dependencies and broader spatial interactions.

In addition to EEG signals and categorical labels, each image is paired with a short natural-language description derived from the ImageNet metadata. These captions serve as semantic text features for multimodal alignment experiments with CLIP. Following the dataset protocol, three sessions are used for training, one for validation, and one for testing, allowing both within-subject and cross-subject evaluation of model generalization.

3.3 Evaluation Metric

Model performance is evaluated using task-specific metrics. For EEG-based image classification, we use **accuracy** as the primary metric to measure the proportion of correctly predicted image categories across trials and subjects. Accuracy provides a straightforward indication of how well the model decodes visual category information from EEG signals.

For multimodal retrieval experiments involving EEG-to-caption or EEG-to-image alignment, we adopt four complementary metrics: **Recall@K**, which measures how often the correct match appears among the top K retrieved results; **CLIPScore**, which quantifies semantic similarity in the CLIP embedding space; **mean Average Precision (mAP)**, which captures overall retrieval ranking quality; and **BERTScore**, which evaluates textual similarity between predicted and ground-truth captions. Together, these metrics provide a comprehensive view of both categorical discrimination and semantic alignment performance.

3.4 Loss Function

For the EEG classification task, we use the **cross-entropy loss**, which measures the difference between predicted category probabilities and the true labels. This objective is well-suited for multi-class classification and provides stable convergence when training the baseline CNN-based models.

For the next phase of the project, we plan to extend the objective to support multimodal alignment between EEG and CLIP embeddings. Potential formulations include a **cosine similarity loss** or **contrastive loss** to encourage paired EEG and text (or image) representations to lie close in the shared embedding space, and a **knowledge distillation loss** based on Kullback–Leibler (KL) divergence with temperature scaling (τ) to transfer semantic structure from CLIP to the EEG encoder. The specific loss design will be finalized once the EEG–CLIP projection framework is implemented.

4 Baseline and Extensions

4.1 Baseline Selection and Evaluation

The baseline models used in this project are convolutional neural network (CNN)-based architectures originally introduced by Wang et al. [5] for EEG-based visual object classification. These models serve as reference points for evaluating the performance of our proposed multimodal framework. The CNN-based baselines extract localized temporal features using one-dimensional convolutions followed by max-pooling, while some variants incorporate recurrent or dense layers to capture sequential and non-linear dependencies in the EEG signal. All models were trained to classify

visual object categories from EEG recordings using supervised learning, with **cross-entropy loss** and **accuracy** as the primary evaluation criterion.

We successfully replicated the baseline performance reported by Wang et al. [5], achieving comparable classification accuracy across subjects. This confirms that our data preprocessing and training pipelines are functioning correctly. The reproduced baseline establishes a solid foundation for subsequent model extensions that aim to improve temporal–spatial feature extraction and cross-subject generalization.

In addition to EEG classification, we also conducted baseline experiments using a pretrained CLIP model to establish a reference for multimodal retrieval. By computing cosine similarity between image and text embeddings, we verified the reliability of the pretrained CLIP representations and evaluated semantic alignment using Recall@K, CLIPScore, mean Average Precision (mAP), and BERTScore. These baselines provide a meaningful reference for the multimodal alignment phase, where EEG representations will be projected into the same semantic embedding space.

4.2 Implemented Extensions / Experiments

Building on the baseline EEG classification results, we are extending the framework toward multimodal alignment between EEG and CLIP embeddings. The following experiments and extensions are planned or under development:

- **EEG encoder refinement:** Replace the baseline CNN with a hybrid architecture that combines convolutional and transformer layers to better capture temporal–spatial dependencies in EEG signals.
- **Knowledge distillation:** Experiment with distillation from CLIP's pretrained encoders, using different temperature values (τ) to analyze their effect on alignment stability and generalization.
- **Parameter-efficient fine-tuning:** Apply Low-Rank Adaptation (LoRA) to selectively fine-tune CLIP's text encoder, improving EEG–text alignment without retraining the full model.
- **Cross-session and cross-subject evaluation:** Test whether the learned EEG representations generalize across different sessions and participants to assess robustness and transferability.
- **Projection and contrastive objectives:** Explore alternative projection heads (e.g., MLP or attention-based layers) and contrastive objectives to enhance EEG-to-text retrieval performance.

5 Results and Analysis

5.1 Results

For the baseline EEG classification, we successfully replicated the performance reported in Wang et al. [5], achieving similar accuracy across subjects. This validates the correctness of our preprocessing pipeline and confirms the reproducibility of the dataset's reference model. The classification results obtained using the proposed model indicate moderate discriminative ability, consistent with the challenging nature of EEG-based visual decoding. Across the test set, the model achieved an overall accuracy of **6.38%**, with a precision of **6.11%**, recall of **6.38%**, and F1-score of **5.73%**.

Per-class performance varied significantly, ranging from a maximum of **25.77%** for the *boat* category to only **0.77%** for the *bird* class. A few other categories, such as *diningtable* and *person*, also demonstrated relatively higher accuracies of around **10%**, while most others remained near chance level (5%).

For the image–caption retrieval experiment using the pretrained CLIP model, the following metrics summarize the observed performance:

- **Retrieval Metrics (Recall@K):** Recall@1 = 0.1982, Recall@3 = 0.3270, Recall@5 = 0.3950. Class-aware Recall@1 = 0.9699, Recall@3 = 0.9825, Recall@5 = 0.9851.
- **Mean Average Precision (MAP):** Caption-level MAP = 0.2980; Class-aware MAP = 0.8305.

- **Semantic Metrics:** CLIPScore (mean) = 0.3223 ± 0.0237; BERTScore F1 (mean) = 0.9243 ± 0.0474; high-similarity rate (> 0.7) = 1.0000.

5.2 Discussion

When compared to the results presented by Wang et al. [5], where baseline convolutional and transformer-based models reached between **8% and 12%** top-1 accuracy depending on architecture and training configuration, our implementation falls within the lower bound of the expected performance range. Our results validate the overall experimental setup and demonstrate that the implemented model captures meaningful but limited discriminative information from the EEG signals, consistent with prior literature.

Qualitatively, the confusion matrix suggests that certain object categories elicit more distinct and consistent neural signatures, potentially due to visual salience or semantic familiarity. The overrepresentation of *boat* and *bottle* among correct predictions could reflect category-specific neural responses that are easier to decode given their distinct visual features or frequency of occurrence in the training set.

Future iterations of this work will focus on enhancing subject generalization and temporal modeling, potentially through contrastive learning or subject-adaptive normalization, to bridge the remaining performance gap relative to the reported benchmarks.

On the other hand, across all metrics, CLIP demonstrates strong semantic alignment between image and caption embeddings, particularly under class-aware evaluation. Detailed per-class MAP scores (ranging from 0.12 to 0.55) further confirm that retrieval performance varies by object category, with higher scores observed for semantically rich classes such as *person*, *dog*, and *cat*. These results provide a solid multimodal baseline for the next phase, where EEG features will be projected into the CLIP embedding space to enable brain-to-language retrieval.

6 Future Directions

Our immediate focus is on improving the performance and generalization of the EEG encoder architecture. We are currently experimenting with deeper and more expressive transformer-based encoders that can better capture long-range temporal–spatial dependencies in EEG data. This includes exploring multi-head attention across channels and time, residual connections, and subject-agnostic normalization layers to enhance cross-subject transferability.

Future experiments will also investigate advanced training strategies to bridge EEG features with the CLIP embedding space more effectively. We plan to implement a distillation-based objective, where the EEG encoder learns to approximate CLIP's image and text representations through softened output distributions. This approach is expected to stabilize training and transfer semantic structure from the pretrained CLIP model to the EEG domain.

Additionally, we aim to incorporate parameter-efficient fine-tuning techniques such as Low-Rank Adaptation (LoRA) to adapt CLIP's text encoder with minimal computational overhead. This would allow us to tailor the semantic space to EEG-specific representations without retraining large portions of the CLIP model.

Other potential directions include experimenting with temporal contrastive learning for unsupervised EEG pretraining, evaluating alternative projection mechanisms (e.g., nonlinear MLP heads or attention-based fusion), and conducting ablation studies on loss components and architecture depth. Together, these steps will guide the development of a more robust and semantically aligned EEG–language representation framework.

7 Administrative Details

7.1 Team Contributions

- **Diego Bermúdez Sierra** is focusing on the **model architecture design**, **training implementation**, and **evaluation pipeline**. He is developing the experimental framework and leading the implementation of baseline and transformer-based EEG encoders.

- **Santiago Bolaños Vega** is focusing on the **report writing**, **literature review**, and **EEG data preprocessing**. He is also conducting background research on multimodal learning and preparing the dataset pipeline for model training.
- **Joint Contributions** Both members collaborate closely on **debugging**, **experiment design**, and overall project direction. The **CLIP integration** component will be defined and implemented collaboratively in the next phase of the project.

7.2 GitHub Repository

The code, notebooks, and experimental results for this project are available at:

`https://github.com/9sntg/exploring-eeg`

7.3 Appendix

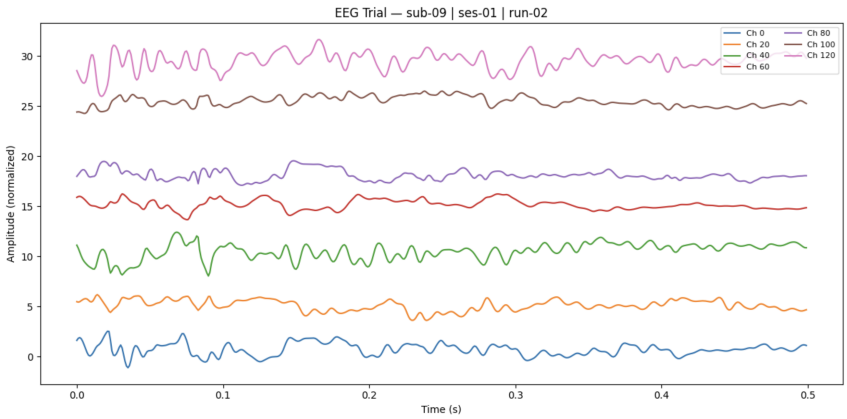

Figure 1: Sample Distribution of the EEG for subject 9, session 1, and run 2.

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[3] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[5] Y. Wang, J. Zhang, C. Li, H. Liu, X. Li, X. Wu, H. Zhao, W. Chen, and J. Liu. A multi-subject and multi-session eeg dataset for modelling human visual object recognition. *Scientific Data*, 12 (1):843, 2025.