Modeling human visual neural representations by combining vision deep neural networks and large language models

Boyan Rong [1,2,3], Alessandro Thomas Gifford [1], Emrah Düzel [2,3], Radoslaw Martin Cichy [1]

1 Department of Education and Psychology,
Freie Universität Berlin, Berlin, Germany
2 Institute of Cognitive Neurology and Dementia Research,
Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
3 German Center for Neurodegenerative Diseases (DZNE),
Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

Abstract:

Human visual perception involves transforming low-level visual features into high-level semantic representations. While deep neural networks (DNNs) trained on object recognition tasks have been promising models for predicting hierarchical visual processing, they often fail to capture higher-level semantic representations. In contrast, large language models (LLMs) encode rich semantic information that aligns with later stages of visual processing. Here we investigated whether combining vision DNN features and semantic embeddings from LLMs can better account for the neural dynamics of visual perception in electroencephalography (EEG) data. We demonstrated that their combination significantly improved the prediction of neural responses compared to either model alone. This approach outperformed multimodal modelling, and model comparison showed that the observed improvement was due to capturing complex information rather than a single factor.

Keywords: visual perception; encoding models; deep neural networks; language models; semantic processing

Introduction

Visual perception is a fundamental cognitive process involving multiple processing stages (Riesenhuber and Poggio, 1999). While DNNs trained on object recognition tasks have successfully captured the hierarchical visual representations (Cichy and Kaiser, 2019), they are limited in representing semantic information (Jozwik et al., 2023). In contrast, LLMs trained on vast corpora of texts represent semantic aspects (Doerig et al., 2024). On this basis, we hypothesized that integrating these complementary features from vision and language models would better explain the neural dynamics during visual perception than either model alone. To test this, we derived vision DNN representations from CORnet-S (Kubilius et al., 2019) and LLM representations from OpenAI's text-embedding-3-large model. We then trained encoding models predicting neural responses in EEG data from either model alone or their convex combination.

Methods

Encoding models of EEG visual responses

We used a large-scale EEG dataset (THINGS EEG2) of 10 participants viewing 16,740 naturalistic object images (Gifford et al., 2022). We trained linear encoding models to predict EEG responses. The encoding models differed by the model representation used as the regression basis (Fig. 1):

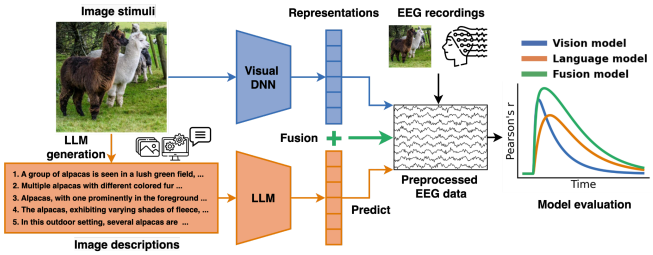

Figure 1: Model pipeline.

i) the vision DNN representations alone (termed **vision model**, in blue). The vision DNN representations were extracted using CORnet-S (Kubilius et al., 2019), a brain-inspired DNN. We extracted feature maps from the last layers of areas V1, V2, V4, IT, and decoder. Then we applied nonlinear PCA to reduce the representations to 1,000 dimensions.

ii) the LLM representations alone (termed **language model**, in orange). To obtain LLM representations, we first used GPT-4V (OpenAI et al., 2024) to generate five distinct versions of descriptions for each image. Then, these descriptions were converted to embeddings using OpenAI's text-embedding-3-large model and averaged across 5 versions. The embeddings were reduced to 1,000 dimensions using PCA.

iii) the convex combination of vision DNN and LLM representations together (termed **fusion model**, in green).

We evaluated model performance using Pearson correlation between predicted and actual EEG responses for each channel and time point. We calculated noise ceilings to estimate theoretical optimal performance given the noise level in the EEG data. To isolate the unique neural contributions associated with each feature type, we conducted partial correlation analyses between fusion model predictions and EEG responses, controlling for either vision DNN or LLM representations. We bootstrapped peak latencies (95% CI, 10,000 iterations).

Results

Combining vision DNN with LLM representations improves neural prediction with distinct spatiotemporal patterns for visual and semantic processing

The vision model (blue curve) peaks at 110 ms (105-115 ms) and the language model (orange curve) peaks at 365

ms (185-370 ms), with a significant latency difference of 255 ms (75-265 ms, $p < 0.001$). This is consistent with the expectation that the vision model predicts earlier visual processing stages, and the language model predicts later semantic processing stages. The fusion model (green curve) combined the advantages of both models, significantly outperforming either model alone (Fig. 2A).

Subtracting individual model predictions from the fusion model revealed distinct temporal contributions: improvements specific to the vision model (purple curve) peaked early at 90 ms (85 – 90 ms), while improvements specific to the language model (red curve) peaked later at 365 ms (360-400 ms) (275 ms (275 – 310 ms, $p < 0.001$ for latency difference) (Fig. 2B). Topographic analysis using partial correlations revealed that the vision model's unique contributions localized to medial occipito-parietal electrodes (Fig. 2C), whereas the language model's unique contributions revealed two distinct semantic processing stages: an early stage (peaks at 200 ms (195 – 260 ms)) in bilateral temporo-occipital electrodes and a later stage (peaks at 365 ms (360-385 ms)) involving additionally parieto-frontal electrodes (Fig. 2D).

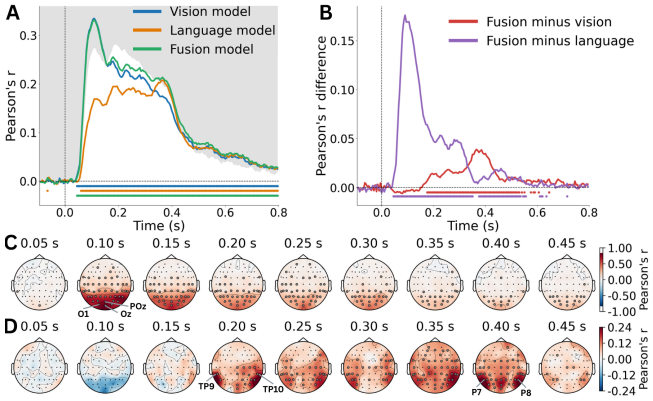

Figure 2: Performance evaluation. A) Pearson correlation **B)** Subtraction **C)** Partial correlation controlled for LLM representations **D)** Partial correlation controlled for DNN vision model representations

The fusion approach integrates complex semantic information and outperforms multimodal models

To determine the nature of the predictive information derived from LLMs we compared various fusion models using different language inputs. We found that the fusion model with full descriptions outperformed those using human-annotated or top 5 DNN-generated category labels (Fig. 3A), as well as those with limited linguistic inputs (noun-only, adjective-only, verb-only in descending order) (Fig. 3B). Those results demonstrated that our fusion model integrated complex semantic information, with most information coming from nouns.

We benchmarked our fusion model approach against CLIP (Radford et al., 2021). We found that our fusion model significantly outperformed CLIP encoding model when deriving features from the image or text encoder of CLIP, or fusing them together (Fig. 3C), suggesting that specialized models optimized for their respective feature types might extract more informative representations than joint training multimodal approaches.

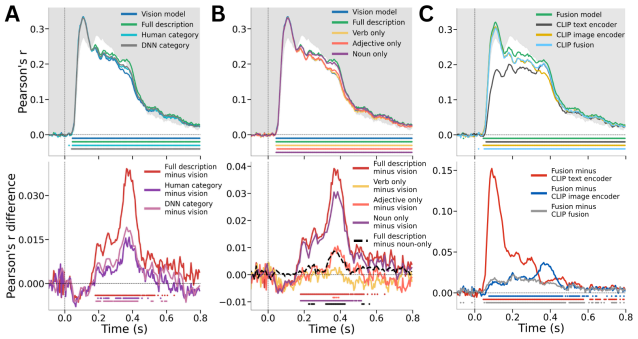

Figure 3: Model comparisons. A) Category-only **B)** Limited linguistic inputs **C)** CLIP

Discussion

Our fusion model demonstrated superior performance compared to unimodal models, confirming the complementary nature of visual and semantic information during visual perception (Enge et al., 2023). We observed distinct spatiotemporal neural dynamics for visual and semantic information processing. Visual effects emerged early (~90 ms) in medial occipito-parietal channels, consistent with feedforward early visual processing models (Cichy et al., 2014). Semantic effects occurred in two stages. Early semantic effects peaked near 200 ms in bilateral temporo-occipital electrodes. The timing and localization align with anterior temporal lobe (ATL) activation (Schendan and Maher, 2009; Clarke et al., 2015), indicating an early semantic activation stage. The late semantic pattern (~365 ms) parallels the N400 component (Kutas and Federmeier, 2011), suggesting a deeper semantic integration stage.

Acknowledgments

We thank the HPC Service of Zedat, Freie Universität Berlin (Bennett et al., 2020) for computing resources. This study was supported by the German Research Council (DFG) grants CI 241/1-3, CI 241/1-7 INST 272/297-2 (R.M.C.), INST

272/297-2 (E.D.), European Research Council (ERC) Consolidator grant (ERC-CoG-2024101123101) (R.M.C.) and a scholarship by the China Scholarship Council (B.R.).

References

Bennett, L., Melchers, B., & Proppe, B. (2020). Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin. http://dx.doi.org/10.17169/refubium-26754

Cichy RM, Kaiser D (2019) Deep Neural Networks as Scientific Models. Trends Cogn Sci 23:305–317.

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neurosci 17:455–462.

Clarke A, Devereux BJ, Randall B, Tyler LK (2015) Predicting the Time Course of Individual Objects with MEG. Cereb Cortex N Y NY 25:3602–3612.

Doerig A, Kietzmann TC, Allen E, Wu Y, Naselaris T, Kay K, Charest I (2024) Visual representations in the human brain are aligned with large language models. http://arxiv.org/abs/2209.11737

Enge A, Süß F, Rahman RA (2023) Instant Effects of Semantic Information on Visual Perception. J Neurosci 43:4896–4906.

Gifford AT, Dwivedi K, Roig G, Cichy RM (2022) A large and rich EEG dataset for modeling human visual object recognition. NeuroImage 264:119754.

Jozwik KM, Kietzmann TC, Cichy RM, Kriegeskorte N, Mur M (2023) Deep Neural Networks and Visuo-Semantic Models Explain Complementary Components of Human Ventral-Stream Representational Dynamics. J Neurosci 43:1731–1741.

Kubilius J, Schrimpf M, Kar K, Hong H, Majaj NJ, Rajalingham R, Issa EB, Bashivan P, Prescott-Roy J, Schmidt K, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ (2019) Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. http://arxiv.org/abs/1909.06161

Kutas M, Federmeier KD (2011) Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). Annu Rev Psychol 62:621–647.

OpenAI et al. (2024) GPT-4 Technical Report. http://arxiv.org/abs/2303.08774

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.

Schendan HE, Maher SM (2009) Object knowledge during entry-level categorization is activated and modified by implicit memory after 200 ms. NeuroImage 44:1423–1438.