Brain-Inspired Embedding Model: Scaling and Perceptual Fine-tuning

Stephen Chong Zhao (chong.zhao.1@vanderbilt.edu)
Data Science Institute, Vanderbilt University

Jason Lee (jason.j.lee@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Andrew Bender (abender@health.ucsd.edu)
Neurosciences Graduate Program, University of California San Diego

Trisha Mazumdar (trisha.mazumdar@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Adaline Leong (jia.yin.leong@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Prince Owusu Nkrumah (prince.owusu-nkrumah@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Mark Wallace (mark.wallace@vanderbilt.edu)
Psychology Department, Vanderbilt University

David A. Tovar (david.tovar@vanderbilt.edu)
Psychology Department, Vanderbilt University

Human perception is complex and multifaceted, making it challenging to quantify the subtle nuances and variations in how individuals perceive and categorize objects. To address this, we propose a novel brain-inspired mental embedding model called CLIP-HBA (Human Behavioral/Brain Analysis), leveraging the multimodal capabilities of the CLIP (Contrastive Language-Image Pre-training) architecture to create generalizable embeddings from human behavioral outputs and neural data. By fine-tuning the CLIP model with a 66-dimensional behavioral embedding derived from the SPoSE (Sparse Positive Similarity Embedding) model and the THINGs dataset, CLIP-HBA demonstrates improvements in behavioral and brain alignment compared to the original CLIP-ViT (Vision Transformer) model. The model's generalizability is validated through external magnetoencephalography (MEG) datasets, consistently outperforming CLIP-ViT in brain alignment. This work opens new avenues for creating personalizable embeddings specific to diverse populations.

Keywords: Object Perception; Embeddings; Transformers

Introduction

Quantifying the subtle nuances and variations in human perception has traditionally been challenging, relying predominantly on behavioral studies or direct brain data measurements. These methods, while informative, are often limited to a set number of participants or stimuli. To bridge this gap, we propose a novel approach leveraging the capabilities of a multimodal vision transformer model—specifically, the CLIP (Contrastive Learning Image Pre-training) architecture developed by OpenAI (Radford et al., 2021). Our model, CLIP-HBA (Human Behavioral/Brain Analysis), integrates multiple modalities (vision and text) to mirror the multifaceted nature of human perception. By leveraging the scalable and transferable nature of the transformer architecture (Pope et al., 2022), CLIP-HBA is designed to align closely with human behavioral outputs and neural data, providing a more comprehensive framework for understanding perceptual differences among individuals.

We fine-tuned the CLIP model with a 66-dimensional behavioral embedding obtained from a large-scale behavioral modeling - THINGS Database (Hebart et al., 2019). CLIP-HBA's performance was evaluated by assessing object dissimilarity alignment with validation behavioral data and comparing its alignment with brain signal cognitive responses, as measured by magnetoencephalography (MEG) data, against the original CLIP-ViT-L/14 (Vision Transformer) and the SPoSE model, a behavioral model trained on the large behavioral triplet studies from THINGS itself.

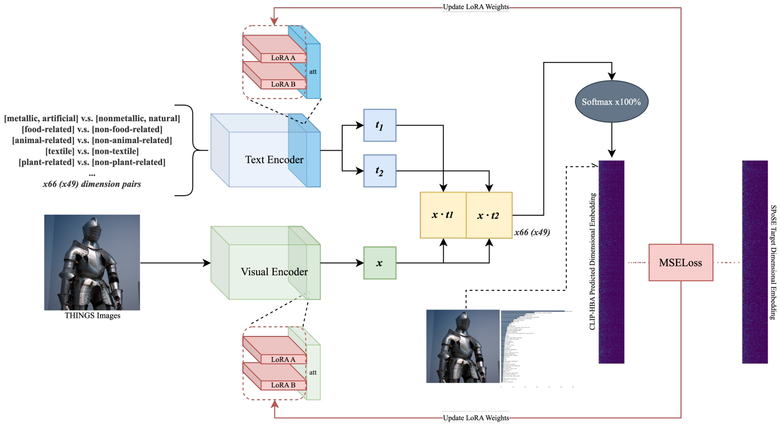
Update LoRA Weights
LoRA A
LoRA B
att
[metallic, artificial] v.s. [nonmetallic, natural]
[food-related] v.s. [non-food-related]
[animal-related] v.s. [non-animal-related]
[textile] v.s. [non-textile]
[plant-related] v.s. [non-plant-related]
...
x66 (x49) dimension pairs
Text Encoder
t_1
t_2
Softmax x100%
$x \cdot t1$
$x \cdot t2$
x66 (x49)
THINGS Images
Visual Encoder
x
CLIP-HBA Predicted Dimensional Embedding
MSELoss
SPoSE Target Dimensional Embedding
LoRA A
LoRA B
att
Update LoRA Weights

Figure 1: Schematic of the CLIP-HBA model training workflow. Fine tuning on the human mental embeddings occurs with the MSELoss function which computes the loss for backpropagation, affecting only the LoRa layers.

Methods

Our training procedure initializes using the model with the ViT L/14 backbone and freezing the entire CLIP model. We integrate LoRA layers into the attention output projection layers of the last layer of the transformer text encoder and the last two attention layers of the vision encoder (rank = 8, dropout rate = 0.1). Only the LoRA layers are unfrozen for updates during training (Hu et al., 2021).

The CLIP architecture processes text and image inputs. Each image input is paired with either 66 or 49 prompts, matching the dimensionality of the target SPoSE embedding (66d or 49d). Each prompt pair is named after a SPoSE model dimension. The image and text inputs are processed through their respective encoders, and the outputs are combined into a representational matrix for each image and dimensional pair, creating a series of n-dimensional binding matrices. During training, an MSELoss function calculates the loss, enabling backpropagation to update the LoRa layers in the targeted attention layers. This focused updating refines our model while maintaining computational and training efficiency.

Behavioral Validation

After fine-tuning the CLIP-HBA model, we conducted inference on a set of 48 object images from the THINGs database whose combination was fully sampled in a behavioral odd one out task (Hebart et al., 2020). These 48 objects were specifically excluded during training to prevent data leakage. We calculated the Pearson r correlation between our model's predicted Representational Dissimilarity Matrix (RDM) (Kriegeskorte et al., 2008) and the fully sampled behavioral RDM, achieving a 0.75 Pearson r correlation with a minimal p-value, in-

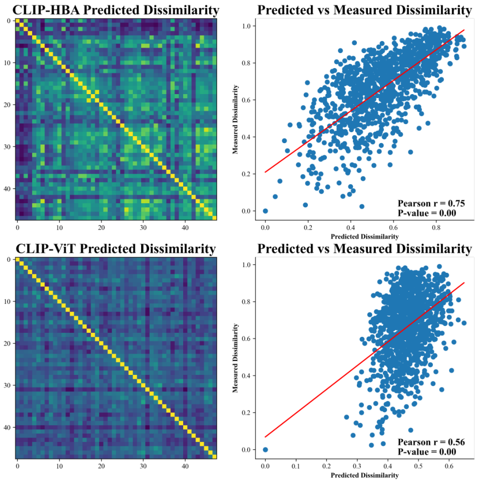

Figure 2: Outcomes of the behavioral validation for the CLIP-HBA model. Left panels show a 48x48 representational dissimilarity matrix (RDM) measured from CLIP-HBA (top) and CLIP-ViT (bottom) models. Right panels compare predicted dissimilarities to the measured data

dicating a 95% confidence interval (CI) of [0.72, 0.77]. We also ran the same process using the original CLIP-ViT model to establish a baseline for comparison. From this model, we extracted a 768-dimensional embedding for each image from the last layer activation (Muttenthaler and Hebart, 2021). The Pearson r correlation against the fully sampled behavioral RDM for CLIP-ViT was 0.56, with a 95% CI of [0.52, 0.60]. This comparison highlights the superior generalizability and behavioral alignment of the CLIP-HBA model, demonstrating approximately a 35% improvement over the baseline CLIP-ViT model.

Brain MEG validation

Initially trained on behavioral data, we explored CLIP-HBA for better alignment with brain data compared to CLIP-ViT. After fine-tuning, we used CLIP-HBA to analyze the 1854 THINGs image set, creating an 1854x1854 object RDM. We then computed time-resolved Spearman correlations with MEG RDMs from four participants exposed to the same 1854 visual stimuli(Hebart et al., 2023). This process was also applied to create a CLIP-ViT 1854x1854 RDM. A lower bound noise ceiling was calculated, and we compared the temporal correlations for the original SPoSE model embeddings. We assessed model alignment with MEG RDMs by calculating the Area under Curve (AUC) for each correlation line (Figure 3). Results show that CLIP-HBA's correlation pattern and AUC are similar to the SPoSE model, while CLIP-ViT shows lower correlations and AUC, indicating that fine-tuned CLIP-HBA improves both behavioral and neural alignment.

To demonstrate generalizability, CLIP-HBA was tested with three external datasets. Figure 3 B1 and B2 show MEG alignment results using data on neural responses to animate and inanimate objects under various image quality conditions (blurriness) (Grootswagers et al., 2017b). Figure 3B1 shows model correlation to 48 clear, monochromatic images of animate and inanimate objects, while Figure 3B2 shows correlation to the same objects with blurriness. In both analyses, CLIP-HBA produces a higher AUC than CLIP-ViT, with both models peaking later than the cross-subject correlation. Figure 3B3 validates our approach using MEG data from a study using color stimuli with backgrounds of humans, animals, fruits, and man-made objects (Grootswagers et al., 2017a), showing similar results with CLIP-HBA outperforming CLIP-ViT in brain alignments.

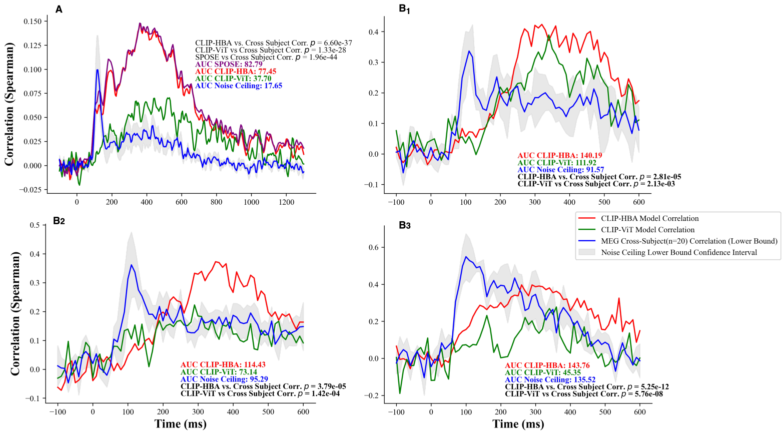

Figure 3: Model Brain Alignment Across MEG Data Sets. A) Temporal correlation of model-predicted object RDMs with THINGs MEG RDMs derived from 1854 visual stimuli. B1-B3) Displays the model correlations with external MEG datasets representing various conditions and object categories.

Discussion

Our findings show promising behavioral and brain alignment, even without training on brain data. This suggests that integrating brain data into our training process—using brain decoding accuracy matrices or RDMs as training targets—could further refine the model's alignment with human perception. As the model better scores dimensions akin to human judgment, it approaches a closer representation of human perceptual processes. This fine-tuning success opens avenues for representing perceptions across demographics, such as infants (Xie et al., 2022) or neurodivergent populations, allowing us to explore perceptual differences across diverse groups. Moving forward, we aim to refine our modeling techniques, expand our data scope, and create more scalable, representative behavioral and neural embedding models.

Acknowledgments

Vanderbilt IDD-Reads Award and the Vanderbilt Data Science Institute for GPU resources

References

Grootswagers, T., Contini, E., and Carlson, T. (2017a). Hyperalignment of dynamic responses using meg. In *Proceedings of the Organization for Human Brain Mapping Annual Meeting*, page 3548, Vancouver, Canada. Organization for Human Brain Mapping. Poster Session presented on June 28.

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., and Carlson, T. A. (2017b). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of cognitive neuroscience*, 29(12):1995–2010.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., and Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24.

Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Muttenthaler, L. and Hebart, M. N. (2021). Thingsvision: A python toolbox for streamlining the extraction of activations from deep neural networks. *Frontiers in neuroinformatics*, 15:679838.

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Levskaya, A., Heek, J., Xiao, K., Agrawal, S., and Dean, J. (2022). Efficiently scaling transformer inference.

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

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M., and Cichy, R. M. (2022). Visual category representations in the infant brain. *Current biology : CB*, 32(24):5422–5432.e6.