ReAlnet: Achieving More Human Brain-Like Vision via Human Neural Representational Alignment

Zitong Lu (lu.2637@osu.edu)
Department of Psychology, The Ohio State University
Columbus, OH 43210 USA

Yile Wang (yile.wang@utdallas.edu)
Department of Neuroscience, The University of Texas at Dallas
Dallas, TX 75080 USA

Julie D. Golomb (golomb.9@osu.edu)
Department of Psychology, The Ohio State University
Columbus, OH 43210 USA

Abstract:

Despite advancements in artificial intelligence, object recognition models still lag behind in emulating visual information processing in human brains. Recent studies have highlighted the potential of using neural data to mimic brain processing; however, these often rely on invasive neural recordings from non-human subjects, leaving a critical gap in understanding human visual perception. Addressing this gap, we present, for the first time, 'Re(presentational)Al(ignment)net', a vision model aligned with human brain activity based on non-invasive EEG, demonstrating a significantly higher similarity to human brain representations. Our innovative image-to-brain multi-layer encoding framework advances human neural alignment by optimizing multiple model layers and enabling the model to efficiently learn and mimic human brain's visual representational patterns across object categories and different modalities. Our findings suggest that ReAlnet represents a breakthrough in bridging the gap between artificial and human vision, and paving the way for more brain-like artificial intelligence systems.

Keywords: Object Recognition; Neural Alignment; Human Brain-Like Model

Introduction

While current vision models in artificial intelligence (AI) are advanced, they still fall short of capturing the full complexity and adaptability inherent in the human brain's information processing. Deep convolutional neural networks (DCNNs) have reached a performance level in object recognition that rivals human capabilities (Lecun et al., 2015), and many studies have identified representational similarities in the hierarchical structure between DCNNs and the ventral visual stream (Cichy et al., 2016; Güçlü & van Gerven, 2015; Kietzmann et al., 2019; Lu & Golomb, 2023; Yamins et al., 2014). However, the current alignment between DCNNs and human neural representations, while promising, still presents significant opportunities for further exploration and enhancement.. Enhancing the resemblance between visual models and the human brain has become a critical concern for both computer scientists and neuroscientists.

How can we leverage our understanding of the human brain to enhance current AI vision models? Previous models have limitations in emulating the complexity of the human brain's visual information processing, even with increased model depth and layers (Rajalingham et al., 2018). Researchers have attempted various strategies to improve AI models, including altering the model's architecture (Bai et al., 2017; Choi et al., 2023; Finzi et al., 2022; Han & Sereno, 2022; Kar et al., 2019; Kietzmann et al., 2019; Kubilius et al., 2019; Lee et al., 2020; Lu et al., 2023; Spoerer et al., 2017; Sun et al., 2017; Tang et al., 2018) and changing the training task (Konkle & Alvarez, 2022; O'Connell et al., 2023; Prince et al., 2023). However, limited studies have focused on directly using neural responses to complex visual information as feedback to improve the model's similarity to human brains.

Can we directly use human brain activity to align ANNs on object recognition and achieve more human brain-like vision models? Several recent studies have begun to let models learn neural representations, obtained from animal invasive neural recordings (mouse V1, monkey V1 or IT) (Dapello et al., 2023; Federer et al., 2020; Li et al., 2019; Pirlot et al., 2022; Safarani et al., 2021).

Here, we propose a more human brain-like vision model, ReAlnet, effectively aligned with human brain representations obtained from EEG recordings, based on a novel and effective encoding-based multi-layer alignment framework. Our representational alignment framework allows us to obtain personalized vision models by aligning with individuals' neural data. Moreover, the human brain-aligned ReAlnet shows improved similarity to human brain representations across different modalities (both human EEG and fMRI) and human behavior.

Methods

In this study, our core focus is to investigate whether aligning the model with individual human neural data can enhance the model's similarity to the human brain.

Alignment framework (ReAlnet training)

We applied a novel image-to-brain multiple-layer encoding alignment framework which lets the model not only accurately classify the object category, but also generate realistic EEG signals via minimizing both classification and generation losses during the training (Figure 1A). Based on this alignment framework, we build ten individual ReAlnets, using the state-of-the-art CORnet-S model (Kubilius et al., 2018, 2019) as the foundational architecture. Each ReAlnet, which has the same architecture as CORnet, is additionally trained on a real human subject's EEG signals, recorded while viewing a massive number of natural images from THINGS EEG2 (Gifford et al., 2022) *training* set.

Similarity measurement (ReAlnet testing)

EEG similarity: We employed an independent test dataset consisting of 200 images and associated EEG activity from the THINGS EEG2 *test* set. These test set images had not been presented at all during the training process, coming from entirely novel (untrained) object categories. For models (ReAlnet and COrnet), we input these 200 images to each model and obtain the feature vectors corresponding to each image for each layer in the model. Then we calculated the temporal similarity between different models and human brain EEG based on the representational similarity analysis (RSA) method.

fMRI similarity: Similar to EEG, we then evaluated the model's similarity to human brain fMRI representations (a completely different modality) from human subjects viewing novel image categories based on Shen fMRI (Shen et al., 2019) test set.

Behavioral similarity: We measured the similarity between the model and human behavior in several object recognition tasks using the Brain-Score platform (Schrimpf et al., 2020) based on two behavioral benchmarks.

Results

Compared to state-of-the-art CORnet, ReAlnet shows significantly higher similarity to human EEG neural dynamics for all visual layers when tested on novel images (Figure 1B). Notably, ReAlnet also produce significantly higher similarity to human brain patterns measured with fMRI (Figure 1C). Importantly, ReAlnet model representations are also significantly more similar to human behavior than original CORnet (Figure 1D). These results suggest that ReAlnet effectively learns not just the patterns of EEG data, but the brain's internal processing patterns of visual information. This leads to ReAlnet exhibiting a higher similarity than the original CORnet not only to within-modality EEG but also to cross-modality fMRI and behavior.

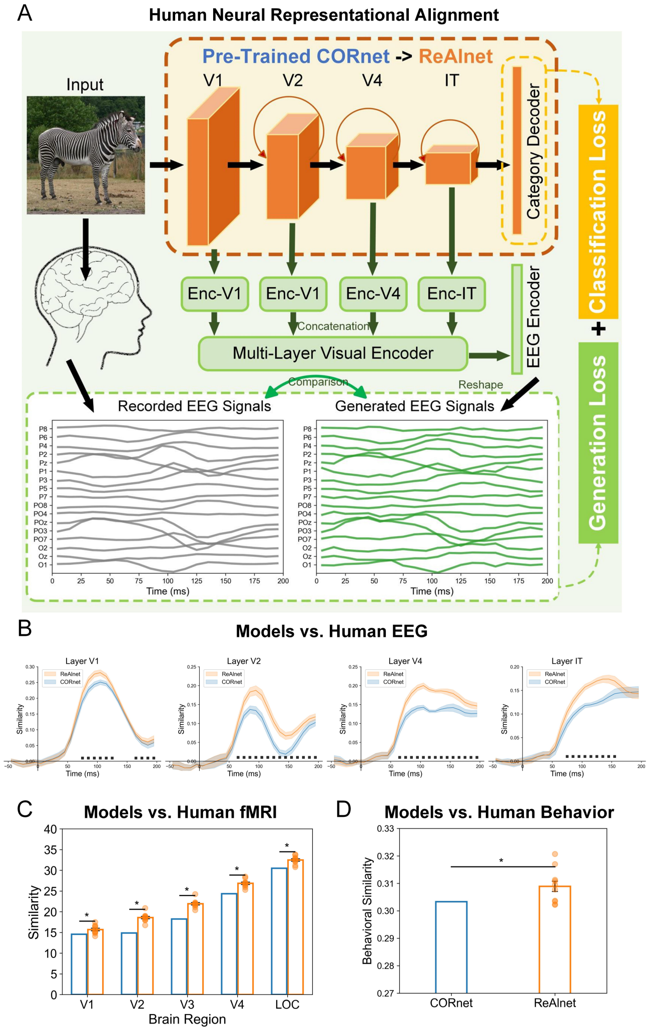

Figure 1: (A) An overview of ReAlnet alignment framework. Adding an additional multi-layer encoding module to an ImageNet pre-trained CORnet-S, the outputs contain the category classification results and the generated EEG signals. Using THINGS EEG2

training dataset, we aim to minimize both classification loss and generation loss, enabling CORnet to not only stabilize the classification performance but also effectively learn human brain features and transform into ReAInet. (B) Representational similarity time courses between human EEG and models for different layers respectively. Black square dots at the bottom indicate the timepoints where ReAInet vs. CORnet were significantly different (*p*<.05). Shaded area reflects ±SEM. (C) Representational similarity between models and human fMRI of five different brain regions based on one subject viewing natural images in Shen fMRI dataset. Asterisks indicate significantly higher similarity of ReAInet than that of CORnet (*p*<.05). Each circle dot indicates an individual ReAInet. (D) Similarity between models and human behavior based on the Brain-Score platform. Each circle dot indicates an individual ReAInet. Asterisks indicate significantly higher similarity of ReAInet than that of CORnet (*p*<.05).

Conclusion

Our study transcends traditional boundaries by employing a groundbreaking alignment framework that pioneers the use of human neural data to achieve a more human brain-like vision model, ReAInet. Demonstrating significant advances in bio-inspired AI, ReAInet not only aligns closely with human EEG and fMRI but also exhibits hierarchical individual variabilities and increased similarity to human behavior, mirroring human visual processing. We hope that our alignment framework stands as a testament to the potential synergy between computational neuroscience and machine learning and enables the enhancement of any AI model to be more human brain-like, opening up exciting possibilities for future research in brain-like AI systems.

Acknowledgments

This work was supported by research grants from the National Institutes of Health (R01-EY025648) and from the National Science Foundation (NSF 1848939).

References

Bai, S., Li, Z., & Hou, J. (2017). Learning two-pathway convolutional neural networks for categorizing scene images. *Multimedia Tools and Applications*, *76*(15), 16145–16162.

Choi, M., Han, K., Wang, X., Zhang, Y., & Liu, Z. (2023). A Dual-Stream Neural Network Explains the Functional Segregation of Dorsal and Ventral Visual Pathways in Human Brains. *Advances in Neural Information Processing Systems (NeurIPS)*, *36*. http://arxiv.org/abs/2310.13849

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13.

Dapello, J., Kar, K., Schrimpf, M., Geary, R. B., Ferguson, M., Cox, D. D., & DiCarlo, J. J. (2023). Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. *International Conference on Learning Representations (ICLR)*, *12*.

Federer, C., Xu, H., Fyshe, A., & Zylberberg, J. (2020). Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, *131*, 103–114.

Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K., & Grill-Spector, K. (2022). Topographic DCNNs trained on a single self-supervised task capture the functional organization of cortex into visual processing streams. *NeurIPS 2022 Workshop SVRHM*.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Han, Z., & Sereno, A. (2022). Modeling the Ventral and Dorsal Cortical Visual Pathways Using Artificial Neural Networks. *Neural Computation*, *34*(1), 138–171.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of*

Sciences of the United States of America, *116*(43), 21854–21863.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 1–12.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N. J., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., & Dicarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *BioRxiv*.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*.

Li, Z., Brendel, W., Walker, E. Y., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F. H., Pitkow, X., & Tolias, A. S. (2019). Learning from brains how to regularize machines. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*.

Lu, Z., Doerig, A., Bosch, V., Krahmer, B., Kaiser, D., Cichy, R. M., & Kietzmann, T. C. (2023). End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions. *ArXiv*.

Lu, Z., & Golomb, J. D. (2023). Generate your neural signals from mine: individual-to-individual EEG converters. *Proceedings of the Annual Meeting of the Cognitive Science Society 45*.

O'Connell, T. P., Bonnen, T., Friedman, Y., Tewari, A., Tenenbaum, J. B., Sitzmann, V., & Kanwisher, N. (2023). Approaching human 3D shape perception with neurally mappable models. *ArXiv*.

Pirlot, C., Gerum, R. C., Efird, C., Zylberberg, J., & Fyshe, A. (2022). *Improving the Accuracy and Robustness of CNNs Using a Deep CCA Neural Data Regularizer*. http://arxiv.org/abs/2209.02582

Prince, J. S., Alvarez, G. A., & Konkle, T. (2023). A contrastive coding account of category selectivity in the ventral visual stream. *BioRxiv*.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Safarani, S., Nix, A., Willeke, K., Cadena, S. A., Restivo, K., Denfield, G., Tolias, A. S., & Sinz, F. H. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems (NeurIPS)*, *34*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*(1), e1006633.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

Sun, T., Wang, Y., Yang, J., & Hu, X. (2017). Convolution Neural Networks With Two Pathways for Image Style Recognition. *IEEE Transactions on Image Processing*, *26*(9), 4102–4113.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., Hardesty, W., Cox, D., & Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8835–8840.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–

8624.