Ecological Data and Objectives for Human Alignment

Akash Nagaraj (akash_n@brown.edu)

Alekh Karkada Ashok (alekh_karkada_ashok@brown.edu)

Drew Linsley (drew_linsley@brown.edu)

Francis E Lewis (francis_lewis@brown.edu)

Peisen Zhou (peisen_zhou@brown.edu)

Thomas Serre (thomas_serre@brown.edu)
Carney Institute for Brain Science, Brown University
Providence, RI, USA

Abstract

As deep neural networks (DNNs) improve on object recognition benchmarks, their representations diverge from those used by human vision. We hypothesized this misalignment arises from the contrasting data and objectives used to train DNNs versus those that shape human visual development. To test this, we developed a framework for training DNNs on rich spatiotemporal image sequences of 3D objects to improve the alignment of DNNs with human vision by training with data and objective functions that more closely resemble those relied on by brains. We evaluated three training objectives: masked autoencoding (MAE), masked vision modeling (MVM), and ”causal vision modeling” (CVM), in which models predict future frames. Remarkably, CVM yielded DNN representations well-aligned with human 3D object recognition psychophysics. CVM-trained DNNs exhibited the same accuracy patterns and reaction time effects as humans for rotated objects. Representational analysis revealed that CVM causes DNNs to learn equivariance to out-of-plane transformations, explaining their human-like behavior. This provides a path towards reverse-engineering biological vision and developing artificial systems that better mimic the brain. Future work could further enrich the data and objectives to capture additional developmental principles shaping human vision.

Keywords: Human vision; spatiotemporal representation learning; representation alignment

Introduction

Deep neural networks (DNNs) have achieved remarkable success on object recognition (Shankar et al., 2020) and segmentation (Linsley, Kim, Ashok, & Serre, 2020) benchmarks on a massive computational scale. However, as DNN accuracy on these benchmarks improves, their representations and behaviors become increasingly misaligned with human vision (Fel*, Rodriguez*, Linsley*, & Serre, 2022; Linsley et al., 2023). For example, current DNNs rely on features that are very different from those diagnostic for humans, and their performance at predicting neural responses in the primate brain has stagnated relative to older models.

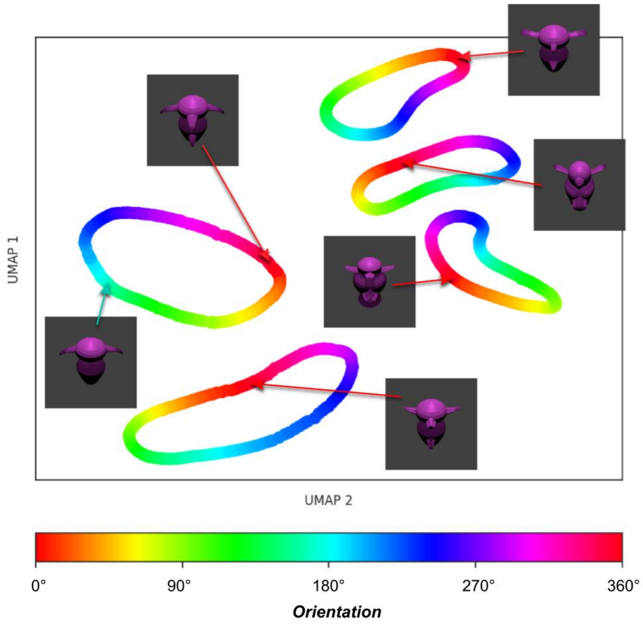

Figure 1: **CVM-trained DNNs learn equivariance to 3-Dimensional (out-of-plane) object transformations.** UMAP was used to decompose model-learned representations of objects into 2-dimensions, which revealed distinct ring-like manifolds for each object.

This growing divergence between DNNs and biological vision implies the current deep learning paradigm must be revised to reverse engineer the brain’s visual intelligence. One partial solution is constraining DNN representations to align with human behavioral data directly (Fel* et al., 2022). However, this approach requires extensive human experiments and does not provide insights into the developmental principles that shape human vision from childhood.

Here, we propose that ecological data diets and objective functions inspired by human visual development can induce DNNs to learn more human-like representations and behaviors. Specifically, we hypothesize that a key distinction missing in current DNN training is the rich spatiotemporal experiences humans accumulate through actively observing objects in the world, often without explicit supervision. We test whether DNNs trained on naturalistic object videos and objectives mirroring this developmental experience can better capture human visual intelligence.

We developed a framework for systematically testing the role of different data diets and objective functions on the representations learned by DNNs. We systematically evaluate how these DNNs best explain human behavior on popular psychophysics stimuli —'Greebles' (Gauthier & Tarr, 1997).

Underlying the alignment of CVM-trained DNNs are representations that exhibit smooth equivariance to 3-dimensional (out-of-plane) object transformations. This capability is not found in DNNs trained on the same data through any other means, showing that it is possible to align the visual behavior of DNNs with humans by constructing them using ecological data and objective functions.

Methods

Training Data Generation

To provide DNNs with more naturalistic visual experiences akin to human development, we generated a large dataset of spatiotemporal image sequences depicting 3D objects. We utilized neural radiance fields (Mildenhall et al., 2020) (NeRFs) trained on the Common Objects in 3D (Reizenstein et al., 2021) (CO3D) dataset to create photorealistic 3D models of 18,619 common objects across 50 categories previously trained and released as part of the PeRFception challenge (Jeong et al., 2022). For each object, we rendered a 50-frame video sequence by moving a virtual camera along a circular trajectory around the object.

The models were trained on short 4-frame snippets extracted from these longer video sequences, with a fixed number of skipped frames between each sampled frame. This allowed the models to be exposed to coherent spatiotemporal dynamics while still leveraging temporal invariance over various timescales.

Experimental Setup

We trained Vision Transformer (Dosovitskiy et al., 2021) (ViT) models on the generated video data. The ViT consisted of two components: 1) A 12-layer frame encoder that processed each individual 224x224 pixel image frame, splitting it into 16x16 pixel patches or "tokens." 2) An 8-layer spatial-temporal decoder that operated on the frame encoder outputs to ultimately reconstruct the target output image(s) during training. The encoder weights were shared across all input frames to learn spatial representations invariant to temporal dynamics.

Learning Objectives

To investigate different inductive biases, we trained ViT models with three distinct self-supervised training objectives:

- **Masked AutoEncoder (MAE):** Randomly masked a subset of image patches and trained the model to reconstruct the missing patches (He et al., 2021).

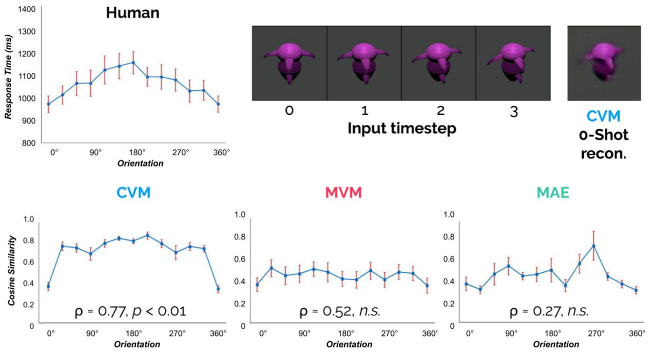

Figure 2: **A CVM-trained model's recognition confidence aligns with human reaction time in a psychophysics experiment.** Human participants were tested on their ability to identify 'Greebles' in various poses. Their reaction time grew as the objects were rotated further away from their canonical poses (Ashworth III et al., 2008). A CVM trained on naturalistic object sequences was able to reliably predict the pose of objects, and its recognition confidence strongly aligned with human reaction time. Neither MVM- nor MAE-trained models exhibited the same behavior.

- **Masked Vision Modeling (MVM):** Inspired by masked language modeling (Devlin, Chang, Lee, & Toutanova, 2018), we randomly mask an entire intermediate frame in the input and trained the model to reconstruct the masked frame.
- **Causal Vision Modeling (CVM):** Taking inspiration from causal language models (OpenAI, 2023), we masked the final frame in every sequence and trained to predict this future state based on the preceding frames.

For all objectives, the decoder was used only during training to reconstruct the masked image regions. For evaluations, we used the frame encoder representations, which captured the spatiotemporal dynamics in a parametric form.

Results

The results showed that the CVM-trained DNNs closely mimicked human perception in terms of recognition confidence and reaction time in psychophysics tasks involving Greebles. We tested the same effect in models trained with CVM, MVM, and MAE training objectives in three steps: *(i)* generated image sequences from a camera revolving around 15 greeble classes. *(ii)* stored each model's representation of the canonical view (at 0° orientation) of every greeble as a template. *(iii)* compared each model's representation of every other view of the Greebles to this stored template. We measured model recognition accuracy by assigning the class to the nearest template and the model reaction time as the cosine similarity of the template to all other views of each greeble. The CVM-trained model's accuracy was unrivaled (Human: 0.88, CVM: 0.64, MVM: 0.51 & MAE: 0.44) and had image representation dissimilarities significantly correlated with human reaction times ($p < 0.01$, Fig 2).

We next investigated why CVM-trained models were significantly more aligned with humans than any other model tested. To do this, we decomposed CVM-trained model representations of Greebles with UMAP into a 2-dimensional embedding to better interpret the structure it contains. Surprisingly, we found that the model grouped all images from any given Greeble into a manifold, in which camera orientations were ordered and linearly decodable (Fig 1). In other words, CVM-trained models learned equivariance to out-of-plane camera rotations during their training, and this equivariance transferred to the Greeble stimuli '0-shot' (*i.e.*, without additional training). Such structure is non-trivial, and we did not observe it in either MVM- or MAE-trained models.

Acknowledgements

This work was funded by ONR grant N00014-24-1-2026. We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program and computing hardware supported by NIH Office of the Director grant S10OD025181 through the Center for Computation and Visualization at Brown University.

References

Ashworth III, A. R., Vuong, Q. C., Rossion, B., & Tarr, M. J. (2008). Recognizing rotated faces and greebles: What properties drive the face inversion effect? *Visual Cognition*, *16*(6), 754–784.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Fel*, T., Rodriguez*, I. F., Linsley*, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Adv. Neural Inf. Process. Syst.*.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision research*, *37*(12), 1673–1682.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv:2111.06377*.

Jeong, Y., Shin, S., Lee, J., Choy, C., Anandkumar, A., Cho, M., & Park, J. (2022). Perfception: Perception using radiance fields. *Advances in Neural Information Processing Systems*, *35*, 26105–26121.

Linsley, D., Feng, P., Boissin, T., Ashok, A. K., Fel, T., Olaiya, S., & Serre, T. (2023, June). Adversarial alignment: Breaking the trade-off between the strength of an attack and its relevance to human perception.

Linsley, D., Kim, J., Ashok, A., & Serre, T. (2020). Recurrent neural circuits for contour detection. *International Conference on Learning Representations*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, *abs/2003.08934*. Retrieved from `https://arxiv.org/abs/2003.08934`

OpenAI. (2023, March). GPT-4 technical report.

Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., & Novotny, D. (2021). Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International conference on computer vision.*

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., & Schmidt, L. (2020). Evaluating machine accuracy on ImageNet. In H. D. Iii & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 8634–8644). PMLR.