Ecological data and objectives align deep neural network representations with humans

Akash Nagaraj, Alekh Karkada Ashok, Drew Linsley,
Francis E Lewis, Peisen Zhou, Thomas Serre
Carney Institute for Brain Science,
Department of Cognitive Linguistic & Psychological Sciences
Brown University, Providence, RI 02912
akash_n@brown.edu

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

The many successes of deep neural networks (DNNs) over the past decade have largely been driven by computational scale rather than insights from biological intelligence. Although DNNs have been surprisingly adept at explaining behavioral and neural recordings from humans, a growing number of reports indicate that DNNs are becoming progressively worse models of human vision as they improve on standard computer vision benchmarks. Here, we provide evidence that one path towards improving the alignment of DNNs with human vision is to train them with data and objective functions that more closely resemble those relied upon by brains. We find that DNNs trained to capture the causal structure of large spatiotemporal object datasets learn generalizable object representations that exhibit smooth equivariance to 3-dimensional (out-of-plane) variations in object pose and are predictive of human decisions and reaction times on popular psychophysics stimuli. Our work identifies novel data diets and objective functions that better align DNN vision with humans and can be easily scaled to generate the next generation of DNNs that behave as humans do.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in object recognition benchmarks through computational and data scale, achieving human-level performance on visual tasks ranging from object classification [1] to segmentation [2]. However, as the accuracy of DNNs on benchmarks has improved in recent years, the alignment of their representations, behaviors, and strategies with humans has decreased precipitously. For example, the most accurate DNNs today rely on features that have a very low correlation with those that humans find diagnostic for object recognition [3, 4] and are as (in)accurate at predicting responses to images evoked by neurons in the inferotemporal cortex as AlexNet [5]. This growing gap between human and DNN vision implies that the current deep learning paradigm needs to be revised to have any hope of reverse engineering biological vision and creating artificial vision systems that can see, behave, and process information like humans.

A partial solution to the misalignment of DNNs and human vision systems is the 'neural harmonizer:' a constraint on DNN optimization that forces a model's image representations to align with those used by human observers to classify the same images [3]. Despite the efficacy of the neural harmonizer for improving DNN alignment with humans visual decisions [3], representations [3], adversarial robustness [4] and neural predictions [5], the method is limited in several fundamental ways. First,

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

the neural harmonizer relies on large behavioral datasets to constrain DNN representations, and these datasets are difficult to collect and scale to the training regimes that have yielded the most accurate models to date. Second, while the neural harmonizer is proof that DNN alignment can be improved with humans without hurting object classification accuracy, it does not get us closer to understanding the developmental principles that shape human vision. Identifying such principles would advance our basic understanding of human vision and offer an eminently scalable way of fixing the alignment problem.

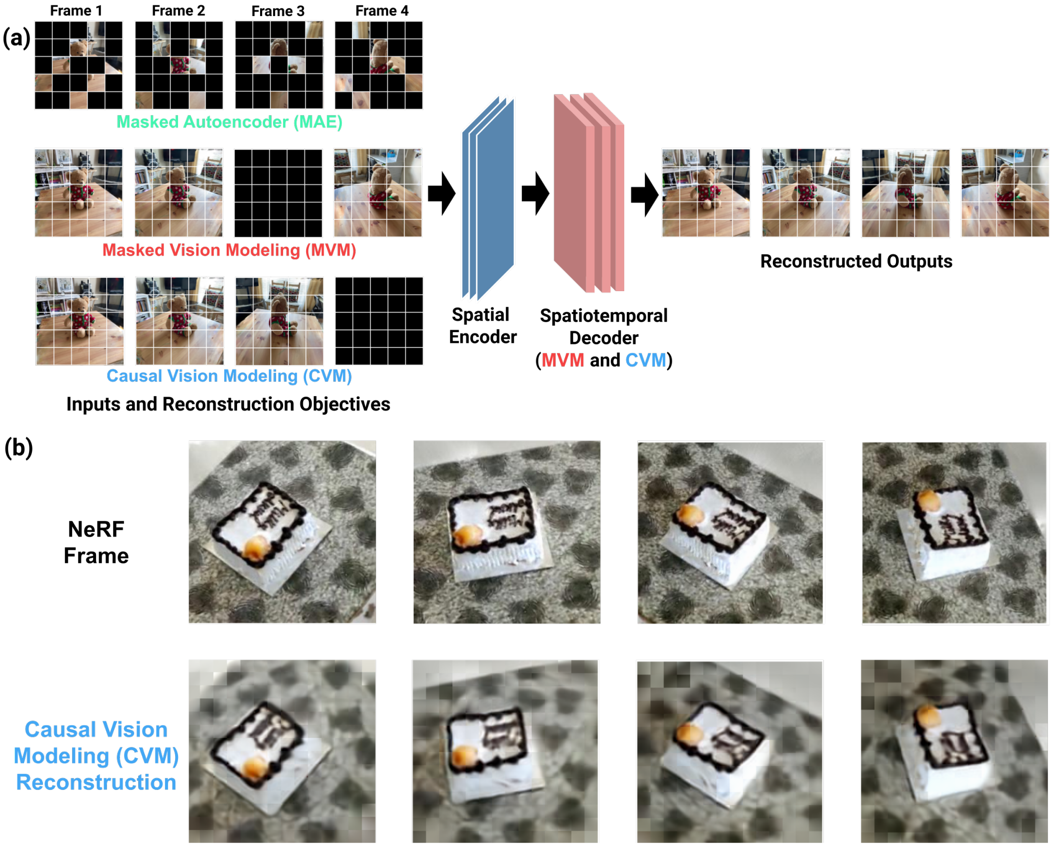
(a)
Frame 1
Frame 2
Frame 3
Frame 4
Masked Autoencoder (MAE)
Masked Vision Modeling (MVM)
Causal Vision Modeling (CVM)
Inputs and Reconstruction Objectives
Spatial Encoder
Spatiotemporal Decoder (MVM and CVM)
Reconstructed Outputs
(b)
NeRF Frame
Causal Vision Modeling (CVM) Reconstruction

Figure 1: **A framework for investigating the effects of ecological data and objective functions on deep neural network (DNN) representations.** (**a**) We train Visual Transformers (ViTs) on spatiotemporal data generated with neural radiance field (NeRF) object models. Each input consists of four frames captured by a camera following a radial trajectory around an object (*e.g.*, a teddy bear, as shown). The ViT consists of an encoder, with weights shared across frames (*i.e.* encoding 2-Dimensional features), and a decoder, with spatiotemporal weights (*i.e.* decoding across space and time). Models are trained on one of three objectives (denoted by the colors), where a part of the image is masked from the input. (1) Causal Vision Modeling (CVM): predict the next frame. (2) Masked Vision Modeling (MVM): predict the intervening frame. (3) Masked Autoencoder (MAE) [6]: predict the content of masked patches in an image. The decoder is only used for training, and the encoder representations of images are used at test time. (**b**) DNNs trained with CVM learn to reconstruct a future frame in a sequence accurately.

Contributions. In this work, we investigate whether ecological data diets and behavioral objectives can shape DNNs to produce more human-like representations and behavior. We focus our efforts on a basic distinction between how state-of-the-art DNNs and humans learn to recognize objects: While DNNs learn to recognize objects through datasets containing millions or billions of object images and explicit categorical supervision, humans do the same by observing objects as they move through the world and often learning about them without explicit supervision. We hypothesize that this difference between the data diets and objective functions of humans vs. DNNs is a key factor driving the growing misalignment of DNNs.

- We developed a framework for systematically testing the role of different data diets and objective functions on the representations learned by DNNs. We generate rich, naturalistic, spatiotemporal image sequences and instruct DNNs to learn from these through a variety of well-controlled objective functions that focus models on orthogonal aspects of the data.
- We discover that DNNs best explain human behavior on popular psychophysics stimuli (‘Greebles’ [7]) when trained to predict the next state of an object — an objective which we refer to as ‘Causal Vision Modeling’ (CVM).
- 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.
- It is therefore possible to align the visual behavior of DNNs with humans by construction through ecological data and objective functions.

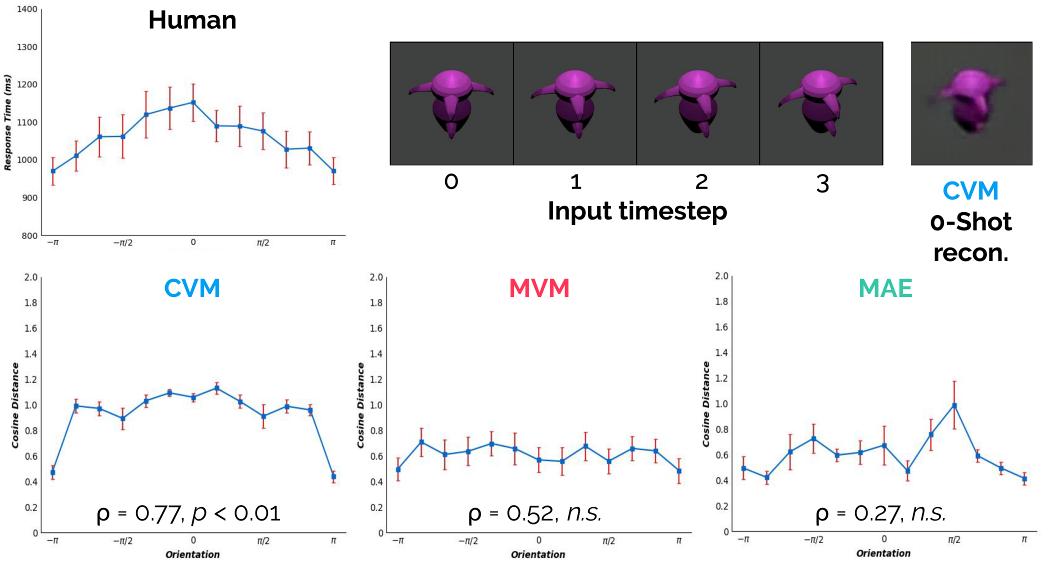

Figure 2: **A CVM-trained model’s recognition confidence aligns with human reaction time in a psychophysics experiment.** Human participants were tested on their ability to identify ‘Greebles’ in various poses. Their reaction time grew as the objects were rotated further away from their canonical poses [8]. A CVM trained on naturalistic object sequences was able to predict the pose of objects reliably, and its recognition confidence strongly aligned with human reaction time. Neither MVM- nor MAE-trained models exhibited the same behavior.

2 Methods

Training datasets. We hypothesized that the internet data diets used to train DNNs today are one of the reasons why these models are growing progressively less aligned with human vision. To address this problem, we devised an approach to generate unbounded amounts of rich spatio-temporal object image data, which we thought might capture similar kinds of experiences that humans have with objects. Specifically, we turned to neural radiance fields [9] (NeRFs) to build 3-dimensional models of individual objects, then created sequences from images taken by a virtual camera as it revolved around the object.

We used NeRFs trained on the Common Objects in 3D (CO3D [10]) dataset that was previously released as part of the PeRFception challenge [11]. Unlike that challenge, however, we investigate the performance of models trained on spatio-temporal sequences of images instead of random views of objects. Our dataset contained 18,619 NeRFs of common objects from 50 MS-COCO [12] categories. We rendered a 50-frame video of each object. Models were trained on randomly selected chunks of four frames from this sequence with a predetermined number of skipped frames in between selected frames.

Models We trained multiple instances of a modified Vision Transformer (ViT) [13] on the spatiotemporal image data generated from NeRFs. The ViT consisted of two parts: a 12-layer frame encoder operating on 224 × 224 pixel images and an 8-layer spatial-temporal decoder [14] that operated on the outputs of the encoder and ultimately generated an image-sized output. Each frame passed into the encoder was split into patches, or 'tokens,' of size 16 × 16 pixels (Fig. 1). The decoder was only used for training and discarded for the experiments discussed in Results.

Objective functions. Inspired by the success of masked auto-encoding approaches for images [15], videos [14] and language [16, 17], we realized that it is possible to investigate a multitude of objective functions by asking DNNs to solve different reconstruction tasks. This work focuses on just three that resemble popular objectives used in machine learning today or are speculated to be important for biological learning. (*i*) The popular masked auto-encoding (MAE) [15] where a proportion of image patches are randomly masked. (*ii*) Masked vision modeling (MVM), inspired by the popular BERT objective [16] from language modeling, where an entire intermediate frame in a sequence is masked. (*ii*) Causal vision modeling (CVM), inspired by the objective of causal language models popularized by the GPT family of models [18], where the final frame in a sequence is masked (Fig 1a).

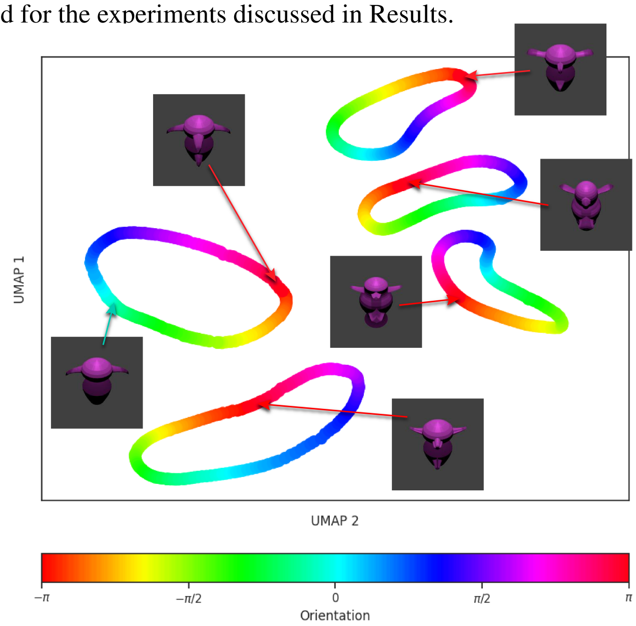

Figure 3: **CVM-trained DNNs learn equivariance to 3-dimensional (out-of-plane) object transformations**. UMAP was used to decompose ViT-encoder representations of objects into 2 dimensions, which revealed distinct ring-like manifolds for each object.

3 Results

Human psychophysics. After training models with the same hyperparameters as [19], we tested the alignment of these models with human behavior on images of 'Greebles:' a popular dataset that has been used to investigate the sensitivity of human recognition capabilities to 3-dimensional (out-of-plane) rotations [8]. In those experiments, it was found that human recognition accuracy worsened and reaction time increased as objects were rotated further away from their canonical, front-facing view.

We tested the same effect in models trained with CVM, MVM, and MAE training objectives. We did this in three steps: First, we generated image sequences from a camera revolving around 15 greeble classes. Next, we stored each model's representation of the canonical view of every greeble as a template. Third, we 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 (Fig 2). 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.

Representational analysis 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. 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.

Acknowledgments

This work was funded by ONR grants N00014-19-1-2029 and N00014-24-1-2026. The computing hardware was supported in part by NIH Office of the Director grant #S10OD025181 via the Center for Computation and Visualization (CCV) at Brown University.

References

[1] Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., Schmidt, L.: Evaluating machine accuracy on ImageNet. In Iii, H.D., Singh, A., eds.: Proceedings of the 37th International Conference on Machine Learning. Volume 119 of Proceedings of Machine Learning Research., PMLR (2020) 8634–8644

[2] Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al.: Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences **115**(24) (2018) 6171–6176

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

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

[5] Linsley, D., Rodriguez, I.F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., Serre, T.: Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. (June 2023)

[6] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv:2111. 06377

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

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

[9] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. CoRR **abs/2003.08934** (2020)

[10] Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: International Conference on Computer Vision. (2021)

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

[12] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer (2014) 740–755

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[14] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35** (2022) 10078–10093

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

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

[17] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. (2018)

[18] OpenAI: GPT-4 technical report. (March 2023)

[19] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022) 16000–16009