Let’s disagree to agree: Model identifiability through disagreeability

Brian Cheung (cheungb@mit.edu)[1]
Center for Brains, Minds, and Machines,
Massachusetts Institute of Technology

Erin Grant (erin.grant@ucl.ac.uk)[1]
Gatsby Unit & Sainsbury Wellcome Centre,
University College London

Helen Yang (helenzy@mit.edu)
Center for Brains, Minds, and Machines,
Massachusetts Institute of Technology

Boris Katz (boris@csail.mit.edu)
InfoLab,
Massachusetts Institute of Technology

Tomaso Poggio (tp@csail.mit.edu)
Center for Brains, Minds, and Machines,
Massachusetts Institute of Technology

[1] Equal contribution

Abstract

Recent advancements in artificial intelligence (AI) have led to the development of vision systems that closely resemble biological visual systems in terms of behavior and neural recordings. However, there is increasing empirical evidence that the representations learned by such systems at scale are *convergent*: AI systems trained on large datasets tend to learn similar representations despite differences in architecture and training procedure. This lack of *identifiability* via representation and behavior presents a challenge to comparison pipelines commonly used to validate AI systems as models of biological vision, as it limits the ability to reason about the unique computational properties of an individual model. We call for a renewed focus on the stimuli that serve as the input to these pipelines and demonstrate that, for standard naturalistic image datasets used to pre-train and validate vision systems, there are a minority of stimuli that cause maximal disagreement among AI systems even if these systems achieve a high degree of agreement with the target function. We address the identifiability challenge by systematically exploring the narrowed space of these *contrastive stimuli* in order to provide the necessary signal to adjudicate between large-scale AI systems as models of biological vision.

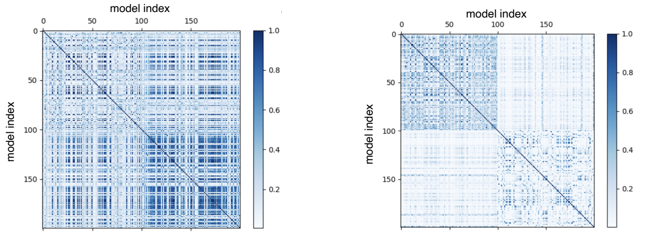
Keywords: deep learning; representational similarity; system identification; neural network architecture

Introduction

The representations learned by artificial intelligence (AI) systems tend to be more similar at scale, making it difficult to discern the unique computational properties of individual systems based solely on their behavior or representations (Han et al., 2023). This representational convergence is a natural consequence of the increasing level of complexity required for machine learning algorithms to solve an ever-broadening array of tasks that are directly analogous to tasks accomplished by their biological counterparts (Cao & Yamins, 2021; Huang et al., 2021; van Rossem & Saxe, 2024). This positions AI systems as essential subjects for comparative analysis.

Standard approaches for comparing AI and biological vision, such as representational similarity analysis (RSA; Kriegeskorte et al., 2008), centered kernel alignment (CKA; Kornblith et al., 2019; Song et al., 2012), and linear regression (Schrimpf et al., 2020), compute a similarity between model and neural activations aggregated over a standardized set of naturalistic or semi-naturalistic stimuli. These approaches may thus be fundamentally limited in their ability to reveal the computational differences between highly optimized AI systems with significant representational overlap (Conwell et al., 2022; Han et al., 2023).

To enable more informative comparisons between AI and biological vision systems, we require techniques that can provide identifiability of individual models even in the regime of convergent representations. In this work, we propose leveraging *disagreement* between high-performing models to generate more sensitive probes for model differences. Even when models converge in overall representation, we demonstrate that they still exhibit subtle but informative differences in their predictions and activations for specific stimuli and task conditions. By focusing on these points of maximal divergence, we generate a targeted set of stimuli that reveals the unique signatures of individual models, allowing us to map the computational differences between AI and biological vision systems.



(a) ImageNet stimuli.

(b) Synthetic stimuli.

Figure 1: Model-to-model similarity matrix using CKA over **(a)** ImageNet and **(b)** synthetically engineered stimuli (see main text). Models at indices 0-99 and 100-199 are convolutional and attention, respectively. Synthetic stimuli separate layer motifs, seen here as block-diagonal structure.

Furthermore, searching for *contrastive stimuli*—stimuli that maximize disagreement among models—does not necessarily require ground truth annotations, broadening the applicability of this approach beyond standard supervised image classification benchmarks. For many tasks, the presence of disagreement itself can reveal common failure cases among models or incongruencies in stimuli (*e.g.*, hard examples). Disagreement highlights how models see the world differently, not just whether they see the world correctly.

Proof-of-concept

We first follow the hypothesis-driven paradigm of manually engineering stimuli based on known architectural differences between model classes. Layers in a convolutional neural network (Fukushima, 1988) have a local receptive field, whereas those of attention layers in vision transformers (Vaswani et al., 2023) are global. We design synthetic stimuli to exploit this difference (images of squares placed apart at random distances). We evaluated the representational similarity of 200 models (100 convolutional, 100 attention) over this synthetic data and stimuli from ImageNet (Deng et al., 2009). Our synthetic stimuli (**Figure 1b**) separate layer motif significantly better than the naturalistic ImageNet image set (**Figure 1a**).

Methods

While synthetic data, as exemplified by the proof-of-concept above, has been successful in revealing differences between the representations of biological and artificial systems (Geirhos et al., 2019), engineering synthetic benchmarks to test individual hypotheses is difficult to scale to the

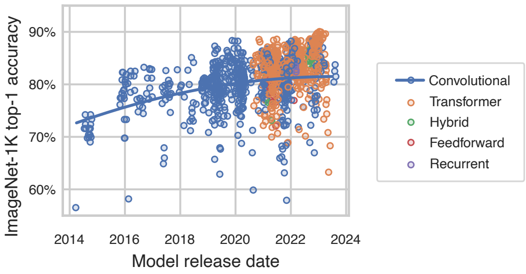


Figure 2: Task performances against release dates among the inventory. The lowest and highest performers are AlexNet and a large-scale self-supervised vision transformer, EVA-02. Quadratic line-of-best-fit is shown for convolutional models.

ever-expanding repository of AI systems. Instead, we propose a method for automated search for naturalistic stimuli that elicit maximal disagreement amongst a candidate set of models. Unlike past work in stimulus optimization (Golan et al., 2020), we target real-world stimuli rather than synthetic or artificially-perturbed stimuli. Naturalistic stimuli elicit naturalistic representations and behaviors, which are of primary interest when validating AI systems against biological vision.

We collect the largest inventory of vision models evaluated for behavioral and neural predictivity to date: 1355 models with distinct parameters across a variety of architectures, training objectives, and pre-training and fine-tuning datasets (from small- to large-scale); see **Figure 2** and refer to Krizhevsky et al. (2012) and Fang et al. (2023) for details of the AlexNet and EVA-02 architectures. We emphasize the variation in performance and architecture class as it allows us to test prior assertions about task performance, architecture, and neural predictivity at scale (cf. Conwell et al., 2022) and to relate these conclusions to model identifiability.

Measuring agreement

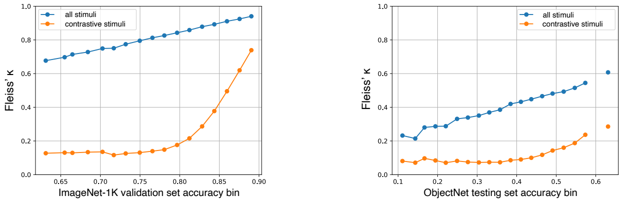
We derive overall and per-stimulus agreement measures from Fleiss' κ (Fleiss, 1971). Let T be a matrix of size $n_{\text{stimulus}} \times n_{\text{cat}}$, where T_{ij}, represents the number of models that predict the categorical response j for stimulus i. Then

$$n_{\text{total}} = \sum_{i=1}^{n_{\text{stimulus}}} \sum_{j=1}^{n_{\text{cat}}} T_{ij} \quad \text{and} \quad n_{\text{rater},i} = \sum_{j=1}^{n_{\text{cat}}} T_{ij}$$

are the total number of category j ratings for stimulus i and the total number of raters (models with a response) for stimulus i, respectively. For each stimulus i, we then compute the per-stimulus agreement as:

$$p_{\text{agree},i} = \frac{\sum_{j=1}^{n_{\text{cat}}} T_{ij}^2 - n_{\text{rater},i}}{n_{\text{rater},i} \cdot (n_{\text{rater},i} - 1)} \,, \tag{1}$$

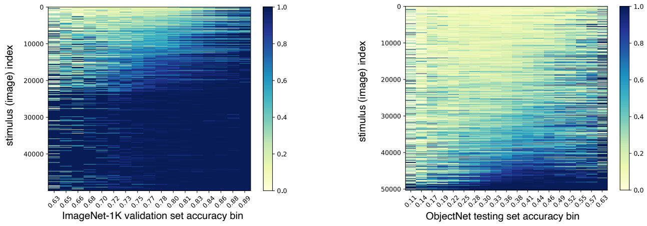
which ranges from 0 (low) to 1 (high) agreement.



(a) ImageNet-1K validation.

(b) ObjectNet testing set.

Figure 3: Overall agreement (Fleiss' κ) for models binned by ImageNet-1K validation set accuracy over all (blue) and contrastive (1000 lowest-agreement) stimuli (orange); see **Figure 4** for disaggregation by stimulus. The contrastive stimuli exhibit lower agreement across model performance levels.



(a) ImageNet-1K validation.

(b) ObjectNet testing set.

Figure 4: Per-stimulus agreement (Eq. 1) sorting stimuli (images) from low (yellow) to high (blue); see **Figure 3** for aggregation. The most contrastive stimuli are contained subsets of the full dataset, seen here as continuity of the yellow region.

Results

Figs. 3a and 3b compare agreement with the full dataset of stimuli versus a set of contrastive stimuli (here, the 1000 images with the lowest agreement across all models) for the ImageNet validation set and the ObjectNet testing set (Barbu et al., 2019). Contrastive stimuli persist for a wide range of system performance levels, even while higher accuracy imposes a floor on agreement (since correct predictions must be in agreement; see Geirhos et al., 2020, for related discussion).

Figs. 4a and 4b display the more fine-grained measure of per-stimulus agreement (Eq. 1), evidencing a wide range of agreeability for different stimuli (images) and datasets (ImageNet *vs.* ObjectNet). **Figure 4b** demonstrate that some stimuli in ObjectNet remain significantly disagreeable for all model sets regardless of performance level constraint.

Conclusion

In this work, we have proposed a novel approach for enhancing the identifiability of AI systems as models of biological vision by leveraging stimulus disagreement. This method holds potential not only for distinguishing between different computational models in artificial intelligence, but also as a process to generate stimulus candidates for subsequent neural response collection in biological studies.

References

Barbu, A, Mayo, D, Alverio, J, Luo, W, Wang, C, Gutfreund, D, Tenenbaum, J, & Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems*, *32*.

Cao, R, & Yamins, D. (2021). Explanatory models in neuroscience: Part 2—constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*.

Conwell, C, Prince, JS, Kay, KN, Alvarez, GA, & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03.

Deng, J, Dong, W, Socher, R, Li, LJ, Li, K, & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Fang, Y, Sun, Q, Wang, X, Huang, T, Wang, X, & Cao, Y. (2023). Eva-02: A visual representation for neon genesis.

Fleiss, JL. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, *1*(2), 119–130.

Geirhos, R, Meding, K, & Wichmann, FA. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

Geirhos, R, Michaelis, C, Wichmann, FA, Rubisch, P, Bethge, M, & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; Increasing shape bias improves accuracy and robustness. *Proceedings of the 7th International Conference on Learning Representations*.

Golan, T, Raju, PC, & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337.

Han, Y, Poggio, T, & Cheung, B. (2023). System identification of neural systems: If we got it right, would we know? *Proceedings of the 40th International Conference on Machine Learning*.

Huang, T, Zhen, Z, & Liu, J. (2021). Semantic relatedness emerges in deep convolutional neural networks designed for object recognition. *Frontiers in Computational Neuroscience*, *15*.

Kornblith, S, Norouzi, M, Lee, H, & Hinton, G. (2019). Similarity of neural network representations revisited. *International conference on machine learning*, 3519–3529.

Kriegeskorte, N, Mur, M, & Bandettini, PA. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Krizhevsky, A, Sutskever, I, & Hinton, GE. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Schrimpf, M, Kubilius, J, Hong, H, Majaj, NJ, Rajalingham, R, Issa, EB, Kar, K, Bashivan, P, Prescott-Roy, J, Geiger, F, Schmidt, K, Yamins, DLK, & DiCarlo, JJ. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like?, 407007.

Song, L, Smola, A, Gretton, A, Bedo, J, & Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, *13*(5).

van Rossem, L, & Saxe, AM. (2024, February 14). *When representations align: Universality in representation learning dynamics*. arXiv: 2402.09142.

Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Gomez, AN, Kaiser, L, & Polosukhin, I. (2023). Attention is all you need. *Advances in Neural Information Processing Systems*.