Evaluating and supervising vision models with multi-level similarity judgments

Frieda Born [†, 1,2,3] **Lukas Muttenthaler** [†, 2,3,4] **Klaus Greff** [4] **Thomas Unterthiner** [4]

Andrew K. Lampinen [4] **Klaus-Robert Müller** [2,3,4,5,6] **Michael C. Mozer** [4]

[1] Max Planck Institute for Human Human Development, Berlin, Germany
[2] Machine Learning Group, Technische Universität Berlin, Germany
[3] BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[4] Google DeepMind
[5] Department of Artificial Intelligence, Korea University, Seoul
[6] Max Planck Institute for Informatics, Saarbrücken, Germany
[†] Equal contribution.

Abstract

Vision foundation models are becoming increasingly pervasive. Despite their incredible success, it remains unclear to what degree they see the world the way humans do. A growing body of recent work investigates the alignment between human and model representations but has not systematically characterized this alignment across levels of conceptual abstraction. Here, we attempt to bridge this gap and collect a large human similarity judgment dataset of triplet odd-one-out choices on three levels of semantic abstraction: *coarse-grained*, *fine-grained*, and *class-boundary*. This multi-level behavioral dataset enables more nuanced comparisons between humans and computer vision models than has previously been possible. Models and people are best aligned on class-boundary and worst aligned on coarse-grained similarity judgments. Human alignment with various model types depends on the level of abstraction: image/text models match people best for superordinate categories, but self-supervised image models match best for fine-grained semantic categories. Our dataset facilitates the evaluation—and potentially the improvement—of vision foundation models.

Keywords: representation learning, similarity, object concepts, human behavior

Introduction

Vision foundation models excel at various object recognition and image segmentation tasks, often reaching human-level performance (Dosovitskiy et al., 2021; Dehghani et al., 2023; Radford et al., 2021). However, their representations tend to be different from human representations. For example, they are less sensitive to the semantic hierarchy of object categories that humans follow to judge object similarity (Peterson et al., 2018; Attarian et al., 2020; Roads & Love, 2021; Muttenthaler, Dippel, et al., 2023; Muttenthaler, Linhardt, et al., 2023). Ideally, a model's representational similarity space would reflect this semantic hierarchy—for example, representing semantically-related concepts (like cats and dogs) more closely than less similar ones (like cats and strawberries). However, while today's computer vision models can identify objects with high accuracy, their representation space does not fully reflect the semantic structure that shapes human concept spaces (Muttenthaler, Linhardt, et al., 2023). This mismatch may impair both the downstream applications of these models in computer vision and their utility as models for cognitive neuroscience.

Recent research has shown that correcting for *coarse-grained semantic* misalignment in the representations of vision models improves their downstream task performance and makes the representations more human-like (Muttenthaler, Linhardt, et al., 2023). Another work shows that supervising vision models with human similarity judgments about objects from the same category can improve their nearest neighbor and image retrieval performance (Fu et al., 2023), thus correcting their *fine-grained semantic* similarity structure. Hence, supervision signals about human fine-grained semantic object similarity can improve model representations in a complementary way to signals about human coarse-grained semantic object similarity.

Here, we attempt to assess model-human semantic agreement more thoroughly. We collect a multi-level human similarity judgment dataset that enables the simultaneous evaluating *coarse-grained semantic*, *fine-grained semantic*, and *class-boundary* object similarity. Assessing varying granularities extends current approaches, such as standard supervised learning, which predominantly probes representational boundaries of clearly defined categories. Thus, our method enables the evaluation of vision models on broader conceptual links between different categories and nuanced variations within a class, probing more closely the complex semantic structure of human concept spaces. Complementarily, participants' reaction times (RTs) provide an additional measure for evaluating (and potentially training) models on human uncertainty signals.

Methodology

To collect human similarity judgments, we conducted a behavioral online experiment with participants ($N = 450$) recruited through Prolific (https://www.prolific.ac/). In a triplet odd-one-out task (Figure 1A), participants are shown three images and are asked to select the image that is the least similar, i.e., the odd-one-out. Our experiment consists of $n = 330$ trials, including diverse classes of triplets. We used triplets with three levels of abstraction: *coarse-grained semantic*, which comprised

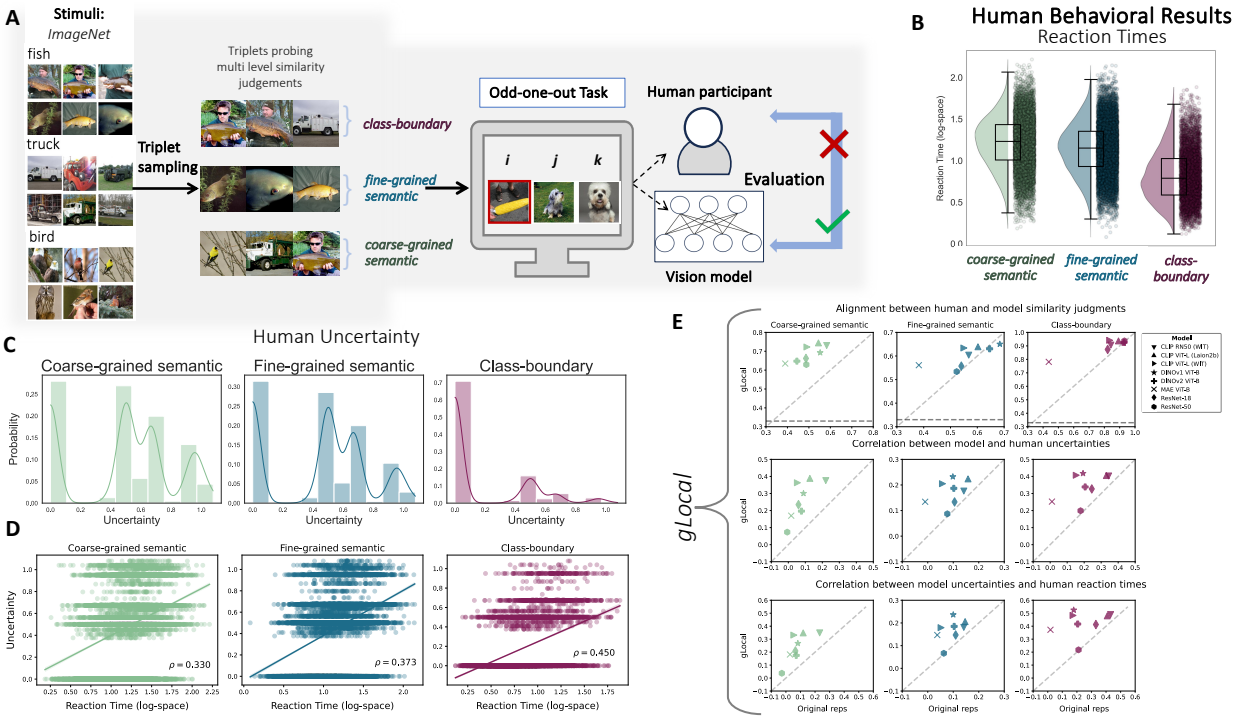

Figure 1: Data collection of multi-level human similarity judgments, model evaluation, and behavioral human results. A. Pipeline of triplet sampling, data collection, and model evaluation., B. RT (in log-space) across the three triplet types, C. Uncertainty distributions across the three triplet classes., D. Correlation of uncertainties with RTs, E. Model analysis about alignment between human and model similarity judgments (top), correlation between model uncertainties and human uncertainties (middle), and correlation between model and human RTs (bottom)

three images from three different categories; *fine-grained semantic*, showing three images from the same category; and *class-boundary*, with two images from the same and one from a different category. Instead of randomly sampling triplets—which would reproduce dataset biases—we stratified sampling by superclasses. ImageNet classes follow the WordNet hierarchy (Deng et al., 2009; Russakovsky et al., 2015), which includes higher-level classes. For example, all dog breeds can be summarized as a single dog superclass. To avoid presenting dogs, birds, and other fine-grained classes that are overrepresented in ImageNet more frequently to the participants than other categories, we grouped the ImagNet classes into 717 coarse-grained WordNet superclasses. We followed a Latin Square Design (LSD)(Grant, 1948) to counterbalance triplet presentation both within each participant and across the entire sample. This method minimizes confounds that might bias results, by balancing variations in triplet types (e.g., fine-grained semantic vs. coarse-grained semantic), presentation timing (e.g., start vs. end of the experiment), and ImageNet classes (such as dogs or birds).

For each triplet, we collected five responses, to measure human consistency. We used these responses to estimate a response probability distribution for each triplet and calculated the discrete (Shannon) entropy of the distribution as a measure of the variability or uncertainty of participant responses. Higher entropy indicates stronger disagreement between participants. Additionally, we computed the arithmetic mean of log RTs across participants for each triplet as a measure of latency (see Figure 1; B-D).

Results and Conclusion

Behavior. We collected a behavioral dataset that systematically sampled human image similarity judgments at multiple levels of abstraction. This framework enabled us to explore human similarity judgments at both global and local scales. To capture the degree of (dis-)agreement between participants, we analyzed the variability of human responses across the three triplet types. Entropy distributions across all triplet classes are below the chance level of $\log(3)$, indicating high consistency.

Participants showed the greatest agreement on class boundary triplets, indicating a strong consensus in identifying the odd-one-out when images distinctly diverged in global semantic features (e.g., animate vs. inanimate objects). In contrast, both coarse-grained semantic triplets—spanning broader, often unrelated categories—and fine-grained semantic triplets (e.g., different dog breeds) showed higher uncertainty. This indicates that in situations where participants have to navigate the similarity space without overt semantic distinctions, they may employ diverse, individualized strategies to make a choice. Results from the RT analysis (Figure 1B) corroborate these findings. Participants reacted faster in the class-boundary than in the fine-grained and coarse-grained settings.

Model analysis. We analyze three different classes of models: *supervised*, *self-supervised*, and *image/text contrastive* models.

From each class, we select two representative models. We find that most models are weakly aligned out-of-the-box. We therefore attempt to align their representation spaces using gLocal (Muttenthaler, Linhardt, et al., 2023)—a recent method for transforming a neural network representation into a space that is more aligned with human similarity judgments while preserving a representation's local similarity structure, to avoid trading off downstream task performance. We find that gLocal does improve alignment overall — most substantially for coarse-grained semantic triplets. Models whose zero-shot alignment is poor benefit more. ImageNet-trained supervised models are the most poorly aligned and benefit the most from applying gLocal to their representation space, across all triplet types. We find that self-supervised image models achieve the strongest alignment with human judgments for fine-grained and class-boundary triplets. In contrast, image/text models show better alignment with human similarity judgment for coarse-grained semantic triplets (see Figure 1E).

Conclusion. Neural-network vision models have become pervasive in multiple aspects of our daily lives, and are increasingly used as models in cognitive neuroscience (Yamins & DiCarlo, 2016). Thus, it appears increasingly important to understand whether these models make judgments that are in line with those of humans. To facilitate this goal, we collected a large similarity judgment dataset on multiple levels of abstraction. Our behavioral analysis of human responses shows that decision entropy varies across participants depending on the level of abstraction and whether judgments pertain to global or local semantic similarities. In cases where category boundaries are either broadly disparate or finely nuanced, the decision variability observed in our dataset highlights its potential for probing neural network models across multiple levels of abstraction. Ascertaining this variability is the first step towards aligning these models with the diverse and nuanced strategies that humans appear to use when assessing image similarities.

Acknowledgements

FB, LM, and KRM acknowledge funding from the German Federal Ministry of Education and Research (BMBF) for the grants BIFOLD22B and BIFOLD23B. We thank Katherine Hermann and Ishita Dasgupta for helpful comments on an earlier version of the manuscript.

References

Attarian, I. M., Roads, B. D., & Mozer, M. C. (2020). Transforming neural network visual representations to predict human judgments of similarity. In *Neurips 2020 workshop svrhm.*

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., ... Houlsby, N. (2023). Scaling vision transformers to 22 billion parameters. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 7480–7512). PMLR.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

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. In *International conference on learning representations.*

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 50742–50768). Curran Associates, Inc.

Grant, D. A. (1948). The latin square principle in the design and analysis of psychological experiments. *Psychological bulletin*, *45*(5), 427.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Human alignment of neural network representations. In *11th international conference on learning representations, ICLR 2023, kigali, rwanda, mai 01-05, 2023.* OpenReview.net.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A., & Kornblith, S. (2023). Improving neural network representations using human similarity judgments. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 50978–51007). Curran Associates, Inc.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between Deep Neural Networks and Human Representations. *Cogn. Sci.*, *42*(8), 2648–2669. doi: 10.1111/cogs.12670

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, 18–24 Jul). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 8748–8763). PMLR.

Roads, B. D., & Love, B. C. (2021). Enriching imagenet with human similarity judgments and psychological embeddings. In *IEEE conference on computer vision and pattern recognition, CVPR 2021, virtual, june 19-25, 2021* (pp. 3547–3557). Computer Vision Foundation / IEEE. doi: 10.1109/CVPR46437.2021.00355

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Li, F. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252. doi: 10.1007/s11263-015-0816-y

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.