Evidence for Shepard's Law in the Representational Spaces of Deep Vision Models

Daniel L. Carstensen (daniel_carstensen@brown.edu)
Department of Cognitive and Psychological Sciences, Brown University

Steven M. Frankland[1] **(steven.m.frankland@dartmouth.edu)**
Program in Cognitive Science, Dartmouth College

Serra E. Favila[1] **(serra_favila@brown.edu)**
Department of Cognitive and Psychological Sciences, Brown University

[1] Denotes equal contribution

Abstract

Shepard's (1987) universal law of generalization states that generalization strength decays as a concave function of stimulus distance in psychological space. While widely supported in biological systems, its relevance to artificial neural networks remains unclear. We tested this law across 26 diverse deep vision models using human similarity judgments of naturalistic images. Across models, embedding distances produced concave generalization gradients and aligned closely with human psychological spaces. To examine the role of semantic content, we analyzed model gradients across network depth and compared gradient shapes to human-derived benchmarks. Language-aligned models most closely resembled human data, suggesting semantic representations contribute to model-human alignment. Our findings extend Shepard's law to modern artificial systems, providing further evidence for its universality. They also highlight deep vision models as compelling proxies for psychological space, providing a novel framework for assessing representational alignment between artificial and human cognition.

Keywords: artificial neural networks; embedding spaces; generalization; perceptual similarity; representational alignment

Introduction

Generalization is a fundamental challenge for information processing systems. Consider a bird that preys on a bumblebee and is stung; to avoid repeating this, it must generalize from this experience to identify similar organisms. This necessity for generalization, however, can be exploited, as seen with harmless hoverflies mimicking bumblebees (Edmunds & Reader, 2014). Similarly, artificial systems like convolutional neural networks face the challenge of generalization when classifying novel images, needing to learn useful representations from varied pixel-level inputs. The broad scope of this challenge motivates the search for unifying principles, notably Shepard's (1987) universal law of generalization. Shepard proposed that generalization strength between two stimuli decays as an invariant, concave function of their distance in "psychological space". While not directly observable, this space is often studied using non-metric multidimensional scaling (NMDS) on similarity data (Shepard, 1962). Extensive empirical work supports Shepard's law in living organisms, including recent studies using naturalistic images (Ghirlanda & Enquist, 2003; Marjieh et al., 2024; Shepard, 1987).

If truly universal, this law should also apply to artificial information processing systems. Although theoretical accounts support this universality (Chater & Vitányi, 2003; Frank, 2018; Shepard, 1987; Sims, 2018; Tenenbaum & Griffiths, 2001), prior empirical tests in artificial systems are scarce and inconclusive (Rustom, Öğmen, & Yazdanbakhsh, 2022; Serrano & Miralles, 2023). Deep neural networks (DNNs) demonstrate high performance in computer vision (LeCun, Bengio, & Hinton, 2015; Kheradpisheh et al., 2016; Russakovsky et al., 2015) and exhibit representational parallels with biological vision (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Muttenthaler et al., 2023; Rajalingham et al., 2018; Sucholutsky et al., 2023; Yamins et al., 2014). Importantly, prior work has shown that the representational geometry of DNN embedding spaces is predictive of human similarity judgments and can even be finetuned (Jha, Peterson, & Griffiths, 2023; Peterson, Abbott, & Griffiths, 2016, 2018). Nevertheless, DNNs substantially differ from human visual cognition in a number of ways, including their respective strategies in object recognition tasks (Bowers et al., 2023; Linsley et al., 2023). Hence, it remains unclear if the internal embedding spaces of DNNs can serve as viable models for psychological space.

Here, we evaluate this possibility, specifically asking whether distances in the embedding spaces of DNNs predict human similarity judgments via a concave generalization gradient. We selected diverse deep vision models and utilized a large dataset of natural images with human similarity judgments (Peterson et al., 2018). By extracting image embeddings and computing pairwise distances, we obtained model-derived distances corresponding to human similarity scores, allowing examination of the resulting generalization gradients. To our knowledge, no prior study has tested Shepard's law across many modern deep vision architectures under large-scale, naturalistic conditions. We present evidence that the internal representations of deep vision models adhere to the universal law of generalization, validating Shepard's law as universal and providing a novel alignment measure between neural network and human representations.

Methods & Results

Dataset and Model Selection

We examined whether internal embedding spaces of deep vision models follow Shepard's law, which states that generalization strength decreases as a concave function of stimulus distance in psychological space. We used a naturalistic image dataset collected by Peterson et al. (2018), comprising human similarity ratings (reported from 0-10 and then scaled to 0–1) for six diverse categories containing 120 images each: animals, automobiles, fruits, furniture, various, vegetables. Prior analyses using NMDS verified Shepard's law for a subset of these category sets (Marjieh et al., 2024). We replicated these results and extended them to the remaining category sets. Then, we selected 26 pretrained vision models covering a diverse set of model families that varied in architecture, training task and data, and parameter count (CLIP, DINO, DINOv2, DreamSim, Open CLIP, ResNet, SimCLRv2, ViT, VGG) to assess representational generalization broadly. We additionally chose pixel-level MSE as a baseline model of low-level perceptual distance.

Embedding Extraction and Gradient Computation

We extracted image embeddings from each model's final hidden layer, selecting classification tokens for transformer-based architectures. Then, we computed the pairwise cosine

distances between embeddings and normalized them to 0-1.[2] Finally, we matched the human-evaluated similarity score for each image pair with its associated cosine distance in model embedding space, yielding a similarity-distance tuple. Given sparse sampling at high similarity levels, we grouped the similarity-distance tuples into 100 equal bins and computed one averaged similarity-distance value within each bin to stabilize gradient estimates.

Curve Fitting and Evaluation

To quantitatively test adherence to Shepard's law, we fit four curve types (linear, quadratic, exponential, Gaussian) to the binned generalization gradients using a 5×5-fold cross-validation procedure. Nonlinear fits consistently outperformed linear fits across models (see Fig. 1 for example). The Gaussian curve yielded the best fits overall, with significantly lower error and higher explained variance ($R^2 = 0.856$, RMSE $= 0.065$, ΔBIC $= -66.9$ compared to linear). Furthermore, quadratic curves independently confirmed concavity (positive second derivatives) in 84% of fits. These results strongly support Shepard's law across models. Notably, no nonlinear curve reliably provided improved fits across all image datasets for gradients computed using the pixel-level MSE model.

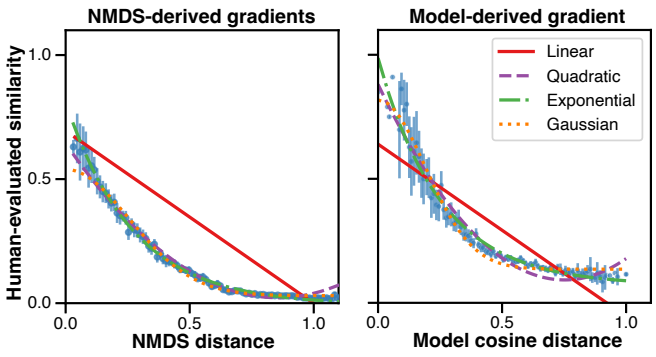

Figure 1: Example gradients for the "fruits" image set. Both the NMDS-derived gradient (left) and the CLIP ViT-B/16-derived gradient (right) clearly show concavity.

Psychological Space Alignment

To determine if embedding spaces aligned structurally with psychological space, we regressed human similarity-derived NMDS distances onto corresponding model-derived cosine distances. Regression analyses, combined across models via random-effects meta-analysis, showed strong representational alignment, with model distances explaining 87.1% of variance (pooled slope $= 0.802$, intercept $= 0.080$, $r^2 = 0.871$). Thus, model embeddings proved effective proxies for psychological spaces underlying human judgments. The pixel-level MSE model did not exhibit high predictive strength, with only 8% variance explained.

Influence of Semanticity

To assess the influence of semantic content on generalization gradients, we first evaluated generalization gradients across all VGG11 layers. Strong nonlinear curve fit improvements only reliably emerged in the final four hidden layers, reflecting diminished influence of low-level perceptual features and potentially increased semantic content relevant to human categorization (Cohen et al., 2020; LeCun et al., 2015; Sucholutsky et al., 2023). To further quantify the role of semanticity, we conducted bootstrapping (10,000 samples, with replacement) for each model-derived gradient and the NMDS-derived gradient, fitting exponential curves to each bootstrap sample.[3] Comparing exponential coefficient distributions (a, b; excluding offset c due to its weaker relevance for gradient shape) via KL divergence, language-aligned models CLIP and Open CLIP consistently ranked among the models closest to human data, suggesting possible contributions of semantic alignment to generalization patterns. However, gradients based solely on word embeddings of category labels (e.g., "Tiger") exhibited concavity but substantially weaker fits and lower regression alignment, indicating that purely semantic category structure may be relevant but insufficient for fully capturing human similarity structure.

Conclusion

We demonstrated that embedding spaces from diverse deep vision models adhere robustly to Shepard's universal law of generalization, with human-evaluated similarity declining as a concave function of model-derived stimulus distance. This alignment holds across multiple model architectures and training paradigms, highlighting that modern artificial systems share fundamental representational properties with human psychological spaces. Moreover, we found that semantic representations may play an important role in this alignment. Our approach provides a novel framework for assessing the representational alignment between deep neural networks and human psychological spaces, opening new avenues for interpretability and cognitive model evaluation.

References

Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., ... et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. doi: 10.1017/S0140525X22002813

Chater, N., & Vitányi, P. M. B. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

[2]While prior work has mainly relied on Euclidean distance as a measure of distance in psychological space (Marjieh et al., 2024; Shepard, 1962, 1987), we used cosine distance in our analyses because it is standard practice in computer vision and has been shown to predict human similarity judgments (Fu et al., 2023; Radford et al., 2021; Roads & Love, 2021). Nevertheless, we also performed all analyses using Euclidean distance, yielding comparable results.

[3]Here, we chose the exponential curve since it best fit the NMDS-derived gradients (Marjieh et al., 2024).

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), 27755.

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, *11*(1), 746.

Edmunds, M., & Reader, T. (2014). Evidence for batesian mimicry in a polymorphic hoverfly. *Evolution*, *68*(3), 827–839.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9803–9806.

Fu, S., Tamir, N. Y., Sundaram, S., Chai, L., Zhang, R., Dekel, T., & Isola, P. (2023). DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 50742–50768). Curran Associates, Inc.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting Low-Dimensional psychological representations from convolutional neural networks. *Cognitive Science*, *47*(1), e13226.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Computational Biology*, *10*(11), e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, *6*(1), 32672.

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

Linsley, D., Rodriguez, I. F. R., FEL, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2023). Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In *Thirty-seventh conference on neural information processing systems.*

Marjieh, R., Jacoby, N., Peterson, J. C., & Griffiths, T. L. (2024). The universal law of generalization holds for naturalistic stimuli. *Journal of Experimental Psychology: General*, *153*(3), 573–589.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Human alignment of neural network representations. In *The eleventh international conference on learning representations.*

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). *Adapting deep network features to capture psychological representations.*

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.

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.

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with human similarity judgments and psychological embeddings. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 3546-3556). IEEE Computer Society.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Rustom, F. B., Öğmen, H., & Yazdanbakhsh, A. (2022). *Object detection, recognition, deep learning, and the universal law of generalization.* arXiv.

Serrano, C., & Miralles, D. (2023). Internal representation and Shepard's law in an artificial haptic system. In *Artificial Intelligence Research and Development - Proceedings of the 25th International Conference of the Catalan Association for Artificial Intelligence* (Vol. 375, pp. 48–58). IOS Press.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219–246.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... Griffiths, T. L. (2023). *Getting aligned on representational alignment.* arXiv.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40; discussion 652–791.

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.