From Pixels to Human Typicality Judgments: Disentangling Category Structure and Neural Network Representations

Itay Inbar
Department of Cognitive & Brain Sciences
Ben Gurion University of the Negev

Tal Golan
Department of Cognitive & Brain Sciences
School of Brain Sciences and Cognition
Data Science Research Center
Ben-Gurion University of the Negev

Abstract

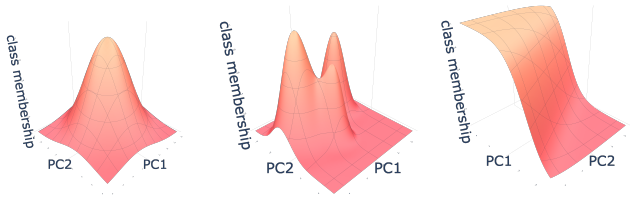
Humans can consistently rate the typicality of objects with respect to basic-level categories, but what do these ratings reveal about the computational mechanisms underlying categorization? We evaluated human typicality judgments against predictions from image-computable models. Each model paired a vision transformer (ViT), trained on one of five tasks, with one of three category structure models—prototype, exemplar, or a linear decision-bound model. This yielded 15 models systematically varying in representational and category structure assumptions. We found that predictions from a prototype model using the representations of a ViT trained on image classification aligned most closely with human judgments. However, this model's advantage over the alternatives was not consistently significant, and its performance remained well below the leave-one-subject-out noise ceiling. Simulations showed that although some models were statistically indistinguishable in prediction accuracy, all 15 made distinct predictions. We discuss experimental design considerations that may enable stronger comparisons among these alternative models.

Introduction

Humans categorize visual stimuli with remarkable speed and ease. Yet the computational processes underlying this perceptual function remain unresolved. The problem is twofold: What is the representational space in which categories are defined, and how are the categories structured within it?

Advances in deep learning have produced a diverse range of candidate representational spaces, but it remains unclear which of them, if any, aligns closely with human categorization (Rajalingham et al., 2018; Battleday, Peterson, & Griffiths, 2020; Golan, Raju, & Kriegeskorte, 2020). As for category structure, two longstanding theories continue to compete: **Prototype theory** holds that for each category, people estimate a mean representation of previously encountered instances and classify new stimuli by comparing them to the resulting prototypes; **Exemplar theory**, by contrast, argues that people store individual category members and classify new stimuli by comparing them to these previously encountered exemplars. Distinct from both theories, artificial neural network classifiers apply multinomial regression to derive class probabilities, a process more akin to the **linear decision bound model** (Ashby & Maddox, 1993).

Typicality judgments, which reflect graded category membership (Rosch & Mervis, 1975), may provide clues about the underlying computational mechanisms. Lake, Zaremba, Fergus, and Gureckis (2015) reported a correlation between human typicality judgments and neural network classification confidence, measured as either logits or softmax outputs. Adapting the design of Battleday et al. (2020), who considered classification boundaries, we compare human typicality judgments to neural network predictions, independently manipulating two factors: (1) the representational model, drawn from contemporary machine learning, and (2) the structure of categories within this space, contrasting the prototype, exemplar, and linear decision bound models (Fig. 1).



(a) Prototype (b) Exemplar (c) Decision bound

Figure 1: Illustration of three alternative category structure models within a shared representational space. **(a)** Prototype model: class membership peaks at the category mean. **(b)** Exemplar model: class membership peaks in regions with many nearby category members. **(c)** Linear decision bound model: membership increases along a weighted feature axis, favoring extreme over central exemplars.

Methods

Behavioral data. We reanalyzed the *ViSpa* typicality experiment (Günther, Marelli, Tureski, and Petilli, 2023, Experiment 3), in which participants were shown sets of five images from the same category and selected the most and least typical items. For each of 1500 categories, a single image set was presented, receiving responses from 30 to 33 participants.[1]

Class membership modeling. We implemented the cognitive theories using mathematical models naturally aligned with each. Prototype theory was realized through Linear Discriminant Analysis (LDA), assuming an isotropic Gaussian distribution centered at the mean features of each category. Exemplar theory was realized through Kernel Density Estimation (KDE) with an isotropic Gaussian kernel. The linear decision-bound model was realized through multinomial regression. Each model was fit to the visual features of the ImageNet-21K examples corresponding to the categories included in ViSpa. The embeddings were obtained from the penultimate layer of one of five Vision Transformers (ViTs) trained on different visual tasks: SLIP, CLIP, SimCLR, and DINOv2 (all from Meta FAIR), and ImageNet-1K classification (Google, pre-trained on ImageNet 21K). This resulted in 15 fitted models (three per ViT). To standardize feature dimensionality, all feature spaces were reduced to 500 dimensions using principal component analysis (PCA), fit separately on 2,000 ImageNet-21K categories not included in the ViSpa experiment.

Response selection modeling. For each set of five ViSpa images, we transformed model-predicted class membership scores ($\log P(\text{image} \mid \text{class})$ for LDA and KDE, classifica-

[1] We retained 785 categories in which all five items were unequivocal members of the named basic-level category, based on visual inspection. The reported results do not qualitatively change when this selection procedure is omitted.

tion logits for multinomial regression) into response probabilities using the maximum-difference model (Marley & Louviere, 2005). Specifically, for each of the 20 possible combinations of an item chosen as most typical and an item chosen as least typical, we subtracted the score of the least typical item from that of the most typical item. The resulting 20 logits were softmax-normalized to obtain response probabilities.

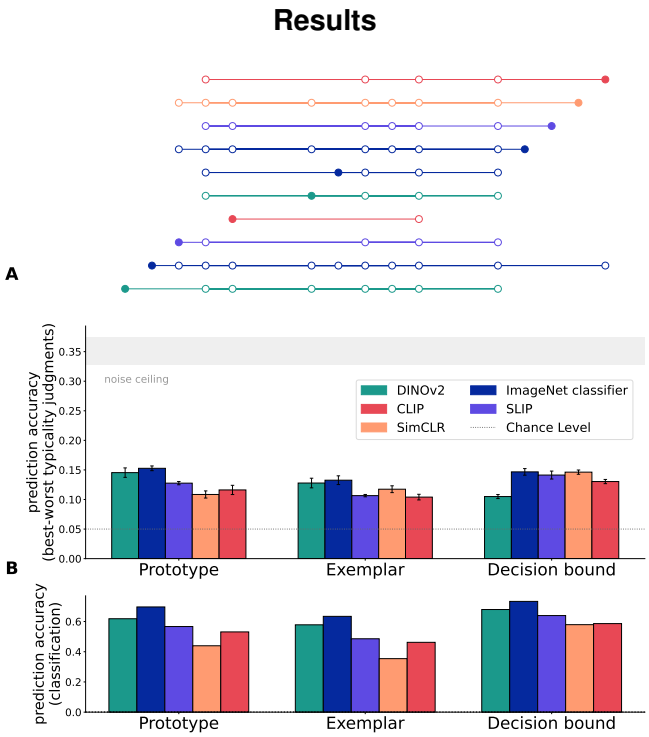


Figure 2: Performance of prototype, exemplar, and decision-bound models across different deep neural network feature spaces. **(A)** Cross-validated accuracies for predicting human typicality judgments. Along each horizontal line, filled dots indicate significantly higher accuracy compared to empty dots. Each dot corresponds to the bar below it. Wilcoxon signed rank test, corrected for 105 pairwise comparisons using Benjamini–Hochberg FDR ($q < 0.05$). **(B)** Test classification accuracies on the corresponding classes in ImageNet-21K. **The decision-bound model maximized classification accuracy, but not necessarily alignment with human typicality judgments.**

Numerically, the model that best predicted human typicality judgments was the prototype category-structure applied to the penultimate-layer representations from the ImageNet classifier (Fig. 2A). However, no single model–feature-space combination significantly outperformed all others. Accuracy reached 15.3%, which is significantly above the chance level of 5% (i.e., the probability of correctly guessing both the most and least typical items out of five). However, it remains well below the human leave-one-subject-out noise ceiling (noise ceiling: 32.75%).

We also evaluated the models' classification performance on the corresponding classes in ImageNet-21K (Fig. 2B). The decision-bound model achieved notably higher classification accuracy than the prototype and exemplar models. This is expected, as multinomial regression is discriminative.

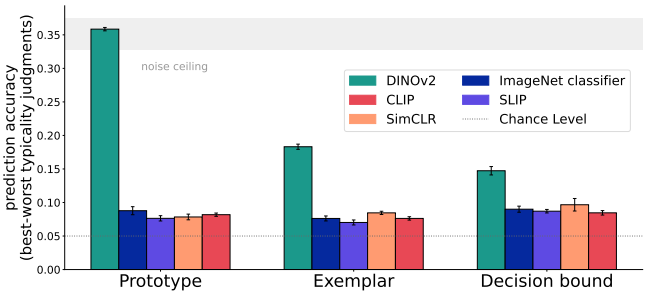


Figure 3: Simulated prediction accuracy on typicality judgments sampled from the DINOv2-prototype model. The data-generating model achieves perfect self-prediction and significantly outperforms all other models. Comparable results were observed when sampling data from each of the other 14 models.

Was the similarity in model performance due to indistinguishable predictions? We simulated human judgments using multinomial sampling based on each model's predictions (Fig. 3). We found that only the model used to generate the data reached the noise ceiling, indicating that all 15 models made distinct predictions.

One way to improve the model–human alignment is to linearly reweight the features. However, the associated performance gains did not generalize to test data, even with optimal regularization. This might be due to the structure of the ViSpa dataset, in which each category has a single five-item test set.

Conclusion

In this work, we have formalized three categorization theories and applied them to five representational spaces, putting to test 15 distinct image-computable models.

While models based on neural network representations predict human typicality judgments at accuracy levels significantly above chance, our results indicate that we are still far from achieving two more ambitious goals: (1) to fully account for human typicality judgments, and (2) to clearly identify the underlying computational mechanisms.

The next steps toward these goals should include collecting more extensive typicality judgment datasets, potentially using model-guided stimulus synthesis; adapting neural network representations to the new empirical data; and developing better models of human classification that may integrate elements of the competing category-structure theories.

Acknowledgments

This was supported by the Israel Science Foundation (grant number 534/24 to T.G.).

References

Ashby, F., & Maddox, W. (1993, September). Relations between Prototype, Exemplar, and Decision Bound Models of Categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400. doi: 10.1006/jmps.1993.1023

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020, October). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 5418. doi: 10.1038/s41467-020-18946-z

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020, November). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337. doi: 10.1073/pnas.1912334117

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023, July). ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, *130*(4), 896–934. doi: 10.1037/rev0000392

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep Neural Networks Predict Category Typicality Ratings for Images. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*.

Marley, A., & Louviere, J. (2005, December). Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology*, *49*(6), 464–480. doi: 10.1016/j.jmp.2005.05.003

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018, August). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, *38*(33), 7255–7269. doi: 10.1523/JNEUROSCI.0388-18.2018

Rosch, E., & Mervis, C. B. (1975, October). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. doi: 10.1016/0010-0285(75)90024-9