Predicting human perceptual similarity and memory false alarms using visual and semantic deep neural networks

Natalia Kurilenko (Natalia.Kurilenko@cshs.org)
Graduate School of Biomedical Sciences
Department of Biomedical Sciences, Cedars-Sinai Medical Center,
8700 Beverly Blvd, Los Angeles, CA 90048, USA

Kevin J. M. Le (kvnjmle@caltech.edu)
Computation and Neural Systems
Division of Biology and Biological Engineering, California Institute of Technology
1200 East California Blvd, Pasadena, CA 91125, USA

Juri Minxha, PhD (jminxha@caltech.com)
Division of Biology and Biological Engineering, California Institute of Technology
1200 East California Blvd, Pasadena, CA 91125, USA

Ueli Rutishauser (Ueli.Rutishauser@cshs.org)
Department of Neurosurgery, Cedars-Sinai Medical Center,
8700 Beverly Blvd, Los Angeles, CA 90048, USA

Abstract:

Despite considerable effort, predicting human similarity judgments and aspects of memory that rely on such judgments remains challenging. In this work, we collected a large set of human similarity judgments and compared the performance of semantic and visual deep neural networks in predicting them. We then examine the effectiveness of the computational similarity metrics in predicting false alarms in a recognition memory task. We show that general visual features best predict perceptual similarity while combined visual and semantic information better explain memory performance.

Keywords: similarity; deep learning; visual memory

Introduction

Visual memory is built upon a trade-off between two seemingly incompatible aims. First, a familiar item should be recognized in different contexts, requiring generalization and invariance. Second, different items should be distinguished even if they are very similar, requiring fine discrimination. Similarity plays a fundamental role in both processes: the probability of generalization increases with the similarity between two items (Shepard, 1987), and the ability to discriminate memories relies on the dissimilarity between brain representations (Leal & Yassa, 2018). How do people perceive similarity, represent it in the brain, and use these representations to guide behavior?

Answering these questions requires establishing a similarity metric for study items (e.g., visual stimuli in memory tasks). Computer vision research has shown that convolutional neural networks (CNN) predict human similarity judgments reasonably well (Jozwik et al., 2017; Zhang et al., 2018), but these models fail in more complex tasks with semantically rich stimuli (Rosenfeld et al., 2018). Notably, even the most sophisticated CNN models that correlate with human judgments are often outperformed by simple categorical models (Jozwik et al., 2017; Shoham et al., 2024). Here we present **a quantitative similarity metric for visual stimuli and explore its utility in predicting performance of human subjects in a recognition memory task**.

Methods

Similarity-judgments task Independent pools of participants (n=45 and n=54) performed 2 variants of a multi-arrangement task (Kriegeskorte & Mur, 2012). The task was to arrange images within an arena based on visual similarity (Fig. 1A). Then, a representational dissimilarity matrix (RDM) was computed for each participant.

Image caption task Image captions for SGPT text embedding were collected from Amazon Mechanical Turk participants who were asked to describe the images. Each image had ~20 captions from different annotators and we chose the caption whose embedding was the medoid.

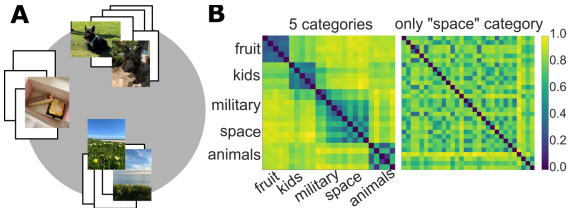

Figure 1 Task and dissimilarity matrices. (A) Example of trial in multi-arrangement task. **(B)** RDMs computed from the similarity judgments collected with the task in A.

Recognition memory task The task was previously described (Rutishauser et al., 2010). Briefly, subjects (n=41) first viewed 100 novel images. Then, subjects were presented with 50 novel and 50 old images and for each, were asked to indicate whether they had seen the image before or not (Fig. 2A). Images belonged to five different categories with the same number of images in each.

Models We use several pre-trained deep neural networks (DNN). DINOv2 is a self-supervised vision transformer (ViT) model that extracts general purpose visual features without training labels. CLIP is a natural language-supervised ViT-based model, trained to match text and images. AlexNet, VGG16, and ResNet50 are deep CNNs trained to categorize images. SGPT is a GPT transformer-based model trained on text to perform semantic search. The "pixel-wise" model is correlation distance between pixels for a pair of images. The "categories" model is an RDM with binary values (0 and 1).

Table 1 Model performance in predicting human RDMs

	5 categories (NC[1]=0.48)	**Only "space" (NC=0.41)**
Model	ρ_a±s.e.m	ρ_a±s.e.m
categories	0.307±0.009	–
pixel-wise	0.004±0.010	0.061±0.007
SGPT	0.367±0.024	0.292±0.022
AlexNet	0.309±0.015	0.241±0.015
VGG16	0.320±0.018	0.248±0.019
ResNet50	0.335±0.017	0.250±0.021
CLIP	0.379±0.022	0.261±0.016
DINOv2	**0.421±0.017**	**0.385±0.019**

[1]Noise ceiling.

Results

DINOv2 model captures human similarity judgments within and across semantic categories. For a systematic evaluation of DNN model performances in explaining human perceptual similarity, we collected similarity judgments for (1) a dataset of 25 images (five images each of five categories) (Fig.1B, left) and (2) a dataset of 30 images in the "space" category (Fig.1B, right) to remove the effects of categorical structure. The stimuli dataset were processed with DNN models to obtain embeddings (see methods; for SGPT embeddings, collected text

captions were used instead of images; a category model was created only for the first dataset). We compared each model RDM with each human subject's similarity RDM with randomized tiebreaking Spearman's ρ_a (Schütt et al., 2023). Table 1 depicts average correlations for each model in both datasets. Next, we performed statistical inference on these evaluations and found that all models except for the pixel-wise model in the first dataset performed significantly better than 0 (one-sided t-test p-value<0.001, Bonferroni-corrected α=0.00125/ 0.001428 for first/second dataset). In the first dataset, SGPT, CLIP, and DINOv2 significantly outperformed the category model with DINOv2 being significantly better than all other models (FDR-corrected pairwise t-test for 28 model-pairs comparisons). As DINOv2 is trained in a self-supervised manner, without labels, this suggests that general visual features best explain similarity ratings. In the absence of category, DINOv2 still significantly outperformed all other models (same inference). Note that in both datasets, only DINOv2 performance did not significantly differ from lower bound noise ceiling (one-sided t-test p-values=0.00135/0.1970 in first/second dataset).
In summary, we found that deep features of the ViT-based model DINOv2, learned from images alone, predict human perceptual similarity judgments better than classic CNN-based features or features of ViT models trained with text-guided supervision.

Deep feature-based image similarity contributes to memory-based false alarm rate. As a next step, we explored if the derived similarity metric could explain memory behavior. We used an independent pool of participants that performed a recognition memory task (Fig. 2A). On average, participants correctly recognized 68.73% of old images and incorrectly recognized 22.44% of new images as "old" ("false alarm rate"). We hypothesized that "false alarms" are at least partially a result of the new images being similar to previously studied images. To test this, for each new image i shown during retrieval, we computed the probability of incorrectly recognizing this image as "old" ($p('FA')_i$). We then regressed these probabilities against the minimum cosine distance (min d_{cos}) between deep feature vectors for an image i (a_i) and all images presented during learning (B). In addition to similarity, false alarms can also result from

Table 2 Linear regression for false alarm rate[2]

Model	Adjusted R^2	*F/p-value*
categories	**0.234**	**4.748/0.0028**
ResMem	0.039	2.997/0.0898
SGPT	0.077	5.089/0.0287
AlexNet	-0.013	0.373/ 0.5440
VGG16	0.123	7.843/0.0073
ResNet50	0.088	5.757/0.0204
CLIP	**0.255**	**17.733/0.0001**
DINOv2	0.110	7.033/0.0108

[2]df for F-stats = 4, 45 (category); 1, 48 (other models)

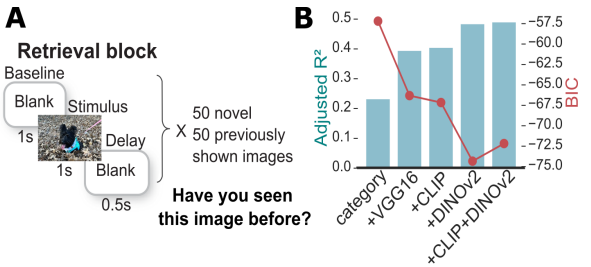

Figure 2 Memory task and performance. (A) Scheme of retrieval block of recognition memory task. **(B)** Comparison between models in predicting false alarms.

an image-intrinsic sense of familiarity as predicted by strength theory (Norman & Wickelgren, 1969). To account for this, we used ResMem, the CNN trained to predict image memorability (Needell & Bainbridge, 2022). We also regressed $p('FA')$ against the category variable to control for the known strong influence of semantic category on memory (Kramer et al., 2023). The regression results for all models are shown in Table 2. We found that by itself, only the CLIP-derived similarity (expressed as min d_{cos}) explains more variance in false alarm rate than simple category membership. Interestingly, image memorability (ResMem predictions) did not explain significant amount of the variance in false alarms. These results are consistent with previously reported contribution of semantic information to memory. But can we explain more variance in false alarms by adding visual features? To answer this, we investigated if best-performing models could contribute to the false alarm rate variance in addition to what is explained by category membership alone. We fitted the following linear model:

$$p('FA')_i \sim 1 + \beta_1 \cdot \text{category} + \beta_2 \cdot \min(d_{cos}(a_i, B)),$$

where "category" is a categorical variable with 5 levels. DINOv2-based cosine similarity together with category membership explained 48,6% of the variance in false alarm rate (Fig. 2B), significantly better than category alone (likelihood-ratio test p-value=0.000004, Bonferroni corrected α= 0.0025). Note that this model also had the lowest Bayes information criterion (BIC, red line in Figure 2B) suggesting that it is preferred over just a category model or a more complex model that also includesd CLIP. To summarize, we confirmed that category strongly influences memory performance but adding a perceptual feature-based metric (expressed as DNN-based cosine similarity) can increase explained variance up to almost 50%.

Discussion

In this work we established similarity metric for visual stimuli and demonstrated that this metric can meaningfully capture memory performance expressed as false alarm rate. We envision the next steps in bridging behavior data with brain similarity representations via the tools established here to shed light on neural underpinnings of visual perception and memory.

Acknowledgments

This project was supported by the BRAIN initiative through U01NS117839.

References

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01726

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981. https://doi.org/10.1126/sciadv.add2981

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00245

Leal, S. L., & Yassa, M. A. (2018). Integrating new findings and examining clinical applications of pattern separation. *Nature Neuroscience*, *21*(2), Article 2. https://doi.org/10.1038/s41593-017-0065-1

Needell, C. D., & Bainbridge, W. A. (2022). Embracing New Techniques in Deep Learning for Estimating Image Memorability. *Computational Brain & Behavior*, *5*(2), 168–184. https://doi.org/10.1007/s42113-022-00126-5

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology*, *6*(2), 192–208. https://doi.org/10.1016/0022-2496(69)90002-9

Rosenfeld, A., Solbach, M. D., & Tsotsos, J. K. (2018). *Totally Looks Like—How Humans Compare, Compared to Machines* (arXiv:1803.01485). arXiv. http://arxiv.org/abs/1803.01485

Rutishauser, U., Ross, I. B., Mamelak, A. N., & Schuman, E. M. (2010). Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature*, *464*(7290), Article 7290. https://doi.org/10.1038/nature08860

Schütt, H. H., Kipnis, A. D., Diedrichsen, J., & Kriegeskorte, N. (2023). Statistical inference on representational geometries. *eLife*, *12*, e82566. https://doi.org/10.7554/eLife.82566

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shoham, A., Grosbard, I. D., Patashnik, O., Cohen-Or, D., & Yovel, G. (2024). Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nature Human Behaviour*, 1–16. https://doi.org/10.1038/s41562-024-01816-9

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric* (arXiv:1801.03924). arXiv. https://doi.org/10.48550/arXiv.1801.03924