Testing a Learning Theory of Aesthetic Appeal Using Category Learning and Deep Neural Networks

Edward A. Vessel (evessel@ccny.cuny.edu)
Department of Psychology, The City College of New York
160 Convent Ave.
New York, NY 11215 USA

Andrew Frankel
Cognitive Neuroscience Program, CUNY Graduate Center
New York, NY USA

Aubrey Valdez
Department of Psychology, The City College of New York
New York, NY USA

Aishwarya Gurung
Cognitive Neuroscience Program, CUNY Graduate Center
New York, NY USA

Colin Conwell
Center for Brains, Minds
Machines, Massachusetts Institute of Technology
Boston, MA USA

Abstract

How do people mentally represent visual art, and how do those representations relate to aesthetic value? The learning theory of aesthetic valuation suggests that the aesthetic appeal we feel from engaging with visual objects is an affective signal for learning, and thus depends on how those objects relate to what we know about the visual world. Yet this theory is hard to test, given the difficulty of directly measuring the relevant aspects of an observer's internal perceptual models. We outline a behavioral and modeling paradigm for training observers in a visual artwork training task and, in parallel, tuning deep neural networks (DNNs) to serve as proxies for internal representations. Here we show that the task successfully modulated observer's knowledge and internal representations about a set of artworks, and we explore how architecture and training target affect the ability of DNNs to capture salient aspects of human observers' behavior.

Keywords: aesthetic value; art; category learning; uniqueness

Introduction

Aesthetic value judgments are a core aspect of cognition, yet the psychological mechanisms supporting aesthetic valuation are poorly understood.

Stimulus-driven (or "universalist") approaches to aesthetics seek to identify stimulus features (e.g. symmetry, contour) that affect everyone's experience in the same way (see (Vessel, Ishizu, & Bignardi, 2022) for a review). In contrast, we take an "interactionist" approach, which acknowledges that not everyone responds to the same stimulus in the same way (Vessel, Maurer, Denker, & Starr, 2018): it is important to consider how a particular stimulus interacts with a particular observer. This approach seeks to understand the internal, subjective constructs that mediate between the internal processing of a stimulus and expressed aesthetic value.

The learning theory of aesthetics (Van de Cruys, Frascaroli, & Friston, 2024; Biederman & Vessel, 2006) suggests that aesthetic value is an affective learning signal that is fundamentally personal: how an object impacts a viewer depends on how that object relates to what a person knows about the visual world: is it familiar? Is it unique?We hypothesize that aesthetic value is highest for stimuli that are on the edge of what a person knows in a ”zone of learning” (Metcalfe, Schwartz, & Eich, 2020): relatable to what we know, but offering the promise of learning something new.

Testing this theory requires measuring the relevant aspects of a person's internal model of the visual world, which in turn depends on their personal biography of visual experience. These internal models are highly detailed and hierarchically structured, and thus not possible to assess directly.

Here we present ongoing work that seeks to combine a category learning task with machine learning to create proxies for internal representations which can then be used to test and refine a theory for how internal aesthetic value judgments relate to internal representations.

Embedding Artworks in a Similarity Space

In a first phase, we used DreamSim (Fu et al., 2023), a DNN optimized to align with human similarity judgments to embed 106,112 images of artworks from wikiart.org packaged for download on Github (Ushio, 2024) into a 1,792-dimensional latent vector perceptual similarity space. In initial explorations, the similarity space for a subset of 8800 artworks was further reduced to 500 dimensions using principal components analysis (PCA; capturing 85% of the variance) and visualized using tSNE (van der Maaten & Hinton, 2008) in 2 dimensions. This resulted in a highly structured space, with paintings clustering by artist and art movement and the degree of intermixing reflecting differences in the confusability of artists (e.g. highly similar impressionists, distinctive surrealists).

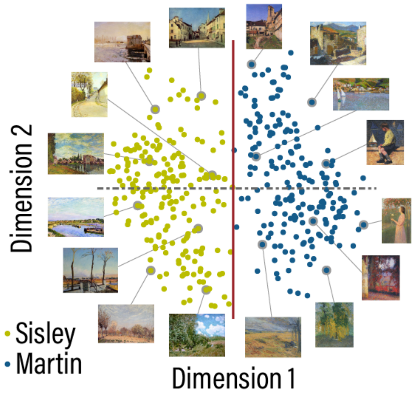

Figure 1: 2-dimensional stimulus space for category learning.

Identification of Two Artists for Category Learning

Through inspection of local regions of this space (using PCA with 2 dimensions) we identified two artists (Henri Martin, Alfred Sisley) whose paintings occupied neighboring regions of the similarity space but were largely separable by a boundary between them. One PC axis discriminated the two artists and a second captured parallel variance in content. The paintings from these artists were then split into two halves, producing a stimulus space with 4 quadrants. Paintings on the wrong side of the classification boundary were removed and the set was further reduced to 100 paintings per quadrant (Fig. 1).

Behavioral Effects of Category Learning

In a second phase, N=24 human participants were trained to distinguish between the two artists. Half the participants were trained on the top half of the space, half on the bottom. Training consisted of four blocks of 90 trials, plus additional ”catch” trials to assess learning. Each trial began with a fixation point (1 s) followed by two images side-by-side, one from each artist (6 s) with the names of each artist underneath.

In a subsequent test session (different day), participants viewed 80 artworks (20 from each quadrant) and made four judgments in separate blocks: familiarity ratings (Block 1); uniqueness ratings (Block 2); categorize paintings by artist (plus confidence judgment, Block 3); ratings of aesthetic appeal (Block 4). Artworks from the two training quadrants could be images shown during training (20 total) or novel ”independent, identically distributed” images (IID; 20 total). Artworks from the generalization quadrants were ”out-of-distribution” stimuli (OOD; 40 total).

Participants successfully learned to categorize artworks by artist (Fig. 2a; average accuracy 82% for training set) and generalized to novel stimuli (76.5% for IID artworks; 72.7% for OOD artworks). Importantly, performance was worse close to the category boundary, a hallmark of category learning. Participant ratings of familiarity and uniqueness (Fig. 2b) were affected by training: familiarity decreased with distance from the training region (not shown) and novel OOD artworks adjacent to the train/generalization boundary were rated as more unique. This shows that training did indeed lead to a change in how observers represented the stimuli.

In turn, aesthetic ratings were correlated with ratings of uniqueness (r = 0.33; Fig. 2c), yet also mildly positively correlated with ratings of familiarity (r = 0.15).

Tuning DNNs using the Same Paradigm

In a third phase, DNNs pretrained to perform object recognition were fine-tuned using the same artworks as human observers. We trained models with 2 different architectures, VGG16 (Simonyan & Zisserman, 2014) pretrained on ImageNet-1k (Deng et al., 2009), and ConvNeXtV2 (Woo et al., 2023) pretrained using masked auto-encoding style self-supervision on the images (but not labels) of ImageNet-21k (Russakovsky et al., 2015). Both networks were fine-tuned using two different targets and loss functions: 1) classification by artist name using cross-entropy loss, and 2) the signed distance to the category boundary derived from the 2D PCA reduction of DreamSim embeddings.

Both architectures with both training targets successfully learned to categorize by artist, generalizing to IID and OOD stimuli (Fig. 2d). Training on distance to category boundary led to more humanlike performance than training on artist alone. Current work is focused on identifying metrics derived from these personalized DNNs that reflect human ratings, and on using them to predict human performance.

Conclusion

By embedding real artworks into a human-aligned similarity space, we constructed a paradigm for systematically modifying human observers’ knowledge about a set of artworks and, in parallel, tune pre-trained DNNs with the same stimuli. We see evidence that training created a ”zone of learning” adjacent to the training region in which familiarity and uniqueness are higher than for stimuli further away. This is a promising direction for studying highly subjective mental states that are dependent on individual learning histories.

Acknowledgments

Thanks to Sunil Ale, Edwarné Woodley-Gift, Luis Morales, August Miller and Michael Grossberg. Funding sources: Am. Psych. Assoc. Div. 10 Micro-Grant, PSC-CUNY Award, Google Arts & Culture sponsorship, CUNY JFRASE.

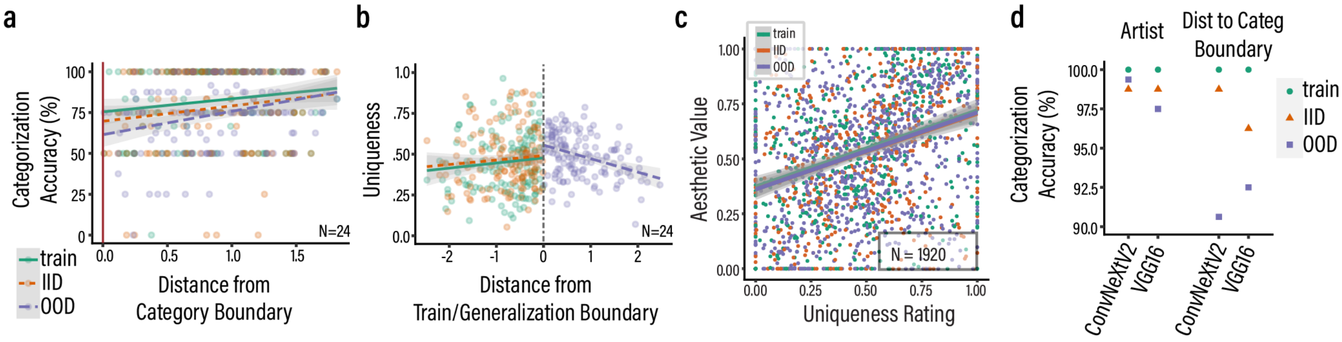

Figure 2: a) After training, human participants generalize categorization to independent, identically-distributed (IID) and out-of-distribution (OOD) artworks, with poorer performance closer to the boundary. b) Ratings of uniqueness are higher for OOD artworks that are near the training set. c) Ratings of aesthetic value increase as a function of uniqueness. d) DNN's tuned using artist category or distance to category boundary perform very well; some configurations show more human-like performance.

References

Biederman, I., & Vessel, E. A. (2006). Perceptual Pleasure and the Brain. *American Scientist*, *94*(3), 247–253. (ISBN: 1040359019) doi: 10.1511/2006.3.247

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). Miami, FL, USA: IEEE. Retrieved from `http://www.image-net.org`. doi: 10.1109/CVPR.2009.5206848

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. *arXiv preprint arXiv:2306.09344*. (arXiv: 2306.09344v2)

Metcalfe, J., Schwartz, B. L., & Eich, T. S. (2020, October). Epistemic curiosity and the region of proximal learning. *Current Opinion in Behavioral Sciences*, *35*, 40–47. (Publisher: Elsevier Ltd) doi: 10.1016/j.cobeha.2020.06.007

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 (IJCV)*, *115*(3), 211–252. doi: 10.1007/s11263-015-0816-y

Simonyan, K., & Zisserman, A. (2014, September). Very Deep Convolutional Networks for Large-Scale Image Recognition.
(arXiv: 1409.1556)

Ushio, A. (2024, June). *Asahi417/wikiart-image-dataset.* Retrieved from `https://github.com/asahi417/wikiart-image-dataset`,

Van de Cruys, S., Frascaroli, J., & Friston, K. (2024, January). Order and change in art: towards an active inference account of aesthetic experience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1895). doi: 10.1098/rstb.2022.0411

van der Maaten, L., & Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579-2605.

Vessel, E. A., Ishizu, T., & Bignardi, G. (2022, October). Neural correlates of visual aesthetic appeal. In *The Routledge International Handbook of Neuroaesthetics* (pp. 103–133). Routledge. doi: 10.4324/9781003008675-7

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*(June), 121–131. (Publisher: Elsevier) doi: 10.1016/j.cognition.2018.06.009

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023, January). ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders.
(arXiv: 2301.00808)