Beyond Learnability:

Understanding Human Visual Development with DNNs

Lei Yuan

Department of Psychology and Neuroscience

University of Colorado Boulder

Corresponding author: Lei Yuan

Email: lei.yuan@colorado.edu

Laboratory website: https://www.colorado.edu/lab/del/

Keywords: Intelligence, Learning, Development, Vision, Machine learning

Abstract

Orhan and Lake demonstrated the computational plausibility that children can acquire sophisticated visual representations from natural input data without inherent biases, challenging the need for innate constraints in human learning. The findings may also reveal crucial properties of early visual learning and inform theories of human visual development.

How to build intelligence? Building intelligence involves three crucial components: the input data, the learning machinery, and the learning outcome. Advances in wearable sensors that can capture learners' input data on the scale of daily life [1], [2], in machine learning algorithms that can simulate human learning [3], and in measured performance across cultures and contexts [4] contribute new insights into the interplay of input learning data, learning machinery, and learning outcomes.

The foundations for building human intelligence begin in infancy, and it is the properties of that foundation that motivated a recent paper by Orhan and Lake [5]. The richness, complexity, and noise of real-life input of young children challenge explanations of how children rapidly learn all that they know. Traditional theory proposed a solution in terms of "inductive biases"—that is, the children's learning machinery was biased to make certain conclusions from the data, driven by innate constraints [6]. Orhan and Lake demonstrated the computational plausibility that children can build sophisticated representations of visual objects from naturalist input data without built-in inductive biases. The study used an input dataset collected from three infants who wore head cameras capturing their point-of-view scenes (or egocentric views) for brief durations (1-2 hours) weekly between the ages of 6-31 months. They then used the collected video frames to train Deep Neural Networks (DNNs) in a self-supervised manner with no names or category labels for objects.

One such model used by Orhan and Lake was a Vision Transformer (ViT) embedding model trained via various self-supervised learning (SSL) algorithms (e.g., DINO). Rather than learning labels for instances (as in supervised learning), embedding models are optimized to represent perceptually similar images closer in the representation space and more different images farther apart. A method known as Data Augmentation plays a critical role [7]. In this

method, the training data are augmented by creating a series of variations of the original input data through augmentation methods, such as rotation or cropping. The models were then optimized by training with variations (i.e., augmented views) as belonging to the same object. The models were tested in a suite of downstream benchmark tasks (e.g., image classification, semantic segmentation). The overall results demonstrated that a relatively small subset of augmented children's input data yielded quite strong performance, achieving about 70% of the performance of models trained through the same unsupervised learning algorithms with the full dataset of clear, photographed images from ImageNet.

Data Augmentation has been argued to mimic the properties of natural human experiences [7]. As an active perceiver moves, uses, and looks at objects in the world, he or she creates many varied ego-centric images of the same objects. As such, natural human experiences are fundamentally constrained by time and space. Images in the same physical space (e.g., sitting at the kitchen table) are more similar than images captured from different contexts (e.g., sitting at the kitchen table versus playing at the park). Images sampled from the same context are likely to be about the same objects or people—e.g., different views of the same spoon pushing around food on a plate. The computational efficiency of the children's view images over that of ImageNet may well stem from the proportion of objects with these "natural" augmentations. Classic theories of human concept learning have posited that similarity plays an important role in concept learning and categorization [8]. These traditional theories contrast with more recent ones, which place more emphasis on how learning the names of objects (in a supervised way) teaches children object categories [9]. Orhan and Lake's results demonstrate that similarity itself may be a powerful pathway to categorization.

The second model employed by Orhan and Lake is generative, with the ability to generate new images or fill in missing portions of an existing image. It works by first learning a discrete, compressed representation of input images using a VQ-GAN (Vector Quantized Generative Adversarial Network). A Transformer is then trained to generate images sequentially from top to bottom within this compressed representation. Models trained with children's data generated completions that exhibited broad consistency with the original scene, regarding the color, texture, orientation, and outline of the objects or people depicted. Interestingly, compared to models trained with clear photograph images from ImageNet, models trained with children's egocentric view data were less successful at generating finer details. A similar result was also obtained in the embedding models trained with children's views, which were less object-centric and showed more sensitivity to global context and background scene information. Do these performance differences between models trained with children's egocentric images and those with photographs unveil relevant principles of visual category learning?

From certain perspectives, both infants' early visual experiences and their developing visual systems can be considered immature; however, these immaturities may be uniquely adaptative for developing a robust visual representation system [3]. Infants are born with low visual acuity and show preferences for simple images with fewer edges and high spatial contrast [10]. These properties of the infants' visual system may be beneficial for similarity-based category learning because they may promote the discovery and learning of similar global structures across images with different views. DNN models combined with realistic ego-centric view data may be useful for simulating individual learning trajectories and testing theoretical hypotheses about the development of visual representations. Conversely, Data Augmentation may approximate but not yield optimal within-object variation. Incorporating properties of

human visual development—e.g., context-dependent, multiple views of the same object, low acuity—may inspire a new generation of machine learning algorithms that mimic human-level visual intelligence.

Acknowledgments: This work was supported by the National Science Foundation DRK12 Grant 2200781 to Lei Yuan. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Reference

[1] L. B. Smith, C. Yu, H. Yoshida, and C. M. Fausey, "Contributions of Head-Mounted Cameras to Studying the Visual Environments of Infants and Young Children," *J. Cogn. Dev.*, vol. 16, no. 3, pp. 407–419, 2015, doi: 10.1080/15248372.2014.933430.

[2] E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, and S. Tor, "Day by day, hour by hour: Naturalistic language input to infants," *Dev. Sci.*, vol. 22, no. 1, p. e12715, 2019, doi: 10.1111/desc.12715.

[3] S. Sheybani, H. Hansaria, J. N. Wood, L. B. Smith, and Z. Tiganj, "Curriculum Learning with Infant Egocentric Videos," in *Advances in Neural Information Processing Systems, 36*, 2024.

[4] M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman, *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press, 2021.

[5] A. E. Orhan and B. M. Lake, "Learning high-level visual representations from a child's perspective without strong inductive biases," *Nat. Mach. Intell.*, vol. 6, no. 3, pp. 271–283, Mar. 2024, doi: 10.1038/s42256-024-00802-0.

[6] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, Mar. 2011, doi: 10.1126/science.1192788.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[8] E. Rosch and C. B. Mervis, "Family resemblances: Studies in the internal structure of categories," *Cognit. Psychol.*, vol. 7, no. 4, pp. 573–605, Oct. 1975, doi: 10.1016/0010-0285(75)90024-9.

[9] C. Yu and L. Smith, “Rapid word learning under uncertainty via cross-situational statistics.,” *Psychol. Sci.*, vol. 18, no. 5, pp. 414–20, May 2007, doi: 10.1111/j.1467-9280.2007.01915.x.

[10] V. Ayzenberg and M. Behrmann, “Development of visual object recognition,” *Nat. Rev. Psychol.*, vol. 3, no. 2, pp. 73–90, Dec. 2023, doi: 10.1038/s44159-023-00266-w.