Fast and robust visual recognition young children

Vladislav Ayzenberg (vayzenb@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA, USA

Stella Lourenco (stella.lourenco@emory.edu)
Department of Psychology, Emory University
Atlanta, GA, USA

Abstract

By adulthood, humans can rapidly identify objects from sparse visual displays and do so across disruptions to the object's appearance. However, little is known about the development of these abilities. Here, we examined the robustness of children's (3 to 5 years) recognition abilities using a challenging object recognition task which required them to identify rapidly presented objects (100 - 300 ms; forward and backward masked) that had complete, perturbed, or deleted contours. To shed light on the mechanisms underlying their recognition abilities, we compared their performance to biologically plausible deep neural networks (DNNs) with feedforward or recurrent architectures which were trained with either curated or variable image sets. We also characterized the gaps between child and machine vision by comparing children to performance optimized models. We found that even the youngest children could identify objects at high speeds when object contours were perturbed or deleted. Analyses of DNN performance revealed that both recurrence and variable visual experience were crucial for improving recognition accuracy, though they generally performed worse than children. These findings suggest that young children's visual recognition abilities are fast and robust, but the mechanisms underlying these abilities are not understood well enough to implement into current models.

Keywords: Development; Deep Learning; Object Classification; Shape Perception; Recurrence; Feedback

Introduction

Humans rapidly extract meaning from sparse, and often incomplete, visual information. Indeed, by adulthood, participants can identify objects presented as quickly as 100 ms (Grill-Spector & Kanwisher, 2005) and can do so even when object features are distorted or partially occluded (Biederman & Cooper, 1991; Murray et al., 2001). When and how do such recognition abilities develop in childhood?

In adults, robust visual processing abilities are supported primarily by the ventral visual pathway (DiCarlo et al., 2012). When the object properties are clearly visible, a single feedforward pass through the ventral hierarchy is sufficient to determine the object's identity or category (Serre et al., 2007). However, in more challenging cases, such as when the object parts are distorted or partially occluded, recurrent processes are needed to disambiguate the object's identity (Kar et al., 2019; Tang et al., 2018). Thus, recurrent processing may be one neural mechanism that is important for the development of robust recognition abilities.

Another, not mutually exclusive, mechanism is that extensive variability in children's visual experience allows them to recognize objects across a range of contexts. Consistent with this possibility, DNNs with more variability in their training data showed more human-like performance on object recognition tasks (Bambach et al., 2017; Geirhos et al., 2018).

However, directly testing which mechanisms underlie children's visual recognition abilities is challenging. First, few studies have tested children in the same challenging conditions typically used with adults (as well as non-human primate and machines), which makes mechanistic comparisons across age (and organism) difficult (Ayzenberg & Behrmann, 2023). Second, young children's limited attentional capacities make it challenging to conduct long or complex neuroimaging experiments, frequently leaving the neural mechanisms underlying visual recognition unclear (Grill-Spector et al., 2008). Finally, researchers are not ethically able to manipulate or restrict a child's visual experience, making it difficult to estimate what kinds of visual experiences are crucial for developing robust recognition abilities.

Thus, in the current study we set out to accomplish three goals. **(1)** Determine the upper-bound of children's abilities using a challenging object recognition task. We did this by presenting 3- to 6-year-olds with a task that required them to identify rapidly presented object outlines (100 - 300 ms; forward and backward masked) with complete, perturbed, or deleted contours (Fig1A-B). **(2)** Identify possible mechanisms that support recognition abilities in young children. To achieve this goal, we compared children to DNNs with either feedforward or recurrent architectures that were trained on either curated images, or images with greater variability. **(3)** Provide a benchmark by which to identify existing gaps between children and DNNs. To this end, we compared children to performance optimized DNNs.

Methods

Participants. We tested 128 children (M_{age} = 4.62, Range = 3.05 – 5.95; 64 females) in one of three (randomly assigned) contour conditions (complete, perturbed, deleted; n = 42 per condition).

Models. For biologically plausible models, we implemented vonenet with either feedforward (vonnet_ff) or recurrent (vonenet_r) architectures (Dapello et al., 2020) and trained them on either ecoset (Mehrer et al., 2021) or a stylized version of ecoset (Geirhos et al., 2018). For the performance optimized models we tested a Vision Transformer (ViT) (Dosovitskiy et al., 2020) and ConvNext (Liu et al., 2022).

Child testing procedure. Children were tested with a two-alternative forced-choice procedure where they had to identify a rapidly presented stimulus that was both forward and backward masked (Fig 1B). Duration of stimulus displays varied between 100-300 ms, which was determined via a titrated procedure.

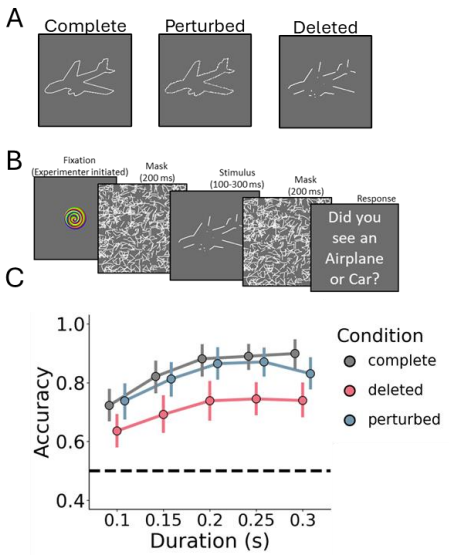

Fig 1. (A) Example stimuli from the three contour conditions. (B) Trial procedure used with children. (C) Child performance in each condition at each stimulus duration. Dotted line indicates chance (0.5).

Model testing procedure. Models were evaluated by, first, training a classifier using the feature activations from the penultimate layer of each model on naturalistic images of each object and, then, testing them on each stimulus display in a pairwise fashion (e.g., trained on photographs of airplanes and cars; tested on the perturbed airplane stimulus; 20-fold cross-validation). To minimize the possibility that our evaluation decisions impacted DNN performance, we tested models with six different classifiers and parametrically varied the number of images used to train the classifier (5 to 300). The best performing model across all classifier variations was compared to children.

Results and Discussion

Children. Overall, children accurately identified objects in all conditions even at the fastest speeds (Fig 1C). When examining performance separately by age and condition, we found that 4- and 5-year-olds performed above chance in every condition and all durations (*ps* < .038, *ds* > 0.63). By contrast, 3-year-olds performed above chance at all durations of the complete condition (*ps* < .007, *ds* > 1.00), but only at 200 ms in the perturbed condition (p = .001, d = 1.27) and 250 ms in the deleted condition (p = .007, d = 0.98). These findings suggest that by 4 years-of-age, children rapidly extract meaning from sparse visual displays – even when information is missing. Even 3-year-olds performed above chance in many cases.

Models. Amongst the biologically plausible models, recurrent models (vonenet_r) generally outperformed feedforward models (vonenet_ff), and models trained on a more variable image set (stylized-ecoset) outperformed those trained on a curated image set (ecoset). These findings are consistent with the hypothesis that recurrence and variable visual experience is crucial for robust visual recognition. Interestingly, models with recurrent architectures, such as vonenet_r_ecoset, performed as well as, or better than, performance-optimized models, such as ConvNext. This finding suggests that smaller models with biologically plausible architectures (vonenet_r: 55m params vs. ConvNext: 198m params) trained on smaller, but more ecologically valid image sets (ecoset: 1.5m images vs. ImageNet: 14m images) can achieve competitive performance on visual recognition tasks.

DNNs vs. Children. Children and models were compared on the basis of overlapping confidence intervals. Overall, models primarily matched the performance of the youngest children, but only when children were tested with the fastest durations or most challenging conditions (deleted contours). They rarely matched the performance of 4- or 5-year-olds at any speed or condition. Thus, although recurrence and variable visual experience improves the performance of DNNs on visual recognition tasks, it is largely insufficient to match the recognition abilities of young children.

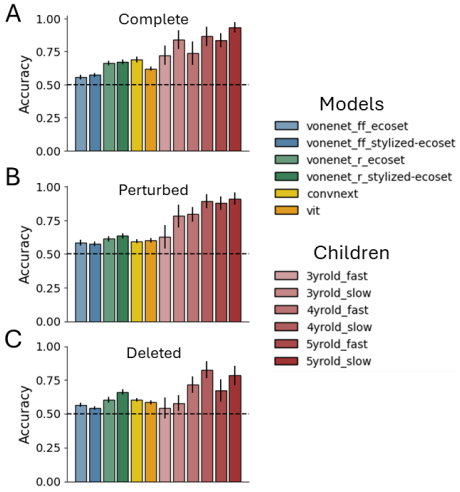

Fig 2. Performance of children and models in the (A) complete, (B) perturbed, and (C) deleted contour conditions. Child data for each age was aggregated into fast (100 ms & 150 ms) and slow (200 ms & 250 ms) stimulus durations.

Conclusion

We sought to understand when and how robust visual recognition abilities develop in young children. Our results showed that young children succeed at identifying objects from sparse visual displays, at speeds as fast as 100 ms even when objects have disrupted contour information. By contrast, biologically plausible and performance-optimized DNNs rarely matched the visual recognition abilities of children. Altogether, these findings suggest that young children already have robust recognition capacities, but there remain large gaps in our ability to approximate these processes in current computational models.

References

Ayzenberg, V., & Behrmann, M. (2023). Development of visual object recognition. *Nature Reviews Psychology*.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2017). An egocentric perspective on active vision and visual object learning in toddlers. 2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob),

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*(3), 393-419.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, *33*, 13073-13087.

DiCarlo, James J., Zoccolan, D., & Rust, Nicole C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415-434. https://doi.org/https://doi.org/10.1016/j.neuron.2012.01.010

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv*. https://arxiv.org/abs/1811.12231

Grill-Spector, K., Golarai, G., & Gabrieli, J. (2008). Developmental neuroimaging of the human ventral visual cortex. *Trends in Cognitive Sciences*, *12*(4), 152-162.

Grill-Spector, K., & Kanwisher, N. (2005). Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychological Science*, *16*(2), 152-160. http://www.jstor.org.proxy.library.emory.edu/stable/40064192

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974-983. https://doi.org/10.1038/s41593-019-0392-5

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Murray, R. F., Sekuler, A. B., & Bennett, P. J. (2001). Time course of amodal completion revealed by a shape discrimination task. *Psychonomic Bulletin & Review*, *8*(4), 713-720.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424-6429. https://doi.org/10.1073/pnas.0700622104

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., & Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835-8840. https://doi.org/10.1073/pnas.1719397115