Impact of a biomimetic training regimen based on early visual experience on neural network organization and behavior

Marin Vogelsang[*]
Department of Brain and Cognitive Sciences
MIT
ozaki@mit.edu

Lukas Vogelsang[*]
Department of Brain and Cognitive Sciences
MIT
lvogelsa@mit.edu

Gordon Pipa
Institute of Cognitive Science
Osnabrueck University
gpipa@uos.de

Sidney Diamond
Department of Brain and Cognitive Sciences
MIT
spdiamon@mit.edu

Pawan Sinha
Department of Brain and Cognitive Sciences
MIT
psinha@mit.edu

Abstract

While deep convolutional neural networks have emerged as powerful computational model systems, they are known to be primarily driven by local features, rather than global shape information, and typically lack robustness to stimulus perturbations that degrade input quality. In this study, we examine the impact of incorporating insights from human development into the training of deep neural networks for vision. Specifically, we focus on the role of experience with initially degraded sensory inputs characteristic of early visual development. Previous work demonstrated that commencing deep network training with initially degraded inputs along isolated perceptual dimensions (specifically, visual blur and color degradations) improved generalization performance. Here, inspired by the joint developmental trajectories of newborns, we examined the consequences of 'biomimetic' training regimens that transitioned from blurry, achromatic to non-blurry, full-color visual inputs. These simulations reveal that such biomimetic training induces a more human-like bias to derive classification decisions based on global shape, rather than local texture. Further, receptive field analyses suggest that the joint development of spatial frequency and chromatic sensitivities can provide a candidate account for the emergence of the division of the visual pathway into parvo- and magnocellular systems. Ablation studies further suggest that magnocellular-like receptive fields are causally driving the shape bias of the biomimetic network. These results have important implications for understanding a key aspect of visual pathway organization and hold applied significance for enhancing deep learning training procedures based on incorporating developmental aspects of human behavior.

[*]Equal contribution

NeurIPS 2024 Workshop on Behavioral Machine Learning.

1 Introduction

While deep neural networks have emerged as powerful computational models, unlike humans, their classification decisions are predominantly driven by local features rather than global shape information [Geirhos et al., 2018a]. Moreover, these models typically lack robustness to stimulus perturbations [Geirhos et al., 2018b, 2020, Taori et al., 2020, Dunn et al., 2021]. Here, we examine the consequences of incorporating insights from human development into the training of machine learning systems. Notably, enabling computational model systems to learn in a manner more akin to newborns can provide important benefits to the field of machine learning [Zaadnoordijk et al., 2022].

More specifically, dating back at least to Turkewitz and Kenny [1982], Newport [1988], and Elman [1993], the notion has been advanced that starting to learn with initially limited inputs may provide important benefits. Most relevant to the investigation described here, as is the case with several aspects of human perceptual development, color sensitivity [Adams and Courage, 2002] and visual acuity[Dobson and Teller, 1978] mature over the months or years following birth from limited to proficient. Past research has demonstrated that starting deep network training with initially blurry imagery leads to spatially extended receptive field structures and increased generalization capabilities [Vogelsang et al., 2018] – at least for faces [Jang and Tong, 2021] – and better category learning [Jinsi et al., 2023]. Blur-trained CNNs have recently also been shown to lead to better predictions of neural responses, greater robustness to noise, and a stronger shape bias [Jang and Tong, 2024]. Similarly, in the domain of color vision, training with initially color-degraded inputs has been shown to result in more robust internal representations and greater generalization to color-removed or hue-shifted images [Vogelsang et al., 2024b]. However, in real biological systems, the developmental progression of visual acuity and color sensitivity occurs together in time, albeit at different rates.

With this motivation, we computationally examined the consequences of such joint developmental progressions. Two primary objectives drove this investigation. First, we aimed to determine whether deep network training that commences with initially blurry and color-degraded inputs leads to more global-shape-based processing. Second, in aiming to study the potential mechanisms enabling such shape bias, we sought to probe whether the joint developmental progression of these two perceptual dimensions could causally influence how they are encoded in the neural representations. As described further below, this could, in addition, provide a candidate account for the emergence of the parvo-magnocellular distinction of the mammalian visual system. This is of particular relevance since global-shape processing in humans may be driven by the magnocellular system, which is generally thought to be involved in coarse-grained visual processing [Livingstone and Hubel, 1987].

Providing some context for this candidate mechanism, cells in the mammalian visual pathway can be broadly classified into magnocellular and parvocellular groups [Livingstone and Hubel, 1988]. Two key characteristics that differentiate these groups are color [Wiesel and Hubel, 1966] and spatial frequency sensitivities [Derrington and Lennie, 1984], with the magnocellular group exhibiting low spatial frequency and low chromatic sensitivity, and the parvocellular group exhibiting both high spatial frequency and high chromatic tuning. Here, we examined whether the joint coding of low spatial frequency and low color information in magnocellular units, and high spatial frequency and high color sensitivity in parvocellular units, could be an outcome of the co-occurrence of these two features at different time points during visual development (both being low early in development, and being high later on). Subsequently, we sought to determine whether it is indeed the magnocellular units that may causally drive a system's shape-based processing.

2 Methods

We trained the Alexnet [Krizhevsky et al., 2017] (slightly modified; see below), ResNet-50 [He et al., 2016], and Inception v3 [Szegedy et al., 2016] architectures on the ImageNet database [Deng et al., 2009], using two progressions of inputs (see Supplemental Methods for details):

1. In a 'standard' regimen, as a non-developmental control, we trained our networks on high-resolution, full-color images for the entire training duration (comprising 200, 40, and 20 epochs for the AlexNet, ResNet-50, and Inception v3 architectures, respectively).
2. As part of a developmentally-inspired 'biomimetic' regimen, we trained the networks on blurry, achromatic images for the first half of epochs, and on high-resolution, full-color images for the second half.

Following training, we tested the standard vs. biomimetic networks' behavior in terms of their tendency to classify images based on global shape vs. local texture features, based on the methodology detailed in Geirhos et al. [2018a] and using their 1280 test images provided online. Each of these images exhibits a shape-texture conflict – for instance, between the global shape of an airplane and the local texture of a cat. If a network were to classify such image as an airplane, the decision would be shape-consistent; classifying the image as a cat would be texture-consistent. All other classifications are judged as entirely incorrect.

To establish linkages between these behavioral signatures and receptive field characterizations, we focussed on the AlexNet, as its large first-layer filters (henceforth, receptive fields) allow for adequate frequency-based analyses. Although qualitatively similar results were obtained with the original AlexNet architecture, to further facilitate the precision of these analyses, we adjusted the network slightly in order to feature fewer but even larger receptive fields (RFs) in the first convolutional layer (including a total of 48 instead of 96 RFs, each increased from 11x11 to 22x22 pixels). For behavioral-representational linkages, we subsequently carried out systematic ablation studies.

3 Results

3.1 Biomimetic training leads to a markedly stronger and more human-like shape bias

We first investigated whether the classification decisions of the standard and biomimetic models were biased toward texture or shape when presented with images that feature texture-shape conflicts [Geirhos et al., 2018a]. As shown in Figure 1 (top panels), the standard networks exhibited no bias toward classifying images based on shape over texture, consistent with Geirhos et al. [2018a]. Notably, across all three networks tested, biomimetic training resulted in a markedly stronger and more human-like (albeit not fully human-level) shape bias, consistent also with Jang and Tong [2024] in the domain of blur. The enhanced shape bias of the biomimetic models is evident across almost all of the categories, as shown in the bottom panels of Figure 1.

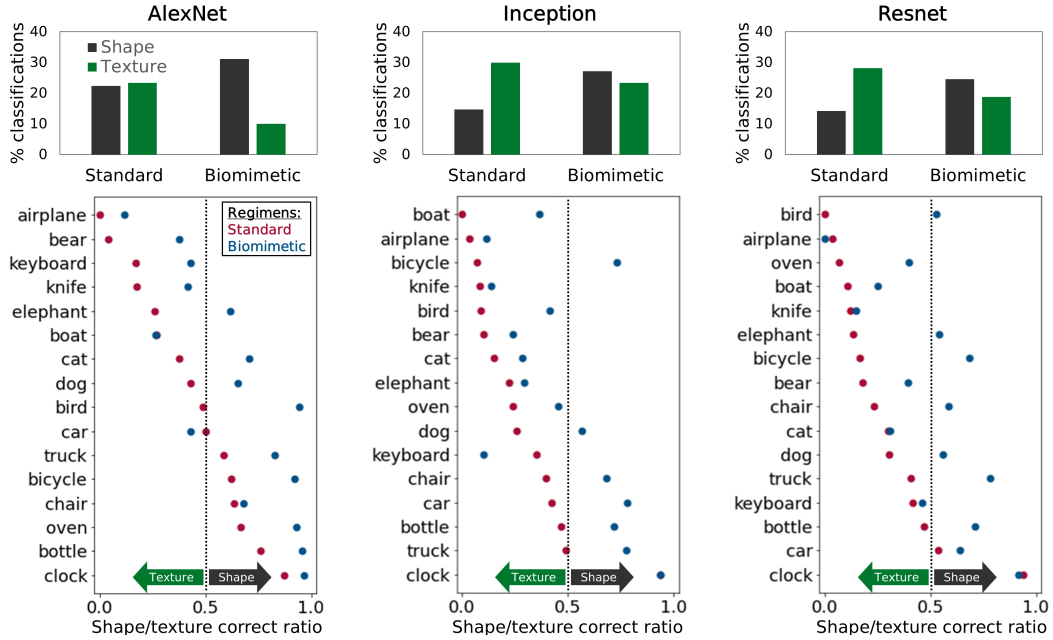

Figure 1: Results of shape/texture bias analysis. Top: Percentage of total classifications that are correct in terms of texture and, separately, correct in terms of shape (the two bars do not add up to 100% as about half of the images are classified entirely incorrectly). Bottom: Percentage of shape-based, as opposed to texture-based, correct classifications (excluding any entirely incorrectly classified images), depicted separately for each of the 16 categories.

3.2 Biomimetic training leads to a sub-population of magnocellular RFs that appear to drive the network's shape bias

Figure 2 depicts the distribution of individual RFs of the AlexNet in terms of their spatial frequency and color coding. It is evident that biomimetic training results in RFs that are markedly less tuned to high spatial frequency and color content. Most relevant for the parvo-magno distinction, we examined the joint distribution of these two features. In contrast to the standard network, whose RFs showed no clear relationship between spatial frequency and color tuning (if any, an anti-correlation), the biomimetic model exhibits a clear cluster of magnocellular-like RFs (highlighted by the ellipse). These magno units are characterized by low spatial frequency tuning and low color sensitivity.

To test whether these magno-like RFs causally contribute to the observed shape bias, we selectively ablated the network's most vs. least color-tuned RFs (with low color, rather than low frequency, here as a simple proxy index for magnocellular units). This investigation revealed that ablating less than 25 % of the least color-tuned (i.e., magno-like) RFs eliminates the shape bias of the biomimetic model entirely, while ablation of the most color-tuned (i.e., parvo-like) RFs did not have such an effect. In contrast, the standard model exhibited no differential ablation outcome. Thus, in keeping with the expected psychophysical correlate of the magnocellular system, the shape bias appears to indeed be driven by the magnocellular sub-population of RFs.

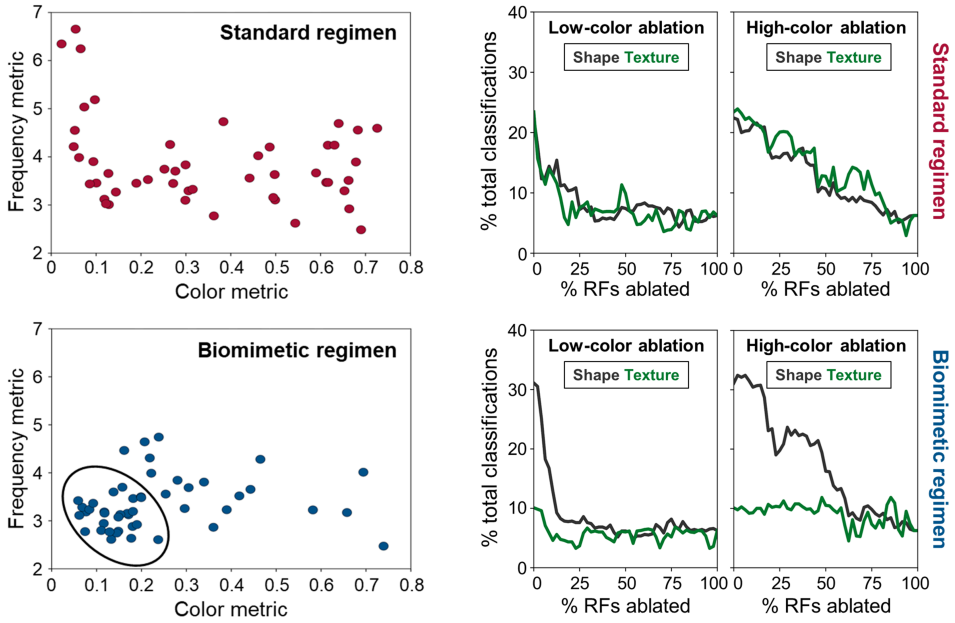

Figure 2: Results of receptive field analysis for the AlexNet. Left: Characterization of first-layer RF properties when the AlexNet was trained in a standard (top) and biomimetic (bottom) fashion. Scatter plots show the joint frequency and color coding of individual RFs. The ellipse marks a group of magno-like RFs in the biomimetic model. Right: Shape-texture bias when the least colorful (i.e., magno-like) and the most colorful (i.e., parvo-like) units are ablated, depicted separately for the standard (top) and biomimetic (bottom) model.

4 Conclusion

We have reported that training with a developmentally-inspired progression from blurry, achromatic to non-blurry, full-color imagery leads to a markedly stronger and more human-like shape bias, albeit not fully reaching human levels [Geirhos et al., 2018a]. This finding is consistent with recent findings by [Jang and Tong, 2024] in the domain of blur, and aligns well with the notion that initial degradations during perceptual development may be adaptive and provide a scaffold, rather than act as hurdles, for the acquisition of later perceptual skills [Turkewitz and Kenny, 1982, Elman, 1993, Newport, 1988, Dominguez and Jacobs, 2003, Vogelsang et al., 2018, 2023, 2024a].

In examining potential mechanisms of shape-based processing, we have also presented a candidate account of the emergence of the parvo- and magnocellular pathway division based on early developmental trajectories of sensory experience. Our computational results provide first support for this account based on joint coding properties of individual receptive fields, with magnocellular-like receptive fields furthermore emerging as drivers of a network's shape bias. Future work could examine the robustness of these findings across additional architectures (including those allowing for state-of-the-art performance) and parameter settings (including learning rates and training duration), along with more extensive representational and functional characerizations of individual units in influencing classification decisions.

In conclusion, the results presented here offer a potential account of the genesis of the parvo/magno distinction and also present a teleological perspective on why normal development progresses as it does. More broadly, this work demonstrates how findings from human development can help improve training procedures of machine learning systems.

Acknowledgements

This work has been supported by NIH grant R01EY020517 to Pawan Sinha. Lukas Vogelsang is supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT. Marin Vogelsang is supported by the Japan Society for the Promotion of Science (JSPS), Overseas Research Fellowship.

References

Russell J Adams and Mary L Courage. A psychophysical test of the early maturation of infants' mid-and long-wavelength retinal cones. *Infant Behavior and Development*, 25(2):247–254, 2002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

AM Derrington and P Lennie. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):219–240, 1984.

Velma Dobson and Davida Y Teller. Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. *Vision research*, 18(11):1469–1483, 1978.

Melissa Dominguez and Robert A Jacobs. Developmental constraints aid the acquisition of binocular disparity sensitivities. *Neural Computation*, 15(1):161–182, 2003.

Isaac Dunn, Hadrien Pouget, Daniel Kroening, and Tom Melham. Exposing previously undetectable faults in deep neural networks. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 56–66, 2021.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018a.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018b.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Hojin Jang and Frank Tong. Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of vision*, 21(12):6–6, 2021.

Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.

Omisa Jinsi, Margaret M Henderson, and Michael J Tarr. Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *Plos one*, 18(1):e0280145, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740–749, 1988.

Margaret S Livingstone and David H Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7(11):3416–3468, 1987.

Elissa L Newport. Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language. *Language sciences*, 10(1):147–172, 1988.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Gerald Turkewitz and Patricia A Kenny. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 15(4):357–368, 1982.

Lukas Vogelsang, Sharon Gilad-Gutnick, Evan Ehrenberg, Albert Yonas, Sidney Diamond, Richard Held, and Pawan Sinha. Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44):11333–11338, 2018.

Lukas Vogelsang, Marin Vogelsang, Gordon Pipa, Sidney Diamond, and Pawan Sinha. Butterfly effects in perceptual development: A review of the 'adaptive initial degradation'hypothesis. *Developmental Review*, 71:101117, 2024a.

Marin Vogelsang, Lukas Vogelsang, Sidney Diamond, and Pawan Sinha. Prenatal auditory experience and its sequelae. *Developmental Science*, 26(1):e13278, 2023.

Marin Vogelsang, Lukas Vogelsang, Priti Gupta, Tapan K Gandhi, Pragya Shah, Piyush Swami, Sharon Gilad-Gutnick, Shlomit Ben-Ami, Sidney Diamond, Suma Ganesh, et al. Impact of early visual experience on later usage of color cues. *Science*, 384(6698):907–912, 2024b.

Torsten N Wiesel and David H Hubel. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *Journal of neurophysiology*, 29(6):1115–1156, 1966.

Lorijn Zaadnoordijk, Tarek R Besold, and Rhodri Cusack. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520, 2022.

A Supplemental methods

Network training

All networks were implemented using Keras / TensorFlow v2. The official train / test set split of the ImageNet database was used, resulting in a test set containing 50,000 images (50 for each of the 1000 ImageNet classes). Training was carried out with a constant learning rate of 0.001, a batch size of 128, the categorical cross-entropy loss function, and Stochastic Gradient Descent as optimizer with a Nesterov momentum of 0.9. Training comprised a total of 200 epochs for the AlexNet, 40 epochs for the ResNet-50, and 20 epochs for the Inception v3 architecture, to ensure fair convergence. Image preprocessing / augmentation only comprised horizontal flipping at random, pixel value rescaling from a [0, 255] to a [-1, 1] distribution, and random cropping of 256x256 pixel images to 227x227 pixel segments (for the AlexNet) or to 224x224 pixel segments (for the ResNet). For the Inception architecture, the full 256x256 pixels were used as inputs. When blurring (for training or testing), a Gaussian blur with a sigma of 4 was applied.

Color metric for receptive field analysis

To quantify the colorfulness of a given first-layer convolutional filter (receptive field) of the minimally modified AlexNet, we first computed the intensity differences, m, across the three color channels, for each individual pixel:

$$x = R\cos(0^\circ) + G\cos(120^\circ) + B\cos(-120^\circ) \tag{1}$$

$$y = R\sin(0^\circ) + G\sin(120^\circ) + B\sin(-120^\circ) \tag{2}$$

$$m = \sqrt{x^2 + y^2} \tag{3}$$

where R, G, and B represent the three channels' pixel-wise intensities. This distribution (over the 22x22 pixels for a given receptive field) was then summarized by computing the average of the top-48 (i.e., approximately the top-10 %) most colorful pixels.

Spatial frequency metric for receptive field analysis

To quantify the spatial frequency content of each receptive field, receptive fields were first transformed to grayscale and upscaled by a factor of 100 (to avoid any noise due to the discreteness of the index). Then, a 2D-FFT was applied, and the presence of different frequencies was summarized using radial averaging. This provided us with a 1D histogram over frequencies, which we further summarized in our spatial frequency metric as a weighted average:

$$\text{weighted average frequency} = \frac{\sum_f (\text{amp}(f) \cdot f)}{\sum_f \text{amp}(f)} \tag{4}$$

where amp refers to the amplitude of a given frequency, and f refers to the frequency itself.