Controlled-rearing studies of newborn chicks and deep neural networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Convolutional neural networks (CNNs) can now achieve human-level performance on challenging object recognition tasks. CNNs are also the leading quantitative models in terms of predicting neural and behavioral responses in visual recognition tasks. However, there is a widely accepted critique of CNN models: unlike newborn animals, which learn rapidly and efficiently, CNNs are thought to be "data hungry," requiring massive amounts of training data to develop accurate models for object recognition. This critique challenges the promise of using CNNs as models of visual development. Here, we directly examined whether CNNs are more data hungry than newborn animals by performing parallel controlled-rearing experiments on newborn chicks and CNNs. We raised newborn chicks in strictly controlled visual environments, then simulated the training data available in that environment by constructing a virtual animal chamber in a video game engine. We recorded the visual images acquired by an agent moving through the virtual chamber and used those images to train CNNs. When CNNs were provided with similar visual training data as chicks, the CNNs successfully solved the same challenging view-invariant object recognition tasks as the chicks. Thus, the CNNs were not more data hungry than animals: both CNNs and chicks successfully developed robust object models from training data of a single object.

1 Introduction

After decades of lagging behind the recognition abilities of even young children, machine-learning systems can now rival human adults on challenging object recognition tasks (Krizhevsky et al., 2012). In addition to powering new technologies, modern machine-learning systems are serving as executable, neurally-mechanistic models in the brain and cognitive sciences (Kriegeskorte & Douglas, 2018). Cognitive scientists have been particularly interested in a class of computer vision models called deep convolutional neural networks (CNNs), which are directly inspired by neurophysiological observations taken from biological visual systems, including a restricted connectivity pattern that resembles the receptive field organization found in the animal visual cortex (Hubel & Wiesel, 1968; Fukushima, 1980; LeCun et al., 1990). These "end-to-end" models learn to recognize objects from raw high-dimensional sensory inputs (pixels) and produce behavioral categorizations as outputs. Notably, these models produce internal unit response properties at each level of the network that are similar to actual neurophysiological unit responses at the corresponding levels in animal visual systems (Yamins & DiCarlo, 2016). CNNs can also be used to control the activity state of individual neurons and populations of neurons, indicating that these models capture rich causal properties of how biological visual systems process information (Bashivan et al., 2019). These models are built at scale (i.e., they can take any retinal image as input) and can rival the behavioral performance of mature animals across many challenging visual recognition tasks (Schrimpf et al., 2020).

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

Despite these strengths, there is a lingering worry about using CNNs as models of the brain. The worry relates to the large amount of data needed to train CNNs, compared to the relatively small amount of training data apparently needed by newborn animals. Specifically, the CNNs that have revolutionized computer vision require massive amounts of training data. These models often have hundreds of layers, tens of millions of parameters, and are trained on millions of images across a thousand different object categories. In contrast, newborn animals require small amounts of training data in order to solve challenging perceptual and motor tasks, with many core abilities emerging within the first few days of life (Held & Hein, 1963; Walk et al., 1957).

One of the most striking differences between newborn animals and CNNs comes in the domain object recognition. In particular, a number of automated controlled-rearing experiments with newborns chicks have demonstrated that chicks rapidly learn to solve challenging object perception tasks, even in the absence of extensive visual experience with objects. Soon after hatching, newborn chicks are capable of object parsing (Wood & Wood, 2021), visual binding (Wood, 2014), view-invariant object recognition (Wood, 2013; Wood & Wood, 2015a), view-invariant face recognition (Wood & Wood, 2015b), rapid object recognition (Wood & Wood, 2017), action recognition (Goldman & Wood, 2015), and object permanence (Prasad et al., 2019). All of these abilities emerge when chicks see their first object, indicating that newborn visual systems can perform "one-shot" learning without extensive training data with objects. From the perspective of CNNs, this is an impressive feat. CNNs typically require thousands to millions of labeled training images to develop object recognition, whereas newborn chicks develop object perception from visual experience with a single object. Consequently, learning in newborn brains appears to be fast and efficient, whereas learning in CNNs appears to be slow and inefficient. If this conclusion were accurate, then it would place significant constraints on the use of CNNs as models of visual development.

Here, we suggest that the learning gap between animals and machines is not as large as it appears. To compare learning across animals and machines, we performed parallel controlled-rearing experiments on newborn chicks and CNNs. First, we raised newborn chicks in strictly controlled virtual environments and measured the chicks' view-invariant object recognition performance. Second, we created a virtual simulation of the controlled-rearing chambers in a video game engine, and then simulated the visual training data available in the chick's environment by recording the first-person images acquired by an agent moving through the virtual animal chamber. Third, we trained CNNs using that simulated training data and tested their object recognition performance with the same recognition tasks used to test the chicks. Accordingly, the chicks and CNNs were trained with the same visual data and tested with the same tasks, allowing for direct comparison of their learning abilities.

2 Method

Animal experiments & stimuli In this work, we focused on the view-invariant recognition task from Wood (2013). After hatching, each chick was moved from the incubator to a controlled-rearing chamber in darkness. The controlled-rearing chamber was equipped with two display walls that were used to display the object animations. The chicks' entire visual object experience was limited to the virtual objects projected on the display walls.

The stimulus set from Wood (2013) consisted two 3D objects, each of which was shown from 12 different viewpoint ranges. Each animation displayed the object rotating through a 60° viewpoint range about an axis passing through its centroid, completing the full back and forth rotation every 6s. During the first week of life (Training Phase), the chicks' visual experience was limited to a single virtual object seen from a single viewpoint range. During the second week (Test Phase), the chicks were tested on their ability to recognize that object from the 12 viewpoints (11 novel, 1 familiar), using a two-alternative forced choice test. The imprinted object was shown on one display wall, and the unfamiliar object was shown on the opposite display wall. Test trials were scored as "correct" when the chicks spent a greater proportion of time with their imprinted object and "incorrect" when the chicks spent a greater proportion of time with the unfamiliar object. In order to succeed on the task, chicks needed to learn invariant object representations that generalize across large, novel, and complex changes in the object's appearance.

The chicks performed well on the task, scoring 79% (chance = 50%) when the object was shown from a familiar view and 69% when the object was shown from novel viewpoints. Thus, the chicks successfully generated view-invariant representations that generalized across substantial variation in

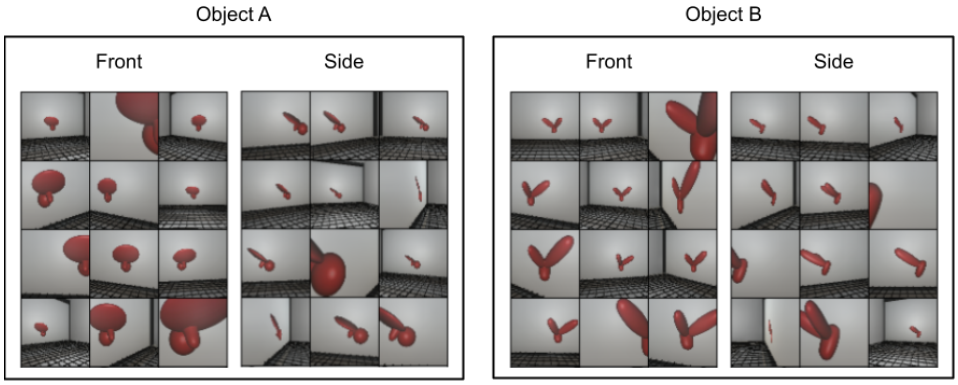

Figure 1: Examples of the simulated training data used to train the CNN models. The dataset was constructed by recording the visual observations of an agent moving within a virtual controlled-rearing chamber.

the object's appearance. This study demonstrates that newborn visual systems are capable of building robust object representations from training data of a single object seen from a limited range of views.

Simulating the training data available to chicks In order to mimic the visual experiences of the chicks raised in the controlled-rearing chambers, we constructed an image dataset (Figure 1) by sampling visual observations of an agent moving through a virtual controlled-rearing chamber. The virtual chamber and agent were created with the Unity 3D game engine and the ML-Agents Toolkit (Juliani et al., 2020). We first sampled the visual observations of an agent moving randomly through the chamber, then programmatically removed images that did not contain an object. The resulting dataset contained 40,000 images (10k images for each of the 4 object animations that the chicks were raised with during the training phase).

Unsupervised training Understanding the origins and development of object recognition requires building computational models that learn like newborn brains, using unsupervised learning algorithms. To this end, we trained CNNs with the images sampled from the virtual controlled-rearing chamber, using unsupervised learning methods. We used two unsupervised learning algorithms to train our CNNs: variational autoencoders (Kingma & Welling, 2014) and contrastive learning (Chen et al., 2020). We used a standard ResNet architecture (He et al., 2016) with 18 layers as the base encoder for all of our CNN models. Importantly, the models only observed a single object rotating through a single viewpoint range during training. For each unsupervised method, we trained models in each of the 4 visual rearing conditions (2 objects presented from 2 viewpoint ranges) presented to the chicks.

Linear evaluation With linear classifiers, we evaluated the classification performance of the unsupervised CNNs using the same recognition task that was presented to the chicks. Task performance was assessed by adding a single fully connected linear readout layer on top of the penultimate layer of each trained CNN and then training only the parameters of that readout layer on the binary object classification task.

For linear evaluation, we collected 24,000 simulated visual observations from an agent moving randomly in the virtual chamber (1,000 images for each of 2 objects in 12 viewpoint ranges). The object identities were used as the ground-truth labels.

To evaluate whether the learned representations could generalize across novel viewpoints, we systematically varied the number of viewpoint ranges used to train the linear classifiers. To do this, we used three different training and evaluation splits: 6-fold, sparse 6-fold, and sparse 12-fold.

- 6-fold: We first divided the dataset into 6 folds such that each fold contained images of each object rotating through 2 viewpoint ranges. The linear classifiers were cross-validated by training on 5 folds (10 viewpoint ranges) and testing on the held-out fold (2 viewpoint ranges).

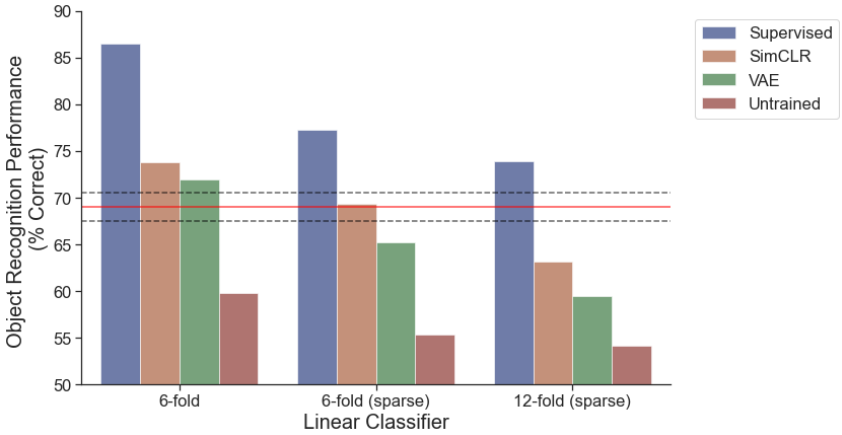

Figure 2: View-invariant object recognition performance of unsupervised and baseline CNNs. In the 3 linear classifier conditions, the number of viewpoints used to train and test the classifiers differed across conditions: "6-fold" (*train*: 10 viewpoints; *test*: 2 viewpoints), "6-fold sparse" (*train*: 2 viewpoints; *test*: 10 viewpoints), and "12-fold sparse" (*train*: 1 viewpoint; *test*: 11 viewpoints). The chicks and unsupervised CNNs were only trained with images of a single object seen from a single viewpoint range, and the linear classifiers were cross-validated with different views in the training versus test sets. Thus, these results reflect the *generalization performance* of the chicks and models across novel views. The red horizontal line shows the chicks' performance on the 2-way forced choice task, with black dashed lines representing standard error.

- Sparse 6-fold: We divided the dataset into 6 folds in the same way as previously described, but importantly, we inverted the ratio of the training and validation data: 1/6 of the data were used for training and 5/6 of the data were used for testing. Thus, the linear classifiers were trained using only 2 viewpoint ranges and tested on the remaining 10 viewpoint ranges.
- Sparse 12-fold: In this extreme case, we split the dataset into 12 folds so that each fold contained images of each object rotating through a single viewpoint range. As such, 1/12 of the data were used for training and 11/12 of the data were used for testing. Thus, the linear classifiers were trained using only 1 viewpoint range of each object and tested on the remaining 11 viewpoint ranges.

For each of the three linear classifier conditions, transfer performance was evaluated by first fitting the parameters of a linear classifier on the training set and then measuring classification accuracy on the held-out test set. We report average cross-validated performance on the held-out images not used to train the linear readout layer.

Baselines We also included two baseline models: untrained and supervised CNNs. The untrained CNN was a randomly initialized ResNet-18 encoder. The supervised CNN had the same network architecture as the untrained baseline, but was optimized for a challenging large-scale object classification task (ImageNet; Deng et al., 2009). CNNs trained on the ImageNet dataset achieve high transfer performance on various downstream visual tasks (Huh et al., 2016) and can serve as accurate models of object recognition in mature visual systems (Battleday et al., 2020; Yamins & DiCarlo, 2016). Thus, the transfer performance of an ImageNet trained CNN can serve as a proxy to quantify the performance of an "ideal" observer (mature visual system) on the task.

3 Results and Discussion

Figure 2 shows the view-invariant object recognition performance of the CNNs. We also report the performance of the newborn chicks from Wood (2013). The baseline ImageNet-trained supervised CNN performed the best among all models across all cross-validation schemes, outperforming the chicks by 4.9% points even with the most sparse linear classifiers (sparse 12-fold). However, this CNN was trained on a massive dataset (millions of images across 1,000 object categories). Can CNNs still perform well on this task when they receive sparse training data, akin to newborn chicks?

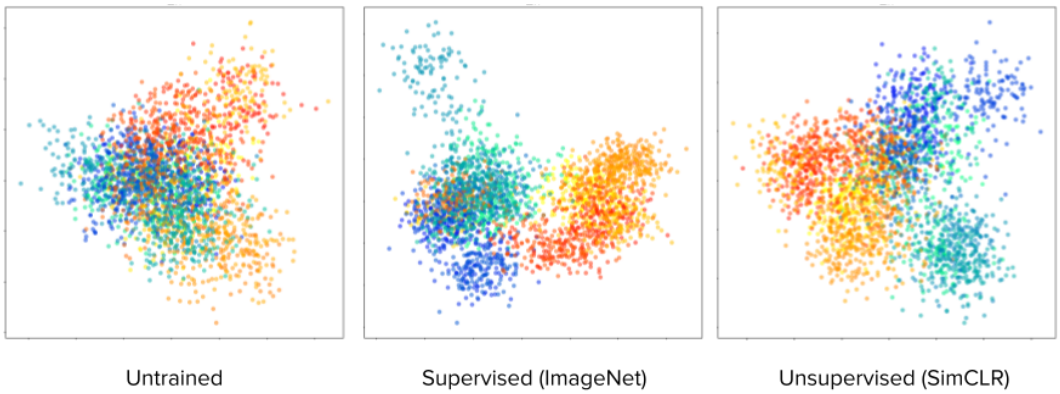

Figure 3: Two-dimensional projections of the feature representations from the untrained, supervised, and unsupervised CNNs. Each point represents a CNN representation of an input image containing a single object. Colors denote the identities and viewpoint ranges of the objects; warm colors (red-yellow) represent Object 1 and cold colors (green-purple) represent Object 2. To create these visualizations, we used linear-discriminant-analysis.

Remarkably, the contrastive learning method (SimCLR) showed comparable performance as the chicks, indicating that CNNs can build invariant object representations from sparse visual input. Across all of the linear classifier conditions, the contrastive learning method (SimCLR) showed better transfer learning than the variational autoencoders. This result is consistent with previous studies reporting that contrastive learning methods outperform other classes of unsupervised methods in downstream classification tasks (Chen et al., 2020; Zhuang et al., 2021).

To visualize the representational space learned by the CNNs, we used two-dimensional linear discriminant analysis (LDA) (Figure 3). We first extracted CNN features of images randomly sampled from the simulated dataset and then fitted an LDA model using those features as inputs. To test whether the underlying CNN representations encoded linearly separable information about object identity, we used the combination of 12 viewpoint ranges and 2 object identities (instead of object identities alone) as labels. Although the ground truth of object identity was not explicitly encoded in the labels, the resulting LDA plots for unsupervised (SimCLR) and supervised (ImageNet) features showed clear distinction between the two object categories (green-blue dots vs red-yellow dots).

Overall, our results show that CNNs can learn to solve view-invariant object recognition tasks from sparse visual input, akin to newborn chicks. For both chicks and CNNs, a visual environment with a single object contains sufficient training data for building a view-invariant object representation. Thus, CNNs might not be as "data hungry" as previously thought. Given that newborn animals and CNNs are capable of similar feats of object recognition after receiving similar sets of training data, we argue that CNNs can serve as models of visual development in animals. Indeed, the initial "proto-architecture" of newborn visual systems shares two core organizational principles with CNNs: both CNNs and newborn visual systems are hierarchically and retinotopically organized (Arcaro & Livingstone, 2021). This finding suggests that the hierarchical and retinotopic architecture of CNNs might be a reasonably good approximation of the "initial state" of newborn visual systems.

There would be many advantages to using CNNs as models of visual development because current theories of visual development are incomplete in many ways. First, they are not image computable, requiring a human in the loop to determine the predictions a theory should make in response to a particular stimulus. Second, current theories do not provide a quantitative account of how newborn visual systems learn and respond to raw visual inputs. Third, current theories do not explain why newborn brains develop the way they do. By using CNNs within the framework of goal-driven modeling, we can tackle these problems by building neurally mechanistic, image computable models of newborn visual systems. CNNs can also be tested on a wide range of novel instances of inputs and rigorously falsified in ways that prior models of visual development cannot. Goal-driven modeling with CNNs has revolutionized the study of mature vision. Likewise, we speculate that this approach can provide a strong computational foundation for the study of visual development.

References

Michael J. Arcaro and Margaret S. Livingstone. On the relationship between maps and domains in inferotemporal cortex. *Nature Reviews Neuroscience*, 22(9):573–583, September 2021.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019.

Ruairidh M. Battleday, Joshua C. Peterson, and Thomas L. Griffiths. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, 11 (1):5418, December 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, June 2020.

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*, pp. 248–255, 2009.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.

Jason G. Goldman and Justin N. Wood. An automated controlled-rearing method for studying the origins of movement recognition in newly hatched chicks. *Animal Cognition*, 18(3):723–731, May 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Richard Held and Alan Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876, 1963.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, March 1968.

Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614 [cs]*, December 2016.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A General Platform for Intelligent Agents. *arXiv:1809.02627 [cs, stat]*, May 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114 [stat.ML]*, December 2014.

Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160, September 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Y. LeCun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, volume 2, 1990.

Aditya Prasad, Samantha M. W. Wood, and Justin N. Wood. Using automated controlled rearing to explore the origins of object permanence. *Developmental Science*, 22(3), May 2019.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, pp. S089662732030605X, September 2020.

R. D. Walk, E. J. Gibson, and T. J. Tighe. Behavior of light-and dark-reared rats on a visual cliff. *Science*, 126:80–81, 1957.

Justin N. Wood. Newborn chickens generate invariant object representations at the onset of visual object experience. *Proceedings of the National Academy of Sciences*, 110(34):14000–14005, August 2013.

Justin N. Wood. Newly Hatched Chicks Solve the Visual Binding Problem. *Psychological Science*, 25(7):1475–1481, July 2014.

Justin N. Wood and Samantha M.W. Wood. Measuring the speed of newborn object recognition in controlled visual worlds. *Developmental Science*, 20(4):e12470, July 2017.

Samantha M. W. Wood and Justin N. Wood. A chicken model for studying the emergence of invariant object recognition. *Frontiers in Neural Circuits*, 9, February 2015a.

Samantha M. W. Wood and Justin N. Wood. Face recognition in newly hatched chicks at the onset of vision. *Journal of Experimental Psychology: Animal Learning and Cognition*, 41(2):206–215, April 2015b.

Samantha M. W. Wood and Justin N. Wood. One-shot object parsing in newborn chicks. *Journal of Experimental Psychology: General*, 2021.

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, January 2021.