LAYERWISE LEARNING RATES FOR OBJECT FEATURES IN UNSUPERVISED AND SUPERVISED NEURAL NETWORKS AND CONSEQUENT PREDICTIONS FOR THE INFANT VISUAL SYSTEM

Anonymous authors
Paper under double-blind review

ABSTRACT

To understand how object vision develops in infancy and childhood, it will be necessary to develop testable computational models. Deep neural networks (DNNs) have proven valuable as models of adult vision, but it is not yet clear if they have any value as models of development. As a first model, we measured learning in a DNN designed to mimic the architecture and representational geometry of the visual system (CORnet). We quantified the development of explicit object representations at each level of this network through training by freezing the convolutional layers and training an additional linear decoding layer. We evaluate decoding accuracy on the whole ImageNet validation set, and also for individual visual classes. CORnet, however, uses supervised training and because infants have only extremely impoverished access to labels they must instead learn in an unsupervised manner. We therefore also measured learning in a state-of-the-art unsupervised network (DeepCluster). CORnet and DeepCluster differ in both supervision and in the convolutional networks at their heart, thus to isolate the effect of supervision, we ran a control experiment in which we trained the convolutional network from DeepCluster (an AlexNet variant) in a supervised manner. We make predictions on how learning should develop across brain regions in infants. In all three networks, we also tested for a relationship in the order in which infants and machines acquire visual classes, and found only evidence for a counter-intuitive relationship. We discuss the potential reasons for this.

1 INTRODUCTION

1.1 THE DEVELOPMENT OF OBJECT RECOGNITION IN INFANTS

Visual discrimination of objects begins to develop early in infancy. Even newborns orient towards faces rather than other stimuli, showing that they can discriminate them from an early age (Johnson et al., 1991). By 3-4 months old, infants can identify statistical regularities across exemplars - for example, when presented a sequence of images from one visual class (e.g., cats) they prefer to look at a subsequently presented animal if it is from a deviating class (e.g., a dog) (Quinn et al., 1993; French et al., 2004). At 6 months, infants start to look at a visual class corresponding to a concurrently spoken label (Bergelson & Swingley, 2012), although vocabulary remains very limited until after the first birthday, when it typically begins to grow rapidly.

In addition to measuring behaviour, neuroimaging can measure activity in the ventral visual stream, the brain system responsible for object vision. In adults, functional magnetic resonance imaging (fMRI) has found that there are regions that are selective for particular visual classes, such as faces, body parts or places (Kanwisher et al., 1997; Epstein & Kanwisher, 1998; Downing et al., 2001). In 4-6 month old infants, selectivity is already present in the ventral visual stream (Deen et al., 2017), although it continues to develop for many years (Gomez et al., 2017).

1.2 Deep learning as a model of the development of vision

In studies of object vision in adults, DNNs have proven valuable (Yamins & DiCarlo, 2016). By examining which visual classes evoke similar or dissimilar patterns of activity, it has been found that the representational geometry in these DNNs can capture some of the representational geoemetry of the ventral visual stream, as measured with functional magenetic resonance imaging (fMRI) (Khaligh-Razavi & Kriegeskorte, 2014; Guclu & van Gerven; Wen et al., 2018), electroencephalography (EEG) and behavioural studies (Cichy et al., 2016).

Thus, fully-trained DNNs are valuable models of the adult ventral visual stream, but this does not imply that they will be valuable models of the developmental process. It is the overarching goal of our work to develop ways to test the value of DNNs in explaining infant behaviour and brain development.

1.3 Aims of the current study

We begin to test the parallel between learning in infants and DNNs by addressing two open questions:

1. Should we expect representations in brain regions of the visual hierarchy to develop simultaneously or asynchronously? A principle of infant development is that brain regions underlying simpler functions develop first, and are followed by those underlying more complex functions [Charles A. Nelson in Shonkoff & Phillips (2000)]. To provide a model of whether this might happen in the visual hierarchy we examined how representations developed in different layers of a DNN during training.
2. Are visual classes that are learned earlier by infants also learned earlier by DNNs and/or represented in shallower layers? In infants, the acquisition of visual classes can be estimated by the onset of the receptive or expressive use of the words for the classes. It has been found that, when measured this way, some visual classes are learned before others, e.g., body parts and vehicles precede food and clothing (Braginsky et al., 2015). We ask whether some of this ordering is attributable to visual complexity, as reflected in the DNN learning rates and/or layers.

2 Methods

2.1 Choice of networks and training

2.1.1 CORnet-S with supervised training

Studies that have shown a parallel between DNNs and the adult brain have used networks trained in a supervised manner (Khaligh-Razavi & Kriegeskorte, 2014; Guclu & van Gerven; Cichy et al., 2016; Wen et al., 2018; Jozwik et al., 2018). For comparison with these studies, we therefore started with a supervised network. Specifically, we used CORnet-S (Kubilius et al., 2018), as this was designed to capture the architectural principles and representational geometry of the ventral visual stream while achieving good classification performance. It has four blocks mapping onto regions in the ventral visual stream (V1, V2, V4 and IT).

2.1.2 DeepCluster with unsupervised training

Infants' access to labels is extremely impoverished and so to develop computational models that capture the dynamics of infant learning it will be necessary to evaluate unsupervised training. One current state-of-the-art unsupervised strategy for learning visual features for object recognition is DeepCluster (Caron et al., 2018). This uses a simple but elegant technique for self-supervised learning, in which at the start of an epoch, each image (from ImageNet) is passed forward through a convolutional network, and the resulting output activations are clustered across all of the images using k-means. These clusters are assigned labels, which are then used to learn the weights of the convolutional network with stochastic gradient descent on batches of images in the typical way.

Like Caron et al. (2018) we used 10,000 clusters. As a convolutional network, we used AlexNet (Krizhevsky et al., 2012) which contains five blocks each with a single convolutional layer. These blocks were modified as in Caron et al. (2018), with the local response normalisation layers removed and batch normalisation used instead (Ioffe & Szegedy, 2015). Also like Caron et al. (2018), a fixed linear transformation based on Sobel filters was used on the input to remove colour and increase local contrast.

2.1.3 AlexNet with supervised training

CORnet was trained in a supervised way, and DeepCluster in an unsupervised way. Any difference in the results might be due to this difference in supervision. However, the convolutional networks at the heart of these two networks also differ, which could also cause differences. To control for this, we therefore repeated the experiments using the same AlexNet variant as in DeepCluster, but trained in a supervised way.

2.2 Learning trajectories

The value for object recognition of the representations was assessed for each of the blocks in the networks (4 blocks for CORnet, and 5 blocks for DeepCluster and Alexnet). Specifically, we quantified the explicit representation (DiCarlo & Cox, 2007) of object class using the method by Zhang et al. (2017) of freezing weights in the convolutional layers, and training a linear decoder on the output of each block to decode the ImageNet categories. This was done across epochs in the learning process, to capture the development of the representations in each of the layers.

2.3 Summarising learning of visual classes

To summarise the learning curves of the individual classes we fitted the performance with the curve

$$p = A(1 - exp(-kt))$$

where p is top-5 precision, t the epoch, A the asymptotic level of performance, and k the learning rate. Fitting minimised least squares with the addition to the cost function of two regularisation terms equal to k^2 (to discourage implausibly high learning rates) and $(A < 100) * (A - 100)^4$ (to discourage A values greater than 100

2.4 AoA in Infants

We also compared the order in which visual classes are acquired in infants and machines. To assess when infants acquire each visual class, we used estimates for the age of acquisition (AoA) of the word for the class (Kuperman et al., 2012). A number of linguistic factors are known to affect when words are first used, including the frequency of the word in language and its number of phonemes, but the second strongest factor is the ”concreteness” of the word (Braginsky et al., 2015). This suggests that the strength of the visual representation of a class has an effect on when its label is acquired. We tested this in two ways: by relating the AoA of a class in infants to the speed of acquisition in a network; and by relating it to the degree to which a class is decoded in early vs. late layers.

Using the Natural Language Toolkit (NLTK), the WordNet synsets for the 1000 ImageNet classes were compared to Kuperman et al. (2012) database of 30,000 English words with age-of-acquisition ratings using Leacock et al. (1998) semantic similarity metric. The classes with the highest similarity score were considered as matching, and manually inspected for any incorrect comparisons or synset definitions. These were deleted, leaving a total of 308 classes in which further analyses were conducted.

2.5 Class Categorisation

To provide a visualisation of when different types of classes were learned, we clustered the 308 classes using Leacock et al. (1998)'s metric and then clustering (scipy.cluster.hierarchy.fcluster) to yield 20 classes. By visual inspection, we then attached a label to each of the classes.

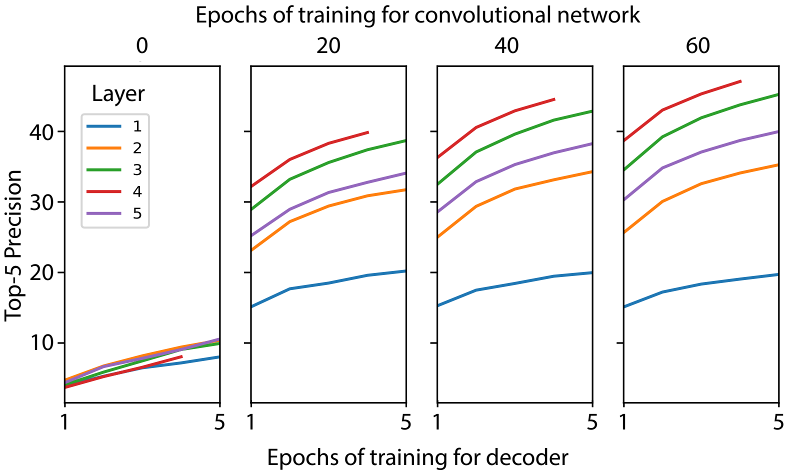

Figure 1: Top-5 precision as a function of training epochs of the top decoding layer for DeepCluster.

2.6 Implementation

Training was run on AWS using the Deep Learning AMI version 24.0 on either a p2.8xlarge instance (8 x NVIDIA K80 GPUs with 488 GB of RAM) or a p3.8xlarge instance type (4 NVIDIA Tesla V100 GPUs and 244 GB RAM), using Python 3.6 with Pytorch 1.1. Spot instances were used to reduce cost. The three networks were trained from scratch. The ISLVRC 2012 set was used for training and validation. The CORnet-S code was obtained from `https://github.com/dicarlolab/CORnet`. DeepCluster, Alexnet and the linear classifier implementation were from `https://github.com/facebookresearch/DeepCluster`.

3 Results

3.1 Determining number of training epochs for the object decoder

Training the object decoders was the most computationally expensive part of this project, as one was trained for every layer across many epochs and models. It was therefore desirable to use as few training epochs as possible. To evaluate how many were needed, we trained decoders for 5 epochs on features from a sample of convolutional training epochs (0, 20, 40, 60) and all layers 1. It was found that while there was a steady increase in decoding performance up to (and presumably beyond) the 5 epochs, the relative performance across different layers, or epochs, was broadly captured by epoch 2. For further analyses we therefore used 2 epochs of training for the decoding layer.

3.2 Aim 1: Development of representations in different layers during training

3.2.1 CORnet

Explicit representation of object class in the four blocks of the CORnet network during training is shown in Fig. 2a. The earlier blocks in the hierarchy (V1, V2 and V4) reached their asymptotic level quickly (around epoch 1), but IT continued to learn until at least epoch 25. However, although IT took longer to reach its asymptote, even after minimal training (epoch 1) it contains greater explicit information than the lower layers.

Extrapolating from this model, therefore, we might expect in the infant brain to see substantially earlier maturation of lower-order visual processing regions (V1, V2 and V4), than higher-order brain regions (e.g., IT). However, even early in development, infant IT would be expected to contain stronger explicit representations of object class.

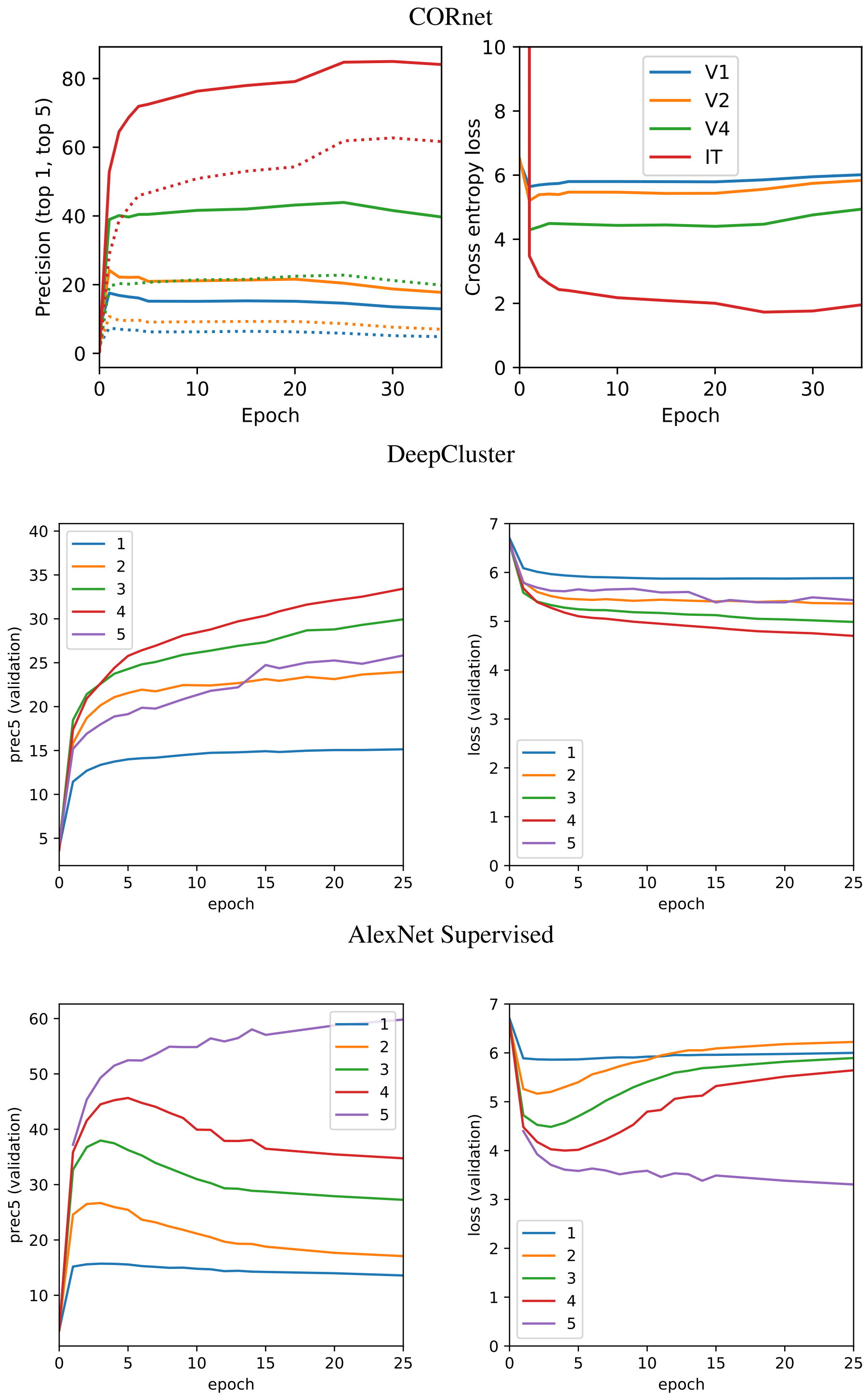

Figure 2: Explicit representation of visual class measured in the three networks during training. The left panels show top-5 precision and the right panels cross-entropy loss.

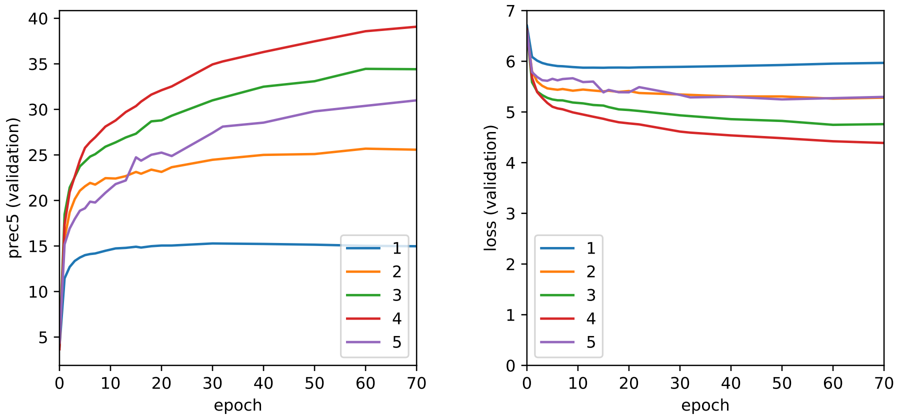

Figure 3: The unsupervised network, DeepCluster, learned more slowly than the supervised networks. To test if the layers with the strongest explicit object representation changed over a longer period of extended learning, we trained the convolutional layers to 70 epochs.

3.2.2 DeepCluster

In contrast, for DeepCluster, representations mature in a more ”bottom-up” manner (Fig. 2b). Specifically, the explicit representation of object class does not monotonically increase with layer - even at the end of 60 training epochs, layer 4 contains stronger representation of object class than layer 5. Furthermore, the order of the layers varies through training, with layer 3 stronger than layer 4 early in training, and layer 2 containing stronger representations of object class than layer 5.

As DeepCluster learned more slowly than the supervised networks, we extended training to 70 epochs 3. It can be seen that it was continuing to learn, particularly in the higher layers, but the order of the layers did not change within this range.

This more developmentally plausible unsupervised model predicts not only that higher-order visual regions will develop more slowly, but that earlier regions may initially lead in the presence of representations of object class.

In supervised training, object labels are provided at the top layer of the network, and so it is perhaps not surprising that even at early epochs the entire network including the upper layers are maturing. In contrast, in unsupervised learning, the only source of information is the visual input, and so it is perhaps not surprising that maturation proceeds in a more bottom up manner: until good representations have developed in the early layers, there is poorer information at higher layers.

3.2.3 AlexNet

However, CORnet and DeepCluster are not just different in their training strategies, but also the convolutional networks at their heart. To control for this, we repeated training with AlexNet. The results, in Fig. 2c show that even when the same convolutional network as DeepCluster is used, but with a supervised training strategy, the bottom-up learning trajectories of DeepCluster are eliminated.

Strikingly, explicit object representation in the lower layers actually reduced from epochs 3-5 onwards, when assessed with top-5 or loss. This was seen much more weakly for CORnet, and not at all for DeepCluster. It appears to be a feature of supervised learning.

3.3 Aim 2: Acquisition of visual classes

The learning curves were fit well by the model (left two columns of Fig. 4). The joint distribution (Fig. 4, right column) showed that classes which are learned quickest are ultimately learned best, as

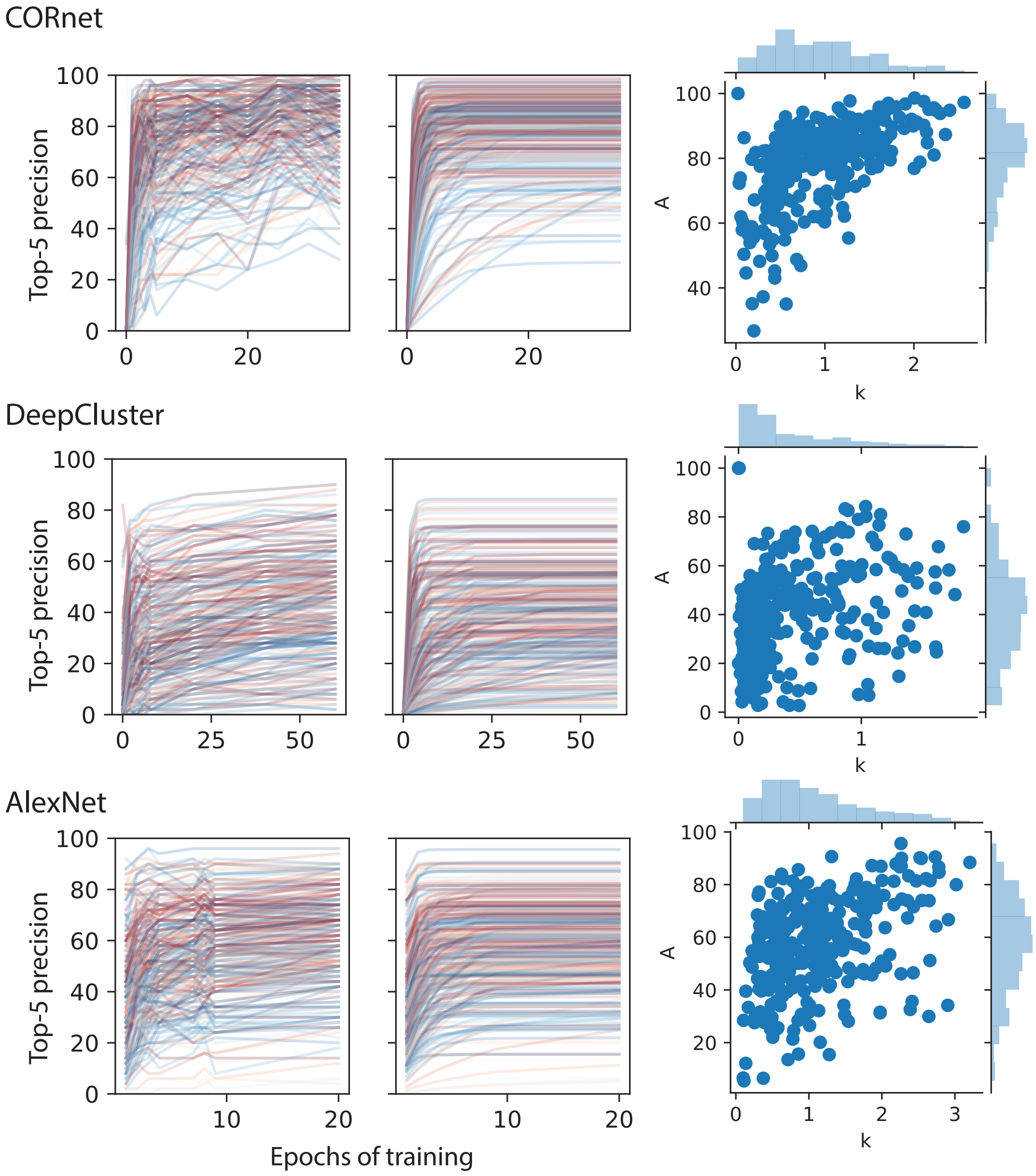

Figure 4: Left: Precision for each visual class during training, in the most informative layer of the network - the top layer for CORnet and AlexNet, and the penultimate layer for DeepCluster. The color of the curves shows AoA of the class's name for infants (blue to red for low to high AoA). Centre: The learning curves were parameterised with a fit, shown here. Right: Distributions of the fit parameters. They were correlated for all networks, showing that classes that were learned more quickly were also converging on a higher asymptote.

the two fit parameters were strongly correlated for all three models (CORnet r(308)=0.62; p¡0.001; DeepCluster r(308)=0.36, p¡0.01; AlexNet r(308)=0.39, p¡0.001)

These fit parameters were then used to compare the machine with human learning. Paradoxically, classes learned more precisely by the model were if anything learned *later* by infants (Fig. 5, correlation of AoA and parameter A, CORnet r(308)=0.11 p=0.06; DeepCluster r(308)=0.10, p=0.09 ; AlexNet r(308)=0.14, p¡0.02). Although classes learned more precisely were in general learned more quickly in the model, there was no relationship observed between learning rate parameter (k) and infant age of acquisition. Examination of the best and worst classes (top-5) suggested that possibly the object's context strongly drove machine categorisation (e.g., presence of water or snow).

Furthermore, from CORnet, although some classes were explicitly well represented as early as V1, there was no evidence that these classes were acquired earlier in infants (r=0.02, 0.01, 0.03 in V1, V2 and V4, respectively, all N.S.). As in the fits, classes with higher precision in IT were acquired later by infants (r=0.15, p¡0.01).

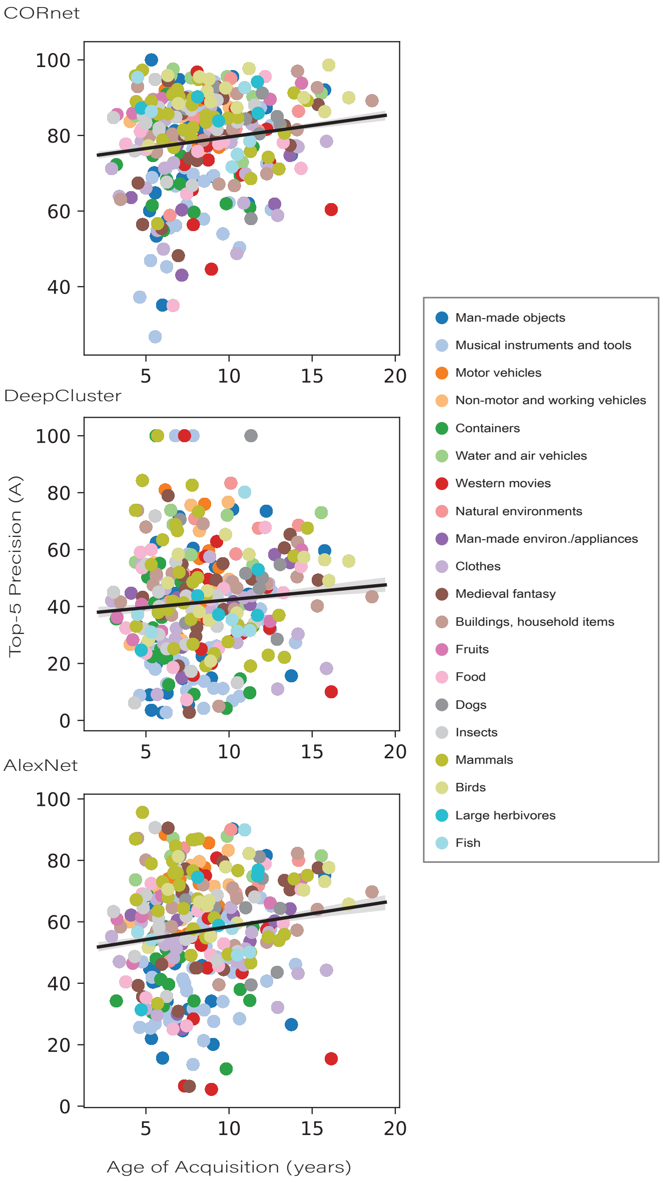

Figure 5: Relationship between infant age of acquisition and asymptote of machine performance for each visual class in the most informative layer of each network. The colours show different semantic groupings.

4 Limitations

The AoA measure used was obtained from self-report in adults. This has been validated in many ways in psycholinguistics, but it is possible that for our goal, it would be preferable to use parent report of AoA (e.g., `wordbank.standford.edu`), measures of category discrimination, or neuroimaging.

The ImageNet classes are esoteric and unecological. More human-like learning will probably require more human-like (or baby-like) training sets.

5 Conclusions

DNNs have the potential to provide a quantitative model of infant learning. In all models, later layers learned more slowly than earlier layers. However, we found qualitatively different trajectories of learning for supervised and unsupervised training. Specifically, unsupervised learning led to a more bottom-up progression in learning, so that earlier in learning, lower layers contained more explicit information about object class than higher layers, but this partially reversed through training. We did not find a correspondence between the order in which infants and machines learn visual classes, but there are a number of ways in which the approach may be developed.

Acknowledgments

Funded by the European Research Council Advanced Grant FOUNDCOG 787981

REFERENCES

E Bergelson and D Swingley. At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1113380109. URL `http://dx.doi.org/10.1073/pnas.1113380109`.

Mika Braginsky, Daniel Yurovsky, Virginia A Marchman, and Michael C Frank. Developmental changes in the relationship between grammar and the lexicon. In *CogSci*, pp. 256–261, 2015.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. July 2018. URL `http://arxiv.org/abs/1807.05520`.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, June 2016. ISSN 2045-2322. doi: 10.1038/srep27755. URL `https://www.nature.com/articles/srep27755`.

Ben Deen, Hilary Richardson, Daniel D Dilks, Atsushi Takahashi, Boris Keil, Lawrence L Wald, Nancy Kanwisher, and Rebecca Saxe. Organization of high-level visual cortex in human infants. *Nature communications*, 8:13995, 2017.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598, 1998.

Robert M French, Denis Mareschal, Martial Mermillod, and Paul C Quinn. The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: simulations and data. *J. Exp. Psychol. Gen.*, 133(3):382–397, September 2004. ISSN 0096-3445. doi: 10.1037/0096-3445.133.3.382. URL `http://dx.doi.org/10.1037/0096-3445.133.3.382`.

Jesse Gomez, Michael A Barnett, Vaidehi Natu, Aviv Mezer, Nicola Palomero-Gallagher, Kevin S Weiner, Katrin Amunts, Karl Zilles, and Kalanit Grill-Spector. Microstructural proliferation in human cortex is coupled with the development of face processing. *Science*, 355(6320):68–71, 2017. ISSN 0036-8075. doi: 10.1126/science.aag0311. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5373008`.

Umut Guclu and Marcel A J van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. doi: 10.6080/K0QN64NG. URL `http://dx.doi.org/10.6080/K0QN64NG`.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Mark H. Johnson, Suzanne Dziurawiec, Hadyn Ellis, and John Morton. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1):1–19, August 1991. ISSN 0010-0277. doi: 10.1016/0010-0277(91)90045-6. URL `http://www.sciencedirect.com/science/article/pii/0010027791900456`.

Kamila Maria Jozwik, Nikolaus Kriegeskorte, Radoslaw Martin Cichy, and Marieke Mur. Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations. In *2018 Conference on Cognitive Computational Neuroscience*, Philadelphia, Pennsylvania, USA, 2018. Cognitive Computational Neuroscience. doi: 10.32470/CCN.2018.1232-0. URL `https://ccneuro.org/2018/Papers/ViewPapers.asp?PaperNum=1232`.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

S M Khaligh-Razavi and N Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 2014. ISSN 1553-734X. URL `http://journals.plos.org/ploscompbiol/article/figures?id=10.1371/journal.pcbi.1003915`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, 2018. doi: 10.1101/408385. URL `https://www.biorxiv.org/content/early/2018/09/04/408385`.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990, 2012.

Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

P C Quinn, P D Eimas, and S L Rosenkrantz. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):463–475, 1993. ISSN 0301-0066. doi: 10.1068/p220463. URL `http://dx.doi.org/10.1068/p220463`.

Jack P Shonkoff and Deborah A Phillips. *From neurons to neighborhoods: The science of early childhood development.* ERIC, 2000.

Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization. *Scientific Reports*, 8(1):1–17, February 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-22160-9. URL `https://www.nature.com/articles/s41598-018-22160-9`.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016. ISSN 1097-6256. doi: 10.1038/nn.4244. URL `http://dx.doi.org/10.1038/nn.4244`.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.