What do CNNs gain by imitating the visual development of primate infants?

Shantanu Jaiswal
jaiswals_shantanu@ihpc.a-star.edu.sg
Dongkyu Choi
choi_dongkyu@ihpc.a-star.edu.sg
Basura Fernando
fernando_basura@ihpc.a-star.edu.sg

Social and Cognitive Computing
Institute of High Performance Computing
Agency for Science and Technology Research (A*STAR), Singapore.

Abstract

Deep convolutional neural networks have emerged as strong candidates for a model of human vision, often outperforming competing models on both computer vision benchmarks and computational neuroscience benchmarks of neural response correspondence. The design of these models has undergone several refinements in recent years drawing on both statistical and cognitive insights and, in the process, shown increasing correspondence to primate visual processing representations. However, their training methodology still remains in contrast to the process of primate visual development, and we believe that it can benefit from being more aligned with this natural process. Primate visual development is characterized by low visual acuity and colour sensitivity as well as high plasticity and neuronal growth in the first year of infancy, prior to the development of specific visual-cognitive functions such as visual object recognition. In this work, we investigate the synergy between the gradual variation in the distribution of visual input and the concurrent growth of a statistical model of vision on the task of large-scale object classification, and discuss how it may yield better approaches to training deep convolutional neural networks. The experiments we performed across multiple object classification benchmarks indicate that a growing statistical model trained with a gradually varying visual input distribution converges to a better generalization at a faster rate than traditional, more static training setups.

1 Introduction

The first models of convolutional neural networks [12] drew inspiration from the work of Hubel and Wiesel [15], introducing a computational architecture for visual processing based on the properties of the simple and complex cells found in the cat's visual cortex. Through successive refinements in the model architecture [13, 14, 21], appropriate application of training methods [19, 23], and collection of large-scale datasets [7], researchers have developed models that not only show human-level performance in object classification, but also serve as tools to study and predict behavioural and neural responses in computational neuroscience [26]. While recent work has sought to refine the model in terms of its architecture to address specific limitations such as few shot learning [11] or compactness for closer correspondence to the human ventral pathway [20], the current training methodology of these

models can be further improved. The conventional training approach for convolutional neural networks (CNNs) assumes a fixed statistical network in terms of capacity and architecture receiving a mostly constant visual input distribution throughout training[1]. Networks for specific tasks such as object classification are trained to minimize a classification loss through backpropagation until convergence. While such a training setup has found great success in computer vision, it often requires a large number of training iterations for convergence, and may benefit from being more aligned to the process of visual development in primate infants.

The study of infant visual development has its roots in the work of Fantz [10], who published key findings on a systematic approach to identify and measure the preferential attention to visual stimuli in human infants. This challenged the common assumption at that time of infants being congenitally blind, and piqued interest in the study of primate visual development. While a number of remarkable findings characterizing visual development have been published since then, in this work, we specifically study two relevant aspects when training statistical models of vision. First, it is established that human infants demonstrate low color sensitivity and poor spatial resolution in processing visual input during the first year of infancy [17]. Thus, the input distribution upon which a developing infant acquires visual abilities is much coarser compared to the distribution processed by adults, and it constantly gets refined until about the first year of infancy. Hence, the development of higher-order visual functions such as object recognition, stereopsis, and figure-ground segregation is postulated to emerge at older ages, once basic visual input processing abilities are sufficiently mature [4, 17].

While this might be a result of the physiological maturation of the infant's retina and photoreceptor development, it is of our interest to identify the possible role that a gradually refining visual input distribution may have on the concurrently developing visual cortex, bringing us to our second relevant aspect of visual development – network growth. Developmental researchers recognize the high rate of synaptic growth in the form of connections or myelination in the first year of growth and the ability of the cortex to wire itself during early development in primate infants [1, 5]. Specifically for the visual cortex, the process of wiring is known to be plastic and dependent on the nature of visual experience, although the interaction between cortical wiring and visual experience remains not well understood [22]. We thus recognize the growth of the visual cortex in early infancy as the second salient aspect of infant visual development relevant to the training of a statistical model for vision. In this work, we do not delve into the physiological aspects of visual cortex development, and regard growth as only being the addition of parameters to a statistical model of post-retinal visual processing.

Based on these two aspects of primate infant visual development discussed above, we aim to investigate a training setup wherein a growing statistical model of vision receives a gradually refining visual input distribution, and compare its performance to other setups wherein either the visual input or the statistical model is fully formed. Through appropriate experiments, we also aim to determine the role that each of these aspects may play in aiding visual learning. Previous research investigating the role of growth in learning includes Elman's work [9] in the domain of language acquisition. He showed how gradually increasing the window size of a simple recurrent network [8] during training, analogous to the increase in working memory and attention span in child development, allows the network to learn the task of processing complex sentences with correct verb agreement prediction significantly better than a static, fully formed network. Elman hypothesized that the relatively slow

[1] Here, "mostly" indicates the possible usage of data augmentation techniques.

process of development in humans, which is often seen in a negative light, may in fact be 'the enabling conditions which allow learning to be most effective' [9]. Our work hopes to uncover a similar finding in the context of visual learning – whether having an originally coarse visual representation that refines over time may, in fact, aid learning when taken in consideration with a concurrently growing statistical model of post-retinal visual processing. In the next sections, we describe the methodology of our approach that includes the computational methods for implementing model growth and refinement in the visual input distribution, followed by experiment details and a discussion of corresponding results.

2 Methodology

We aim to implement a statistical model of vision that grows in parameters during training with the concurrent refinement of the visual input distribution in terms of spatial resolution, saturation and contrast. While primate infant visual development is most likely to be a continuous process depending on physiological maturation and nature of visual experience, we take inspiration from Elman [9] and propose to train the network in a discrete number of stages. In each of these stages, we grow the statistical network and vary the visual input distribution in accordance to stage specific settings (see Figure 1). It is to note that while our CNN training strategy draws inspiration from cortical development in primate infants, it is not intended to be physiologically equivalent at the neural level.

2.1 Network growth resembling cortical growth

The notion of growth in CNN models of vision has been captured in previous approaches such as greedy layer-wise training [3] where new layers are added to an existing network in an unsupervised or a supervised setup. Recent work has also investigated how growing model capacity in terms of network depth or width during training or fine-tuning can lead to better classification and transfer performance [2, 25].

For our work, we adopt a similar approach to greedy layer-wise addition in a supervised setup, where we increase the depth of the network by adding a new block of convolutional layers at each stage. Formally, the architecture $\mathcal{A}_s$ at stage $s \in \{1..,\mathcal{S}\}$ appends a new feature extraction module $\mathcal{F}_s$ to the prior architecture at stage $s-1$. The feature extraction module $\mathcal{F}_s$ consists of a sequence of convolutional blocks, each with an input down-sampling operator P_s and a sequence of convolutional operators $W^s_\Theta = \{W_{\theta_k}\}$ (k denoting the convolutional layer depth in a block comprising j convolutional operations followed by non-linearity). The classifier module $\mathcal{C}$ comprises of a single linear layer W_c followed by softmax σ to output the predicted class distribution z over n classes. Hence, the architecture at stage s denoted by $\mathcal{A}_s$ is given by:

$$\mathcal{A}_s = F_s(\cdot, W^s_\Theta) \circ P_s \circ \mathcal{A}_{s-1} \tag{1}$$

$$z = \sigma \circ W_c \circ P_c \circ \mathcal{A}_s \in \mathbb{R}^n \tag{2}$$

As shown in Equation 1, the system appends the previous stage's architecture $\mathcal{A}_{s-1}$ with a new convolutional block $\mathcal{F}_s$ at each stage. The output of $\mathcal{A}_s$ is downsampled through an adaptive pooling operator P_c and passed to $\mathcal{C}$ (i.e., W_c). The parameters at each stage are tuned through backpropagation for given iterations. $\mathcal{A}_0$ is the initial convolutional block.

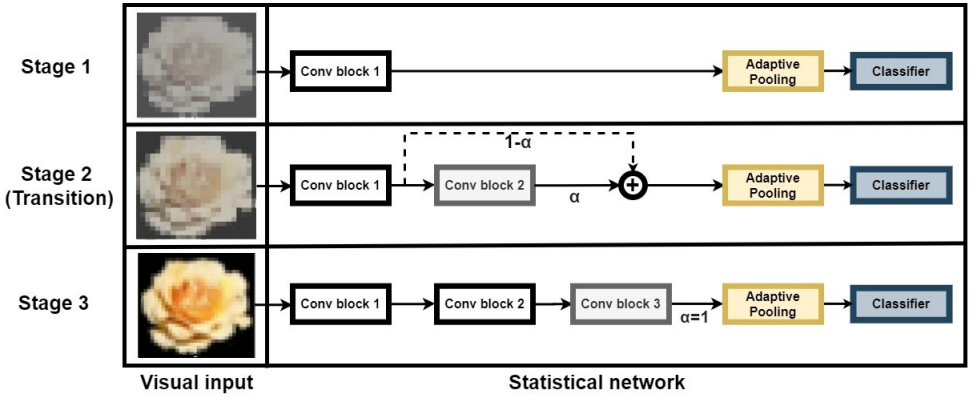

Figure 1: Illustration of our methodology for model growth and visual input refinement intended to simulate infant visual development. Across stages, we progressively refine the visual input in terms of its contrast, saturation, and resolution, while we grow the model in depth by adding new convolutional blocks. To enable continuity in learning when transitioning across stages, we use a decaying weighted sum of the old and new convolutional block (as shown for Stage 2) as input to the classification layer that is shared across stages.

While the above setup captures the conception of statistical network growth across stages, it suffers from discontinuities in learning when transitioning across stages that could effectively increase the number of training iterations to reach convergence besides possibly inhibiting generalization performance.

First, when transitioning across stages, if the number of output channels of the previous convolutional block is different from the new convolutional block, the previous stage classifier layer is effectively discarded, thereby requiring additional training iterations to reinitialize besides losing information for the classifier from the previous stage. Hence, in our work, we adopt the same number of output channels for each convolutional block, and retain the classifier across stages.

Second, when adding a new convolutional block, merely adding a new convolutional block in between the old convolutional block and classifier can cause a sudden shift in the distribution of value of both the input features to the classifier and the backpropagated gradient to the previous stage block. This can lead to a rapid and possibly unstable change in the previously tuned parameters for both the classifier and the previous stage blocks. To avoid this, we gradually shift the gradient flow by utilizing a decaying weighted sum of the activations of the old and new convolutional blocks as the feature extractor output $\mathcal{F}_s^{'}$ fed to the classifier, as shown in:

$$\mathcal{F}_s^{'} = (1-\alpha) * (P_s \mathcal{F}_{s-1}) + \alpha * \mathcal{F}_s \tag{3}$$

Although a similar approach was used in the progressive training of GANs [16], in our implementation α is uniformly scaled from 0 to 1 for only an initial preset number of training iterations for every new stage. Future work could also investigate a greedy strategy in the variation of α wherein it is not uniformly scaled per training iteration.

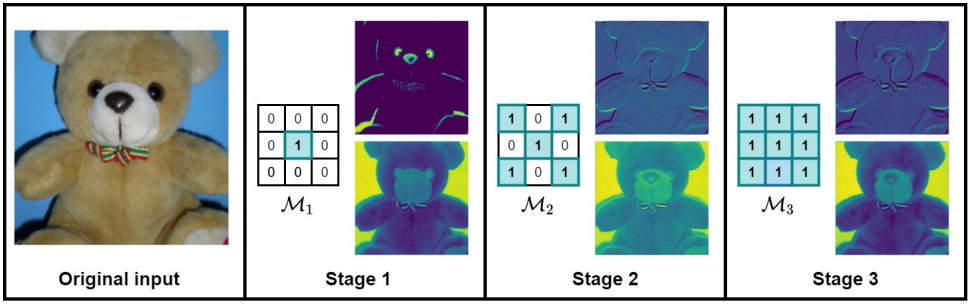

Figure 2: Stage-wise development of the receptive field by applying masks on 3×3 convolution filters across three stages of growth. The effect of masking is shown for selected input convolution filters (12 and 55) of a pretrained Resnet50. At the first stage, edges are more coarse- for example, the overall body sketch is less well formed for filter 12 and the nose is not discriminated for filter 55 in comparison to later stages.

2.2 Refining visual input distribution resembling retinal growth

In relation to visual input processing and development in infant primates, we study two factors – spatial resolution and color sensitivity in terms of saturation and contrast. While refinement of the visual input in primates may be a result of gradual physiological maturation of the retina [17], we do not implement this in our statistical model as part of the current work, and instead represent changes directly in the visual input. As shown in Figure 1, the input images are gradually transformed across stages in terms of contrast, saturation and resolution. While one may draw a parallel to this process of input refinement as being a data augmentation technique, the key difference here is that the transformations are applied sequentially in a developmentally plausible timeline, providing a course visual representation at the onset of training and gradually becoming more refined over later stages. For evaluation, we performed additional control experiments wherein the transformations are applied randomly during training at a similar probability, and did not find the performance to be equivalent. A more detailed illustration of the visual input refinement process is provided in the supplemental materials.

Although the above approach to input refinement mimics the gradual growth of visual input, it requires setting stage-specific values of the input refinement factors (contrast, saturation and resolution) prior to training and these values may vary across datasets. Hence, an alternative approach that we investigate to realize visual input refinement is to grow the receptive field of convolutional filters incrementally across stages. Specifically, given a NxN convolutional kernel, we apply a mask $\mathcal{M}$ of same size at each stage that selectively activates filter values as shown in Figure 2 for a 3x3 filter. We use a simple formulation that allows us to grow the receptive field in 3 stages. At the first stage, $\mathcal{M}$ only activates the centre element (when N is odd) or the middle two principle diagonal elements (when N is even) of the filter. At the second stage, $\mathcal{M}$ activates alternating element indices, while the third stage corresponds to the fully grown receptive field wherein $\mathcal{M}$ is an all-ones matrix. This strategy is inspired by the gradual development of rods and cones and retinal ganglion cells in the retina and resembles the incremental development of high resolution visual processing.

3 Experiments

In this work, we perform two primary experiments. First, we study the performance of a block-style CNN architecture as introduced in Section 2.1 in four setups each varying in refinement of the visual input distribution or growth of the statistical network across stages. We then perform relevant ablation experiments to determine the role of each input refinement factor (saturation, contrast and resolution) and the usage of the decaying weighted sum to gradually transition across stages. Second, we investigate the impact of training conventional variants of CNN models of vision in stages wherein only the visual input distribution is gradually refined across stages, and compare performance to the conventional model training approach. We then analyze the performance when the receptive field is grown over stages, and as before, perform ablation analysis to determine the role of input refinement factors.

3.1 Datasets and training setup

We perform experiments on the CIFAR10, CIFAR100 [13] and a subset of the ImageNet dataset [7] comprising of 200 classes (hereafter referred to as ImageNetH200) selected by traversing the ImageNet (originally WordNet) synset hierarchy. A reduced subset for ImageNet is used due to computational limitations, and we expect our results to scale on the full 1000-class dataset. The CIFAR10 and CIFAR100 datasets comprise of 32x32 RGB images with 50k training and 10k test samples. CIFAR10 comprises of 10 labels whereas CIFAR100 comprises of 100 classes. ImageNetH200 comprises of 256x256 RGB images with approximately 1300 training samples per class. Standard data augmentation techniques of random cropping and random horizontal flips with probability of 0.5 were used for training. Specifically for ImageNetH200, training is done on 224x224 random crop while evaluation is done on the centre crop. Experiment details and hyperparameter settings are provided in the supplemental materials. Code to recreate experiments is available at the following link.

3.2 Evaluating growing networks with gradually refining inputs

In this experiment, we evaluate the training methodology for block style CNN models introduced in Section 2.1, wherein across stages, the visual input distribution is progressively refined while the model is grown in parameters. We compare the learning performance against 3 alternate training setups differing in either the visual input distribution being fully formed or the model being static throughout training. Our hypothesis is that having an initially small network and a coarse visual input that both grow during training may enable stronger hierarchical learning allowing the model to first discriminate global visual patterns before proceeding to finer patterns, and thereby aid learning performance. The model is based on the VGG-13 architecture [24] with batch normalization comprising of 5 convolutional blocks but with all output channels set to 256, and a single fully-connected classification layer. For stage-wise training, we train the model over 5 stages and choose stage-specific parameters to be reflective of the documented infant visual development trajectory [17] as shown in Table 1. We report results on CIFAR10 and CIFAR100 for this experiment.

Do growing models generalize better and faster?
As shown in Table 2, we find that the developmentally inspired setup wherein the model is gradually grown along with a refining input distribution across stages outperforms the baseline approach wherein the model and input are both static throughout training. Specifically,

	CIFAR10 and CIFAR100				
Stage number	1	2	3	4	5
Total epochs	5	5	10	15	315
Stage transition epochs	-	2	4	6	20
Saturation ratio	0.00	0.33	0.67	0.90	1.00
Contrast ratio	0.50	0.60	0.80	0.90	1.00
Resolution	24	26	28	32	32

Table 1: Stage-specific parameters for stage-wise training on CIFAR datasets.

Model	CIFAR10	CIFAR100
Static model, Static input (baseline)	93.76± 0.09	72.53± 0.10
Static model, Refining input	93.55± 0.10	71.91± 0.13
Growing model, Static input	94.10± 0.11	73.92± 0.22
Growing model, Refining input	**94.16± 0.13**	**74.01± 0.24**
Growing model (non-gradual), Static input	93.90±0.11	73.42± 0.11
Growing model (non-gradual), Refining input	93.64±0.15	73.28± 0.13

Table 2: Top-1 test accuracies (%) of a block style CNN architecture in 4 different training setups for CIFAR10 and CIFAR100 reported over 5 trials. Setups with a growing model have a better convergence performance across both datasets, even though all setups have the same model size at the final stage. Additionally, gradually growing the models with a decaying weighted sum performs better than directly adding a new block (non-gradual).

we find an improvement of 0.40% and 1.48% in test accuracy on CIFAR10 and CIFAR100 respectively. Additionally, the developmentally inspired setup has a faster rate of learning, reaching the best accuracy of the baseline approach in 156 epochs in comparison to the baseline's 241 for CIFAR10 (a relative speedup of 35.27%) and 151 epochs in comparison to the baseline's 237 for CIFAR100 (a relative speedup of 36.28%). These results are significant considering both setups had the same model architecture and size at the last stage of training.

It is interesting to analyze the difference in impact of applying a refining input distribution on a static fully formed model and a growing model. When the model is static, training with a refining input distribution degrades performance in comparison to a static distribution with a drop in test accuracy of 0.21% and 0.62% for CIFAR10 and CIFAR100 respectively. However, the same is not true when the model is gradually grown, wherein applying a refining input results in an improvement of 0.06% (94.10% vs 94.16%) and 0.09% (73.92% vs 74.01%) for respective datasets. This is interesting from a statistical learning perspective. Given a static fully formed model at the onset of training, providing an initially coarse input distribution might induce spurious generalizations due to model over-parameterization at early stages of training. This may provide an incorrect initialization for later stages of training wherein the input is more refined, and thereby lead to an overall worse optimum. In contrast, given an initially small model (in parameters) and a coarse input distribution, the model may require lesser iterations for parameterization at earlier stages, and with concurrent growth of both the model and the input may converge to a better generalization.

Importance of gradual growth vs non-gradual growth of networks

The importance of selecting an appropriate model growth mechanism is reflected in the performance difference between non-gradual growth (wherein a new convolutional block is di-

Model	CIFAR10	CIFAR100
Static model, Static input (baseline)	93.76± 0.09	72.53± 0.10
Growing model, Refining input	**94.16± 0.13**	**74.01± 0.24**
Growing model, Refining Input = [Sat,Con]	93.85± 0.11	73.55±0.18
Growing model, Refining Input = [Sat,Res]	94.10± 0.09	73.45±0.16
Growing model, Refining Input = [Con,Res]	94.07± 0.07	73.52±0.11

Table 3: Comparison of "Growing model and Refining input" setup when one of the three input refinement factors- saturation (Sat), contrast (Con) and resolution (Res)- is not altered during training. Removing any one of the factors leads to decline in performance.

rectly added to the existing architecture) and gradual growth (wherein a decaying weighted sum is utilized). The difference is especially pronounced in the refining input setup, wherein non-gradual addition drops accuracy from 94.16% to 93.64% (below baseline performance) in the case of CIFAR10, and 74.01% to 73.28% for CIFAR100.

Ablation analysis of visual input refinement factors

To quantify the role of each visual input refinement factor, we perform ablation experiments in the "Growing model and Refining input" setup. Specifically, we "deactivate" (keep fully formed) one of saturation, contrast or resolution, and compare the performance to the original setup. As shown in Table 3, we find that removing even one of the input refinement factors leads to a drop in generalization performance, indicating that all three factors may contribute to better convergence. For CIFAR100, the performance drop is similar for all three factors, with removal of contrast the most pronounced (0.56%), whereas for CIFAR10, the performance drop is less significant, with removal of resolution most pronounced (0.31%).

3.3 Impact of training popular CNN models with a refining input distribution applied in stages

In this experiment, we evaluate a training methodology for popular CNN models wherein the visual input distribution is progressively refined across stages from an initially coarse representation to a fully formed representation in terms of saturation, contrast and resolution. Here we do not consider statistical network growth across stages and study 3 CNN architectures – ResNet, DenseNet and VGG. As before, we trained the model over 5 stages and chose similar stage parameters as shown in Table 4. The results are reported in Table 5.

	CIFAR10 and CIFAR100					ImageNetH200				
Stage number	1	2	3	4	5	1	2	3	4	5
Total epochs	2	3	5	10	330	2	2	2	4	90
Saturation ratio	0.00	0.25	0.5	0.75	1.00	0.00	0.25	0.5	0.75	1.00
Contrast ratio	0.50	0.60	0.80	0.90	1.00	0.50	0.60	0.80	0.90	1.00
Resolution	24	26	28	30	32	184	194	204	214	224

Table 4: Stage parameters for training on CIFAR10/100 and ImageNetH200 datasets.

We find that training ResNet models (ResNet18 and ResNet50) with a gradually refining input distribution leads to higher validation accuracy across all three datasets. Specifically for ResNet50, we find an increment of 1.7% on CIFAR100 and about 1.0% for CIFAR10

Model	CIFAR10	CIFAR100	ImageNetH200
ResNet18 (Static input)	93.50 ± 0.10	72.68 ± 0.22	73.37
ResNet18 (Refining input)	**94.20 ± 0.25**	**73.93 ± 0.18**	**74.16**
ResNet50 (Static input)	93.62 ± 0.23	73.05 ± 1.01	76.80
ResNet50 (Refining input)	**94.64 ± 0.21**	**74.74 ± 0.95**	**77.76**
VGG13bn (Static input)	**92.84 ± 0.10**	**70.98 ± 0.31**	**74.46**
VGG13bn (Refining input)	92.55 ± 0.12	70.26 ± 0.28	74.02
DenseNet121 (Static input)	94.93 ± 0.14	76.33± 0.27	77.60
DenseNet121 (Refining input)	**95.11 ± 0.18**	**76.48± 0.39**	**77.83**

Table 5: Top-1 test (CIFAR) and validation (ImageNetH200) accuracies (%) for models based on different training strategies. ResNet models respond most favorably to a refining input distribution across all 3 datasets.

Model	CIFAR10	CIFAR100	ImageNetH200
ResNet50 (Static input)	93.62 ± 0.23	73.05 ± 1.01	76.80
ResNet50 (Refining input=[Res,Sat,Con])	**94.64 ± 0.21**	**74.74 ± 0.95**	**77.76**
ResNet50 (Refining input=[Sat,Con])	94.08 ± 0.10	73.62 ± 1.61	76.83
ResNet50 (Refining input=[Res,Con])	94.60 ± 0.12	74.07 ± 1.29	75.74
ResNet50 (Refining input=[Res,Sat])	94.55 ± 0.08	73.93 ± 1.01	77.17
ResNet50 (Receptive field growth)	94.50 ± 0.18	73.89 ± 0.64	75.40

Table 6: ResNet50 input refinement ablation study and receptive field growth results. All three visual input refinement factors contribute to learning. Growing the receptive field in stages performs better than static input setups for CIFAR10/100 without requiring stage parameter settings that can vary across datasets.

and ImageNetH200 respectively. However, we do not see similar improvement for VGG13. Based on the results, we conjecture that the usage of skip connections may enable ResNet models to implicitly capture statistical growth across stages, and thereby respond better to a gradual refinement in the input distribution. To better understand how learning differs when trained with a gradually refining input distribution, we visualize filters induced at salient training points, and provide comparative analysis in supplemental materials.

Analysis of refining input distribution and receptive field growth on ResNet50
As done before, we first perform an ablation analysis of the input refinement factors by deactivating one of the input refinement factors for ResNet50 stage-wise training. Second, we evaluate whether growing the receptive field across three stages can serve as an alternative strategy for capturing gradual visual input refinement.

As shown in Table 6, we find that across all three datasets, removing even one of the input refinement factors leads to a drop in generalization performance. For ImageNetH200, keeping saturation unaltered leads to the highest decline from 77.76% to 75.74% (performing below static input setup). Gradually growing the receptive field performs better than static input setups with an improvement of about 0.9% for CIFAR 10 and CIFAR100 and does not require setting dataset specific stage parameters. However, for ImageNetH200, wherein the input convolutional layer comprises of 7x7 filters (which is 3x3 for CIFAR10/100), applying masking at earlier stages may result in extreme information loss which could inhibit learning at future stages and thereby result in an overall lower performance.

4 Conclusion

In this work, we identified two aspects of visual development in primate infants relevant to the training of CNNs – a gradually refining input visual distribution and a concurrently developing statistical network. Our first experiment analyzed the performance of a block-style CNN architecture when it is gradually grown in stages during training along with the progressive refinement of the visual input distribution. We found that such a training setup attains a better generalization at a faster rate than alternate setups, pointing to a possible synergy between the process of input refinement and statistical network growth. In our second experiment, we evaluated the impact of training conventional CNN variants with a refining visual input distribution applied in stages, and found that it significantly benefits learning in ResNet models, besides some improvements in DenseNet models.

For future work, we wish to draw on physiological insights in visual development such as the formation of top-down feedback in the visual cortex [6] to inform computational mechanisms of network growth, and further investigate other approaches to model retinal and photoreceptor growth. We also hope to study the impact of the stage-wise training approach in the context of few-shot and contrastive learning techniques and extend our work to finer visual tasks such as object recognition and semantic segmentation.

Acknowledgments This research is supported by A*STAR under its *Human-Robot Collaborative AI for Advanced Manufacturing and Engineering* (Award A18A2b0046) and the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2019-010). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of A*STAR.

References

[1] Sandra Ackerman et al. *Discovering the brain*. National Academies Press, 1992.

[2] Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Greedy layerwise learning can scale to imagenet. *arXiv preprint arXiv:1812.11446*, 2018.

[3] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[4] Angela M Brown, Delwin T Lindsey, PremNandhini Satgunam, and Jaime A Miracle. Critical immaturities limiting infant binocular stereopsis. *Investigative ophthalmology & visual science*, 48(3):1424–1434, 2007.

[5] J Le R Conel. *The postnatal development of the human cerebral cortex. Vol. 1. The cortex of the newborn.* Harvard Univ. Press, 1939.

[6] 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(1):1–10, 2017.

[7] 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.

[8] Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225, 1991.

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

[10] Robert L Fantz. Pattern vision in young infants. *The psychological record*, 1958.

[11] Reuben Feinman and Brenden M Lake. Learning a smooth kernel regularizer for convolutional neural networks. *arXiv preprint arXiv:1903.01882*, 2019.

[12] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[13] 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.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[15] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] Lynne Kiorpes. The puzzle of visual development: behavior and neural limits. *Journal of Neuroscience*, 36(45):11384–11393, 2016.

[18] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6, 2009.

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

[20] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

[21] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[22] Tomaso Poggio and Thomas Serre. Models of visual cortex. *Scholarpedia*, 8(4):3516, 2013.

[23] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480, 2017.

[26] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624, 2014.