Modeling Infant Visual Development Yields More Robust Computer Vision

Rohan Agarwal

Abstract

Convolutional neural networks (CNNs) have become ubiquitous for image recognition, with applications including tools for the visually impaired, augmented reality, autonomous vehicles, digital assistants, search engines, infrastructure maintenance, and military target recognition. However, in real-world situations, CNNs frequently encounter images degraded by distance, motion, or camera quality, in contrast to the properly photographed datasets on which they were trained. Though CNNs are designed to process data similarly to the visual cortex, there is little research training CNNs in the way the eye learns to process images. Hypothesizing that it is biologically advantageous for infants to learn to recognize objects through blurred vision, this study explored training CNNs on datasets processed to mimic infant vision. Two datasets with different object similarity were blurred at multiple biologically-accurate levels. Three different-sized networks were trained on images of one level of blur and tested on images of all levels, addressing human biology and real-world data variance through this novel approach. The overall trend was that CNNs trained on blurred images were more robust and had superior overall performance than those trained on clear images, with the strongest results in the largest network tested on low-similarity objects. With this network, blurred training outperformed clear training in identifying blurred images by up to 45%, with losses of no more than 0.6% in identifying clear images. These findings indicate that in applications that value robustness, classifying blurred images, and the recognition of general objects, it is beneficial to adopt biologically-based blurred CNN training.

1. Introduction

Convolutional neural networks (CNNs) have become ubiquitous for image recognition, with applications that include tools for the visually impaired, augmented reality, autonomous vehicles, digital assistants, search engines, infrastructure maintenance, and military target recognition. Most image recognition applications with CNNs require the recognition of general objects or patterns. For example, a computer vision-based approach to autonomous driving would require a CNN to identify large, generic objects such as people, cars, buildings, and roads. (Huval, 2015). However, CNNs lack the robustness to perform well at a large range of distances and visibility conditions (Huval, 2015). Similarly, with aerial drones (used for military purposes, fire detection, infrastructure inspection, vegetation monitoring, glacial observations, etc.), bounding box suggestions are based on categorizing the objects in the image (Radovic, 2017). Furthermore, many tools to help the visually impaired rely on a computer to recognize obstacles before alerting the user; Poggi and Mattoccia used a CNN to recognize eight general classes of obstacles that a visually impaired person would encounter on the streets (Poggi, 2016). In a similar vein, image search algorithms also function by classifying generally similar images into one group (Li, 2015). Additionally, augmented reality requires real-time CNN recognition of general objects to identify surfaces in semantic segmentation (Paszke, 2016). Overall, some of the most important research areas in computer vision rely on or require more robust CNNs to classify general objects.

By training on datasets such as ImageNet (Deng, 2009), CNNs can become very successful at general object recognition (Krizhevsky, 2012). However, popular CNNs such as AlexNet (Krizhevsky, 2012) and VGG (Simonyan et al., 2015) have significantly worse performance when classifying degraded, blurry images (Karahan, 2016; Geirhos, 2018), which are frequently encountered due to the variables and limitations in real-world scenarios (Karahan, 2016; Dodge, 2016). Limitations that can degrade images in nearly any application of CNNs occur when capturing and/or processing images (Dodge, 2016), affected by factors such as distance, motion, camera quality, and environmental elements. One instance of this is that both drones and self-driving cars must handle low-resolution images due to the need to identify objects at far distances, capping their abilities (Radovic, 2017; Huval, 2015). Image processing can be affected further by the potentially limited computational resources required for the application. For example, augmented reality and mobility aids for the blind mostly rely on simpler (smaller-sized) CNNs due to mobile hardware limitations and the necessity of real-time processing (Paszke, 2016; Poggi, 2016). Overall, making CNNs more robust is important for a wide range of applications.

CNNs were designed to mimic the computations of the brain to solve problems in computer vision (Fukushima, 1980). The success of CNNs in computer vision has been mirrored by studies successfully using CNN activity to predict brain activity. CNNs can successfully model fMRI, EEG, and electrode recordings in human and monkey brains (Guclu, 2015; Horikawa, 2017; Yamins, 2016). However, CNNs are still only partial models of the brain, accounting for under 50% variance (Yamins, 2016), leaving opportunities to better capture the encoding of the brain. In addition, instead of only focusing on the brain and image processing, the same concept of modeling human biology could also be applied to eyesight and image input.

Recently, studies have experimented with using more human-like training methods to improve CNN performance. For example, the way toddlers handle unfamiliar objects creates much more image-level object variability, meaning that training neural networks on datasets with similar variability could possibly improve results (Slone, 2019). Results with a single network and dataset show improvements from 23% to 40% accuracy using this idea (Raz, 2019). Biologically, infants have much more limited vision than adults, with lower resolution and lower contrast sensitivity, so they learn to recognize objects through blurred vision for around a year (Banks, 1978; Banks and Crowell, 1993). For facial recognition, training on lightly blurred images increases classification accuracy on blurred images, but lowers accuracy on clear images (Vogelsang, 2018), rendering it impractical for most common applications. However, blurred training does decrease the overall range of performance (Vogelsang, 2018), which shows promise in getting more robust, practical results by modeling infant vision. Since previous results have been mostly impractical or preliminary, however, the current standard method of training CNNs involves large datasets of clear, focused images (Deng, 2009; Goyal, 2018).

The many different applications of CNNs correspond to different types of architectures as well. One of the most popular architectures used as a general standard for CNNs is AlexNet

(Krizhevsky, 2012; Alom, 2018). The mobile, real-time requirements of augmented reality and drones would benefit from a small, efficient network (Paszke, 2016; Radovic, 2017), which is what networks like SqueezeNet are meant to solve (Iandola, 2017). Other networks are popular for their complexity and high accuracy, such as VGG (Simonyan, 2015; Alom, 2018). All of these characteristics could be useful depending on the application, and are therefore worth studying.

This study models infant visual learning using CNNs and blurred images of objects to test the hypothesis that visual learning through blurred infant vision is biologically advantageous for robust vision later in life. Three neural network architectures of various sizes were analyzed using two datasets of varying object similarity. This approach addressed human biology, real-world limitations, and real-world applications through a widely applicable and novel training method. This also furthered the goal of improving CNNs' approximation of human biology and aided in the understanding of human visual learning and processing in CNN studies, revealing avenues for future research in robust CNN training.

2. Methods

2.1. Dataset Selection and Analysis

Two datasets, Imagenette-320 and Imagewoof-320 (https://github.com/fastai/imagenette), were used in this study. These are subsets of the larger ImageNet dataset (Deng, 2009), the former with ten general object classes and the latter with ten highly similar object classes--different dog breeds. Each dataset is split into a training set with 1300 images per class and a testing set with 50 images per class. Each image was resized so that the smallest dimension would be 320 pixels, while the aspect ratio of the image was preserved. The last three classes of both datasets were removed due to limited time and memory resources.

As Imagewoof contained only images of dogs while Imagenette contained pictures of diverse objects, there was increased visual similarity between objects in the Imagewoof dataset. The diversity for each image set was quantified by constructing a 7 class by 7 class distance matrix. The i^{th} row and j^{th} column of the matrix measured the average distance between pictures in class i and pictures in class j. To compute each distance, all images were resized to 75 by 75 pixels and converted to grayscale. Within one dataset, the first 10 images from class i were compared to the first 10 images of class j by the Euclidean distance between two images. The mean of the distances per pair of classes represented the similarity between the two classes.

2.2. Modeling Infant Visual Acuity for Image Processing

To create the blurred datasets, each image was resized to 224 x 224 pixels. The dataset was then copied five times, each with a different kernel size for Gaussian blur using OpenCV 2. Gaussian blur was selected due to it being used as a common representation of optical blur (Uchida, 2013). Based on studies such as Cichy 2016 and Kay 2008, a common range of the space an image takes in a human's field of view is four degrees to twenty degrees. Since the images used were 224 by 224 pixels, dividing 224 px by the number of degrees the image occupies yields the number of pixels per degree at that distance. These values, 11 px/deg and 56 px/deg, were then divided by the visual acuity values of infants at 1 month, 2 months, and 3 months. These values were 2.4 cyc/deg, 2.8 cyc/deg, and 4.0 cyc/deg, respectively (Banks, 1978). The resulting kernel sizes were 2.8 px, 4 px, 4.6 px, 14 px, 20 px, and 23 px. These kernel sizes were were approximated to the nearest odd integer to meet the required parameters of the Gaussian blur function in OpenCV. They were then adjusted to evenly sample blurring resolutions following a logarithmic scale. The final Gaussian blur kernel sizes used were 1 px (equivalent to no blur), 3 px, 5 px, 11 px, and 23 px.

2.3. CNN Selection and Training Parameters

Three networks were used: AlexNet, SqueezeNet 1.1, and VGG-16, all imported from the PyTorch torchvision models subpackage. These networks were selected for their popularity, range of complexity, and known range of accuracy (Krizhevsky, 2012; Iandola, 2017; Simonyan, 2015; Alom, 2018) in order to consider a wide range of use cases. The networks were imported as pretrained and the final classification layer was resized to fit the size of the two datasets. Additional training was conducted using Stochastic Gradient Descent (SGD) optimization, a cross entropy loss function, a learning rate of 0.001, and a momentum of 0.9, motivated by success with similar settings by (Vogelsang, 2018). The batch-size of the datasets was 128. Every ten mini-batches during training, the program would check if the loss value had decreased from the last check; if not, the training would terminate.

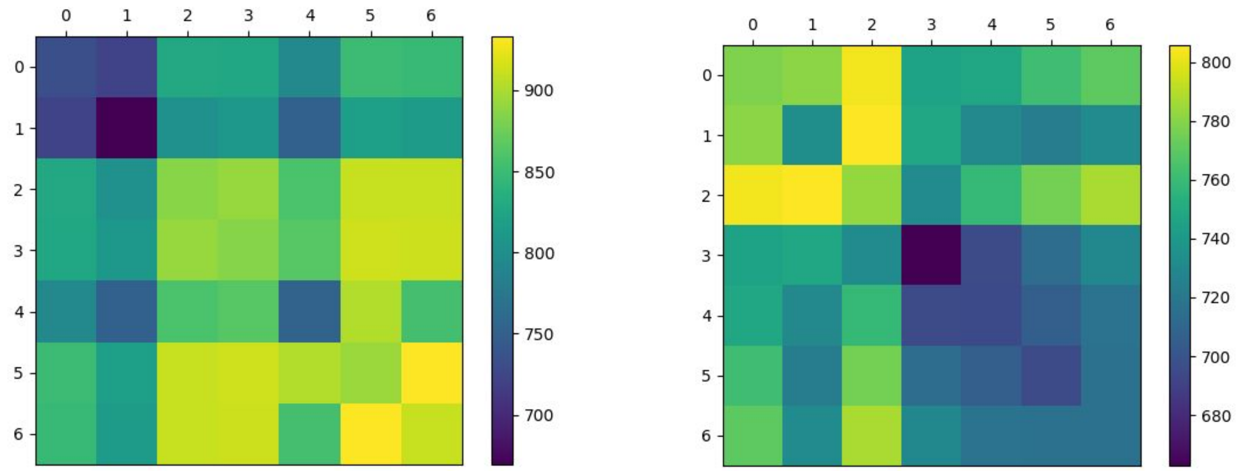
2.4. Testing and Analysis

Each dataset corresponding to one level of blur was used to train each model, then being tested on the datasets for all five levels, recording percent accuracy. The model being tested was reinitialized after the five tests to be retrained for the next five tests, repeating until all 25 combinations of training and testing levels had been completed. This procedure was then repeated for each of the remaining models and the second dataset. Summary statistics (minimum, maximum, mean, and standard deviation) were created for all six sets of values to help analyze the visual trends. Treating the curve where the model was trained on no blur as a control, the percent difference in accuracy to each of the other curves at the same testing level was also calculated (repeated for all six sets of values) in order to compare levels of blurred training to clear training in individual test cases. The correlations between the level of training blur and the mean and median of the percent improvement were calculated as well.

3. Results

3.1. Dataset Similarity Analysis

To verify that the datasets to be experimented on had different levels of similarity between their own object classes, the distance matrices in **Figure 1** were computed and visualized. Differing levels of object similarity and diversity were important in considering potential types of data in various applications. The higher proportion of darker colors in **Figure 1B** than **Figure 1A** shows that the Imagewoof dataset had classes with lower distances than the Imagenette dataset. The Imagenette dataset, since it had higher Euclidean distances between its object classes than the Imagewoof dataset, therefore had lower similarity than the Imagewoof dataset. Numerically, the distances between the Imagenette classes had a mean of 831.5 and a standard deviation of 61.1; for Imagewoof, the mean was 733.3 and the standard deviation was 33.2. These values also support the fact that the classes in the Imagewoof dataset were much more visually similar and difficult to differentiate.



A: Imagenette

B: Imagewoof

Figure 1: The matrices showing the Euclidean distances between the seven classes of Imagenette (A) and the matrix for the seven classes in Imagewoof (B) are presented here. Darker colors show lower distances and higher similarity.

3.2. Performance on Low-Similarity Images

The accuracies of the AlexNet, VGG16, and SqueezeNet networks on the low-similarity Imagenette data are shown in **Figure 2**. On the largest two networks, VGG and AlexNet, all training levels had near-identical performance when classifying unblurred images (**Figures 2A, 2C, 3A, 3C).** As the amount of blur increased in the training data, the performance on blurred images remained at high levels, showing more consistent, robust performance overall when trained on blurred images. The smallest network, SqueezeNet, which had 50x fewer parameters than AlexNet (Iandola, 2017), had a drop in performance on clear images (i.e. 1 px blur) when trained on highly-blurred images (**Figures 2B**, **3B**). The performance with mid-blur training however still followed the trends in the larger two networks, with near-identical performance to clear training at low blur and higher, more robust performance at high blur (**Figure 2B**). The robustness increased as the training blur increases, shown by the flattening curves.

The statistics shown in **Table 1** confirmed these observations in overall performance and robustness. Blurred training yielded accuracies with lower ranges and lower standard deviations, all showing superior robustness and consistency in all network sizes. **Figure 3** shows larger accuracy gains on blurred images than losses on clear images (and nearly no losses for larger networks). VGG, the largest network tested, saw gains of up to 45% and losses of no more than

1%. SqueezeNet, the smallest network tested, experienced losses of no more than 19% and gains of up to 30%. From **Figure 3**, it is clear that, especially for larger networks, the potential gains from blurred training outweigh the minimal losses. Overall performance was superior for larger networks on blurred training, seen by the higher means for these networks (**Table 1**). At the same time, this evidently did not hold true for SqueezeNet, which was superior with blurred training in specific cases, not overall. SqueezeNet's mean performance was highest at mid levels of blur.

Figure 2: Each network was trained on blurred Imagenette images, each of the five blur levels in separate instances. Each instance of training was tested on all five levels of blur. The percent accuracies are shown in these graphs. AlexNet results are shown in 2A, SqueezeNet in 2B, and VGG16 in 2C.

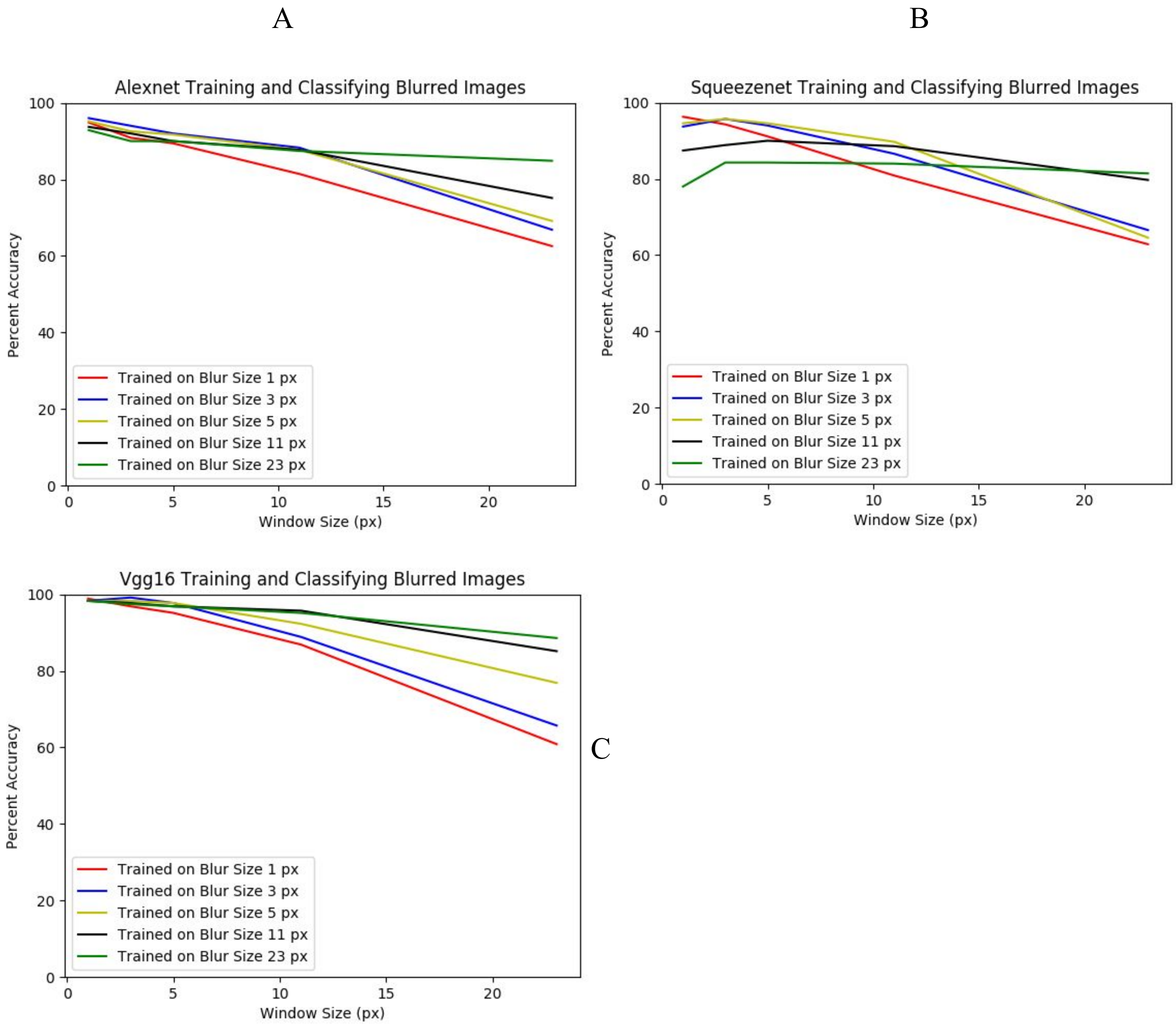


Table 1: Summary statistics for the performance of the three networks on Imagenette.

A: AlexNet

Training Blur Level	***1px***	***3px***	***5px***	***11px***	***23px***
Mean	83.82857	87.42857	87.25714	87.71429	89.02857
Standard Deviation	12.84539	11.84767	10.47173	7.376189	3.021015
Minimum	62.57143	66.85714	69.14286	75.14286	84.85714
Maximum	94.85714	96	95.14286	93.71429	92.85714

B: SqueezeNet

Training Blur Level	***1px***	***3px***	***5px***	***11px***	***23px***
Mean	85.08571	87.31429	87.82857	86.91429	82.4
Standard Deviation	13.77279	12.11442	13.20575	4.127558	2.737495
Minimum	62.85714	66.57143	64.57143	79.71429	78
Maximum	96.28571	95.71429	95.71429	90	84.28571

C. VGG16

Training Blur Level	***1px***	***3px***	***5px***	***11px***	***23px***
Mean	87.71429	89.94286	92.68571	94.74286	95.25714
Standard Deviation	15.69219	14.16722	9.202041	5.453327	3.910217
Minimum	60.85714	65.71429	76.85714	85.14286	88.57143
Maximum	98.85714	99.14286	98.28571	98.28571	98.28571

Table 2 shows that for all networks, the mean improvement is strongly correlated with increasing training blur, again showing that the improvements in blurred training greatly outweigh the losses when encountering a range of image quality. As network size decreases, the median improvement becomes negatively correlated with increasing training blur. Smaller networks tend to lose some accuracy with higher quality images when trained on highly blurred images. However, for larger networks, the correlation between this change and blurred training is weak, even being positively correlated for VGG16, showing their robustness.

Figure 3: Percent improvement in accuracy compared to clear training tested at the same blur level (Imagenette). The "X" represents the mean of the percent improvements.

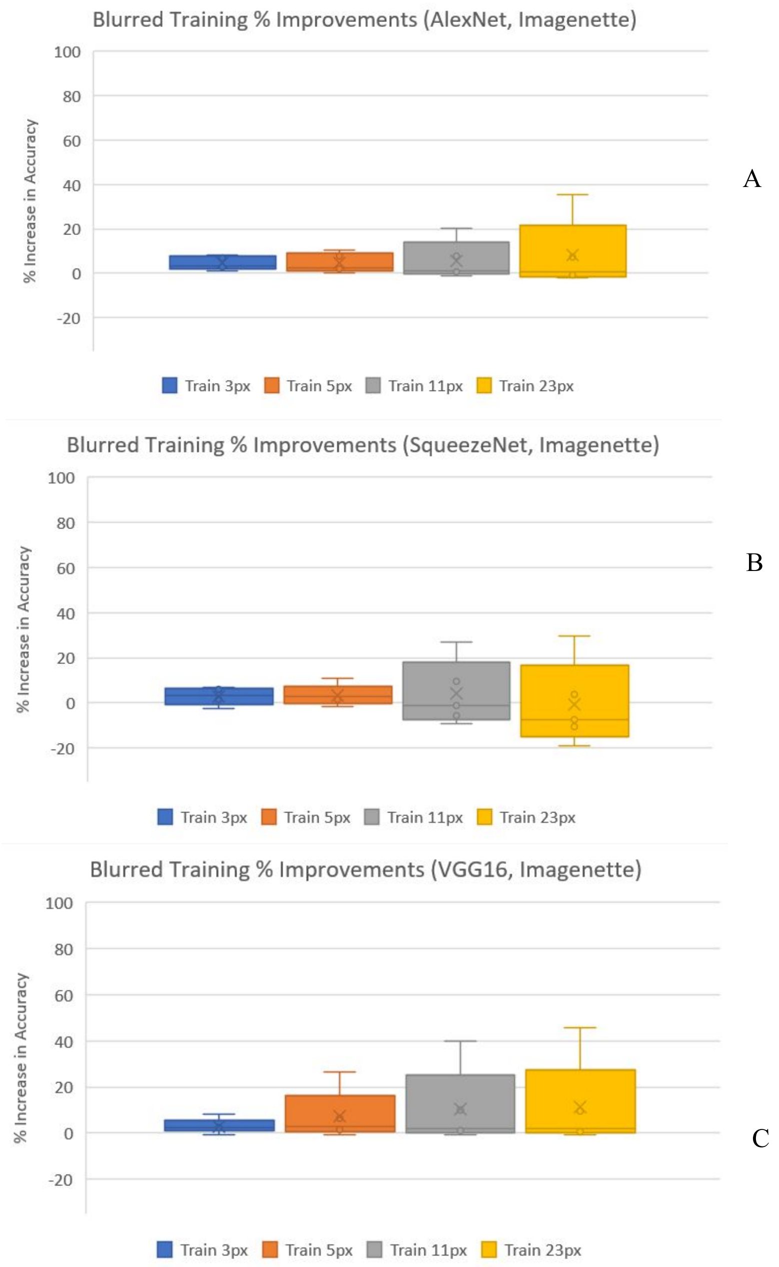


Table 2: Correlation between the level of training blur and the mean/median percent accuracy improvement for the Imagenette (low-similarity, high-diversity) dataset.

(Imagenette)	*VGG16*	*AlexNet*	*SqueezeNet*
Mean	0.84731	0.84543	0.80398
Median	0.19791	-0.32627	-0.9099

3.3. Performance on High-Similarity Images

With the high-similarity Imagewoof dataset and smaller network sizes, blurred training had substantially lower performance on clear images and substantially higher performance on blurred images when compared to clear training. The peaks in the curves became more apparent as well with SqueezeNet and AlexNet, suggesting that with smaller networks, the best performance on images blurred with a kernel of a certain size was achieved by training the model on images blurred with a kernel of the same (**Figure 4**). Similar to the results on Imagenette, higher levels of training blur still yielded more consistent performance on all networks; statistics showed accuracies with lower ranges and standard deviations (**Table 3**). However, the mean performance peaked at lower levels of training blur, moreso for the smaller networks, indicating overall performance fared better with less extreme blur training as network size and image diversity decreased. Percentage improvements in accuracies are represented in **Figure 5**, showing large accuracy increases in classifying blurred images and decreases in classifying clear images when using blurred training. The largest network, VGG, still had minimal losses of 2% and substantial gains up to 45%. SqueezeNet, however, had a large range from losses of 20% to gains of 99%. This, and the difference in trends between large networks and small networks on Imagenette (Section 3.2), suggested that the range between gains and losses became more extreme with smaller networks and harder to differentiate object classes. From **Figure 5**, it is again evident that for larger networks, the potential gains from blurred training outweigh any losses. The losses on clear images and gains on blurry images also become more extreme with smaller networks such as Squeezenet. Overall performance based on the means seemed to be best with moderate amounts of blur, which was consistent with the idea that the effects of blurred training remain effective, but less effective than on easily differentiable classes.

Table 4 shows that the mean improvement is correlated with increasing training blur, moreso with larger networks, again showing that the improvements in blurred training outweigh the losses. However, using the less diverse dataset, the degree to which the gains outweigh the losses decreases, especially with small networks. The negative correlation between the median improvement and increasing training blur makes it clear that blurred training does lower performance on sharper images, even if not by much, as seen by comparing **Figure 5** to **Table 4**.

Figure 4: Each network was trained on blurred Imagewoof images, each of the five blur levels in separate instances. Each instance of training was tested on all five levels of blur. The percent accuracies are shown in these graphs. AlexNet results are shown in 4A, SqueezeNet in 4B, and VGG16 in 4C.

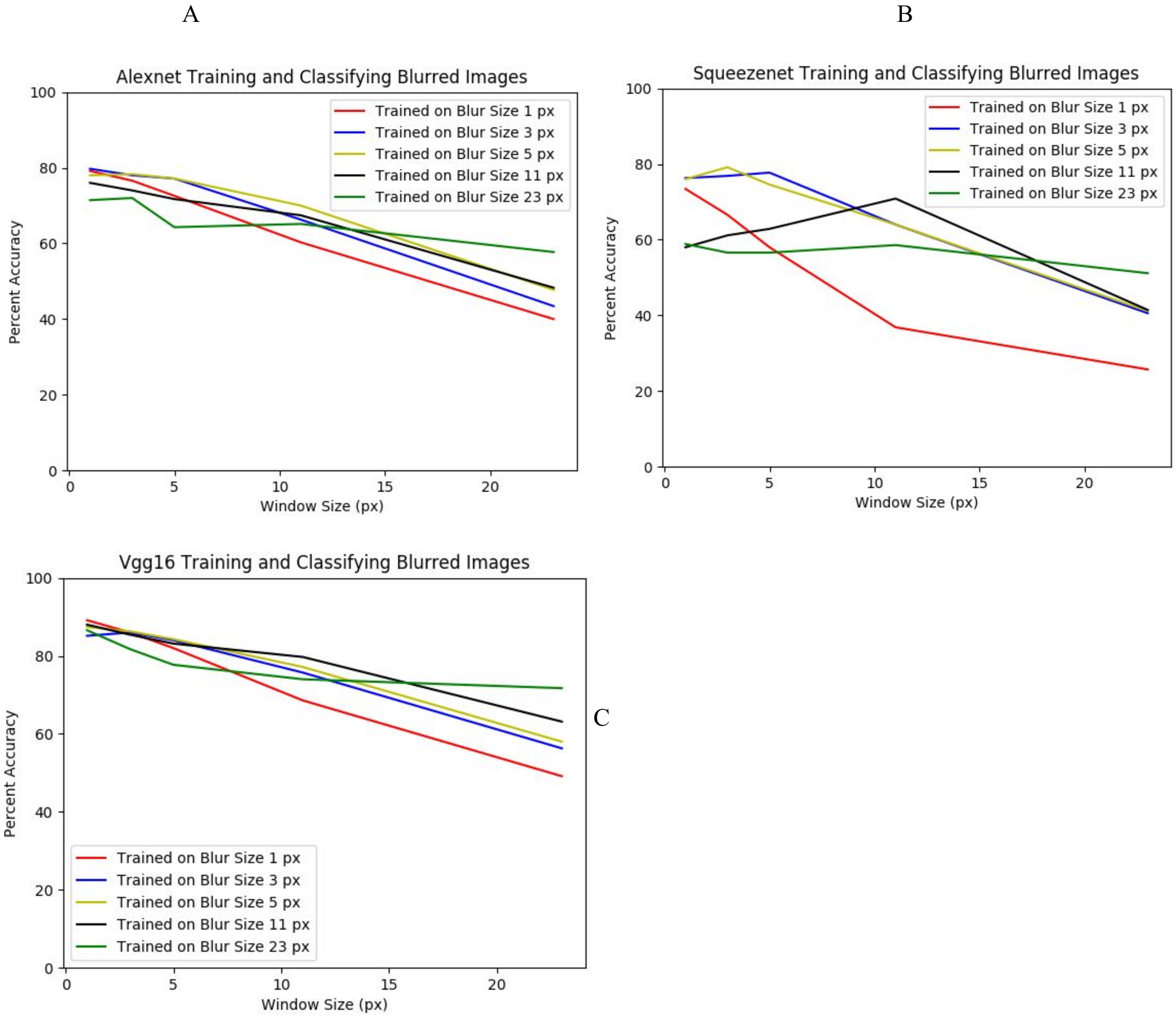


Table 3: Summary statistics for the performance of the three networks on Imagewoof.

A. AlexNet

Training Blur Level	*1px*	*3px*	*5px*	*11px*	*23px*
Mean	65.71429	68.91429	70.22857	67.48571	66.11429
Standard Deviation	16.09284	15.19318	13.03903	11.19657	5.867934
Minimum	40	43.42857	47.71429	48.28571	57.71429
Maximum	79.14286	79.71429	78.28571	76	72

B. SqueezeNet

Training Blur Level	*1px*	*3px*	*5px*	*11px*	*23px*
Mean	52.11429	67.08571	66.97143	58.85714	56.34286
Standard Deviation	20.17232	15.85573	15.51905	10.83645	3.099704
Minimum	25.71429	40.57143	41.14286	41.42857	51.14286
Maximum	73.42857	77.71429	79.14286	70.85714	58.85714

C. VGG16

Training Blur Level	*1px*	*3px*	*5px*	*11px*	*23px*
Mean	74.97143	77.42857	78.62857	79.88571	78.34286
Standard Deviation	16.43118	12.51122	12.20472	9.842847	5.963837
Minimum	49.14286	56.28571	58	63.14286	71.71429
Maximum	89.14286	86	87.42857	88	86.57143

Figure 5: Percent improvement in accuracy compared to clear training tested at the same blur level (Imagewoof).

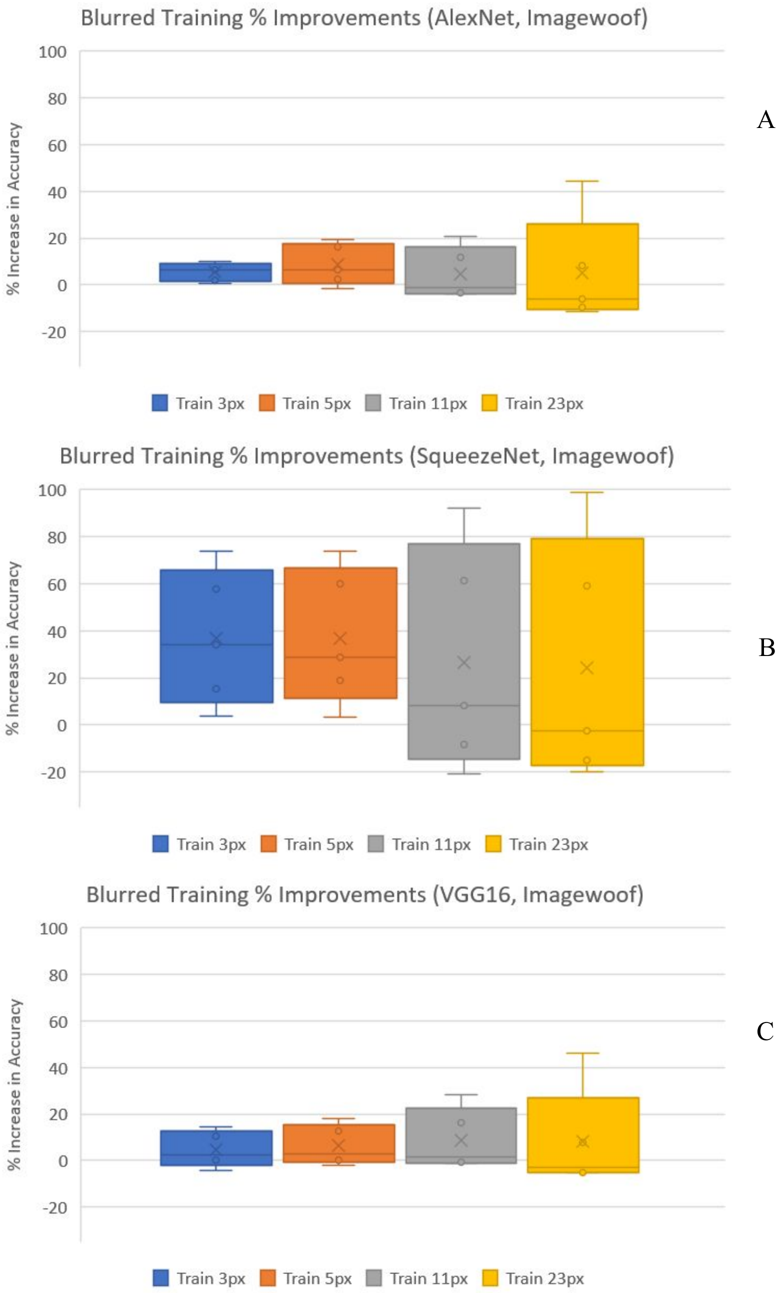


Table 4: Correlation between the level of training blur and the mean/median percent accuracy improvement for the Imagenette (high-similarity, low-diversity) dataset.

(Imagewoof)	*VGG16*	*AlexNet*	*SqueezeNet*
Mean	0.71806	0.21089	0.13191
Median	-0.75723	-0.7865	-0.52164

4. Discussion

Training CNNs on blurred images gives the best accuracy on blurred images in all cases. In terms of peak accuracy with smaller networks, this is most true when the training and testing blurs are equal. However, with larger networks, training on the highest levels of blur is the best throughout, and both clear and blurred training are nearly equal for clear images. The gains and losses become more extreme with higher-similarity data and smaller networks, but the larger networks mostly kept losses negligible. As expected, blurred training helps classify blurred images, which is necessary for any application analyzing blurred images alone. With standard and large networks like AlexNet (Krizhevsky, 2012) and VGG (Simonyan, 2015), the improvements on blurred images exist with minimal to no losses in any other area compared to standard training procedures. Together, these two results allow blurred training to improve the robustness and overall performance of a CNN with negligible drawbacks.

The observation that blurred training resulted in more consistent performance is useful in that the results of any application of that network will not vary as much based on the input data, keeping results predictable for any size network. For large networks with already high performance on clear images, this allows a similar level of performance to be maintained no matter the blurriness of the input data, making for robust performance in real-world applications. The effects of high-similarity datasets in this experiment are similar to those of smaller networks, meaning that applications that require the recognition of small details, such as facial recognition (Ramanathan, 2006), are benefited by this approach when prioritizing consistency with degraded images. A large network is required to maintain similar performance to clear training. To summarize, training on blurred images is useful for robust performance, applications with blurred images, achieving consistent and stable performance, especially with larger networks and more general objects to classify.

One possible reason for blur-trained networks' high accuracy on clear images is that this study loaded and trained pretrained models (i.e. already intended to perform well on clear images). If this is the case, then it is notable that larger networks did not lose this advantage easily like smaller networks, making them more robust for continuous learning. This would also give more evidence to support mixing images of different levels of blur into training data.

A previous study on training CNNs with blurred images focused on facial recognition, concluding that blurred training was best at recognizing images of the same level of blur, and seeing slightly more generalized accuracies (Vogelsang, 2018). This study saw evidence of that pattern in the high-similarity Imagewoof dataset along with other conclusions, suggesting that the Vogelsang study's testing on facial data was an extreme case of high-similarity testing, and fit the results of this study.

5. Conclusions

This research demonstrates that training CNNs on blurred images that are biologically accurate to infant vision is superior for applications valuing robustness, consistency, classifying blurred images, and the recognition of general objects. With larger networks demonstrating even greater benefits. While the differences between clear and blurred training became more extreme with higher-similarity datasets and smaller networks, blurred training remained valuable for more consistent and robust performance throughout and for classifying blurrier images. Many applications in the real-world could utilize the superior performance of training CNNs on blurred images in order to accomplish their tasks more effectively. The specific effects depend on the exact network and type of data, as shown in these results, and therefore the observed trends will manifest to different extents depending on the use case.

These results provide evidence that the blurred vision of a human infant is an advantage for visual learning. It is possible that it gives humans the ability to consistently identify objects, whether later in life with deteriorated vision, from a distance, in peripheral vision, or other non-ideal scenarios. If this is true, perhaps the better results in the larger networks indicate that large CNNs can more accurately represent human abilities and behavior, which is useful for further research into modeling human biology and applying it to CNNs.

One limitation is that with high-similarity data, blurred training drops in accuracy compared to standard clear training. Blurred training may not be useful for applications that rely on this aspect of performance on high-similarity data such as human faces, but since it simultaneously yields other improvements with the same data, the approach's value depends on the specific use case. Since all CNN experiments become more precise with larger datasets, this study should be repeated on larger datasets, both in terms of number of images and number of classes, that also have differing levels of similarity. Future work should build on this study by modeling infant vision even more deeply, such as by training one model on images with progressively decreasing blur, using intervals based on infant visual development. This could possibly improve the observed results even further and expand them to more use cases with more robustness. To assess how well models trained on blurred images represent humans, brain decoding could be used to compare human brain response to CNNs viewing the same images; models of human visual development could be refined with this technique. Another effect of blurred training to be explored is overfitting prevention, which is crucial for any CNN, especially those training on less data; this study did not attempt to explore this, but preliminary and extraneous experiments suggested potential in this area. Eventually, this approach to CNN training needs to be tested with other forms of blur and degradation, and then on data from real-world situations, all in order to apply this training approach to real-world devices.

This study and its implications impact the field of computer vision by demonstrating a training methodology that both models human biology and improves performance, robustness, and consistency. This study’s conclusions on human-like image processing, training, networks, and types of data can guide future research into modeling CNNs off of human vision, processing degraded images, and the development of CNN-based tools. For the broader area of modeling CNNs off of human biology/neuroscience, especially for studies involving brain decoding, this approach establishes a new baseline for trained models to be compared to the human brain since it is a way to bring all models, especially large, human-like ones, closer to human behavior and ability on the same sets of images (Yamins, 2016). This novel training approach and the future research it will guide can make devices in real-world applications more able to handle a variety of situations with robust performance and superior performance on blurred images. Addressing the real-world issue of CNN devices frequently encountering low-quality images and performing poorly on them (Dodge, 2016; Karahan, 2016), tools built for numerous industries, projects, and facets of daily life will be able to mitigate limitations in CNN robustness and accomplish their tasks more effectively with this novel approach. With a better ability to handle the conditions of the real world, computer vision will be able to take a stronger place in devices in areas such as visual impairment assistance, infrastructure, transportation, nature conservancy, and consumer technology, improving lives in many ways.

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Banks, M. S., & Crowell, J. A. (1993). Front-end limitations to infant spatial vision: Examination of two analyses. *Early visual development: Normal and abnormal*, 91-116.

Banks, M. S., & Salapatek, P. (1978). Acuity and contrast sensitivity in 1-, 2-, and 3-month-old human infants. *Investigative Ophthalmology & Visual Science*, *17*(4), 361-365.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*, 27755.

Courage, M. L., & Adams, R. J. (1990). Visual acuity assessment from birth to three years using the acuity card procedure: cross-sectional and longitudinal samples. *Optometry and vision science: official publication of the American Academy of Optometry*, *67*(9), 713-718.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE.

Dodge, S., & Karam, L. (2016, June). Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)* (pp. 1-6). IEEE.

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

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 7538-7550).

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... & He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005-10014.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, *8*, 15037.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., ... & Mujica, F. (2015). An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Karahan, S., Yildirum, M. K., Kirtac, K., Rende, F. S., Butun, G., & Ekenel, H. K. (2016, September). How image degradations affect deep cnn-based face recognition?. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)* (pp. 1-5). IEEE.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Li, J. Y., & Li, J. H. (2015, August). Fast image search with deep convolutional neural networks and efficient hashing codes. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 1285-1290). IEEE.

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Poggi, M., & Mattoccia, S. (2016, June). A wearable mobility aid for the visually impaired based on embedded 3d vision and deep learning. In *2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 208-213). IEEE.

Radovic, M., Adarkwa, O., & Wang, Q. (2017). Object recognition in aerial images using convolutional neural networks. *Journal of Imaging*, *3*(2), 21.

Ramanathan, N., & Chellappa, R. (2006). Face verification across age progression. *IEEE Transactions on Image Processing*, *15*(11), 3349-3361.

Raz, H. K., Abney, D. H., Crandall, D., Yu, C., & Smith, L. B. How do infants start learning object names in a sea of clutter?.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self‐generated variability in object images predicts vocabulary growth. *Developmental science*, e12816.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Uchida, S. (2013). Image processing and recognition for biological images. *Development, growth & differentiation*, *55*(4), 523-549.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, *115*(44), 11333-11338.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356.