Study of Human Shape and Texture biases in Convolutional Neural Networks and Transformers through Data Augmentation

Shikhar Tuli (stuli@princeton.edu)
Department of Electrical Engineering, Engineering Quadrangle, 41 Olden Street
Princeton, New Jersey 08544 USA

Abstract

In the Computer Vision domain, as dataset sizes are increasing, so is the accuracy and complexity of Machine Learning (ML) models. The state-of-the art Convolutional Neural Networks (CNNs) have already crossed humans in accuracy on recognition tasks on datasets like ImageNet. In this high accuracy regime, it no longer remains a question just about accuracy, but also about 'correctly' modelling the brain. In this work, we explore the error consistency of different vision models in ML with humans. We also probe different inductive biases, to go beyond accuracy, and compare these models more rigorously. We show that recently proposed attention-based networks outperform traditional CNNs in terms of consistency with humans.

Introduction

Convolutional Neural Networks (CNNs) are currently the defacto standard for many computer vision tasks including object detection (Ren, He, Girshick, & Sun, 2015), image classification (Krizhevsky, Sutskever, & Hinton, 2012), segmentation (Girshick, Donahue, Darrell, & Malik, 2014), facial recognition (Schroff, Kalenichenko, & Philbin, 2015), and captioning (L. Chen et al., 2017). Recently, work has also been done in Transformers for vision-based tasks, and some have even outperformed state-of-the-art CNNs (M. Chen et al., 2020; Dosovitskiy et al., 2020). However, studies have shown that CNNs trained on popular datasets like ImageNet, inherently gain texture-bias, i.e., such networks tend to classify images by texture rather than by shape (Baker, Lu, Erlikhman, & Kellman, 2018). On the other hand, humans preferentially use shape information for classification (Kucker et al., 2019; Landau, Smith, & Jones, 1988).

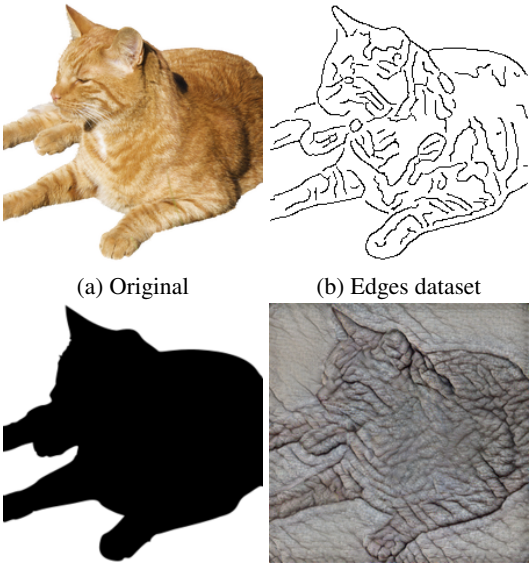
Further, there has been an increasing trend of model accuracy in computer vision tasks with a proportionate increase in the training dataset size. However, as dataset sizes increase, the computation required for training escalates (Sun, Shrivastava, Singh, & Gupta, 2017). For context, state-of-the-art CNNs need thousands of days of GPU/TPU training (Dosovitskiy et al., 2020). Forcing such vision systems to learn based on cognitively-inspired inductive biases should substantially reduce training time and computational cost.

Related Work

Many popular datasets have been used to train CNNs including CIFAR-100 (Krizhevsky, 2009), VTAB (Zhai et al., 2019), and ImageNet (Deng et al., 2009). However, none of these datasets probe into the inductive biases of such networks. To improve the robustness of trained networks, adversarial perturbations have been used (Nguyen, Yosinski, & Clune, 2015). Even though adversarial training has proved to improve accuracy on out-of-distribution data, such models are still not necessarily robust enough to other forms of imperceptible perturbations (Xiao et al., 2018; Tramèr & Boneh, 2019). In addition, there is no evidence that suggests if such examples "improve" the inductive biases of these systems. Nevertheless, recently proposed natural adversarial examples could be used to further test this (Hendrycks, Zhao, Basart, Steinhardt, & Song, 2019). Data augmentation has also been used to improve both generalization and out-of-distribution robustness of CNNs (Yun et al., 2019; Rusak et al., 2020). However, such methods often tend to decrease model accuracy (Lopes, Yin, Poole, Gilmer, & Cubuk, 2019) on traditional CNNs as we show in this paper.

A recent work by Hermann et al. shows that naturalistic data augmentation involving color distortion, noise, and blur substantially decreases texture bias, whereas random-crop augmentation increases texture bias in ImageNet-trained CNNs (Hermann & Kornblith, 2019). With the use of specially designed datasets (namely Stylized ImageNet (Geirhos et al., 2018), Navon dataset (Navon, 1977) and ImageNet-C dataset (Hendrycks & Dietterich, 2018)), Hermann et al. also show that CNNs can learn shape bias as easily as texture bias. However, these datasets have limitations that could be countered only with new datasets (Hermann & Kornblith, 2019). Moreover, this work does not consider other datasets for pre-training and their effect on bias. Drawing out a trade-off between the amount of data needed to "blindly" train a network and the use of appropriate inductive biases in the pre-train process is also worth exploring. It has also been shown that the use of attention in place of convolution appears to have little effect upon texture bias (Hermann & Kornblith, 2019). However, this does not translate to state-of-the-art Transformer models, as we show in this paper. Further, recently proposed error-consistency results between humans and CNNs have also been exploited for better evaluation and in-depth comparison of different vision systems (Geirhos, Meding, & Wichmann, 2020) [1].

[1] All experiments and results for this project are avaiable at https://github.com/shikhartuli/cnn_txf_bias



(a) Original (b) Edges dataset

(c) Silouettes dataset (d) SIN (Cue-conflict)

Figure 1: Input stimuli for Error Consistency tests. Adapted from (Geirhos et al., 2018)

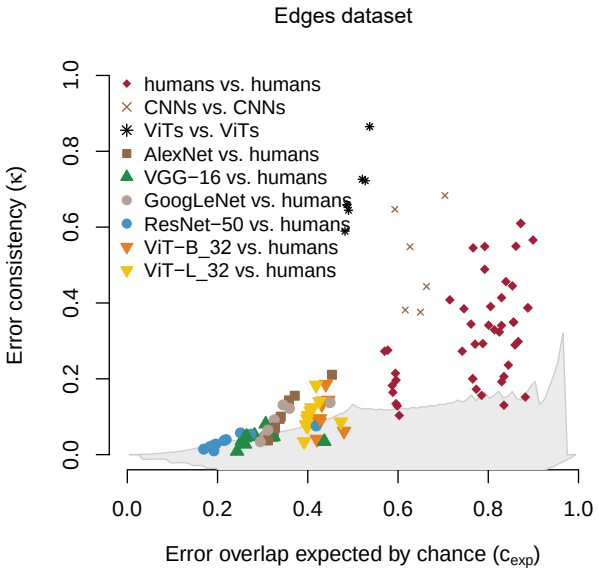


Figure 2: Error Consistency results for Edges dataset

Error Consistency in CNNs and ViT

A central problem in Machine Learning (ML) and Artificial Intelligence (AI), and also in cognitive science and behavioural neuroscience research is to establish whether two decision makers (be they humans or ML-networks) use the same strategy. Accuracy alone cannot distinguish between strategies: two systems may achieve similar accuracy with very different strategies. Hence, it is important to test the strategy of different algorithms and compare them with that of humans to find out which among them is more "brain-like". This is especially essential now as more complex models reach human-level accuracies on recognition tasks.

For this, we conduct trail-by-trail *error consistency* tests to measure if two systems make errors to the same stimulus. We do this as follows: we analyze how many of the decisions (either correct or incorrect) to individual trials are identical - *observed error overlap*. However, this metric alone is not a proxy to the similarity in strategies of two systems. As the accuracy of such systems escalate, so will the probability of this error overlap (*error overlap expected by chance*). Thus we need to normalize this overlap with that of chance. This measure of error consistency is given by the Cohen's κ (Cohen, 1960).

Mathematically, the observed error overlap is defined as $c_{obj_{i,j}} = \frac{e_{i,j}}{n}$ where $e_{i,j}$ is the frequency of equal responses by the two systems. The error overlap by chance is calculated by comparing independent binomial observers i and j with their accuracies as the respective probabilities: $c_{exp_{i,j}} = p_i p_j + (1-p_i)(1-p_j)$. This gives us the error consistency measured by Cohen's κ (Geirhos et al., 2020):

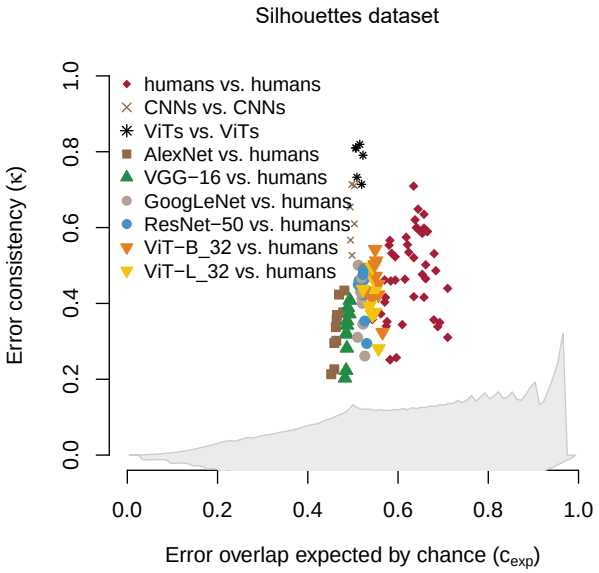


Figure 3: Error Consistency results for Silhouettes dataset

$$\kappa_{i,j} = \frac{c_{obs_{u,j}} - c_{exp_{i,j}}}{1 - c_{exp_{i,j}}}$$

Bounds on this measure and confidence intervals have also been defined theoretically (Geirhos et al., 2020).

Hence, we implemented this error consistency test on different algorithms - four popular CNNs and the recently proposed attention based Vision Transformer (ViT) (Dosovitskiy

et al., 2020). The ViT models used were pre-trained on ImageNet-21K and ILSVRC-2012 datasets (Kolesnikov et al., 2019; Russakovsky et al., 2015). The error consistency tests were implemented on the Edges dataset (Canny edge extractor implemented in MATLAB), the Silhouettes dataset (manual segmentation using domain knowledge) and the Stylized ImageNet (or the Geirhos Style-Transfer) dataset (cue-conflicts generated by texture-based style transfer) provided in (Geirhos et al., 2018). Figure 1 shows a sample input image from each of these datasets. Error consistency was compared between these models and humans, whose responses were obtained from psychophysical experiments (Geirhos et al., 2018). Figures 2, 3 and 4 show the results for the three datasets. Corroborated by recent research (Geirhos et al., 2020), we observe that CNNs indeed have very different strategies compared to humans. On the other hand, ViT has much better error consistency with humans, out of the box. It can also be seen that ViTs have much higher error consistency among themselves than CNNs. This makes sense since we are essentially comparing the same model (four different architectural sizes giving 6 comparisons).

To explain these higher error consistency results for ViT, we tested its shape bias on the Stylized ImageNet (SIN) dataset. The 1,000 ImageNet class predictions were mapped to 16 categories (for the 16-class SIN dataset) using the WordNet hierarchy (Miller, 1995). The results for this test are presented in Figure 5. Within the responses that corresponded to either the correct texture or correct shape category, the fractions of texture and shape decisions have been depicted (averages visualised by vertical lines). As is clear from the figure, ViT has a higher shape bias than traditional CNNs. This could possibly explain the higher error consistency with humans who primarily categorise objects by shape rather than texture (Geirhos et al., 2018).

Shape/Texture-bias with Data Augmentation

Now that we know that Transformers inherently have higher shape bias than traditional CNNs, we set out to test how much benefit they can extract from data augmented fine-tuning.

Inspired by augmentations presented in (T. Chen, Kornblith, Norouzi, & Hinton, 2020; Hermann & Kornblith, 2019), we fine-tuned two models on an augmented dataset - an attention based (ViT_B-32) and a convolution based (ResNet-50x1). For fair comparisons, both these datasets were pre-trained on ImageNet-21K and the ILSVRC-2012 datasets. The standard ResNet-50 network was sequentially trained on the two datasets using transfer learning (Kolesnikov et al., 2019). To limit the amount of time it takes to train the networks, we use the CIFAR-10 dataset. CIFAR-10 is a reasonably small dataset and thus, to counter the effects of random initializations and random augmentations during training, we ran the experiments 3 times and took the mean of the results. For more robust comparisons on larger-scale datasets, these experiments could easily be translated to ImageNet dataset.

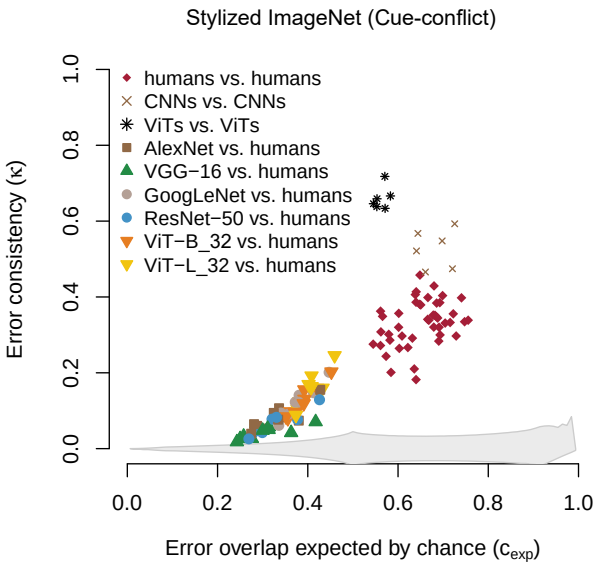


Figure 4: Error Consistency results for Stylized ImageNet

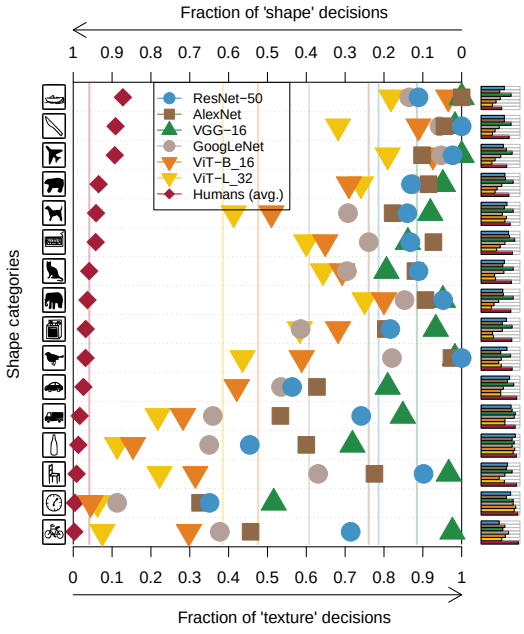
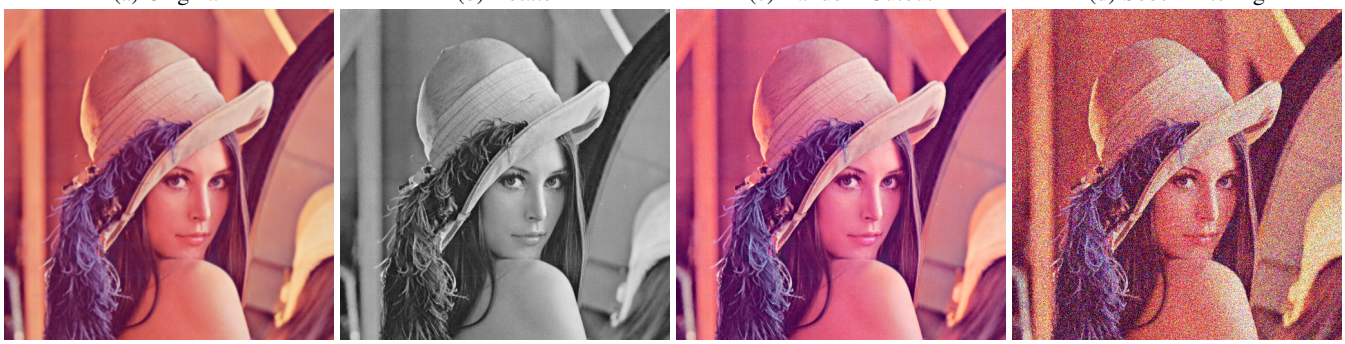


Figure 5: Shape bias for different networks for Stylized ImageNet

Figure 6 presents different augmentations that were used on the CIFAR-10 dataset to fine-tune the networks. After initializing the final layer for 10 classes, the networks were trained on the original CIFAR-10 dataset and then on these augmentations - Rotation ($\pm 90°$, $180°$ randomly), Random Cutout (rectangles of size 2×2 px to half the image width), Sobel filtering, Gaussian blur (kernel size = 3×3 px), Color

(a) Original (b) Rotate (c) Random Cutout (d) Sobel Filtering



(e) Gaussian Blur (f) Color Distortion (color drop) (g) Color Distortion (jitter) (h) Gaussain Noise

Figure 6: Data augmented images for fine-tuning

Augmentation	Shape Bias	Shape Match	Texture Match	CIFAR-10 Accuracy
Baseline	41.46%	19.08%	26.94%	93.91%
+ Rotate	41.83%	22.63%	31.47%	91.75%
+ Cutout	44.72%	22.45%	27.76%	92.27%
+ Sobel Filtering	44.80%	22.86%	28.16%	89.92%
+ Gaussian Blur	44.04%	23.45%	29.80%	90.90%
+ Color Distortion	49.59%	24.49%	24.90%	90.34%
+ Gaussain Noise	49.58%	24.08%	24.49%	88.01%

Table 1: ResNet-50 (BiT-M-R50x1) fine-tuned on augmented CIFAR-10

Augmentation	Shape Bias	Shape Match	Texture Match	CIFAR-10 Accuracy
Baseline	55.25%	33.27%	26.94%	97.29%
+ Rotate	58.97%	37.55%	26.12%	96.60%
+ Cutout	61.14%	43.67%	27.76%	97.16%
+ Sobel Filtering	55.11%	41.08%	33.47%	92.22%
+ Gaussian Blur	62.68%	54.84%	32.65%	97.76%
+ Color Distortion	64.53%	58.65%	32.24%	97.75%
+ Gaussain Noise	65.34%	59.33%	31.47%	97.37%

Table 2: ViT-B_32 fine-tuned on augmented CIFAR-10

Distortion (color jitter with probability 80% and color drop with probability 20%) and Gaussian noise (standard deviation of 0.196 for normalized image). Some of these were directly used from (T. Chen et al., 2020) and others were created from scratch on Tensorflow.

Table 1 shows the results for ResNet-50. The augmentations applied subsequently after fine-tuning on the original CIFAR-10 are additive, to reduce texture bias in the network. Training for every case was implemented for 500 epochs with linear step decay. Further details about hyper-parameter tuning can be found at `https://github.com/shikhartuli/cnn_txf_bias`. As was expected from recent research, increasing shape bias decreases 'clean' accuracy on CIFAR-10 (Hermann & Kornblith, 2019). For bias tests, only the categories common to both CIFAR-10 and Stylized ImageNet were used. Shape match is the percentage of the time the model correctly predicted probe items' shapes. Texture match is the percentage of the time the model correctly predicts probe items' textures. Finally, shape bias is the percentage of the time the model predicts shape for trials on which either shape or texture prediction is correct. From the table, it is clear that shape bias increases with training on successive augmentations.

Table 2 shows the results for ViT-B_32, the smallest Transformer model in the ViT family on augmented CIFAR-10. For every case, the network was trained for 50 epochs with cosine weight decay. Though the accuracy of ViT on CIFAR-10 is higher in this case, it should be noted that only the change in the accuracy is important for us in this study (since accuracy can be improved with further optimization of hyper-parameters). Surprisingly, the accuracy on CIFAR-10 for ViT remains about the same. This effect alone shows that Transformers are superior to CNNs with increasing shape bias. The final shape bias is also higher for ViT compared to ResNet after the networks have been fine-tuned on all types of augmentations.

Conclusions

In this work, we explore the extent of human-bias in different vision models from an error-consistency point-of-view. We see that recently proposed Transformer networks for vision tasks not only have a higher shape-bias like humans, but their responses are also more consistent with those of humans on different tasks. Further, we fine-tuned two models - a Transformer and a traditional Convolutional Neural Network (CNN) on augmented datasets to probe increasing shape-bias that is akin to humans. We observe that Transformers maintain the 'clean' accuracy while also gaining equivalently (if not more) in their shape-bias when compared to CNNs. This could possibly be explained by the basic nature of attention models - humans focus more on the part of the image that is important for the given task and neglect the otherwise noisy background to make predictions.

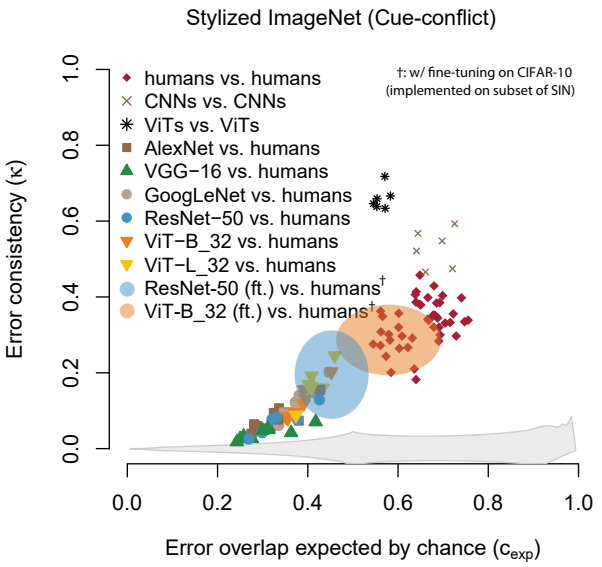


Figure 7: Error Consistency results for fine-tuned networks for Stylized ImageNet

Future Work

Many more tests can still be performed on Transformer models. For example, we could also compare ViT with iGPT on these tasks to see how the architecture within this family could affect shape and texture biases. This could help us formulate architectural "features" that help in modelling better brain-like networks. The currect networks have been fine-tuned on CIFAR-10. This could easily be extended to the ImageNet dataset and the error consistency analysis could be rerun on these models to test how much benefit (in terms of error consistency) could be extracted by ViTs when compared to traditional CNNs. The networks that were fine-tuned on augmented CIFAR-10 were tested for this on a subset of the SIN dataset (7 common categories and not 16 as is the case with 16-class ImageNet). The results are plotted in Figure 7. Precise points have not been plotted since direct comparison would not be appropriate. However, this plot gives us an idea about how ViT could gain more with fine-tuning. The gains for ViT would probably be higher than those for ResNet. Though training on augmented ImageNet is in the pipeline, it would take weeks to get results, and is thus beyond the scope of this course project.

Acknowledgments

I would like to thank Erin Grant and Ishita Dasgupta for their valuable comments.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on

global object shape. *PLoS computational biology*, *14*(12), e1006613.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T.-S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5659–5667).

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., ... Sutskever, I. (2020). Generative pretraining from pixels. In *Proceedings of the 37th international conference on machine learning*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A simple framework for contrastive learning of visual representations.*

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *arXiv preprint arXiv:2006.16736*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 580–587).

Hendrycks, D., & Dietterich, T. G. (2018). Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). Natural adversarial examples. *arXiv preprint arXiv:1907.07174*.

Hermann, K. L., & Kornblith, S. (2019). Exploring the origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071*.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2019). Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* (Tech. Rep.).

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).

Kucker, S. C., Samuelson, L. K., Perry, L. K., Yoshida, H., Colunga, E., Lorenz, M. G., & Smith, L. B. (2019). Reproducibility and a unifying explanation: Lessons from the shape bias. *Infant Behavior and Development*, *54*, 156–165.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299 - 321. doi: https://doi.org/10.1016/0885-2014(88)90014-7

Lopes, R. G., Yin, D., Poole, B., Gilmer, J., & Cubuk, E. D. (2019). Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*.

Miller, G. A. (1995, November). Wordnet: A lexical database for english. *Commun. ACM*, *38*(11), 39–41. Retrieved from https://doi.org/10.1145/219717.219748 doi: 10.1145/219717.219748

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, *9*(3), 353–383.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 427–436).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., & Brendel, W. (2020). Increasing the robustness of dnns against image corruptions by playing the game of noise. *arXiv preprint arXiv:2001.06057*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 815-823).

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the ieee international conference on computer vision* (pp. 843–852).

Tramèr, F., & Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. In *Advances in neural information processing systems* (pp. 5866–5876).

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., & Song, D. (2018). Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y.

(2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the ieee international conference on computer vision* (pp. 6023–6032).

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., … others (2019). A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.