Exploring Corruption Robustness: Inductive Biases in Vision Transformers and MLP-Mixers

Katelyn Morrison [*1] **Benjamin Gilby** [*1] **Colton Lipchak** [1] **Adam Mattioli** [1] **Adriana Kovashka** [1]

Abstract

Recently, vision transformers and MLP-based models have been developed in order to address some of the prevalent weaknesses in convolutional neural networks. Due to the novelty of transformers being used in this domain along with the self-attention mechanism, it remains unclear to what degree these architectures are robust to corruptions. Despite some works proposing that data augmentation remains essential for a model to be robust against corruptions, we propose to explore the impact that the architecture has on corruption robustness. We find that vision transformer architectures are inherently more robust to corruptions than the ResNet-50 and MLP-Mixers. We also find that vision transformers with 5 times fewer parameters than a ResNet-50 have more shape bias. Our code is available to reproduce.

1. Introduction

Research indicates that humans tend to classify objects based on shape rather than color or texture while convolutional neural networks are more biased towards texture (Ritter et al., 2017). Developing and deploying reliable, accurate computer vision models is integral to the success and trust of vision-based technologies such as self-driving cars or assistive technologies. Significant errors in these technologies could be fatal, which is why it is important to understand the limitations of different models.

In the past decade, convolutional neural networks (CNNs) have been the state-of-the-art for computer vision tasks such as image classification. However, recent research has shown the limitations of CNNs in domain generalization tasks. Several recent works seeking to develop and train models that can successfully achieve domain adaptation and/or domain generalization investigate the intricacies that contribute to corruption robustness in a model (Mummadi et al., 2021; Hermann & Kornblith, 2019; Geirhos et al., 2019; Brochu, 2019; Feinman & Lake, 2018). Recently, novel architectures have achieved exceptional performance on ImageNet and CIFAR baselines. While robustness of CNNs have been studied, it is vital to explore the robustness of these new architectures, including how well they perform when presented corrupted images.

Contributions. In this work, we investigate models with three different types of architectures: CNNs, Vision Transformers, and MLP-Mixers. In total, we compare and contrast twenty different pre-trained models. Our findings reveal how various pre-trained vision transformer architectures and MLP-Mixers make decisions (i.e., based on shape or texture) and how well they handle corruptions. Our contributions include the following:

- Comparing corruption robustness and shape bias across state-of-the-art vision transformer architectures and the MLP-Mixer.
- Showing that vision transformers seem to be inherently more robust to common corruptions than CNNs and the MLP-Mixer.

2. Related Works

Investigating inductive biases, such as shape bias and texture bias, and how these biases can improve the robustness of a model have been extensively explored within CNNs. We highlight several advancements in the past few years ranging from data augmentation techniques to novel architectures designed to improve top-1 and top-5 accuracy on image classification tasks.

2.1. Data Augmentation & Training Techniques

Geirhos et al. (2019) conduct an empirical study to understand the inductive biases learned by CNNs trained on ImageNet. After creating several augmented ImageNet data sets,

[*]Equal contribution [1]Department of Computer Science, University of Pittsburgh, Pittsburgh, United States. Correspondence to: Katelyn Morrison <kmorrison@pitt.edu>, Benjamin Gilby <beg59@pitt.edu>, Colton Lipchak <cjl99@pitt.edu>, Adam Mattioli <afm45@pitt.edu>, Adriana Kovashka <kovashka@cs.pitt.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

they show that CNNs are more texture-biased during object recognition tasks while humans are more shape biased. These results are contradicted soon after showing that CNNs can learn shape bias as easily as texture bias (Hermann & Kornblith, 2019). Hermann & Kornblith (2019) indicate that the inductive biases that the CNN learns may be solely dependent on the data it sees instead of the architecture itself.

A more recent empirical study investigates if shape bias and corruption robustness have a direct correlation (Mummadi et al., 2021). Mummadi et al. (2021) compares the accuracy and corruption robustness of CNNs trained on ImageNet with standard images, standard and stylized images, and a combination of edge maps of ImageNet and standard images. They show that the model trained on standard images and edge maps resulted in having the greatest shape bias. However, the network trained on standard and stylized images performed the best on common corruptions. They concluded that the stylized images caused increased shape bias, but corruption robustness was increased by the stylized images, not the shape bias directly.

An alternative approach explains an algorithm for shape-texture debiased learning by augmenting images in the training set with conflicting shapes and textures (Li et al., 2021). This algorithm is still based on using CNNs, but their algorithm proves to achieve improvements on ImageNet-C and Stylized-ImageNet among others. The augmentation in this algorithm consists of using conflicting shape and texture information on the original image.

2.2. Architectures for Better Accuracy

Different convolutional neural network architectures have been modified and reconstructed to achieve a higher accuracy on image classification tasks. Most recently, transformers have been modified and adapted for vision tasks such as image classification. We will only introduce the vision transformer architectures that we included in our experiments, but there are several other variations of vision transformers in the literature.

An architecture called the Vision Transformer (ViT) uses layers of multi-headed self attention and multi-layer perceptrons (Dosovitskiy et al., 2021). They conduct image classification by splitting an image into a fixed number of patches and embedding each image patch. This architecture achieves excellent results compared to CNNs on numerous baselines. Bhojanapalli et al. (2021) investigate several different ViT and ResNet models to understand the robustness of the ViT models. They also show how the two architectures perform when faced against different adversarial attacks such as PGD and FGSM. Overall, their results reveal that the ViT is as least as robust to corruptions as ResNets (Bhojanapalli et al., 2021).

A variation of the ViT vision transformer, called the Swin Transformer, calculates self-attention of a window of image patches to compute predictions for tasks such as image classification (Liu et al., 2021). The windows of image patches shift after calculating the self-attention of the previous windows. This shift results in a hierarchical feature map that provides a better global representation of the image.

Two other variations of the vision transformer architectures are the Data-efficient Image Transformers (DeiT) (Touvron et al., 2020) and Class-Attention in Image Transformers (CaiT) (Touvron et al., 2021). DeiT uses a custom distillation procedure and no convolutional layers and CaiT features class-attention layers.

A recent architecture, called MLP-Mixer, is designed to exclude convolutional and self-attention layers, and instead mix per-location features and spatial information through two MLP-based layers (Tolstikhin et al., 2021). This architecture also incorporates significant augmentation within the pre-processing pipeline to increase accuracy of the model.

To our knowledge, no current research has been published on how numerous different vision transformers compare to one another in terms of corruption robustness or shape bias. No previous research has explored how robust the MLP-Mixer is to corruption either.

3. Method

To explore how robust vision transformer architectures and MLP-Mixers are to corruptions, we perform several experiments consisting of four pre-trained CNNs, fourteen vision transformers, and two MLP-Mixers.

3.1. Pre-trained Models

Convolutional Neural Networks. The convolutional neural networks we chose were inspired by the models used in Geirhos et al. (2019). Specifically, Geirhos et al. (2019) evaluated the shape bias of ResNet-50 (He et al., 2015), AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan & Zisserman, 2015), and GoogLeNet (Szegedy et al., 2014). We evaluated the and corruption robustness of these models to use as a baseline when determining how the vision transformers and the MLP-Mixers perform.

MLP-Mixers. We evaluated two different variations of the MLP-Mixer architecture: the base and large variations. These pre-trained models were provided by the `timm library` (Wightman, 2019).

Vision Transformers. We evaluated a total of four state-of-the-art, competing vision transformers. Due to limited resources and ease of access, we choose to use the pre-trained models provided by the `timm library` (Wight-

man, 2019). Specifically from Wightman (2019), we included the Swin-T, the ViT, and the CaiT pre-trained models in our evaluation. Each of these architectures have multiple pre-trained models available. We used four different pre-trained Swin Transformers, two different pre-trained ViT models, and two different pre-trained CaiT models. We obtained pre-trained DeiT models directly from Facebook Research's GitHub (Touvron et al., 2020). We used six different pre-trained DeiT models.

3.2. Data Sets

ImageNet-C. We evaluated all of our pre-trained models on ImageNet-C to determine corruption robustness (Hendrycks & Dietterich, 2018). ImageNet-C is a benchmark dataset used to assess how robust a model is to common corruptions. This dataset consists of nineteen different corruption types that are categorized within five general corruption categories (*blur, weather, noise, digital, and extra*) for five different severity levels. The dataset is built off of the ILSVRC 2012 validation set which has $1,000$ classes and fifty validation images for each class totalling $50,000$ validation images. In terms of ImageNet-C, each corruption type (i.e., blur $\rightarrow$ motion_blur) has $50,000$ images for each severity level.

Texture-Cue Conflict. We used the Texture-Cue Conflict dataset from Geirhos et al. (2019) to evaluate the shape bias of our models. The Texture-Cue Conflict dataset consists of images that have the shape of one class combined with the texture of another. This results in conflicting shape and texture in each image. Two labels are included to identify ground-truth for *both* the shape and the texture of an image. The dataset includes 16 classes and 80 images per class for a total of 1280 images.

3.3. Evaluation Metrics

The top-1 accuracy and top-1 error from each model is used to understand how robust the model is to different corruptions. Since we are evaluating architectures that are significantly different than CNNs, we decided to deviate away from using the corruption error from AlexNet as a normalization factor (Hendrycks & Dietterich, 2019) when calculating the corruption error. Instead, we obtain the corruption error, CE, by summing the top-1 error for that corruption from severity 2 and severity 5 where f is the given model, s is the severity, and c is the corruption:

$$CE_c^f = (E_{2,c}^f + E_{5,c}^f)$$

To calculate mean corruption error, **mCE**, we take the average of all the corruption errors calculated for a given model. Typically, mCE is calculated by averaging over corruption errors from all severity levels, but we chose to only include the corruption errors from severity 2 and severity 5 in our mCE calculation. We use these two severity levels to represent an average of the overall mCE for a given model. Even though our resulting mCE will not be directly comparable to previously published mCEs, it still provides enough evidence to draw conclusions about the models we evaluated.

We provide the top-1 accuracy on ILSVRC 2012 validation images (Russakovsky et al., 2015) because this is the dataset used for creating ImageNet-C. This metric will help us understand how the model performs on a dataset without corruptions.

Each pre-trained model was also evaluated on the texture-cue conflict dataset from Geirhos et al. (2019) to calculate shape bias. The shape bias of a model is how much the model depends on shape when classifying images while texture bias is how much the model depends on the texture. Shape bias, as stated by Geirhos et al. (2019), is calculated by the following formula:

$$SB = shape_{correct}/(shape_{correct} + texture_{correct})$$

4. Results

By evaluating twenty different pre-trained models on a subset of ImageNet-C and on the Texture-Cue Conflict dataset, we expose the robustness and inductive biases for each of these models. Please refer to our appendix for more in depth results from our experiments and evaluation setup.

4.1. Corruption Robustness

We evaluated every pre-trained model on ImageNet-C and calculated the mCE to understand how each model performed against common corruptions in Table 6. When referring to Table 6, a lower mCE is more favorable and a higher top-1 accuracy is more favorable.

We observe that MLP-Mixer models perform similarly to the CNNs when tested on ImageNet-C. All of the vision transformer models we evaluated achieved a significantly better mCE than the MLP-Mixers and CNNs. One significant observation is that the tiny DeiT vision transformer with only five million parameters achieves an mCE of 60.08% while a ResNet50 with approximately five times the parameters has an mCE of 65.54%. Comparing the tiny DeiT to a GoogLeNet with a similar top-1 accuracy and similar number of parameters, the tiny DeiT outperforms the GoogLeNet by 8.74%. Overall, the architecture that achieved the lowest mCE was Swin Transformer. This model also performed the best on the ILSVRC 2012 validation set. We suspect the Swin transformer performed the best out of all of the vision transformers because of its shifting windows feature providing a global representation of the image.

Table 1. Evaluating Convolutional Neural Networks against Vision Transformer Architectures and MLP-Mixers on ImageNet-C. mCE is calculated using only severity 2 and 5. Top-1 accuracy is calculated for ILSVRC 2012 validation set.

Convolutional Neural Networks			
Blur			
Model	top-1(%)	mCE(%)	#params(M)
ResNet-50	76.02	65.54	26
AlexNet	56.44	83.18	61
GoogLeNet	71.70	68.82	7
VGG-16	69.63	75.10	138

MLP-Mixers			
Model	top-1(%)	mCE(%)	#params(M)
MLP-Mixer_B	72.53	65.54	60
MLP-Mixer_L	68.25	69.65	208

Vision Transformer Architectures			
Model	top-1(%)	mCE(%)	#params(M)
ViT_base	75.73	58.55	86
ViT_large	79.16	49.02	304
DeiT_base	81.84	42.30	86
DeiT_base-dist.	83.16	41.19	87
DeiT_small	79.68	47.79	22
DeiT_small-dist.	81.05	46.25	22
DeiT_tiny	71.92	**60.08**	**5**
DeiT_tiny-dist.	74.38	57.45	6
CaiT_s24	83.28	40.59	47
CaiT_xxs24	78.38	49.28	11
Swin-T_tiny	80.85	50.70	28
Swin-T_small	82.96	45.51	50
Swin-T_base	84.90	38.52	88
Swin-T_large	85.92	**34.63**	**197**

4.2. Shape Bias

We evaluated every pre-trained model on the Texture-Cue Conflict dataset and calculated shape bias to understand whether models were biased towards shape or texture when making decisions. When referring to Table 2, a higher shape bias is more favorable.

We observe that the MLP-Mixers and vision transformers are more biased towards shape than CNNs, and many of the vision transformer models perform similarly to the MLP-Mixers. Notably, the tiny Data-efficient Image Transformer architecture with approximately five times fewer parameters than a ResNet-50 achieves a slightly better shape bias than the ResNet-50.

Figure 3 in our appendix highlights a general inverse relationship between shape bias and mean corruption error. As a model is more robust to common corruptions (smaller mCE), its shape bias increases. However, 4 of only vision transformers presents a weaker inverse relationship. The number of parameters seems insignificant to shape bias/mCE.

Table 2. Evaluating shape bias of Convolutional Neural Networks against Vision Transformer Architectures and MLP-Mixers on Texture-Cue Conflict dataset.

Convolutional Neural Networks		
Model	Shape Bias (%)	# params (M)
ResNet-50	26.17	26
AlexNet	29.80	61
GoogLeNet	28.52	7
VGG-16	16.12	138

MLP-Mixers		
Model	Shape Bias (%)	# params (M)
MLP-Mixer_base	36.90	60
MLP-Mixer_large	38.64	208

Vision Transformer Architectures		
Model	Shape Bias (%)	# params (M)
ViT_base	49.10	86
ViT_large	**55.35**	**304**
DeiT_base	42.32	86
DeiT_base-dist.	39.62	87
DeiT_small	38.26	22
DeiT_small-dist.	36.65	22
DeiT_tiny	**29.37**	**5**
DeiT_tiny-dist.	31.06	6
CaiT_s24	38.65	47
CaiT_xxs24	34.24	11
Swin-T_tiny	25.21	28
Swin-T_small	27.43	50
Swin-T_base	36.39	88
Swin-T_large	40.20	197

5. Conclusion and Future Work

We compare several state-of-the-art vision transformers against CNNs and MLP-Mixers to better understand how these different architectures handle corruptions and if they rely on shape or texture more when classifying images. We find that vision transformers are more biased towards shape than traditional CNNs when classifying out of distribution data, and they are more robust to corruptions. It remains an open question as to how much the augmentation in the vision transformers and MLP-Mixer training pipelines impacts their corruption robustness and shape bias.

We are currently investigating how CNNs that incorporate augmentations in the training pipeline compare and contrast to the MLP-Mixers and vision transformer architectures. Other future directions include incorporating the rest of the severity levels from ImageNet-C to calculate the final mean corruption error for each model. It would also be beneficial to investigate different datasets such as ImageNet-A (Hendrycks et al., 2021), ImageNet-P (Hendrycks & Dietterich, 2019), and ImageNet-R (Hendrycks et al., 2020).

References

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification, 2021.

Brochu, F. Increasing shape bias in imagenet-trained networks using transfer learning and domain-adversarial methods. *CoRR*, abs/1907.12892, 2019. URL http://arxiv.org/abs/1907.12892.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Feinman, R. and Lake, B. M. Learning inductive biases with simple neural networks. *CoRR*, abs/1802.02745, 2018. URL http://arxiv.org/abs/1802.02745.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1807.01697, 2018. URL http://arxiv.org/abs/1807.01697.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples, 2021.

Hermann, K. L. and Kornblith, S. Exploring the origins and prevalence of texture bias in convolutional neural networks. *CoRR*, abs/1911.09071, 2019. URL http://arxiv.org/abs/1911.09071.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Li, Y., Yu, Q., Tan, M., Mei, J., Tang, P., Shen, W., Yuille, A., and cihang xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Db4yerZTYkz.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Mummadi, C. K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V., and Metzen, J. H. Does enhanced shape bias improve neural network robustness to common corruptions? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=yUxUNaj2Sl.

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pp. 2940–2949. PMLR, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions, 2014.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., and Jégou, H. Going deeper with image transformers, 2021.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Evaluation Notes. All models were evaluated using the same preprocessing pipeline and normalization values (imagenet mean/std). Only models for input size 224 were evaluated and compared. **Note:** You can visit our Notion to view an interactive database of our reported numbers for every subclass in ImageNet-C for severity 2 and severity 5. All numbers are reported in tables at the end of the appendix.

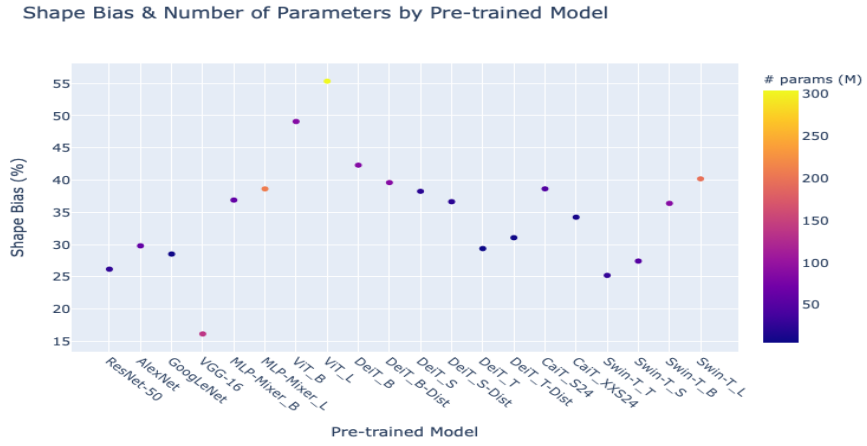

Figure 1. Shape Bias & Number of Parameters by Pre-trained Model

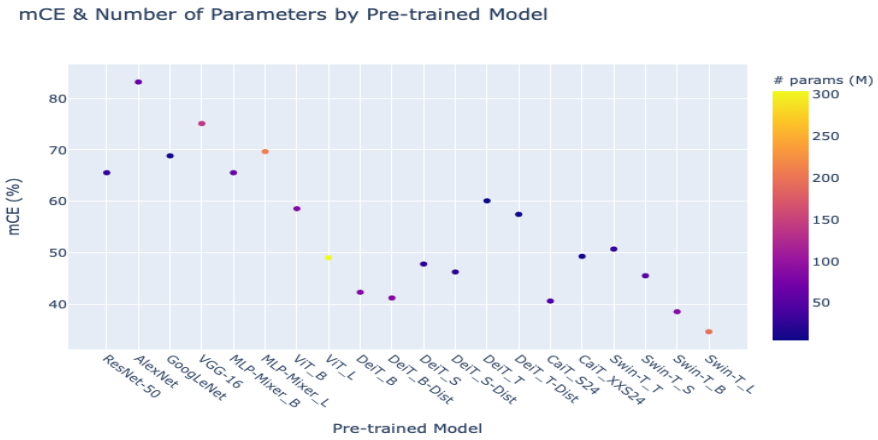

Figure 2. Mean Corruption Error & Number of Parameters by Pre-trained Model

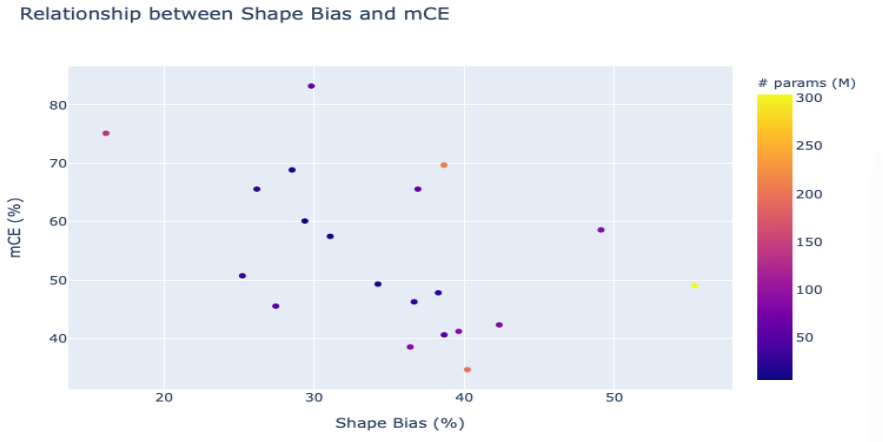

Figure 3. Relationship between mean Corruption Error and Shape Bias

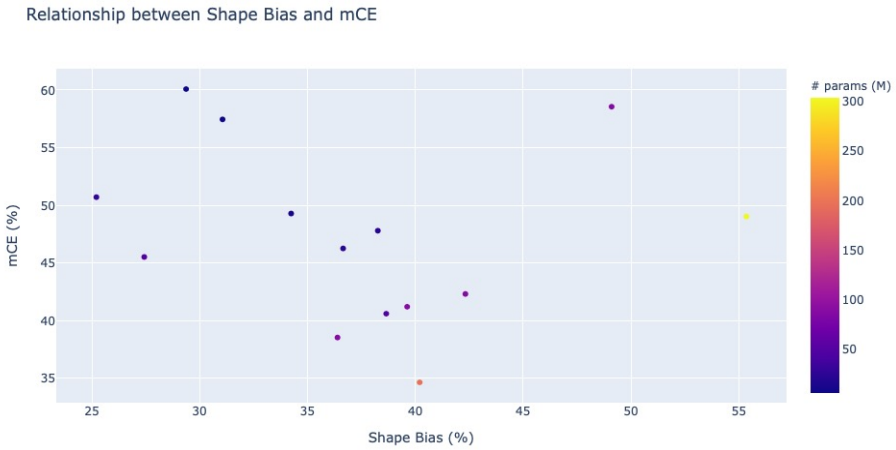

Figure 4. Relationship between Mean Corruption Error and Shape Bias for Vision Transformer architectures only. Two outliers are present in the upper right half of the graph. These two points are the base and large Vision Transformer (ViT). Across the vision transformers we evaluated, there exists an inverse relationship between shape bias and mean corruption error with the exception of the ViT models.

Table 3. Top-1 errors for each corruption from each corruption category for each model within the CNNs and MLP-Mixers for severity level 2.

Convolutional Neural Networks								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
ResNet-50	51.01	60.97	57.33	48.32	59.03	46.91	69.54	30.09
AlexNet	77.35	77.14	76.87	73.20	80.65	78.36	88.70	51.41
GoogLeNet	62.94	64.72	67.81	57.42	62.12	50.35	71.82	35.22
VGG-16	64.956	72.28	76.87	61.80	69.46	56.29	79.36	36.23

Convolutional Neural Networks							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
ResNet-50	52.75	57.46	62.42	47.71	42.06	47.43	39.49
AlexNet	84.01	86.36	91.19	74.82	63.97	74.67	68.16
GoogLeNet	53.72	56.58	69.11	50.19	48.16	58.78	42.13
VGG-16	67.88	72.83	78.24	60.45	48.65	61.19	47.75

Convolutional Neural Networks				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
ResNet-50	44.60	55.68	38.62	39.24
AlexNet	78.97	71.83	56.25	59.55
GoogLeNet	41.50	63.08	40.24	45.73
VGG-16	55.02	61.12	52.98	50.86

MLP-Mixers								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
MLP-Mixer_B	60.83	64.20	68.46	51.55	51.98	41.40	70.56	31.30
MLP-Mixer_L	67.13	69.53	71.39	57.96	57.51	50.32	77.54	36.84

MLP-Mixers							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
MLP-Mixer_B	50.38	55.61	61.42	46.57	43.18	56.62	47.48
MLP-Mixer_L	57.71	60.31	65.94	52.13	49.09	63.29	56.82

MLP-Mixers				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
MLP-Mixer_B	32.82	60.91	39.48	45.16
MLP-Mixer_L	40.56	65.25	45.32	51.94

Table 4. Top-1 errors for each corruption from each corruption category for each model within the CNNs and MLP-Mixers for severity level 5.

Convolutional Neural Networks								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
ResNet-50	84.71	91.59	77.13	86.41	79.36	77.11	83.79	44.01
AlexNet	95.77	94.39	86.73	93.31	93.69	95.24	95.10	75.19
GoogLeNet	92.98	94.63	85.05	91.87	80.51	81.81	82.59	47.48
VGG-16	95.77	94.39	86.73	93.31	79.36	84.47	91.21	54.45

Convolutional Neural Networks							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
ResNet-50	95.52	94.95	95.38	86.17	76.16	87.75	55.42
AlexNet	99.60	99.28	99.65	97.55	90.39	96.66	83.93
GoogLeNet	95.94	95.72	96.78	88.66	79.12	94.87	53.49
VGG-16	99.21	98.96	99.21	94.96	84.38	95.79	64.99

Convolutional Neural Networks				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
ResNet-50	95.68	85.17	77.09	66.37
AlexNet	99.35	82.31	91.94	77.25
GoogLeNet	94.94	86.10	62.77	68.44
VGG-16	97.89	91.91	92.71	80.00

MLP-Mixers								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
MLP-Mixer_B	88.22	94.23	85.02	84.56	71.73	69.22	82.73	38.30
MLP-Mixer_L	89.11	94.87	85.30	86.00	76.74	73.30	85.95	43.76

MLP-Mixers							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
MLP-Mixer_B	98.50	97.57	99.20	86.95	75.50	89.95	49.53
MLP-Mixer_L	95.34	94.60	96.96	83.67	80.87	90.83	55.01

MLP-Mixers				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
MLP-Mixer_B	80.35	83.08	71.33	64.64
MLP-Mixer_L	82.12	83.65	80.60	71.54

Table 5. Top-1 errors for each corruption from each corruption category for each model within the Vision Transformers for severity level 2.

Vision Transformers								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
ViT_Base	43.15	45.43	52.91	37.66	60.47	32.88	69.93	29.43
ViT_Large	34.25	34.28	42.21	30.16	48.40	29.32	55.81	24.81
DeiT_Base	35.94	38.25	44.44	30.94	32.45	30.90	40.33	22.54
DeiT_Base-Dist.	32.71	37.13	42.85	28.27	32.62	25.38	38.46	20.85
DeiT_Small	38.97	43.71	51.37	34.19	36.55	34.16	45.28	24.99
DeiT_Small-Dist.	38.52	43.33	49.34	31.76	37.15	29.31	44.76	23.28
DeiT_Tiny	51.76	56.13	62.67	46.81	49.88	45.66	60.88	33.46
DeiT_Tiny-Dist.	50.30	53.91	61.29	43.28	47.19	39.36	58.08	54.45
CaiT_S24	32.43	38.03	41.77	27.59	32.23	25.52	37.99	20.56
CaiT_XXS24.	39.20	46.99	52.70	34.87	40.04	32.66	47.62	25.78
Swin-T_Tiny	42.44	50.47	51.69	36.27	39.45	32.65	48.62	24.95
Swin-T_Small	37.57	45.51	45.45	31.34	35.67	30.03	43.84	22.68
Swin-T_Base	30.65	37.96	37.60	27.26	32.12	25.21	35.34	20.25
Swin-T_Large	27.20	33.51	33.32	24.50	29.01	23.76	31.11	18.83

Vision Transformers							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
ViT_Base	44.58	48.53	50.56	38.69	34.14	40.17	36.95
ViT_Large	33.40	35.34	37.86	29.16	29.71	32.09	31.36
DeiT_Base	26.47	28.28	28.94	26.69	26.39	33.16	24.59
DeiT_Base-Dist.	25.98	27.25	28.55	24.93	24.67	30.03	22.80
DeiT_Small	31.03	33.07	35.19	30.09	28.96	36.54	27.63
DeiT_Small-Dist.	30.31	31.57	33.37	28.434	27.99	36.05	25.87
DeiT_Tiny	43.65	45.40	48.36	40.65	39.64	49.12	37.22
DeiT_Tiny-Dist.	40.76	42.26	45.86	37.88	36.34	47.38	34.31
CaiT_S24	25.70	27.01	28.16	24.78	23.64	30.27	22.37
CaiT_XXS24.	33.70	35.78	38.00	32.40	30.02	36.61	28.45
Swin-T_Tiny	34.08	35.77	38.04	32.58	30.37	39.77	28.00
Swin-T_Small	29.94	32.18	32.97	29.76	27.46	34.87	24.99
Swin-T_Base	25.61	26.86	27.46	25.07	22.75	28.68	21.84
Swin-T_Large	23.93	25.14	24.93	23.53	21.30	25.91	20.16

Vision Transformers				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
ViT_Base	26.29	51.87	28.23	34.47
ViT_Large	23.49	46.95	23.39	28.10
DeiT_Base	25.24	43.03	26.15	29.54
DeiT_Base-Dist.	23.20	41.26	24.62	27.46
DeiT_Small	28.10	46.62	29.38	31.84
DeiT_Small-Dist.	25.85	44.58	28.27	30.02
DeiT_Tiny	38.32	56.11	41.75	41.53
DeiT_Tiny-Dist.	34.05	52.98	39.65	38.86
CaiT_S24	22.57	40.78	24.94	26.62
CaiT_XXS24.	28.30	47.62	32.59	32.86
Swin-T_Tiny	28.46	46.66	32.47	33.35
Swin-T_Small	25.48	43.73	27.88	30.15
Swin-T_Base	21.83	41.31	23.03	24.26
Swin-T_Large	20.39	37.77	20.61	22.31

Table 6. Top-1 errors for each corruption from each corruption category for each model within the Vision Transformers for severity level 5.

Vision Transformers								
Model	Blur				Weather			
	Defocus	Glass	Zoom	Motion	Frost	Fog	Snow	Brightness
ViT_Base	73.99	80.83	73.94	71.16	79.40	62.12	88.72	56.42
ViT_Large	64.38	70.26	63.92	59.19	66.02	54.15	76.46	43.99
DeiT_Base	69.45	80.52	66.35	65.48	43.55	38.04	49.74	28.60
DeiT_Base-Dist.	67.34	81.39	65.21	64.75	46.51	39.78	49.23	26.21
DeiT_Small	73.40	84.43	71.37	70.90	49.01	50.78	56.34	32.23
DeiT_Small-Dist.	72.46	85.09	70.73	69.14	53.42	45.77	56.23	29.83
DeiT_Tiny	82.20	89.39	78.58	79.41	64.46	70.04	71.34	42.78
DeiT_Tiny-Dist.	82.43	89.95	79.28	78.80	63.16	58.85	69.73	39.18
CaiT_S24	66.066	82.82	64.91	62.21	45.65	39.78	49.56	25.67
CaiT_XXS24.	75.38	87.11	74.27	72.90	54.33	48.72	58.54	32.97
Swin-T_Tiny	78.31	89.83	72.63	74.93	53.88	53.99	62.13	32.70
Swin-T_Small	71.89	89.039	66.78	66.45	48.69	40.86	56.66	29.64
Swin-T_Base	59.87	79.99	58.86	55.93	44.65	35.01	47.42	26.47
Swin-T_Large	54.41	75.31	52.91	49.09	40.21	31.65	40.09	24.27

Vision Transformers							
Model	Noise				Extra		
	Gaussian	Shot	Impulse	Speckle	Spatter	Gaussian Blur	Saturate
ViT_Base	95.87	96.31	96.52	83.24	76.18	68.30	83.79
ViT_Large	87.44	88.29	86.88	65.55	53.15	68.43	52.28
DeiT_Base	51.82	56.10	50.67	46.12	45.76	73.46	32.48
DeiT_Base-Dist.	58.35	59.26	56.67	45.54	43.33	71.57	29.61
DeiT_Small	66.29	68.23	66.03	56.08	51.68	76.88	37.07
DeiT_Small-Dist.	66.63	66.89	65.86	52.95	49.71	75.66	34.41
DeiT_Tiny	84.11	84.44	84.07	70.28	64.41	84.43	49.00
DeiT_Tiny-Dist.	82.64	82.63	82.81	68.42	61.84	84.77	45.34
CaiT_S24	57.05	56.79	56.55	44.31	39.16	69.77	29.39
CaiT_XXS24.	69.25	71.21	68.69	60.03	50.21	77.63	38.02
Swin-T_Tiny	69.64	70.80	72.09	58.10	49.47	81.54	38.05
Swin-T_Small	61.80	63.05	62.71	52.47	44.44	76.20	34.85
Swin-T_Base	53.51	52.74	53.14	42.38	35.94	64.41	30.66
Swin-T_Large	49.50	48.70	48.02	38.70	29.68	59.45	27.63

Vision Transformers				
Model	Digital			
	Contrast	Elastic Transform	Pixelate	Jpeg Compression
ViT_Base	77.01	66.91	55.11	54.96
ViT_Large	69.69	56.34	40.25	45.92
DeiT_Base	51.29	63.59	56.76	43.40
DeiT_Base-Dist.	46.89	66.21	47.38	40.97
DeiT_Small	61.40	66.85	61.91	47.38
DeiT_Small-Dist.	54.46	68.54	55.11	45.00
DeiT_Tiny	76.43	78.05	81.31	59.43
DeiT_Tiny-Dist.	67.13	77.05	77.16	58.01
CaiT_S24	44.60	68.21	47.29	39.65
CaiT_XXS24.	56.05	69.73	62.01	49.50
Swin-T_Tiny	65.69	79.04	67.84	49.68
Swin-T_Small	60.55	73.34	54.45	44.19
Swin-T_Base	52.55	65.19	35.85	34.06
Swin-T_Large	47.47	47.47	32.44	31.78