Investigating the Generalization of Image Classifiers with Augmented Test Sets

Connor Shorten
Florida Atlantic University
Boca Raton, Florida
cshorten2015@fau.edu

Taghi M. Khoshgoftaar
Florida Atlantic University
Boca Raton, Florida
khoshgof@fau.edu

***Abstract*—Adding prior knowledge about the task or domain being learned can greatly facilitate learning. Two of the most common examples of injecting prior knowledge into Deep Learning systems are architecture design and data augmentation. For example, the convolutional architecture biases the model to learn local features. This emphasis on local features has been a useful prior for image processing. Data augmentation is full of examples that utilize prior knowledge. For example, cropping an image and preserving the original label gives the inductive bias of local feature importance. In this study, we aim to see how the priors in architecture interplay with augmentations. We begin by showing the overall benefit of training with data augmentation, improving the Vision Transformer's test accuracy from 74.4% to 84.3% and improving the ResNet's test accuracy from 79.3% to 86.7%. We focus on the distinction between global and local priors such as the difference between the convolution and attention layers and cropping versus noise addition augmentations. These tests are not yet able to find complementing flaws in architectures and augmentations. We find that neither the ResNet or the Vision Transformer is robust to distribution shifts controlled with data augmentation. The performance of both models degrades heavily even with moderate augmentation strengths. Although remedied by explicitly training with the augmentation used to construct the test set, we still see a notable decrease in performance. This study illustrates the utility of generalization testing with data augmentation and the challenge of measuring the impact of global and local priors in architecture.**

***Index Terms*—Computer Vision, Data Augmentation, Generalization, Inductive Bias, Big Data**

I. Introduction

One of the critical tasks in Deep Learning research is the search for prior knowledge, or inductive biases, that can facilitate effective learning. Priors can be integrated into the architecture of neural networks, loss functions, or big data processing schemes, to name a few. Computer vision applications have benefited from the success of prior knowledge in convolutional layers. The convolution layer utilizes the prior knowledge that local neighborhoods of pixels tend to share information about the image. This prior knowledge is integrated into the network through weight sharing in local kernels. Recently, a new architecture design has gained popularity in computer vision, using a much different type of inductive bias. The Vision Transformer [5], building on the attention layer, looks to aggregate features globally across the entire image, rather than in local windows.

Data augmentation [11] is another popular strategy to introduce prior knowledge into a learning system. Augmentations such as rotations, horizontal flips, or crops provide the inductive bias that labels are invariant to these transformations. Different augmentations are built on different priors. For example, heavily cropping images signals that the local features included in the crop preserve the global label. The goal of this study is to understand how these different priors in augmentations interface with priors in architecture design.

More particularly, we study the distinction between global and local priors. We define global priors to be information that is communicated across the entire data point. Local priors are defined to be information communicated in local windows. In images, local priors may focus on (3x3) neighborhoods of pixels. Global priors in images transfer knowledge from pixels as far as the (0,0) top-left coordinate of a pixel grid to the bottom-right, (height,width) position. Convolutions utilize local priors where information is shared in (3x3) windows. Attention is a global mechanism sharing information between all pixel locations. Similarly, the crop augmentation [27] emphasizes truncated windows of the image, whereas the gaussian noise augmentation preserves the global features and introduces local noise. Figure 1 illustrates the differences between global and local transformations of images.

This study aims to bridge the inductive biases in architecture and augmentation design. This is useful for understanding the strengths and weaknesses of different models and training strategies. This is similar to corruption analysis [28] or adversarial example tests [29]. Understanding the performance of different architectural biases will also be useful for curating functional diversity for ensemble learning. Our study is also aimed at understanding how data augmentation policies optimized for convolutional networks will transfer to the Vision Transformer, and generally understanding the performance of the new Vision Transformer model. Better understanding how priors injected into different components of the Deep Learning pipeline interact with each other will be an important tool for future design of big data systems.

We highlight the interplay between architecture and augmentation through generalization tests constructed with augmentations. We first train ResNet [8] and Vision Transformer [5] models, with and without data augmentation. We then evaluate them on test sets composed with monotonically

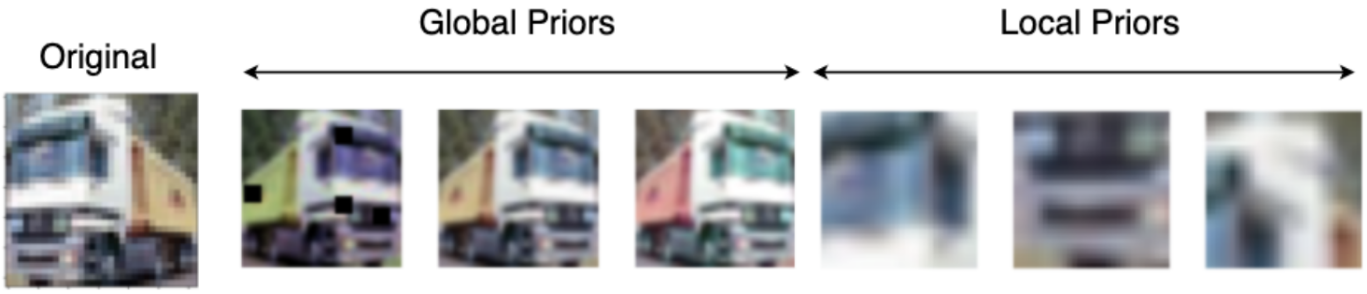

Fig. 1. An illustration of global and local data augmentations.

increasing magnitudes of augmentation. The augmentations tested include cropping, gaussian noise, and rotation. Cropping is an example of a local augmentation. It re-arranges the image into a local neighborhood of pixels. Gaussian noise is a global augmentation. It preserves the distant spatial configurations that would relate the top-left pixel to locations as far as the bottom-right pixel while adding local corruptions. These local noise additions should require the model to possess global representations to pass the corruption tests. We also test rotation as a baseline that represents a hybrid between global and local priors. Compared to cropping or gaussian noise, it is more challenging to define rotation as a global or local prior.

We continue to explore how training with augmentation relates to the performance on these test sets. We begin by showing how the models trained with RandAugment [14] generalize to different parameterizations of RandAugment used to construct test sets. This also shows a fairly narrow window around the training parameters for successful generalization. We also train models solely with the cropping augmentation. The Vision Transformer does outperform the ResNet in this test, hinting at a relationship with the local priors in crop augmented training. However, this Vision Transformer model does not improve over the Vision Transformer trained with RandAugment when evaluated on test sets also derived from cropping. Our contributions are as follows:

- We show the impact of augmentation on ResNets and Vision Transformers.
- We highlight the challenge of training Vision Transformers due to the requirement of long training times.
- We evaluate performance on test sets constructed from data augmentations on ResNets and Vision Transformers trained either with RandAugment or with no augmentation.
- We show that ResNets and Vision Transformers are not robust to distribution shifts controlled with data augmentation.
- We show that training with the augmentations used in testing does not prevent degradation at high magnitudes of augmentation.

The remainder of the paper is organized as follows. We highlight related work (Section II) such as robustness and corruption testing, distribution shift tests, vision transformers, comparative research on convolutions and attention, and data augmentation. In Section III, our experiments first illustrate the impact of augmentation on the performance of the Vision Transformer and ResNet. We then show that the Vision Transformer relies on longer training schedules in order to rival the performance of the ResNet. We then show the degradation of the Vision Transformer and ResNet models when evaluated on test sets constructed with data augmentation. We further investigate how including these augmentations in the training distribution aids in these evaluations. We continue with discussions (Section IV) around the categorization of global and local data augmentations, corruption testing, the impact of augmentation during training, and using generative models to construct test sets. In Section V, we conclude with the key findings of the study and the most promising ideas for future work. Experiment code and model weights are available at https://www.github.com/CShorten/AugmentationZoo.

II. Related Work

A. Robustness and Corruption Testing

Testing generalization with corrupted datasets is a common research area of Deep Learning. A model's ability to generalize to these corruptions is commonly referred to as robustness. Paul and Chen [4] find that Vision Transformers are more robust than CNNs. Morrison et al. [1] also study the impact of the Vision Transformer architecture on robustness. Many of these tests are concerned with adversarial robustness, in which the test distribution is constructed by an adversarial controller [9]. Shao et al. [3] find that the Vision Transformer contains less low-level information and is more generalizable than CNNs. However, we do not find a significant performance difference between the ResNet and Vision Transformer with noise transformations of the original images. Aside from adversarial tests, our experiments show that even standard augmentations significantly decrease test performance.

B. Distribution Shift Tests

Modeling performance is typically evaluated with independent and identically distributed (i.i.d.) test sets. These test sets are sampled from the same distribution as the training data. Researchers are also looking to explore test sets that represent particular kinds of distribution shifts. Koh et al. [45]

present the WILDS benchmark for simulating real-world data distribution shifts. These tests in WILDS span across several domains, such as tumor identification in medical images, camera traps for wildlife monitoring, and satellite imaging, to provide a few examples. In addition to real-world distribution shifts, many researchers turn to augmentations to simulate distribution shift. Geirhos et al. [2] construct the Stylized-ImageNet dataset in which the Neural Style Transfer algorithm [21] is used to test the impact of texture on predictions. This is used to control the texture and shape of images. For example, elephant skin texture can be placed on a cat image and retain the shape of the cat. Geirhos et al. find that ResNets are heavily biased towards the texture of images, whereas humans are much more reliant on the shape, also described as the shape bias. Morrison et al. [1] find that the Vision Transformer has more shape bias than the ResNet model, even with five times fewer parameters in the network.

C. Vision Transformers

The Vision Transformer from Dosovitskiy et al. [5] has seen a surge of interest in computer vision. At the time of this publication, Zhai et al. [6] hold the state-of-the-art in ImageNet [25] classification at 90.45% by scaling the Vision Transformer up to two billion parameters. Steiner et al. [7] explore details about the Vision Transformer such as the impact of augmentation and regularization, concluding that strong augmentation can achieve the same performance as models trained with more data.

D. Convolutions and Attention

Understanding the relationship between convolutions and attention layers in computer vision is a very active area of research. This research is mostly focused on hybrid architecture designs that interleave convolutional and attention layers. Srinivas et al. [16] presented the BoTNet and a taxonomy of these hybrid models. Dai et al. [17] present the CoAtNet which rivals the ImageNet state-of-the-art with a merged convolution and attention layer. Similarly, efficient transformers [20] try to emulate the local structure of convolutions with strided attention patterns. Notably, Gray et al. [19] use this strategy to achieve efficiency improvements with Sparse Transformers. The Transformer in Transformer (TNT) architecture [18] uses local computation within the Vision Transformer's patch projection. Similar to these works, we aim to understand the relationship between convolutions and attention.

E. Data Augmentation

Data augmentation has been a large contributor to success in computer vision as early on as AlexNet in 2012 [13]. Since the use of random cropping in AlexNet, many more augmentation strategies have been developed such as rotations, horizontal flips, and brightness alterations, to name a few. We refer interested readers to [11] for a complete list of these augmentations in computer vision and to [12] for a list of these augmentations in natural language processing, another popular application area of Deep Learning. With a large list of available augmentations, researchers have developed optimization algorithms to find the best strategy of applying them. Cubuk et al. [15] present AutoAugment, which uses reinforcement learning to find an optimal policy, requiring 5,000 GPU hours on CIFAR-10 and 15,000 GPU hours for ImageNet. Cubuk et al. [14] later illustrated that simple heuristics were competitive with expensive optimization. These studies were conducted with convolutional networks and it remains to be seen how these policies transfer to the Vision Transformer architecture. We are motivated by the heuristic design of RandAugment and hypothesize that the Vision Transformer may benefit more from local priors that are missing in the attention layers.

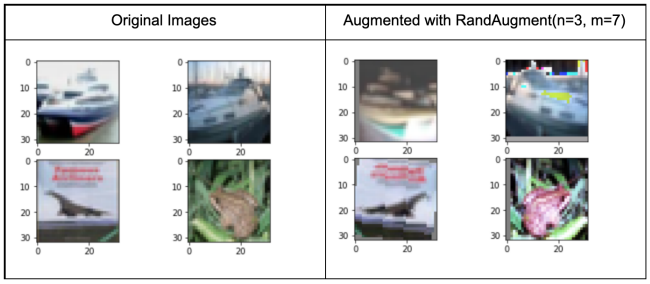

Fig. 2. Images before and after augmentation with RandAugment.

III. Experiments

The following experiments report classification accuracies of ResNet [8] and Vision Transformer [5] models evaluated on the CIFAR-10 dataset. The models are implemented and trained using the Keras framework. The tested Vision Transformer contains 8 layers and 3 attention heads per layer, totaling 21.6 million parameters. The ResNet model contains 50 layers, totaling 25.6 million parameters. We have chosen to put the models on a similar parameter count for comparison. The much larger layer depth of the ResNet can also be viewed as a global image prior guiding convolutions to aggregate local features through several layers. We conclude that this detail may be a key confounding factor for studying global and local augmentation generalization. The models are additionally trained with the RandAugment augmentation implemented in the "imgaug" library. Figure 2 illustrates examples of images before and after the RandAugment transformation. The augmentations used for test set construction are also sourced from the imgaug library. The test sets are constructed by augmenting the CIFAR-10 test set, as opposed to viewing how performance would degrade when augmenting the train distribution the models have been fitted to. These experimnts were run on a Google Colab runtime. We are looking into scaling this up with HPCC systems for future work [26].

A. Impact of Augmentation on the ResNet and Vision Transformer

Table I illustrates the impact of Data Augmentation on the ResNet architecture versus the Vision Transformer. We see that

TABLE I
IMPACT OF DATA AUGMENTATION ON TRAINING RESNET AND VISION TRANSFORMER MODELS.

Model	Augmentation	Train Accuracy	Test Accuracy
ResNet	No	98.9%	79.3%
ResNet	Yes	92.1%	86.7%
Vision Transformer	No	97.5%	74.4%
Vision Transformer	Yes	79.8%	84.3%

TABLE II
PERFORMANCE OF LONGER TRAINING TIMES WITH THE VISION TRANSFORMER AND RANDAUGMENT AUGMENTATION POLICY.

Epoch	Train Accuracy	Test Accuracy
100	79.8%	84.3%
200	83.4%	85.3%
300	84.2%	85.7%
400	85.1%	85.9%
500	85.8%	86.1%

both models overfit the training data without augmentation. Without augmentation, the ResNet is better at simultaneously overfitting this training data and generalizing to the test set, achieving 4.9% higher test accuracy than the Vision Transformer without augmentation. When adding the RandAug data augmentation scheme, both models show a significant improvement in test set generalization. Interestingly, they have a similar test accuracy, but the Vision Transformer only achieves 79.8% accuracy on the train set augmented with RandAugment. This hints that it may be beneficial to continue training the Vision Transformer on this augmented distribution.

B. Longer Vision Transformer Training

Table II explores the longer training schedule of the Vision Transformer with RandAugment, as suggested by the results in Table I. The Vision Transformer's training accuracy improves significantly every 100 epochs of training. However, from epoch 400 to 500, the Vision Transformer test accuracy only improves by 0.2%. We note that 100 epochs of training the Vision Transformer with the RandAugment preprocessing takes 2 hours, so we are limited to unrolling several augmentation configurations into 500 or more epochs. We present this limitation further and how additional tooling will help understand the impact of Data Augmentation on the Vision Transformer in the Discussion section.

C. Generalization Testing with Augmentation Sets

The following experiments test the degradation of the Vision Transformer and ResNet models on test sets constructed with increasing levels of data augmentations. Figure 3 visualizes how these augmentations transform the original images. Figures 4, 5, and 7 provide a performance visualization of all 4 models. These experiments test the Vision Transformer trained with 500 epochs because it starts at a similar test accuracy as the ResNet with 100 epochs. We provide another view of the degradation to different augmentations in Tables III, IV, and V. Note that the models illustrated in the tables are trained without augmentation to attempt to solely illustrate the relationship between the bias in architectures and augmentations.

TABLE III
ACCURACY OF RESNET AND VISION TRANSFORMERS TRAINED WITHOUT AUGMENTATION EVALUATED ON TEST SETS CONSTRUCTED FROM DIFFERENT CROP SIZES.

Crop %	Vision Transformer	ResNet
100%	74.4%	79.4%
95%	73.2%	75.3%
90%	71.3%	72.9%
75%	57.6%	56.6%
50%	30.9%	28.2%
25%	19.5%	18.5%

TABLE IV
ACCURACY OF RESNET AND VISION TRANSFORMERS TRAINED WITHOUT AUGMENTATION EVALUATED ON TEST SETS CONSTRUCTED FROM DIFFERENT MAGNITUDES OF GAUSSIAN NOISE.

Noise Range	Vision Transformer	ResNet
Original	74.4%	79.4%
(10, 20)	36%	32.8%
(10, 30)	36.9%	33.8%
(10, 50)	36%	33.3%
(30, 60)	36.4%	33.6%

Figure 4 shows the accuracy degradation across increasing magnitudes of cropping. A 25% crop with (32x32) CIFAR-10 images results in an (8x8) patch of the image. This patch is then up-sampled to 32x32 with bilinear interpolation. At this crop percentage, all models perform poorly, even those trained with RandAugment that includes cropping in the training distribution. Generally, we see a monotonic performance decrease across all models with decreasing crop size. We had hypothesized that the ResNet would be more robust to this test because of the complementing local priors, however, our results do not confirm this. Figure 7 also illustrates the trend of decreasing performance across stronger augmentations, but does not highlight a significant difference between ResNets and Vision Transformers. We note that future work looking at higher resolution images or more salient cropping may better illustrate differences between the ResNet and Vision Transformer.

Figure 5 shows the performance across increasing magnitudes of added Gaussian Noise. Oddly we do not see further

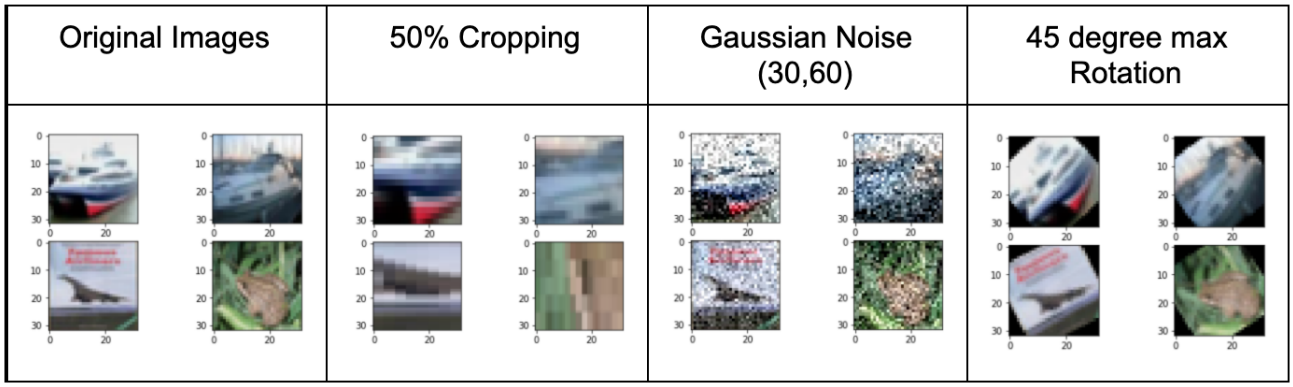

Fig. 3. Images before and after augmentations used for constructing the test sets.

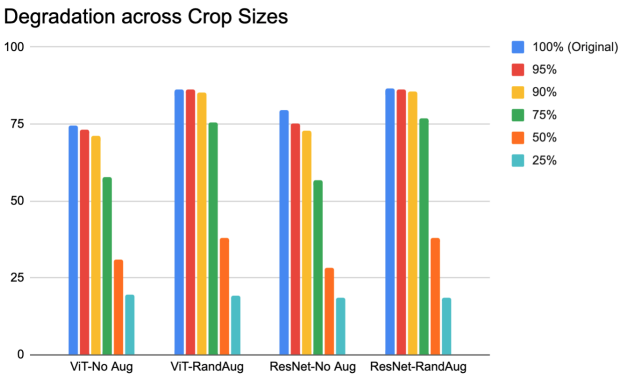

Fig. 4. Performance decrease with increasing crop sizes. We do not see a notable difference between the ResNet and Vision Transformer. We do see the models trained with RandAugment perform better than those trained without any augmentation.

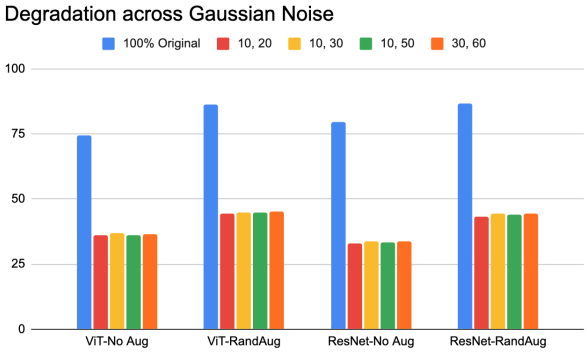

Fig. 5. Performance change with increasing magnitude of added Gaussian Noise.

performance decreases by increasing the level of noise. We further cannot confirm our hypothesis that the global priors in attention complement the global prior of an added noise map. We visualize the impact of stronger gaussian noise in Figure 6 to assure the experiment has been set up correctly.

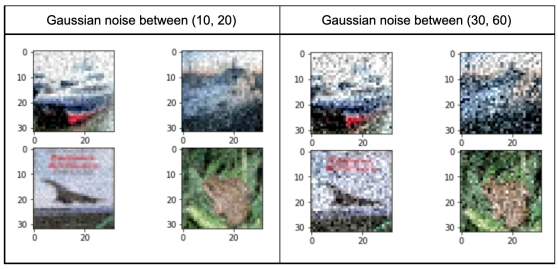

Fig. 6. Difference in images augmented with stronger levels of gaussian noise. We do not see any performance difference across the tested models.

TABLE V
ACCURACY OF RESNET AND VISION TRANSFORMERS TRAINED WITHOUT AUGMENTATION EVALUATED ON TEST SETS CONSTRUCTED FROM DIFFERENT LEVELS OF ROTATION.

Rotation	Vision Transformer	ResNet
Original	74.4%	79.4%
(-10, 10)	69.3%	72%
(-25, 25)	60.6%	58.8%
(-50, 50)	46.7%	43.6%
(-90, 90)	36.3%	32.7%

D. Training with the Tested Augmentations

Even though crop, rotation, and noise augmentations are included in the RandAugment augmentation policy, the Vision Transformer and ResNet models trained with RandAugment still fail at high magnitude augmentation tests. The following experiments probe further into this behavior, looking at test sets constructed from RandAugment and training models solely with the crop augmentation. We had hypothesized that the local prior of cropping would complement the ResNet. However, we alternatively see the Vision Transformer's test accuracy decrease slower across increasing augmentations than the ResNet when both models have been trained with 90% cropping.

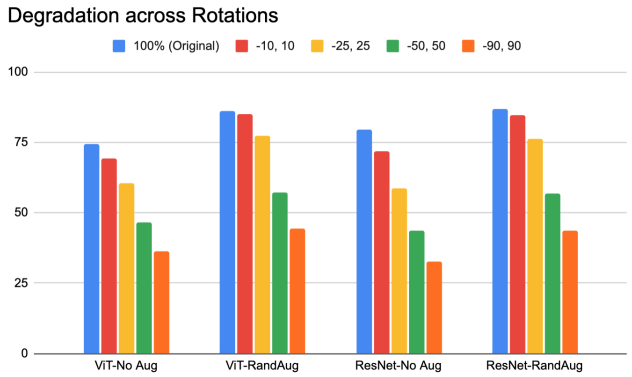

Fig. 7. Performance across different levels of Rotation magnitude. We do not see a notable difference between the ResNet and Vision Transformer. We do see the models trained with RandAugment perform better than those trained without any augmentation.

TABLE VI
ACCURACY OF VISION TRANSFORMER AND RESNET MODELS TRAINED WITH RANDAUGMENT WHEN EVALUATED ON TEST SETS WITH VARYING MAGNITUDES OF RANDAUGMENT.

RandAugment (n,m) %	Vision Transformer	ResNet
Original	86.1%	86.7%
(3,7)	81.4%	80.3%
(3,5)	82.2%	81.5%
(3,9)	79.8%	79.4%
(1,7)	84.8%	84.5%
(5,7)	77.1%	74.8%

Table VI shows the results of RandAugment generalization testing. Similarly to the previous augmentation tests, when we turn up the magnitude to 5 sequential augmentations (n) and an augmentation strength of 7 (m), performance degrades from the original and weakly augmented test sets. Another interesting observation from Table VI is the difference in performance from the original and (3,7) test set. These models were each trained with the (3,7) RandAugment parameterization, yet they generalize better to the unaugmented test set. This is somewhat surprising since the (3,7) configuration produces images fairly far from the original CIFAR-10 data distribution, as illustrated in Figure 2.

Table VII illustrates the degradation of the Vision Transformer and ResNet models with increasingly small crop percentages when trained with 90% cropping. Unlike the RandAugment testing, we do see a slightly better performance at the 90% crop test distribution for the Vision Transformer, but not the ResNet. Similar to the previous tests, training with some level of cropping does not help with generalization to the test sets. Table VIII further illustrates this by comparing the Vision Transformer trained with cropping to RandAugment and no augmentation. Training with augmentation does significantly improve over training without augmentation, but the degradation remains steep with increasing magnitudes of cropping.

TABLE VII
ACCURACY OF VISION TRANSFORMER AND RESNET MODELS WHEN TRAINED WITH 90% CROPPING AND THEN EVALUATED ON TEST SETS CONSTRUCTED WITH VARYING LEVELS OF CROPPING.

Crop %	Vision Transformer	ResNet
100%	79.4%	80.1%
95%	80.1%	78.8%
90%	80.1%	78.1%
75%	72.2%	67.8%
50%	40.9%	36.1%
25%	23.7%	20.3%

TABLE VIII
ACCURACY OF VISION TRANSFORMER MODELS TRAINED WITH DIFFERENT AUGMENTATIONS AND EVALUATED ON TEST SETS CONSTRUCTED FROM CROPPING.

Crop %	Trained with 90% Crop	RandAug	NoAug
100%	79.4%	86.1%	74.4%
95%	80.1%	86.1%	73.2%
90%	80.1%	85.3%	71.3%
75%	72.2%	75.5%	57.6%
50%	40.9%	37.8%	30.9%
25%	23.7%	19.3%	19.5%

IV. DISCUSSION

A. Global and Local Priors in Augmentation

Our tests mainly focus on cropping and adding noise because we think these are the best available augmentations to demonstrate strong local and global biases, respectively. In this section, we describe some other augmentations used in computer vision and how they may be categorized as a local or global prior. All the following augmentations are examples of global priors. We hypothesize that the success of these augmentations with convolutional networks could be due to this complement between the priors in architecture and augmentation.

- **Translational Shift:** Translational shifting describes moving the image along the x or y-axis. Similar to the added noise map, this generally preserves the global structure of the image, biasing the model away from local features. Therefore, we think translational shifting is an example of a global prior.
- **Horizontal Flip:** Horizontal flipping describes constructing a mirror image around the horizontal axis, such that an image of a dog facing left would now face right. As a dog is horizontally flipped, distinguishing features such as the head and tail are spatially much farther away than the original image. Similar to the added noise map, this

also preserves the global structure of the image and we would classify horizontal flipping as a global prior.

- **Rotation:** Rotations change the orientation of visual features similar to horizontal flipping. At large magnitudes, this does significantly change the global structure, but it does not rely on local information such as cropping.
- **Brightness:** The brightness augmentation adds or subtracts from each pixel location to simulate brighter or darker lighting. This is similar to the gaussian noise and preserves the global structure of the image.
- **Cutout:** Cutout is a data augmentation that masks out a contiguous region of the image. This is another example of a global prior, requiring the model not to be overly reliant on local features.
- **Inverse Cutout:** Inverse cutout describes only keeping a contiguous region of the image and blurring out the rest of the image. Inverse cutout is similar to cropping except that the local crop is not up-sampled to the original image size and the non-included part of the image still somewhat remains, although masked out with blurring. This is a novel augmentation that is not commonly used for training computer vision models. However, it would be an example of a local prior augmentation for the sake of these experiments, which are hard to find.

B. Corruption Testing

Unfortunately, our early experiments did not illustrate a connection between the priors in architecture and augmentation. We think another potential direction for this could be to explore the subsets of unaugmented data where each model fails. These subsets may demonstrate local or global priors such as a distant picture of a cat compared to a portrait of the cat's face. Further, we might be able to simulate more semantic augmentation tests through the use of generative modeling. Particularly, models like DALL-E [22] enable natural language prompts to create image datasets. These artificial image datasets may help us further understand the differences between the ResNet and Vision Transformer and maybe see if the priors are illustrated then.

C. Impact of Augmentation during Training

A promising direction for this research could be to further explore how augmentation during training impacts the test set performance. Additionally, we could use consistency training losses to further reinforce the augmentation bias. These consistency losses further penalize the model for having different representations of images before and after augmentation. Large-scale experiments like this may further help understand how augmentation differently impacts ResNets versus Vision Transformers.

D. Generative Model Tests

In addition to using generative models to produce more distribution shifts, we might also be able to identify global and local priors in the outputs of generative models with convolutional and attention-based architectures. For example, the DCGAN architecture [23] was shown to have checkerboard artifacts attributed to the inductive biases in the up-sampling deconvolutional layer. Lee et al. [24] have laid the foundation for training GANs with Vision Transformer architectures. We think it could be interesting to further explore differences in the data sampled from ResNet-based GANs versus Vision Transformer-based GANs.

V. Conclusion

In conclusion, we demonstrate that the ResNet and Vision Transformer models do not generalize to test sets crafted with data augmentation. Even when trained with the RandAugment augmentation policy, the models still fail to pass these tests. We think this is a promising strategy for testing distribution shift generally due to the ease of controlling the test with augmentation magnitude parameters. Our study aimed to illustrate if the global priors in attention cause the model to generalize better to global augmentation tests such as adding noise, or if the local priors in convolution help generalization to local augmentation tests such as cropping. Although the experiments of training with cropping show some performance difference between Vision Transformers and ResNets, we do not gather enough evidence to isolate the contribution of the priors in architecture and augmentation. We leave it to future work to see if the results on crop tests hold for higher resolution images and larger models trained with big data. We think passing these generalization tests should be a good signal for designing inductive biases in future neural architectures. We hope that future work with tools such as a consistency loss to increase the impact of augmentation during training and a more exhaustive search through augmentations will make the relationship between priors clearer. We have explored the relationship of inductive bias in architecture and augmentation. We leave it to future work to explore how inductive bias is related between other areas of Deep Learning performance such as sensitivity to class imbalance [37]–[40], catastrophic forgetting [46], out-of-distribution generalization [45], or other data domains [49]. Further, we can explore components of neural network training other than data augmentation, such as activation functions [33]–[36], normalization layers [44], ensemble design [41]–[43], or loss functions [47]. Understanding the different areas to inject prior knowledge into Deep Learning systems, and how they interact with each other, is a promising direction to achieve the grand visions of Deep Learning [48], such as aid in COVID-19 [32], climate change [31], and scientific discovery [30].

References

[1] K. Morrison, B. Gilby, C. Lipchak, A. Mattioli, and A. Kovashka, "Exploring Corruption Robustness: Inductive Biases in Vision Transformers and MLP-Mixers". CoRR, abs/2106.13122, 2021. URL http://arxiv.org/pdf/2106.13122.pdf.

[2] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-Trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness". In International Conference on Learning Representations, 2019.

[3] R. Shao, Z. Shi, J. Yi, P. Chen, and C. Hsieh, "On the Adversarial Robustness of Visual Transformers". In International Conference on Learning Representations, 2019.

[4] S. Paul and P. Chen, "Vision Transformers are Robust Learners". CoRR, abs/2105.07581, 2021. URL: http://arxiv.org/abs/2105.07581.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale". In International Conference on Learning Representations, 2021.

[6] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling Vision Transformers". CoRR, abs/2106.04560, 2021. URL http://arxiv.org/2106.04560.

[7] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers". CoRR, abs/2106.10270, 2021. URL http://arxiv.org/2106.10270.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition". In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples". CoRR, abs/1412.6572, 2015. URL http://arxiv.org/1412.6572.

[10] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In International Conference on Machine Learning, 2021.

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning". In Journal of Big Data, 2019.

[12] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning". In Journal of Big Data, 2021.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", In Neural Information Processing Systems, 2012.

[14] E. Cubuk, B. Zoph, J. Shlens, and Q. Le, "RandAugment: Practical automated data augmentation with a reduced search space". In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[15] E. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. Le, "AutoAugment: Learning Augmentation Strategies from Data". In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[16] A. Srinivas, T. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck Transformers for Visual Recognition". In IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[17] Z. Dai, H. Liu, Q. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes". CoRR abs/2106.04803, 2021. URL http://arxiv.org/2106.04803.

[18] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. "Transformer in Transformer". CoRR abs/2103.00112, 2021. URL http://arxiv.org/2103.00112.

[19] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating Long Sequences with Sparse Transformers". CoRR abs/1904.10509, 2021. URL http://arxiv.org/1904.10509.

[20] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey". CoRR abs/2009.06732, 2021. URL http://arxiv.org/2009.06732.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks". In IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[22] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation". In Proceedings of Machine Learning Research, 2021.

[23] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". CoRR abs/1511.06434, 2021. URL http://arxiv.org/1511.06434.

[24] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "ViTGAN: Training GANs with Vision Transformers". CoRR abs/2107.05489, 2021. URL http://arxiv.org/2107.05489.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Haung, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge". In IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[26] R. K. L. Kennedy and T. M. Khoshgoftaar. "Accelerated Deep Learning on HPCC Systems." In IEEE International Conference on Machine Learning Applications, 2020.

[27] R. Takahashi, T. Matsubara, and K. Uehara, "Data Augmentation using Random Image Cropping and Patching for Deep CNNs". In IEEE Transactions on Circuits and Systems for Video Technology, 2020.

[28] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In International Conference on Learning Representations, 2019.

[29] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural Adversarial Examples". In IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[30] M. Raghu and E. Schmidt, "A Survey of Deep Learning for Scientific Discovery". CoRR abs/2003.11755, 2021. URL http://arxiv.org/2003.11755.

[31] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, Y. Bengio, "Tackling Climate Change with Machine Learning". CoRR abs/1906.05433, 2019. URL https://arxiv.org/1906.05433.

[32] C. Shorten, T. M. Khoshgoftaar, B. Furht. "Deep Learning applications for COVID-19". In Journal of Big Data, 2021.

[33] G. Castaneda, P. Morris, T. M. Khoshgoftaar. "Evaluation of maxout activations in deep learning across several big data domains." In Journal of Big Data, 2019.

[34] G. Castaneda, P. Morris, J. D. Prusa, T. M. Khoshgoftaar. "Investigation of Maxout Activations on Convolutional Neural Networks for Big Data Text Sentiment Analysis". In The Thirty-Second International Florida Artificial Intelligence Research Society Conference, 2019.

[35] G. Castaneda, P. Morris, T. M. Khoshgoftaar, "Maxout neural network for big data medical fraud detection". In IEEE Fifth International Conference on Big Data Computing Service and Applications, 2019.

[36] G. Castaneda, P. Morris, T. M. Khoshgoftaar, "Maxout Networks for Visual Recognition". In International Journal of Multimedia Data Engineering and Management, 2019.

[37] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, N. Seliya. "A survey on addressing high-class imbalance in big data". In Journal of Big Data, 2018.

[38] J. M. Johnson and T. M. Khoshgoftaar. "Survey on deep learning with class imbalance". In Journal of Big Data, 2019.

[39] J. M. Johnson and T. M. Khoshgoftaar. "Thresholding strategies for deep learning with highly imbalanced big data." In Deep Learning Applications, Volume 2, 2021.

[40] J. M. Johnson and T. M. Khoshgoftaar. "The effects of data sampling with deep learning and highly imbalanced big data." In Information Systems Frontiers, 2020.

[41] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, "Using ensemble learners to improve classifier performance on tweet sentiment data." in IEEE International Conference on Information Reuse and Integration, 2015.

[42] J. D. Prusa, T. M. Khoshgoftaar, N. Seliya. "Enhancing ensemble learners with data sampling on high-dimensional imbalanced tweet sentiment data." In The Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, 2016.

[43] B. Heredia, T. M. Khoshgoftaar, J. Prusa, M. Crawford. "An investigation of ensemble techniques for detection of spam reviews." In International Conference on Machine Learning and Applications, 2016.

[44] H. Liu, A. Brock, K. Simonyan, Q. V. Le, "Evolving Normalization-Activation Layers." CoRR abs/2004.02967, 2020. URL https://arxiv.org/2004.02967.

[45] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, P. Liang, "WILDS: A Benchmark of in-the-Wild Distribution Shifts." CoRR abs/2012.07421, 2021. URL https://arxiv.org/2012.07421.

[46] S. Beaulieu, L. Frati, T. Miconi, J. Lehman, K. O. Stanley, J. Clune, N. Cheney. "Learning to Continually Learn." CoRR abs/2002.09571, 2020. URL https://arxiv.org/2002.09571.

[47] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan. "Supervised Contrastive Learning." In Neural Information Processing Systems, 2020.

[48] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, "Deep learning applications and challenges in big data analytics." In Journal of Big Data 2015.

[49] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks." In Journal of Big Data, 2020.