A Fourier Perspective on Model Robustness in Computer Vision

Dong Yin [1] **Raphael Gontijo Lopes** [2] **Jonathon Shlens** [3] **Ekin D. Cubuk** [3] **Justin Gilmer** [3]

Abstract

Achieving robustness to distributional shift is a longstanding and challenging goal of computer vision. Data augmentation is a commonly used approach for improving robustness, however robustness gains are typically not uniform across corruption types; for example, increasing performance in the presence of random noise is often met with reduced performance on other corruptions such as contrast change. Understanding when and why these sorts of trade-offs occur is a crucial step towards mitigating them, and towards this end, we investigate recently observed trade-offs caused by Gaussian data augmentation and adversarial training. We find that both methods improve robustness to corruptions that are concentrated in the high frequency domain while reducing robustness to corruptions that are concentrated in the low frequency domain. This suggests that one way to mitigate these trade-offs via data augmentation is to use a more diverse set of augmentations. Towards this end we observe that AutoAugment (Cubuk et al., 2018), a recently proposed data augmentation policy optimized for clean accuracy, achieves state-of-the-art robustness on the CIFAR-10-C and ImageNet-C benchmarks.

1. Introduction

Although many deep learning computer vision models have been shown to achieve remarkable performance on many standard i.i.d benchmarks, these models lack the robustness of the human vision system when the train and test distributions differ. For example, it has been observed that commonly occuring image corruptions, such as random noise, contrast change, and blurring, can lead to significant performance degradation (Dodge & Karam, 2017; Azulay & Weiss, 2018). Improving distributional robustness is an important step towards safely deploying models in complex, real-world settings.

Data augmentation is a natural and sometimes effective approach to learning robust models. Examples of data augmentation include adversarial training (Goodfellow et al., 2014), applying image transformations to the training data, such as flipping, cropping, adding random noise, and even stylized image transformation (Geirhos et al., 2018a).

However, data augmentation rarely improves robustness across all corruption types. Performance gains on some corruptions may be met with dramatic reduction on others. As an example, in (Ford et al., 2019) it was observed that Gaussian data augmentation and adversarial training improve robustness to noise and blurring corruptions on the common corruption benchmark (Hendrycks & Dietterich, 2019), while significantly degrading performance on the fog and contrast corruptions. This begs a natural question

What is different about the corruptions for which augmentation strategies improve performance vs. those which those which performance is degraded?

Understanding these tensions and why they occur is an important first step towards designing robust models. Our operating hypothesis is that the frequency information of these different corruptions offers an explanation of many of these observed trade-offs. Through extensive experiments involving perturbations in the Fourier domain, we demonstrate that these two augmentation procedures bias the model towards utilizing low frequency information in the input. This low frequency bias results in improved robustness to corruptions which are more high frequency in nature while degrading performance on corruptions which are low frequency.

Our analysis suggests that more diverse data augmentation procedures could be leveraged to mitigate these observed trade-offs, and indeed this appears to be true. In particular we demonstrate that the recently proposed AutoAugment data augmentation policy (Cubuk et al., 2018) achieves state-of-the-art results on the CIFAR-10-C and ImageNet-C benchmarks, with improvements on 14 out of 15 of the corruptions.

Some of our observations could be of interest to research

[1]Department of EECS, UC Berkeley, Work done while internship at Google Research, Brain Team. [2]Google Research, Brain Team, Work done as a member of the Google AI Residency program `g.co/airesidency`. [3]Google Research, Brain Team. Correspondence to: Dong Yin <dongyin@berkeley.edu>, Justin Gilmer <gilmer@google.com>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning.

on security. For example, we observe perturbations in the Fourier domain which when applied to images cause model error rates on ImageNet to exceed 95% while preserving the semantics of the image. These qualify as simple, few query black box attacks that satisfy the content preserving threat model (Gilmer et al., 2018).

Finally, we extend our frequency analysis to obtain a better understanding of worst-case perturbations of the input. In particular adversarial perturbations of a naturally trained model are more high-frequency in nature while adversarial training encourages these perturbations to become more concentrated in the low frequency domain.

2. Preliminaries

We denote the ℓ_2 norm of vectors (and in general, tensors) by $\|\cdot\|$. For a vector $x \in \mathbb{R}^d$, we denote its entries by $x[i]$, $i \in \{0, \ldots, d-1\}$, and for a matrix $X \in \mathbb{R}^{d_1 \times d_2}$, we denote its entries by $X[i,j]$, $i \in \{0, \ldots, d_1 - 1\}$, $j \in \{0, \ldots, d_2 - 1\}$. We omit the dimension of image channels, and denote them by matrices $X \in \mathbb{R}^{d_1 \times d_2}$. We denote by $\mathcal{F} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{C}^{d_1 \times d_2}$ the 2D discrete Fourier transform (DFT) and by $\mathcal{F}^{-1}$ the inverse DFT. When we visualize the Fourier spectrum, we always shift the low frequency components to the center of the spectrum.

We define high pass filtering with bandwidth B as the operation that sets all the frequency components outside of a centered square with width B in the Fourier spectrum to zero, and then applies inverse DFT. The low pass filtering operation is defined similarly with the difference that the centered square is applied to the Fourier spectrum with low frequency shifted to the center.

We assume that the pixels take values in range $[0, 1]$. In all of our experiments with data augmentation we always clip the pixel values to $[0, 1]$. We define Gaussian data augmentation with parameter σ as the following operation: In each iteration, we add i.i.d. Gaussian noise $\mathcal{N}(0, \widetilde{\sigma}^2)$ to every pixel in all the images in the training batch, where $\widetilde{\sigma}$ is chosen uniformly at random from $[0, \sigma]$. For our experiments on CIFAR-10, we use the Wide ResNet-28-10 architecture (Zagoruyko & Komodakis, 2016), and for our experiment on ImageNet, we use the ResNet-50 architecture (He et al., 2016). When we use Gaussin data augmentation, we choose parameter $\sigma = 0.1$ for CIFAR-10 and $\sigma = 0.4$ for ImageNet. All experiments use flip and crop during training.

Fourier heat map We will investigate the sensitivity of models to high and low frequency corruptions via a perturbation analysis in the Fourier domain. Let $U_{i,j} \in \mathbb{R}^{d_1 \times d_2}$ be a real-valued matrix such that $\|U_{i,j}\| = 1$, and $\mathcal{F}(U_{i,j})$ only has up to two non-zero elements located at (i, j) and the its symmetric coordinate with respect to the image center; we call these matrices the 2D *Fourier basis* matrices (Bracewell & Bracewell, 1986).

Given a model and a validation image X, we can generate a perturbed image with Fourier basis noise. More specifically, we can compute $\widetilde{X}_{i,j} = X + rvU_{i,j}$, where r is chosen uniformly at random from $\{-1, 1\}$, and $v > 0$ is the norm of the perturbation. For multi-channel images, we perturb every channel independently. We can then evaluate the models under Fourier basis noise and visualize how the test error changes as a function of (i, j), and we call these results the Fourier heat map of a model. We are also interested in understanding how the outputs of the models' intermediate layers change when we perturb the images using a specific Fourier basis, and these results are relegated to the Appendix.

3. The Robustness Problem

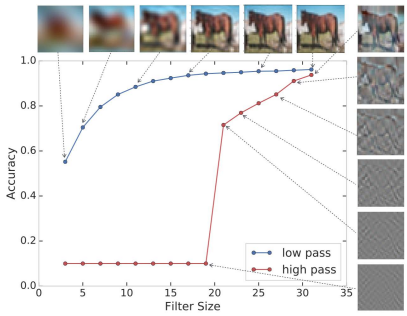


Figure 1: Models can achieve high accuracy using information about the input that would be unrecognizable to humans. Shown above are models trained and tested with aggressive high and low pass filtering applied to the inputs. In the case of high-pass filtering, models can achieve above 70% accuracy despite the fact that the resulting images are unrecognizable to humans. With aggressive low-pass filtering, the model is still above 50% when the images appear to be simple globs of color.

How is it possible that models achieve such high performance in i.i.d settings while performing so poorly in the presence of even subtle distributional shift? There has been substantial prior work towards obtaining a better understanding of the *robustness problem*. While this problem is far from being completely understood, perhaps the simplest explanation proposed is that models lack robustness to distributional shift simply because there is no reason for them to be robust (Jo & Bengio, 2017). In naturally occurring data there are trivial correlations between the input and target that models can utilize to generalize well. However, utilizing such sufficient statistics will lead to dramatic reduction in model performance should these same statistics become corrupted at test time.

In the image domain, there is a plethora of correlations between the input and target. Simple statistics such as colors, local textures, shapes, even unintuitive high frequency patterns (Ilyas et al., 2019) can all be leveraged in a way to achieve remarkable i.i.d generalization. To demonstrate, we experimented with training and testing of models when

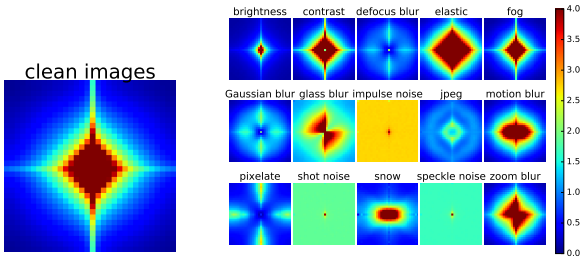


Figure 2: Left: Fourier spectrum of natural images; we estimate $\mathbb{E}[|\mathcal{F}(X)[i,j]|]$ by averaging all the CIFAR-10 validation images. Right: Fourier spectrum of the corruptions in CIFAR-10-C at severity 3. For each corruption, we estimate $\mathbb{E}[|\mathcal{F}(C(X)-X)[i,j]|]$ by averaging over all the validation images.

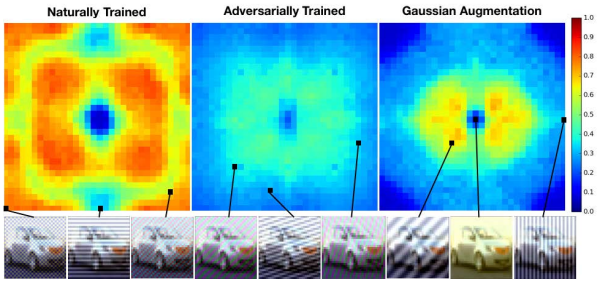


Figure 3: Model sensitivity to additive noise aligned with different Fourier basis vectors on CIFAR-10. We fix the additive noise to have ℓ_2 norm 4 and evaluate three models: a naturally trained model, an adversarially trained model, and a model trained with Gaussian data augmentation. Error rates are averaged over 1000 randomly sampled images from the test set. In the bottom row we show images perturbed with noise along the corresponding Fourier basis vector.

severe filtering is performed on the input in the frequency domain. The results are shown in Figure 1. When low-frequency filtering is applied on CIFAR-10, models can achieve over 50% test accuracy even when the image appears to be simple globs of colour. Even more striking, models achieve 70% accuracy in the presence of the severe high frequency filtering, despite the fact that the filtered image appears uninterpretable to humans.

4. Trade-off and Correlation Between Corruptions: a Fourier Perspective

The previous section demonstrated that both high and low frequency features are useful for classification. A natural hypothesis is that data augmentation may bias the model towards utilizing different kinds of features in classification. What types of features models utilize will ultimately determine the robustness at test time. Here we adopt a Fourier perspective to study the trade-off and correlation between corruptions when we apply several data augmentations.

4.1. Gaussian Data Augmentation and Adversarial Training Encourage Low Pass Filtering

Ford et al. (Ford et al., 2019) investigated the robustness of three models on CIFAR-10-C: a naturally trained model, a model trained by Gaussian data augmentation, and an adversarially trained model. It was observed that Gaussian data augmentation and adversarial training improve robustness to all noise and many of the blurring corruptions, while degrading robustness to fog and contrast. For example adversarial training degrades performance on the most severe contrast corruption from 85.66% to 55.29%. Similar results were reported on ImageNet-C.

We hypothesize that some of these trade-offs can be explained by the Fourier statistics of different corruptions. Denote a (possibly randomized) corruption function by $C : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$. In Figure 2 we visualize the Fourier statistics of natural data as well as the average delta of the common corruptions. Natural images have higher concentrations in low frequencies, thus when we refer to a "high" or "low" frequency corruption we will always use this term on a relative scale. Additive noise is uniformly distributed across the Fourier frequencies and thus has much higher frequency statistics relative to natural data. Many of the blurring corruptions remove or change the high frequency content of images. As a result $C(X) - X$ will have a higher fraction of high frequency energy. For corruptions such as contrast and fog, the energy of the corruption is concentrated more on low frequency components.

The observed differences in the Fourier statistics suggests an explanation for why the two augmentation methods improve performance in additive noise but not fog and contrast — the two augmentation methods encourage the model to become invariant to high frequency information while relying more on low frequency information. We investigate this hypothesis via several perturbation analyses of the three models in question. First, we test model sensitivity to perturbations along each Fourier basis vector. Results on CIFAR-10 are shown in Figure 3. The difference between the three models is striking. The naturally trained model is highly sensitive to additive perturbations in all but the lowest frequencies, while Gaussian data augmentation and adversarial training both dramatically improve robustness in the higher frequencies. For the models trained with data augmentation, we see a subtle but distinct lack of robustness at the lowest frequencies (relative to the naturally trained model). Figure 7 in the appendix shows similar results for three different models on ImageNet.

This is consistent with the hypothesis that the models trained with the noise augmentation effectively learn to apply a low pass filter. As a final test, we analyzed the performance of models with a low/high pass filter applied to the input (we call the low/high pass filters the *front end* of the model). Consistent with prior experiments we find that applying a low pass front-end degrades performance on fog and con-

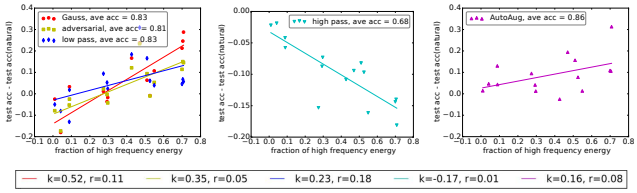


Figure 4: Relationship between test accuracy and fraction of high frequency energy on CIFAR-10-C. Each scatter point in the plot represents the evaluation result of a particular model on a particular corruption type. The x-axis represents the fraction of high frequency energy of the corruption type, and the y-axis represents change in test accuracy compared to a naturally trained model. The legend at the bottom shows the slope (k) and residual (r) of each fitted line.

trast while improving performance on additive noise and blurring. Applying a high-pass front end degrades performance on all corruptions (as well as clean test error), but performance degradation is more severe on the high frequency corruptions. These experiments again confirm our hypothesis about the robustness properties of models with a high (or low) frequency bias.

To better quantify the relationship between frequency and robustness for various models we measure the ratio of energy in the high and low frequency domain. For each corruption C, we apply high pass filtering with bandwidth 27 (denote this operation by $H(\cdot)$) on the delta of the corruption, i.e., $C(X) - X$. We use $\frac{\|H(C(X)-X)\|^2}{\|C(X)-X\|^2}$ as a metric of the fraction of high frequency energy in the corruption. For each corruption, we average this quantity over all the validation images and all 5 severities. We evaluate 6 models on CIFAR-10-C, each trained differently — natural training, Gaussian data augmentation, adversarial training, trained with a low pass filter front end (bandwidth 15), trained with a high pass filter front end (bandwidth 31), and trained with AutoAugment (see a more detailed discussion on AutoAugment in Section 4.3). Results are shown in Figure 4. Models with a low frequency bias perform better on the high frequency corruptions. The model trained with a high pass filter has a forced high frequency bias. While this model performs relatively poorly on even natural data, it is clear that high frequency corruptions degrade performance more than the low frequency corruptions. Full results, including those on ImageNet, can be found in the appendix.

4.2. Limitations of the Fourier analysis

While Figure 4 shows a clear relationship between frequency and robustness gains of several data augmentation strategies, the Fourier perspective is not predictive in all situations of transfer between data augmentation and robustness. In Appendix A, we demonstrate an experiment showing that even if we have the exact Fourier spectrum of the corruption, during training, if we add Gaussian noise that matches the Fourier spectrum of the corruption, we may not gain improvements in robustness. We hypothesize that the story is more complicated for low frequency corruptions because of an asymmetry between high and low frequency information in natural images. Given that natural images are concentrated more in low frequencies, a model can more easily learn to "ignore" high frequency information rather than low frequency information. Indeed as shown in Figure 1, model performance drops off far more rapidly when low frequency information is removed than high.

4.3. More Varied Data Augmentation Offers More General Robustness

The trade-offs between low and high frequency corruptions for Gaussian data augmentation and adversarial training lead to the natural question of how to achieve robustness to a more diverse set of corruptions. One intuitive solution is to train on a variety of data augmentation strategies. Towards this end, we investigated the learned augmentation policy AutoAugment (Cubuk et al., 2018). AutoAugment applies a learned mixture of image transformations during training and achieves the state-of-the-art performance on CIFAR-10 and ImageNet. In all of our experiments with AutoAugment, we remove the brightness and constrast sub-policies as they explicitly appear in the CIFAR-10-C and ImageNet-C benchmarks. Despite the fact that this policy was tuned specifically for clean test accuracy, we found that it also dramatically improves robustness on CIFAR-10-C and ImageNet-C. Here, we discuss the results for ImageNet-C in Table 3 in the appendix in detail, and the full results for CIFAR-10-C can also be found in the appendix.

On ImageNet-C, we compare the robustness of the naturally trained model, AutoAugment (Auto) and the Stylized-ImageNet augmentation strategy (SIN+IN) (Geirhos et al., 2018a). We observe that among the three models, AutoAugment achieves the best average corruption test accuracy of 48%, whereas SIN+IN achieves 45%. Using the mean corruption error (mCE) metric proposed by Hendrycks & Dietterich (2019), we observe that AutoAugment achieves the best mCE of 66, and in comparison, SIN+IN achieves an mCE of 69 (see a formal definition of mCE in the appendix). Interestingly, Figure 7 shows that Autoaugment improves robustness at the low to mid frequencies while sacrificing robustness at some of the higher frequencies.

We provide additional experimental results and observations in the appendix. In particular, in Appendix B, we demonstrate that adversarial examples are not necessarily high frequency phenomenon, and in Appendix E, we make conclusions and discuss future work.

Acknowledgements

We would like to thank Nicolas Ford for helpful discussions.

References

Aydemir, A. E., Temizel, A., and Temizel, T. T. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418*, 2018.

Azulay, A. and Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Bracewell, R. N. and Bracewell, R. N. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Li, S., Chen, L., Kounavis, M. E., and Chau, D. H. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. *arXiv preprint arXiv:1802.06816*, 2018.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7. IEEE, 2017.

Ford, N., Gilmer, J., Carlini, N., and Cubuk, D. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

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. *arXiv preprint arXiv:1811.12231*, 2018a.

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

Gilmer, J., Adams, R. P., Goodfellow, I., Andersen, D., and Dahl, G. E. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Jo, J. and Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

Liu, Z., Liu, Q., Liu, T., Wang, Y., and Wen, W. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. *arXiv preprint arXiv:1803.05787*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *stat*, 1050:11, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Appendix

A. Additional Experiments on Fourier Analysis

A.1. Perturbation with Different Frequency Bandwidths

To test this further, we added noise with fixed ℓ_2 norm but different frequency bandwidths centered at the origin. We consider two settings, one where the origin is centered at the lowest frequency and one where the origin is centered at the highest frequency. As shown in Figure 5, for a low frequency centered bandwidth of size 3, the naturally trained model has less than half the error rate of the other two models. For high frequency bandwidth, the models trained with data augmentation dramatically outperform the naturally trained model.

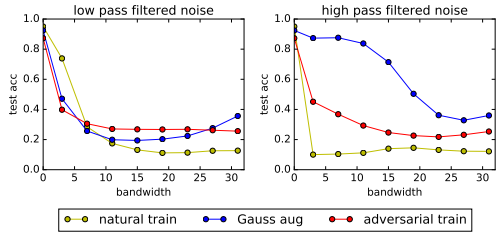


Figure 5: Robustness of models under additive noise with fixed norm and different frequency distribution. For each channel in each CIFAR-10 test image, we sample i.i.d Gaussian noise, apply a low/high pass filter, and normalize the filtered noise to have ℓ_2 norm 8, before applying to the image. We vary the bandwidth of the low/high pass filter and generate the two plots.

A.2. Limitation of Fourier Analysis

We experimented with applying additive noise that matches the statistics of the fog corruption in the frequency domain. We define "fog noise" to be the additive noise distribution $\sum_{i,j} \mathcal{N}(0, \sigma_{i,j}^2) U_{i,j}$ where the $\sigma_{i,j}$ are chosen to match the typical norm of the fog corruption on basis vector $U_{i,j}$ as shown in Figure 2. In particular, the marginal statistics of fog noise are identical to the fog corruption in the Fourier domain. However, data augmentation on fog noise *degrades* performance on the fog corruption (Table 1). This occurs despite the fact that the resulting model yields improved robustness to perturbations along the low frequency vectors (see the Fourier heat maps in the Appendix).

B. Adversarial Examples Are Not Strictly a High Frequency Phenomenon

Adversarial perturbations remain a popular topic of study in the machine learning community. A common hypothesis is that adversarial perturbations lie primarily in the high frequency domain. In fact, several (unsuccessful) defenses have been proposed motivated specifically by this hypothesis. Under the assumption that compression removes high frequency information, JPEG compression has been proposed several times (Liu et al., 2018; Aydemir et al., 2018; Das et al., 2018) as a method for improving robustness to small perturbations. To obtain a better understanding of adversarial perturbations, we performed a similar frequency analysis of them. In particular, for several models we construct adversarial perturbations for every image in the test set. We then analyze the delta between the clean and perturbed images and project these deltas into the Fourier domain. By aggregating across the successful attack images, we obtain an understanding of the frequency properties of adversarial perturbations. The results are shown in Figure 6.

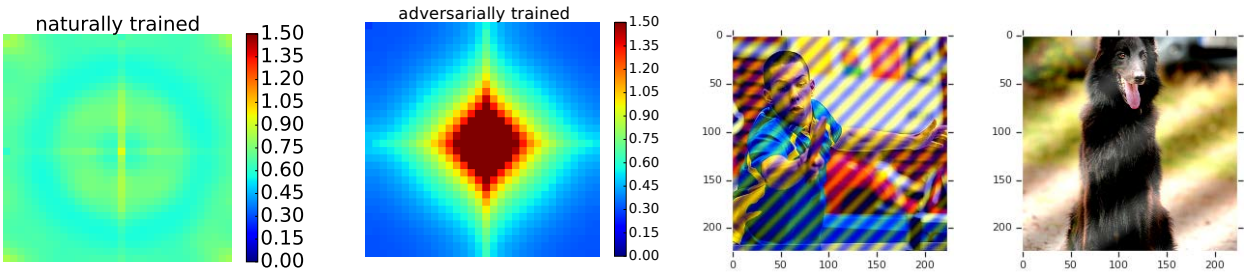
For the naturally trained model, the measured adversarial perturbations do indeed show higher concentrations in the high frequency domain (relative to the statistics of natural images). However, for the adversarially trained model this is no longer the case. The deltas for the adversarially trained model resemble that of natural data. Our analysis provides some additional understanding on a number of observations in prior works on adversarial examples. First, while adversarial perturbations for the naturally trained model do indeed show higher concentrations in the high frequency domain, this does not mean that removing high frequency information from the input results in a robust model. Indeed as shown in Figure 3, the naturally trained model is not worst-case or even average-case robust on any frequency (except perhaps the extreme low frequencies). Thus, we should expect that if we adversarially searched for errors in the low frequency domain, we will find them easily. This explains why JPEG compression, or any other method based on specifically removing high frequency content, should not be expected to be robust to worst-case perturbations.

Second, the fact that adversarial perturbations are more concentrated at lower frequencies suggests an intriguing connection between adversarial training and the DeepViz (Olah et al., 2017) method for feature visualization. In particular, optimizing the input in the low frequency domain is one of the strategies utilized by DeepViz to bias the optimization in the image space towards semantically meaningful directions. Perhaps the reason adversarially trained models have semantically meaningful gradients (Tsipras et al., 2018) is because gradients are biased towards low frequencies in a similar manner as utilized in DeepViz.

As a final note, we observe that adding certain Fourier basis vectors with large norm (40 for ImageNet) degrades test accuracy to less than 5% while preserving the semantics of the image. Two examples of the perturbed images are shown in Figure 6. If additional model queries are allowed, subtler perturbations will suffice — the perturbations used in

fog severity	1	2	3	4	5
naturally trained	0.9606	0.9484	0.9395	0.9072	0.7429
fog noise augmentation	0.9090	0.8726	0.8120	0.7175	0.4626

Table 1: Training with fog noise hurts performance on fog corruption.



(a) (b) (c) (d)

Figure 6: (a) and (b): Fourier spectrum of adversarial perturbations. For any image X, we run the PGD attack (Madry et al., 2017) to generate an adversarial example $C(X)$. We estimate the Fourier spectrum of the adversarial perturbation, i.e., $\mathbb{E}[|\mathcal{F}(C(X) - X)[i, j]|]$, where the expectation is taken over the perturbed images which lead the model to wrong predictions. (a) naturally trained; (b) adversarially trained. The adversarial perturbations for the naturally trained model are uniformly distributed across frequency components. In comparison, adversarial training biases these perturbations towards the lower frequencies. (c) and (d): Adding Fourier basis vectors with large norm to images is a simple method for generating unrestricted black box adversarial examples.

Figure 7 can drop accuracies to less than 30%. Thus, these Fourier basis corruptions can be considered as unrestricted black box attacks, and could be of interest to research on security. The Fourier heat maps with large perturbation (ℓ_2 norm 40) are included in the appendix.

C. Comparison of Model Robustness on All the Corruptions in CIFAR-10-C and ImageNet-C

We first define *mCE*, a quantity that we use to measure the robustness improvement of the models compared to a baseline model. Consider a total of K corruptions, each with S severities. Let f be a model, and $E_{k,s}(f)$ be the model's test error under the k-th corruption in the benchmark with severity s, $k = 1, \ldots, K$, $s = 1, \ldots, S$. Let f_0 be the baseline model. We define mCE as the following quantity:

$$\text{mCE} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{s=1}^{S} E_{k,s}(f)}{\sum_{s=1}^{S} E_{k,s}(f_0)}.$$

For our results on ImageNet-C in the main paper, we use the same baseline model as in (Hendrycks & Dietterich, 2019), i.e., the naturally trained AlexNet model. For our CIFAR-10-C results in Table 2, we use the naturally trained WideResNet model as the baseline model. We present the full test accuracy results on CIFAR-10-C and ImageNet-C in Tables 2 and 4, respectively.

D. Fourier Heat Maps

In this section, we provide the Fourier heat maps for the intermediate layers of the model. We first define the Fourier heat map of the output of a layer. Recall that the H-layer feedforward neural network is a function that maps X to a vector $z \in \mathbb{R}^K$, known as the logits. Let W_h be the weights and ρ_h be the possibly nonlinear activation in the h-th layer. We let

$$z_h(X) = \rho_h(\cdots \rho_2(\rho_1(X, W_1), W_2) \cdots, W_h) \in \mathbb{R}^{p_h}$$

be the output of the h-th layer and thus the logits $z(X) = z_H(X)$. The model makes prediction by choosing $y = \arg\max_k z(X)[k]$. Recall that for a validation image X, we can generate a perturbed image with Fourier basis noise, i.e., $\widetilde{X}_{i,j} = X + rvU_{i,j}$. We then compute layers' outputs $z_h(X)$ and $z_h(\widetilde{X}_{i,j})$, given the clean and perturbed images, respectively, and obtain $\|z_h(X) - z_h(\widetilde{X}_{i,j})\|$ as the model's output change at the h-th layer. We conduct this procedure for n validation images $X^{(1)}, \ldots, X^{(n)}$, compute the average output change, and use this average as a measure of the model's stability to the Fourier basis noise. More specifically, we generate the Fourier heat map of the h-th layer, denoted by $Z_h \in \mathbb{R}^{d_1 \times d_2}$, as a matrix with entries $Z_h[i, j] = \frac{1}{n} \sum_{\ell=1}^{n} \|z_h(X^{(\ell)}) - z_h(\widetilde{X}_{i,j}^{(\ell)})\|$.

In Figure 8, for 5 different models, we demonstrate the Fourier heat maps for the outputs of 5 layer outputs in the WideResNet architecture: the output of the initial convolutional layer, the outputs of the first, second, and third residual block, and the logits, and we also provide the test

	natural	Gauss	adversarial	low pass	high pass	AutoAugment	all-but-one
clean images	0.9626	0.9369	0.8725	0.9235	0.9378	**0.9693**	0.9546
brightness	0.9493	0.9244	0.8705	0.8996	0.9275	**0.9635**	0.9407
contrast	0.8225	0.5703	0.7700	0.6917	0.7806	**0.9526**	0.9015
defocus blur	0.8456	0.8371	0.8355	0.9063	0.7489	**0.9229**	0.9495
elastic transform	0.8600	0.8429	0.8175	**0.8838**	0.7870	0.8726	0.9221
fog	0.8997	0.7194	0.7263	0.8191	0.8811	**0.9463**	0.9061
Gaussian blur	0.7273	0.7907	0.8213	**0.8929**	0.6453	0.8840	0.9448
glass blur	0.5677	0.8046	0.8017	**0.8770**	0.4735	0.7621	0.8503
impulse noise	0.5428	0.8308	0.6881	0.5999	0.3619	**0.8560**	0.9016
jpeg compression	0.8009	**0.9078**	0.8541	0.8405	0.6395	0.8142	0.8807
motion blur	0.8079	0.7715	0.8045	**0.8605**	0.7206	0.8491	N/A
pixelate	0.7317	0.8983	0.8531	**0.9156**	0.6234	0.7066	0.9369
shot noise	0.6773	**0.9233**	0.8275	0.7447	0.5374	0.7834	0.9342
snow	0.8505	0.8835	0.8258	0.8688	0.7929	**0.8939**	N/A
speckle noise	0.7041	**0.9171**	0.8183	0.7502	0.5603	0.8125	0.9352
zoom blur	0.8046	0.8163	0.8279	0.8987	0.6514	**0.8994**	0.9412
average	0.7728	0.8292	0.8095	0.8299	0.6754	**0.8613**	N/A
mCE	1.000	0.9831	1.0825	0.8924	1.4449	**0.6376**	N/A

Table 2: Test accuracy on clean images and all the 15 corruptions in CIFAR-10-C. We compare 6 models: the naturally trained model, Gaussian data augmentation with parameter 0.1, adversarially trained model, low pass filter front end with bandwidth 15, high pass filter front end with bandwidth 31, and AutoAugment without brightness and contrast. Every test accuracy for the corruptions is obtained by averaging over 5 severities. The "average" row provides the average test accuracy over all the corruptions. We also present the results for the all-but-one training. More specifically, for a given corruption type and severity, we train on all the other corruptions at the same severity and evaluate on the given one. Due to some software dependency issue, we were not able to implement two of the corruptions on the training data, therefore, we only report the all-but-one results for 13 of the 15 corruptions. The test accuracy on clean images of all-but-one is averaged over all the "all-but-one" models. Since there test accuracies are not achieved by a single model, we do not compare them with other models, nor do we calculate the average corruption test accuracy and mCE.

error heat map in the last column. In Figure 9, we plot the test error Fourier heat map for two ImageNet models.

E. Conclusions and Future Work

We obtained a better understanding of trade-offs observed in recent robustness work in the image domain. By investigating common corruptions and model performance in the frequency domain we establish connections between frequency of a corruption and model performance under data augmentation. This connection is strongest for high frequency corruptions, where Gaussian data augmentation and adversarial training bias the model towards low frequency information in the input. This results in improved robustness to corruptions with higher concentrations in the high frequency domain at the cost of reduced robustness to low frequency corruptions and clean test error.

Solving the robustness problem via data augmentation alone feels quite challenging given the trade-offs we commonly observe. Naively augmenting on different corruptions often will not transfer well to held out corruptions (Geirhos et al., 2018b). However, the impressive robustness of AutoAugment gives us hope that data augmentation done properly can play a crucial role in mitigating the robustness problem.

Care must be taken though when utilizing data augmentation for robustness to not overfit to the validation set of held out corruptions. The goal is to learn *domain invariant features* rather than simply become robust to a specific set of corruptions. The fact that AutoAugment was tuned specifically for clean test error, and transfers well even after removing the contrast and brightness parts of the policy (as these corruptions appear in the benchmark) gives us hope that this is a step towards more useful domain invariant features. The robustness problem is certainly far from solved, and our Fourier analysis shows that the AutoAugment model is not strictly more robust than the baseline — there are frequencies for which robustness is degraded rather than improved. Because of this, we anticipate that robustness benchmarks will need to evolve over time as progress is made. These trade-offs are to be expected and researchers should actively search for new blindspots induced by the methods they introduce. As we grow in our understanding of

			noise			blur				weather				digital			
model	acc	mCE	Gauss	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bright	contrast	elastic	pixel	jpeg
natural	41	75	37	35	34	35	24	37	35	28	37	53	67	38	52	44	55
Auto	**48**	**66**	**50**	**50**	**49**	39	31	37	34	36	**42**	**62**	**71**	**47**	**54**	**58**	**60**
SIN+IN	45	69	41	40	37	**43**	**32**	**45**	**36**	**41**	**42**	47	67	43	49	56	58

Table 3: Comprison between naturally trained model, AutoAugment (Auto), and training with a combination of Stylized-ImageNet and original ImageNet data (SIN+IN). We use the compressed ImageNet-C images provided in (Hendrycks & Dietterich, 2019). We remove all corruptions that appear in this benchmark from the AutoAugment policy. All numbers are in percentage. The first column shows the average top1 test accuracy on all the corruptions; the second column shows the mCE; the rest of the columns show the average top1 test accuracy over the 5 severities for each corruption. We observe that AutoAugment achieves the best average test accuracy and the best mCE. In most of the noise, weather, and digital corruptions, AutoAugment achieves better performance than Stylized-ImageNet, while the latter is better on blurring corruptions.

	natural	Gauss	low pass	high pass	AutoAugment
clean images	0.7623	0.7425	0.7082	0.7500	**0.7725**
brightness	0.6975	0.6687	0.6214	0.6923	**0.7406**
contrast	0.4449	0.3578	0.3473	0.4911	**0.5656**
defocus blur	0.5023	0.5294	**0.5803**	0.4414	0.5414
elastic transform	0.5637	0.6000	**0.6211**	0.5255	0.5846
fog	0.5715	0.4736	0.4031	0.6459	**0.6534**
frosted glass blur	0.4187	0.5217	**0.6000**	0.3460	0.5073
Gaussian noise	0.4492	**0.6956**	0.4897	0.3979	0.5798
impulse noise	0.4210	**0.6785**	0.4736	0.3737	0.5832
jpeg compression	0.6630	**0.6997**	0.5688	0.6388	0.6893
pixelate	0.5826	0.6173	0.6790	0.5237	**0.6814**
shot noise	0.4294	**0.6820**	0.4894	0.3837	0.5845
zoom blur	0.3663	0.3653	**0.4177**	0.2826	0.3398
average	0.5092	0.5741	0.5243	0.4785	**0.5876**

Table 4: Test accuracy on clean images and 12 corruptions in ImageNet-C. Instead of using the compressed ImageNet-C images provided in (Hendrycks & Dietterich, 2019), the models are evaluated on the corruptions applied in memory (thus the results differ from those in Table 3). Due to some software dependency issue, we were not able to implement 3 of the 15 corruptions in memory, and thus we only the report test accuracy for 12 corruptions. We compare 5 models: the naturally trained model, Gaussian data augmentation with parameter 0.4, low pass filter front end with bandwidth 45, high pass filter front end with bandwidth 223, and AutoAugmentation. Every test accuracy for the corruptions is obtained by averaging over 5 severities.

these trade-offs we can design better benchmarks to obtain a more comprehensive perspective on model robustness.

While data augmentation is perhaps the most effective method we currently have for the robustness problem, it seems unlikely that data augmentation *alone* will provide a complete solution. Towards that end it will be important to develop orthogonal methods — e.g. architectures with better inductive biases or loss functions which when combined with data augmentation encourage extrapolation rather than interpolation.

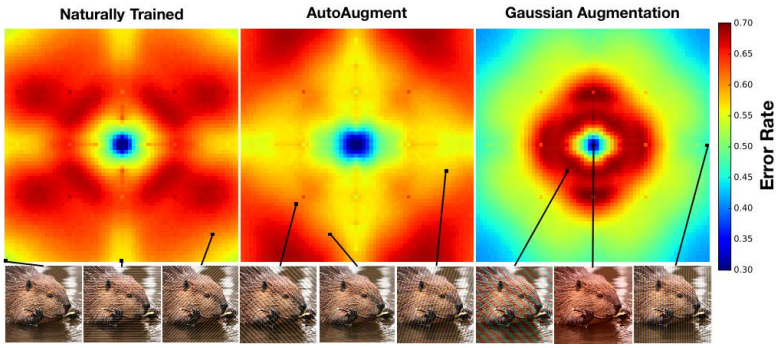


Figure 7: Model sensitivity to additive noise aligned with different Fourier basis vectors on ImageNet validation images. We fix the basis vectors to have ℓ_2 norm 15.7. Error rates are averaged over the entire ImageNet validation set. We present the 63×63 square centered at the lowest frequency in the Fourier domain. Again, the naturally trained model is highly sensitive to additive noise in all but the lowest frequencies. On the other hand, Gaussian data augmentation improves robustness in the higher frequencies while sacrificing the robustness to low frequency perturbations. For AutoAugment, we observe that its Fourier heat map has the largest blue/yellow area around the center, indicating that AutoAugment is relatively robust to low to mid frequency corruptions.

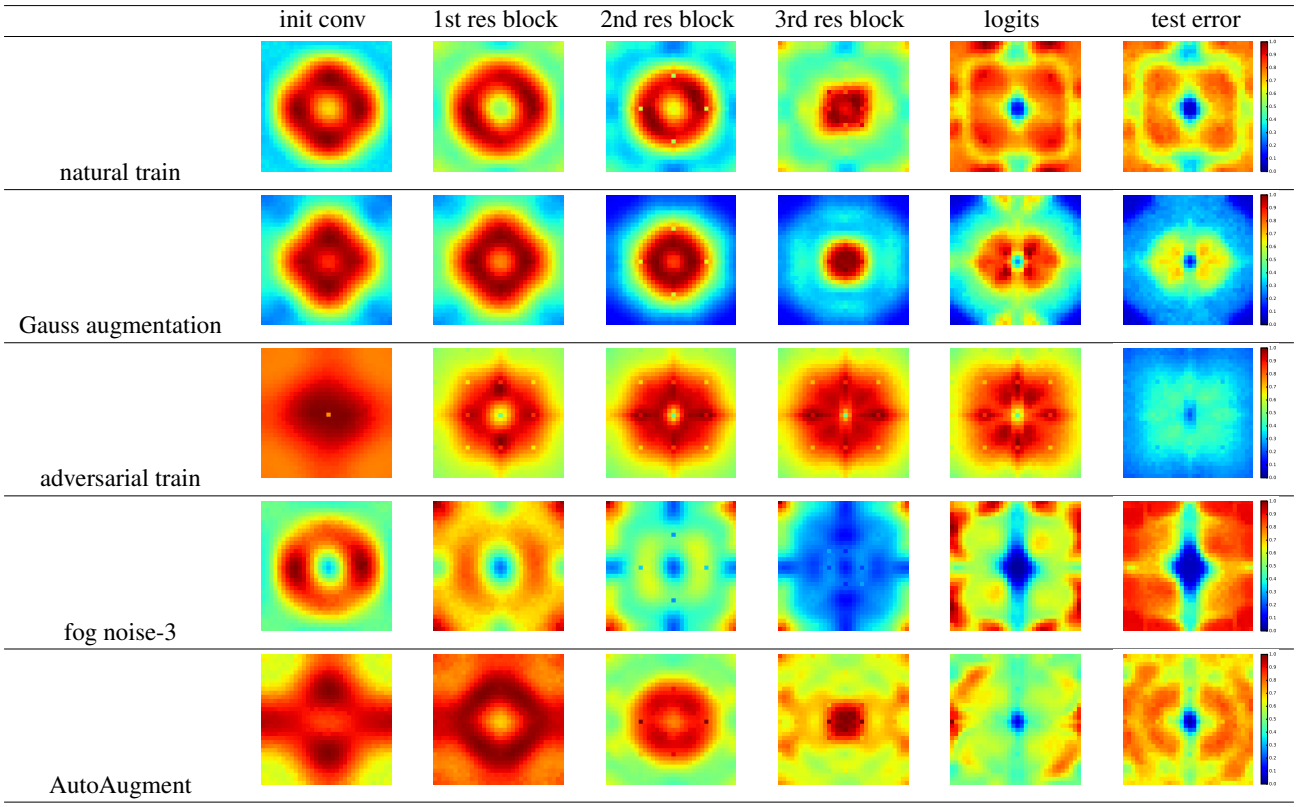


Figure 8: Model heat maps for naturally trained model, Gaussian data augmentation, adversarially trained model, data augmentation with "fog noise" at severity 3 (additive noise that matches the Fourier statistics of fog-3 corruption), and AutoAugment.

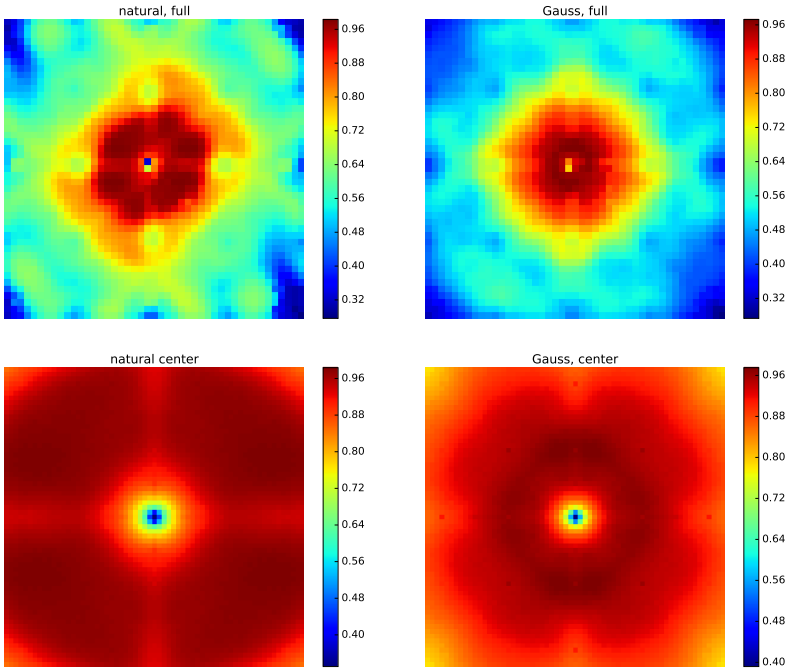


Figure 9: Fourier heat map of ImageNet models with perturbation ℓ_2 norm 40. In a large area around the center of the Fourier spectrum, the model has test error at least 95%. First row: heat map of the full Fourier spectrum (224×224); second row: heat map of the 63×63 low frequency centered square in the Fourier spectrum.