Models Out Of Line: A Fourier Lens On Distribution Shift Robustness

Anonymous Author(s)
Affiliation
Address
email

Abstract

Improving the accuracy of deep neural networks (DNNs) on out-of-distribution (OOD) data is critical to an acceptance of deep learning (DL) in real world applications. It has been observed that accuracies on in-distribution (ID) versus OOD data follow a linear trend and models that outperform this baseline are exceptionally rare (and referred to as "effectively robust"). Recently, some promising approaches have been developed to improve OOD robustness: model pruning, data augmentation, and ensembling or zero-shot evaluating large pretrained models. However, there still is no clear understanding of the conditions on OOD data and model properties that are required to observe effective robustness. We approach this issue by conducting a comprehensive empirical study of diverse approaches that are known to impact OOD robustness on a broad range of natural and synthetic distribution shifts of CIFAR-10 and ImageNet. In particular, we view the "effective robustness puzzle" through a Fourier lens and ask how spectral properties of both models and OOD data influence the corresponding effective robustness. We find this Fourier lens offers some insight into why certain robust models, particularly those from the CLIP family, achieve OOD robustness. However, our analysis also makes clear that no known metric is consistently the best explanation (or even a strong explanation) of OOD robustness. Thus, to aid future research into the OOD puzzle, we address the gap in publicly-available models with effective robustness by introducing a set of pretrained models—*RobustNets*—with varying levels of effective robustness.

1 Introduction

Deep learning (DL) holds great promise for solving difficult real-world problems in domains such as healthcare, autonomous driving, and cyber-physical systems. The major roadblock in adopting DL for real-world applications is that real-world data often deviates from the training data due to noise, corruptions, or other changes in distribution that may have temporal or spatial causes [3, 14, 29, 31]. DL models are known to be highly brittle under such distribution shifts, which raises the risk of making incorrect predictions with harmful consequences. Designing approaches to learn models that achieve both high accuracy on in-distribution data and high robustness to distribution shifts is paramount to the safe adoption of DL in real-world applications.

Although a performance drop under distribution shifts is expected, several recent works [1, 23] have noticed an intriguing phenomenon: accuracy on in-distribution (ID) versus accuracy on distribution-shifted data obeys a linear trend across a wide range of models, corruption benchmarks, and prediction tasks. Unfortunately, as nearly all current DL models bear a substantial loss under distribution shifts, this linear relationship implies that our current training strategies are insufficient to bridge the gap between ID and out-of-distribution (OOD) performance. However, despite being exceedingly rare, models that break this linear trend to obtain effective robustness [1] exist—such models defy explanation by the wisdom that OOD accuracy and ID accuracy should correlate strongly, they are

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

"out of line". Identifying when and why such models arise is key to overcoming OOD brittleness in DL models.

Related research efforts have focused on improving OOD robustness via new techniques such as model pruning [7], data augmentation [15], and ensembling large pre-trained models [25, 34]. While these efforts have created significant excitement in the ML robustness community, it remains unclear (a) under which conditions on OOD data these methods can improve effective robustness (ER), and (b) which underlying model properties make a DNN effectively robust. A more in-depth understanding of these phenomena is likely to help in bridging the gap between ID and OOD performance.

Towards achieving this goal, we carry out a comprehensive empirical study leveraging models trained using each of the aforementioned state-of-the-art approaches to improving OOD robustness. We analyze the impact of these robustness improvement methods on various model architectures, two clean datasets (CIFAR-10 and ImageNet), and a broad range of natural and synthetic distribution shifts. In particular, we view the OOD robustness puzzle through a Fourier lens and ask how spectral properties of OOD data and models influence their effective robustness. In this process, we design new metrics that capture Fourier sensitivity of models as test data moves farther away from the training data manifold. We make the following contributions:

- *A new perspective on the state of OOD robustness:* across diverse models, metrics, and distribution shifts, we observe that no one single metric rules them all, not even in-distribution accuracy. This accents the potential need for a multi-faceted approach to understanding OOD robustness, in contrast to ID generalization where single properties like model flatness can enjoy predictive success across an array of models [18].
- *Design of Fourier sensitivity metrics:* our new metrics quantify various model properties, including spectral ones that can explain the effective robustness of ensembled CLIP models [34] better than ID accuracy.
- *The first collection of publicly-available effectively robust models:* to help crack the OOD robustness puzzle, we make available (upon publication) a dataset of CIFAR-10 models with effective robustness that stems from training under a variety of data augmentation and pruning schemes.

2 Related work

Distribution shift fragility is a well-documented phenomenon among neural networks [28, 29, 14, 13, 22] in which a model trained on an ID dataset performs markedly worse when evaluated on OOD data, even when humans find the OOD data just as easy to classify. These OOD accuracy gaps often follow a linear relationship between ID and OOD accuracy, but the slope and intercept of the linear trendline varies depending on the specific ID and OOD datasets [22, 10]. In our work, we focus on two popular image classification benchmark ID datasets, CIFAR-10 [20] and ImageNet [5], and two OOD test datasets for each, CIFAR-10.1 [28], CIFAR-10-C [14], ImageNetV2 [29], and ImageNet-C [13].

While most models follow the linear ID-OOD trendlines of Miller et al. [22], some models are "effectively robust" and appear above the trendline, with OOD accuracy higher than expected given ID accuracy [1]. These robust models can arise through various methods: by pruning a model as a sort of regularization [7], by training with data augmentation intended to mimic the expected OOD test data [15], or by pretraining on a larger and more diverse dataset and evaluating zero-shot [25], partially finetuning [1], or interpolating the weights of a zero-shot and a finetuned model [34].

Although as a community we are beginning to uncover robust training methods, it remains a mystery why these particular methods achieve robustness on particular distribution shifts. We study this mystery via a Fourier lens based on prior work involving both *image frequency* and *function frequency* analyses. Ortiz-Jimenez et al. [24], Yin et al. [36], and Sun et al. [33] study model sensitivity to Fourier perturbations of the input images and analyze how different data augmentations produce different Fourier sensitivities (and correspondingly robustness to image perturbations of different frequencies). Another line of research focuses on spectral bias [9, 2, 35, 26], wherein models prioritize learning low frequency functions over the input space. This bias towards low function frequencies can also affect robustness and is influenced by training hyperparameters like data augmentation and weight decay [9].

Among the model properties we study as potential predictors of OOD accuracy are ID accuracy [22], model Jacobian norm [16], the linear pixel-space interpolation metrics of Fridovich-Keil et al. [9], and our own Fourier amplitude and phase interpolation metrics. We study these metrics over a wide

range of robust and nonrobust models, including sparse models from [7] on CIFAR-10 and pretrained models from CLIP [25, 34] and RobustBench [4] on ImageNet.

3 Methods

Datasets. We consider two standard image classification tasks: CIFAR-10 [20] and ImageNet [5]. CIFAR-10 consists of low-resolution (32 by 32 pixels) color images from 10 animal and object classes, divided into 50000 training and 10000 test images; ImageNet consists of higher-resolution (cropped to 224 by 224 pixels) color images from 1000 animal and object classes, with more than a million training images and 50000 validation images. For each of these "in-distribution" (ID) datasets we consider two "out of distribution" (OOD) datasets: one defined by a set of synthetic corruptions of the original test/validation images (CIFAR-10-C [14] and ImageNet-C [13], respectively) and another defined by a re-collection of new test images following as closely as possible the original dataset creation procedure (CIFAR-10.1 [28] and ImageNetV2 [29]).

Models. In our CIFAR-10 experiments we use 3 different model architectures: Conv8 [8], ResNet18 [12], and VGG16 [32]. For each model architecture, we consider a range of model pruning strategies that have been studied for their effect on OOD robustness [7, 17] as some of these pruned models were demonstrated to have higher OOD robustness than dense models. Namely, lottery-ticket style pruning methods [8, 30, 27, 6] were able to provide robustness gains on CIFAR-10-C [7]. We outline the various pruning techniques and sparsity levels we utilized in detail in Section 4.2.

In our ImageNet experiments we use 28 standard (nonrobust) pretrained models from Torchvision [21], 5 pretrained models from the ImageNet-C leaderboard on RobustBench [4], and 11 models obtained by robustness-enhancing weight-space interpolation [34] between zero-shot and ImageNet-finetuned CLIP ViT-B/16 [25].

Previously-proposed metrics: ID accuracy, Jacobian norm, pixel interpolation. Because OOD robustness is a complex and multi-faceted problem, we make use of multiple previously-proposed model measurements and apply them to study effectively robust models. To the best of our knowledge, our work is the first application of most of these metrics (except for in-distribution accuracy) to the study of effective robustness. Specifically, we evaluate four previously-proposed metrics that have been observed to correlate with some notion of generalization: accuracy on unseen in-distribution data ("ID accuracy") [22], Jacobian norm [16], and within-class and between-class pixel interpolation high frequency fraction [9]. These latter metrics are analogous to the high frequency fractions we compute for amplitude and phase interpolation (described below), except that interpolating paths simply interpolate in pixel space between two images, where the two images are from either the same class ("within-class") or different classes ("between-class") [9]. Prior work observed that in-distribution generalization typically improves as within-class high frequency fraction decreases and between-class high frequency fraction increases [9]. Reducing the norm of the Jacobian of model outputs with respect to input data can push the decision boundary away from the training points, providing robustness to random perturbations of the input data [16]. Thus, models with smaller Jacobian norms may perform better on corrupted/shifted data. We estimate the norm of the Jacobian using a random-projection-based approach [16]; further details can be found in the supplement.

Fourier interpolation metrics. We also introduce novel Fourier interpolation metrics to capture more information about model behavior that may help answer the robustness puzzle. Inspired by the pixel-space interpolation procedure of Fridovich-Keil et al. [9], we propose evaluating the smoothness of a model's predictions along image paths that perturb the Fourier amplitude or phase information of one test image towards another, while preserving the Fourier phase or amplitude (respectively) information of the original image. The intuition for this type of interpolation is that most of the semantic content of an image is contained within its Fourier phases, which encode structural information like edges. Accordingly, we can perturb the Fourier amplitude of an image without destroying its semantic meaning, and we can reliably destroy semantic content by perturbing Fourier phase. Example Fourier amplitude and phase interpolating paths on ImageNet are shown in Figure 1; example paths on CIFAR-10 are included in the appendix.

More precisely, we compute an interpolating path by first randomly selecting two images $\mathbf{x}_0$ and $\mathbf{x}_1$ from the ID test/validation set. We perform standard preprocessing on each image, which involves

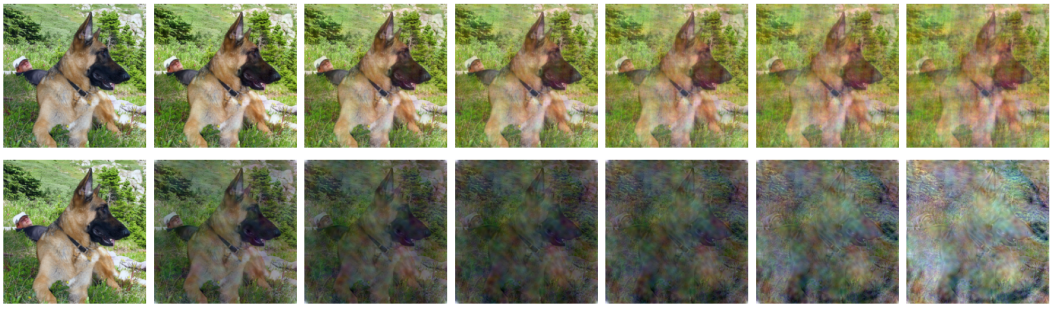

Figure 1: Example Fourier amplitude (top) and phase (bottom) interpolating paths from ImageNet. Each path includes 100 images; every 15th image is visualized here. The first image along each path is an unaltered image from the original test set; the last image has the same Fourier phases (top) or amplitudes (bottom) as the original but has some of the Fourier amplitudes (top) or phases (bottom) of a random other image from the validation set. Amplitude interpolation produces a corruption that preserves semantic content, whereas phase interpolation destroys semantic content.

cropping to a standard size (for ImageNet) and normalizing each pixel by the training mean and standard deviation (for both CIFAR-10 and ImageNet). We then compute the two-dimensional discrete Fourier transform (DFT) of each image and separate the complex-valued results into amplitude and phase components $\mathbf{a}_i$ and $\mathbf{p}_i$ for $i \in \{0, 1\}$. To construct an image along an amplitude interpolating path, we retain the original phase $\mathbf{p}_0$ and interpolate the low-frequency amplitude as $(1 - \lambda)\mathbf{a}_0 + \lambda\mathbf{a}_1$ (while preserving the high frequency amplitude), where $\lambda \in [0, 1]$, and produce an image via the inverse discrete Fourier transform. Phase interpolation is defined analogously; we retain the original amplitude and interpolate the low-frequency phase. We choose a frequency threshold for what is considered low frequency for interpolation based on visual inspection to ensure that amplitude interpolation preserves semantic meaning and phase interpolation destroys it. We interpolate the lowest 40% of image frequencies for both amplitude and phase on CIFAR-10, the lowest 20% of image frequencies for phase on ImageNet, and all image frequencies for amplitude on ImageNet. We choose 100 evenly-spaced λ values to define each path, and choose 5000 random paths for CIFAR-10 and 1000 for ImageNet. The images in Figure 1 are rescaled for visualization (essentially undoing the initial pixel-wise normalization).

We evaluate each model on each path, producing a probability vector over the available classes for each image along the path. We compute two metrics based on these path predictions: *high frequency fraction* (HFF) and *consistent distance* (CD). For high frequency fraction, we take the one-dimensional DFT of the 100 model predictions along the path, average the Fourier amplitudes among all the classes, and compute the fraction of the total amplitude above a frequency threshold. This metric is always between 0 and 1 (usually between 0.1 and 0.3). The higher the average *high frequency fraction*, the more sensitive a model is to the given Fourier perturbation of its input images.

We also consider *consistent distance*, which is defined as the index of the first image along the path that the model classifies differently (by highest softmax prediction) than the original image. This metric is always between 1 and 100, and is often at least 40. Consistent distance is intended as a more intuitive metric to capture the same notion of robustness to corruptions of image Fourier content. In particular, Fourier amplitude corruptions change image statistics but do not hamper human semantic identification; a more robust model in this sense should have a *lower* high frequency fraction and a *higher* consistent distance on average compared to a less robust model.

4 Results

We are interested in understanding *when* (under what conditions on the distribution shift) and *why* (via model properties) neural networks can exhibit effective robustness. We perform a systematic study of three tunable "knobs" that are available to the neural network practitioner and have been shown to impact OOD robustness: pruning [7], data augmentation [15], and weight ensembling [34]. For each knob, we evaluate a benchmark set of pretrained CIFAR-10 or ImageNet models on the original in-distribution test set, several OOD test sets, and a suite of model property metrics capturing

local smoothness (via model Jacobian norm), frequency response to pixel-space interpolation within and between classes [9], and Fourier amplitude and phase sensitivity (via *high frequency fraction* and *consistent distance*). Representative results from our analysis of these three knobs are presented in Sections 4.2, 4.3, and 4.4. For each dataset we consider one natural distribution shift (CIFAR-10.1 or ImageNetV2) as well as six synthetic corruptions (from CIFAR-10-C or ImageNet-C) comprised of two low-frequency corruptions, two mid-frequency corruptions, and two high-frequency corruptions (in that order). Full results on all the corruptions in CIFAR-10-C and ImageNet-C are deferred to the supplement. In all figures, we show 95% confidence intervals around each measurement (Clopper-Pearson for accuracy measurements and Gaussian bounds for averages of other metrics).

We begin by comparing the spectral properties of different distribution shifts in Section 4.1. In the following sections we find that the relationship between model properties and OOD robustness is dependent upon these spectral statistics of the specific distribution shift. This analysis highlights the difficulty of using a single model property to understand OOD performance but also indicates that our proposed metrics may play an important role in solving the OOD puzzle.

4.1 Spectral characterization of distribution shifts

To explore the behavior of models on OOD data, it is beneficial to characterize the nature of the OOD data with respect to the ID test data. Specifically, considering the difference in power spectral density (PSD) between each of the OOD test sets and the original ID test set allows us to categorize each OOD test set as being concentrated on low, mid, or high frequency perturbations with respect to the ID test set. We perform this characterization for CIFAR-10.1 and each of the 15 corruptions in CIFAR-10-C; computational details for PSD are provided in the appendix. Figure 2 shows the distribution shift PSDs for CIFAR-10.1 together with two low (brightness, contrast), mid (defocus blur, pixelate), and high (gaussian noise, impulse noise) frequency CIFAR-10-C corruptions. Of note, this PSD characterization illustrates that the OOD shift encountered on CIFAR-10.1 data is composed of low-frequency information, much like the brightness and contrast CIFAR-10-C corruptions.

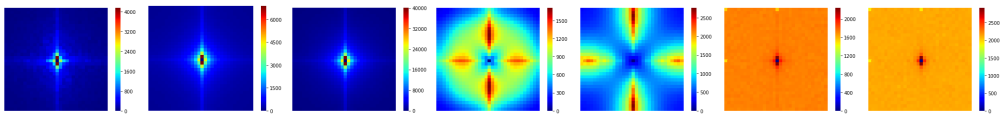

Figure 2: Power spectral densities for a selection of low (CIFAR-10.1, brightness, contrast), mid (defocus blur, pixelate), and high (gaussian noise, impulse noise) frequency shifts w.r.t. CIFAR-10.

4.2 When and why does model pruning confer effective robustness?

Diffenderfer et al. [7] demonstrated the nuanced behavior that model pruning has on OOD robustness, specifically with respect to the distribution shift from CIFAR-10 to CIFAR-10-C. Notably, pruned models are capable of providing superior OOD robustness compared to dense, or unpruned, models. Furthermore, the effect, either positive or negative, and degree of robustness arising from model pruning is dependent on both the model architecture and the pruning algorithm. It remains unknown which properties of these pruned models contribute to their robustness on OOD data.

In an effort to better understand these findings, we further investigate this behavior by considering the three categories of pruning used in [7]: *traditional* (fine-tuning [11], gradual magnitude pruning [37]), *rewinding lottery-tickets* (weight-rewinding [8], learning-rate rewinding [30]), and *initialization lottery-tickets* (edgepopup [27], biprop [6]). Note that we often abbreviate "lottery tickets" as LT. We provide details on each of the pruning techniques in the supplement. For each pruning strategy and each architecture, we prune models to 50%, 60%, 80%, 90%, and 95% sparsity. For all pruning methods, pruning is performed in an unstructured (i.e., individual weights are pruned) and global manner (i.e., prune to a given sparsity across the entire network). Additionally, for traditional and initialization LTs, pruning was performed in a layerwise fashion (i.e., where each network layer is pruned to the given sparsity level). In Figure 3 we investigate *when* (on which types of distribution shifts) and in Table 1 we study *why* (via corresponding model property measurements) these differently-pruned CIFAR-10 models achieve their varying degrees of robustness.

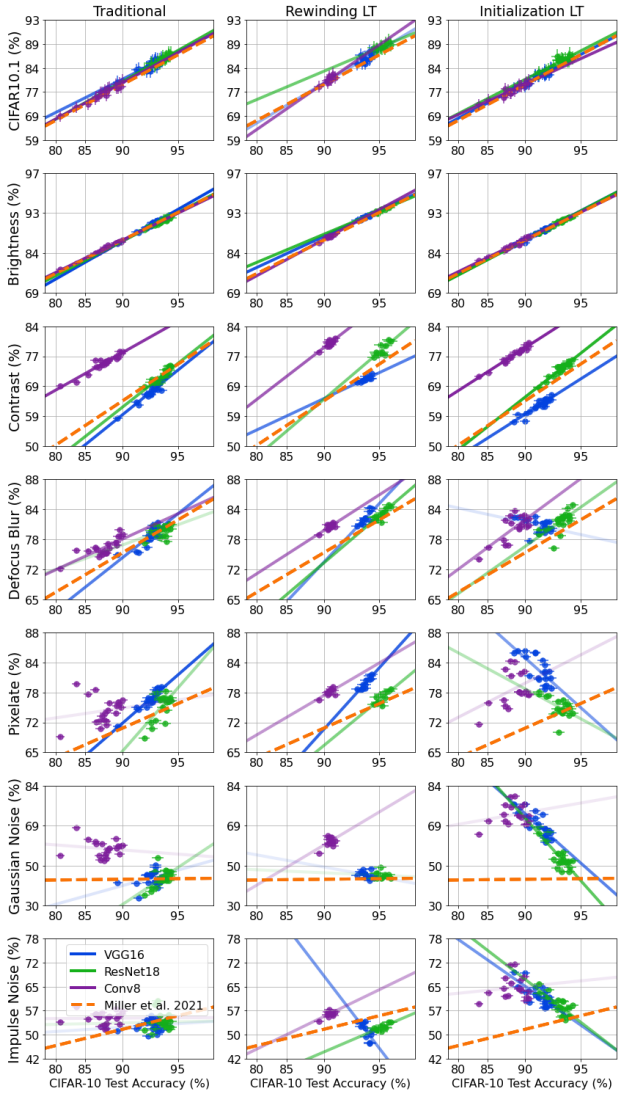

Figure 3: When does pruning affect OOD robustness?

		Traditional		Rewinding LT		Initialization LT	
		m	R^2	m	R^2	m	R^2
CIFAR-10.1	ID accuracy	0.85	**0.85**	0.9	**0.59**	0.79	**0.87**
	Jacobian	-0.0	0.16	-0.01	0.36	0.0	0.19
	Pixel - within	-6.42	**0.79**	-8.54	**0.66**	-6.7	**0.78**
	Pixel - between	8.45	0.48	-1.29	0.25	7.45	0.45
	Amp - HFF	-6.11	**0.72**	-5.97	0.31	-4.73	**0.75**
	Amp - CD	0.04	**0.6**	0.05	**0.52**	0.04	**0.75**
	Phase - HFF	-8.09	0.26	-5.66	0.24	-6.78	0.41
	Phase - CD	0.03	0.33	0.01	0.26	0.02	0.4
Brightness	ID accuracy	1.0	**0.96**	0.9	**0.89**	0.97	**0.97**
	Jacobian	-0.0	0.2	-0.01	0.49	0.0	0.29
	Pixel - within	-7.41	**0.88**	-7.47	**0.79**	-8.09	**0.85**
	Pixel - between	9.73	**0.51**	0.6	0.34	9.32	**0.57**
	Amp - HFF	-6.73	**0.7**	-6.47	**0.53**	-5.9	**0.88**
	Amp - CD	0.04	**0.63**	0.04	**0.67**	0.05	**0.84**
	Phase - HFF	-7.7	0.18	-8.02	0.31	-6.66	0.35
	Phase - CD	0.03	0.27	0.02	0.23	0.02	0.4
Contrast	ID accuracy	0.91	**0.85**	0.89	**0.66**	0.87	**0.91**
	Jacobian	-0.0	0.15	-0.01	0.4	0.0	0.25
	Pixel - within	-6.59	**0.74**	-8.13	**0.64**	-7.4	**0.81**
	Pixel - between	8.08	0.39	3.92	0.3	8.46	**0.51**
	Amp - HFF	-6.13	**0.63**	-5.9	0.4	-5.45	**0.86**
	Amp - CD	0.04	**0.65**	0.04	**0.54**	0.05	**0.83**
	Phase - HFF	-8.83	0.26	-9.17	0.3	-7.18	0.4
	Phase - CD	0.03	0.29	0.02	0.23	0.02	0.42
Defocus Blur	ID accuracy	0.55	**0.51**	0.86	**0.63**	0.42	0.39
	Jacobian	-0.0	0.17	-0.01	**0.64**	0.01	0.45
	Pixel - within	-4.54	**0.55**	-7.82	**0.61**	-4.49	0.41
	Pixel - between	6.35	0.36	8.82	0.26	7.84	**0.52**
	Amp - HFF	-4.01	0.42	-5.96	0.43	-2.79	0.45
	Amp - CD	0.03	0.37	0.04	**0.55**	0.02	0.37
	Phase - HFF	-4.78	0.16	-6.51	0.34	1.13	0.18
	Phase - CD	0.02	0.19	0.01	0.37	-0.0	0.12
Pixelate	ID accuracy	0.73	0.44	0.85	**0.61**	-0.31	0.33
	Jacobian	-0.0	0.11	-0.01	**0.75**	0.0	0.13
	Pixel - within	-6.8	0.46	-8.09	**0.6**	2.67	0.22
	Pixel - between	10.92	0.45	14.0	0.25	0.01	0.23
	Amp - HFF	-3.92	0.3	-7.48	**0.57**	2.11	0.44
	Amp - CD	0.02	0.31	0.05	**0.62**	-0.02	0.48
	Phase - HFF	-0.17	0.26	-8.55	0.31	11.15	**0.61**
	Phase - CD	0.01	0.25	0.02	0.39	-0.03	0.5
Gaussian Noise	ID accuracy	0.55	0.16	0.24	0.1	-1.16	0.47
	Jacobian	-0.0	0.05	-0.0	0.15	-0.0	0.05
	Pixel - within	-6.17	0.24	-2.89	0.14	10.65	0.38
	Pixel - between	7.48	0.11	9.49	0.14	-10.94	0.31
	Amp - HFF	-2.28	0.09	-5.56	0.11	7.09	0.47
	Amp - CD	0.01	0.12	0.03	0.17	-0.06	0.47
	Phase - HFF	-0.52	0.18	-11.92	0.25	16.98	0.49
	Phase - CD	0.0	0.13	0.03	0.18	-0.04	0.42
Impulse Noise	ID accuracy	0.04	0.0	-0.17	0.5	-0.58	0.39
	Jacobian	0.0	0.05	0.01	0.5	-0.0	0.05
	Pixel - within	-1.26	0.03	1.31	**0.54**	5.47	0.33
	Pixel - between	3.21	0.07	-17.81	0.35	-5.25	0.28
	Amp - HFF	0.13	0.01	0.48	0.36	3.43	0.38
	Amp - CD	-0.01	0.04	-0.0	0.47	-0.03	0.38
	Phase - HFF	4.05	0.1	-7.43	0.41	8.61	0.43
	Phase - CD	-0.01	0.06	0.02	0.26	-0.02	0.38

Table 1: Can model metrics explain why pruning affects OOD robustness?

Because different model architectures respond differently to different pruning methods, in each subplot of Figure 3 we color-code models by their architecture and fit a separate probit-domain OOD vs. ID accuracy regression line to each architecture, so that each regression captures a range of different pruning amounts (what fraction of weights are pruned). We display these regressions with opacity proportional to the fit quality R^2, so that regression lines that fit the data well are dark and easily seen, while lower-quality linear fits are more transparent. In the corresponding Tables 1, each reported slope (m) and R^2 value is the average of the corresponding values among the three architectures we consider. For ease of reading, we bold average R^2 values above 0.5, and among these entries we highlight the best (bright green) and second-best (faded green) R^2 values for each pruning method on each OOD dataset. Although these results are nuanced and raise additional open questions, we summarize a few key takeaways from our CIFAR-10 model pruning experiments:

- As reported in Miller et al. [22], ID accuracy is often a strong predictor of OOD accuracy; this trend holds across different pruning methods. However, the linear trend breaks down for mid and high frequency image corruptions like pixelate, Gaussian noise, and impulse noise.
- Even for low-frequency corruptions, different model architectures might lie on either the same or different robustness lines (as evidenced by comparing brightness and contrast).

- Initialization LT pruning confers robustness to high-frequency corruptions and can induce a negative correlation between ID and OOD accuracy for these corruptions.
- Jacobian norm is rarely a good predictor of OOD accuracy, but it does correlate negatively (lower norm is better) for models pruned by rewinding methods on mid-frequency corruptions.
- As reported in [9], lower high frequency fraction HFF (i.e. smoother predictions) for within-class pixel interpolation is a good predictor of OOD accuracy, at least for low-frequency distribution shifts where ID accuracy is also predictive.
- Amplitude and phase interpolation are sometimes predictive of OOD accuracy, particularly for initialization LT pruning, and in the expected direction of correlation (a better model should have lower HFF and higher CD). However, even these metrics fail to explain most of the robustness behavior for high (and some mid) frequency distribution shifts on CIFAR-10.

4.3 When and why does data augmentation confer effective robustness?

A common practice to protect against distribution shift is to train with data augmentation [15]. Accordingly, we revisit each of the CIFAR-10 models considered in Section 4.2 with three different variations of data augmentation used during training: no augmentation (clean), AugMix [15], and Gaussian noise augmentation [19]. In Figure 4 we investigate *when* (on which types of distribution shifts) and in Table 2 we study *why* (via corresponding model property measurements) these differently-augmented CIFAR-10 models achieve varying degrees of robustness. As in Figure 3 and Table 1, we fit a separate regression line to each model architecture and average the resulting slope and R^2 values. The models that contribute to each regression line therefore share an architecture and an augmentation strategy, but vary by pruning method and amount. Our findings in this experiment echo those in Section 4.2, but we make the following additional observations:

- The closer the match between the training data and the OOD evaluation data in terms of their spectral profile, the better ID accuracy, within-class pixel interpolation, and amplitude interpolation HFF and CD are at predicting OOD accuracy. This is evident from the high R^2 values for these metrics when predicting defocus blur accuracy for models trained with AugMix, and when predicting Gaussian noise or impulse noise accuracy for models trained with Gaussian noise.
- The closer the match (in terms of PSD) between the training data and the OOD data, the better the OOD accuracy. This unsurprising result is evident from the higher accuracies of AugMix-trained models on contrast and defocus blur corruptions and the higher accuracies of Gaussian-noise-trained models on Gaussian noise and impulse noise corruptions.

4.4 CLIP: Weight ensembling a robust model on ImageNet

Perhaps the most stunning example of OOD robustness for image classification is CLIP [25], which showed that a large model pretrained on a massive dataset and evaluated zero-shot on ImageNet achieves effective robustness on many OOD test sets simultaneously. Andreassen et al. [1] found that finetuning these robust models towards ImageNet improves ID accuracy on ImageNet but erodes effective robustness, and Wortsman et al. [34] offered a "best of both worlds" solution by interpolating in weight space between a pretrained zero-shot model and its finetuned counterpart, improving ID accuracy on ImageNet without sacrificing robustness. Accordingly, we study a set of 11 CLIP models obtained by interpolating the pretrained and finetuned versions of a CLIP ViT-B/16 model [25, 34] with various relative weights on the two models. We compare this CLIP model set to a benchmark set of 28 standard (nonrobust) ImageNet pretrained models from Torchvision [21] as well as 5 pretrained models from the RobustBench [4] ImageNet-C leaderboard; these models typically use data augmentation to improve corruption performance but are not nearly as effectively robust as the CLIP models. We make the following observations in Figure 5 and Table 3:

- Although in-distribution accuracy is a reasonable predictor of OOD accuracy for most models and distribution shifts, it performs markedly worse when predicting OOD accuracy for CLIP models (and a couple of effectively robust RobustBench models on some OOD shifts).
- Fourier interpolation metrics, particularly HFF and CD for amplitude interpolation, are strong predictors (often $R^2 \geq 0.9$) of OOD performance for CLIP models across OOD shifts. This observation helps answer *why* CLIP models are more robust: they are less sensitive to perturbations of Fourier amplitude, which preserve image semantics for humans but confuse nonrobust models.

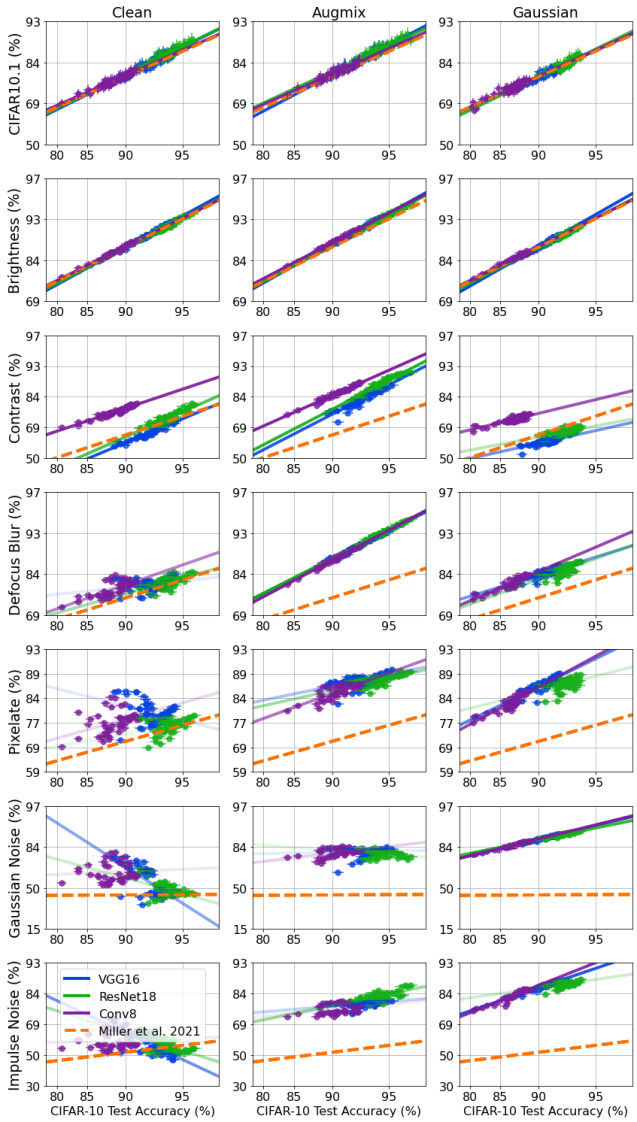

Figure 4: When does augmentation affect OOD robustness?

		Clean		AugMix		Gaussian	
		m	R^2	m	R^2	m	R^2
CIFAR-10.1	ID accuracy	0.91	**0.92**	0.91	**0.93**	0.89	**0.86**
	Jacobian	0.01	0.14	0.01	0.35	-0.0	0.05
	Pixel - within	-8.45	**0.79**	-6.88	**0.86**	-7.13	**0.73**
	Pixel - between	10.68	**0.57**	8.55	**0.67**	6.95	0.45
	Amp - HFF	-5.43	**0.79**	-5.93	**0.83**	-5.64	**0.71**
	Amp - CD	0.04	**0.74**	0.06	**0.66**	0.04	**0.68**
	Phase - HFF	-7.36	0.36	2.63	0.06	4.18	0.09
	Phase - CD	0.02	0.45	0.02	0.19	0.03	0.29
Brightness	ID accuracy	0.99	**0.98**	1.02	**0.99**	1.01	**0.97**
	Jacobian	0.0	0.12	0.01	0.37	0.0	0.06
	Pixel - within	-9.36	**0.87**	-7.7	**0.91**	-7.51	**0.74**
	Pixel - between	11.52	**0.61**	9.73	**0.73**	8.54	**0.6**
	Amp - HFF	-5.94	**0.84**	-6.68	**0.88**	-6.32	**0.78**
	Amp - CD	0.04	**0.78**	0.06	**0.7**	0.04	**0.69**
	Phase - HFF	-7.38	0.33	3.32	0.07	7.44	0.18
	Phase - CD	0.02	0.44	0.02	0.19	0.03	0.24
Contrast	ID accuracy	1.01	**0.87**	1.25	**0.93**	0.55	**0.54**
	Jacobian	0.0	0.08	0.02	0.44	-0.0	0.05
	Pixel - within	-9.88	**0.79**	-9.79	**0.9**	-5.22	**0.61**
	Pixel - between	11.55	**0.5**	12.76	**0.78**	3.12	0.16
	Amp - HFF	-6.55	**0.86**	-8.23	**0.84**	-3.64	0.44
	Amp - CD	0.05	**0.83**	0.07	**0.6**	0.02	0.47
	Phase - HFF	-9.91	0.44	8.98	0.11	0.25	0.03
	Phase - CD	0.03	**0.53**	0.02	0.12	0.02	0.22
Defocus Blur	ID accuracy	0.47	0.34	0.98	**0.98**	0.7	**0.6**
	Jacobian	-0.0	0.04	0.01	0.39	-0.0	0.12
	Pixel - within	-5.33	0.45	-7.44	**0.94**	-6.77	**0.72**
	Pixel - between	5.81	0.24	9.54	**0.76**	5.42	0.31
	Amp - HFF	-3.27	0.42	-6.39	**0.88**	-5.19	**0.64**
	Amp - CD	0.02	0.36	0.06	**0.67**	0.03	**0.6**
	Phase - HFF	-2.38	0.11	4.59	0.08	3.9	0.08
	Phase - CD	0.01	0.13	0.02	0.18	0.03	0.28
Pixelate	ID accuracy	0.11	0.13	0.42	0.45	0.68	**0.55**
	Jacobian	-0.0	0.13	0.0	0.04	-0.01	0.14
	Pixel - within	-2.12	0.14	-3.27	0.49	-6.33	**0.64**
	Pixel - between	2.03	0.16	3.7	0.29	5.31	0.3
	Amp - HFF	-0.4	0.18	-3.0	0.49	-5.1	**0.57**
	Amp - CD	-0.0	0.13	0.03	0.47	0.03	**0.54**
	Phase - HFF	5.4	0.2	2.31	0.05	5.28	0.1
	Phase - CD	-0.01	0.15	0.01	0.08	0.03	0.25
Gaussian Noise	ID accuracy	-1.11	0.23	0.08	0.11	0.88	**0.94**
	Jacobian	-0.03	0.31	-0.01	0.11	-0.0	0.08
	Pixel - within	6.83	0.09	-0.81	0.12	-7.08	**0.85**
	Pixel - between	-17.36	0.23	-0.34	0.07	7.32	**0.57**
	Amp - HFF	5.63	0.25	-1.26	0.08	-5.74	**0.79**
	Amp - CD	-0.04	0.23	0.03	0.15	0.03	**0.7**
	Phase - HFF	10.47	0.26	0.8	0.05	6.39	0.16
	Phase - CD	-0.03	0.27	0.02	0.15	0.03	0.25
Impulse Noise	ID accuracy	-0.64	0.32	0.33	0.26	0.75	**0.67**
	Jacobian	-0.01	0.27	0.0	0.13	-0.0	0.1
	Pixel - within	4.67	0.16	-2.64	0.26	-6.54	**0.71**
	Pixel - between	-9.44	0.26	2.62	0.23	5.93	0.42
	Amp - HFF	3.08	0.3	-2.23	0.27	-5.39	**0.61**
	Amp - CD	-0.02	0.31	0.03	0.25	0.03	**0.53**
	Phase - HFF	5.28	0.25	0.32	0.03	4.36	0.16
	Phase - CD	-0.01	0.27	0.02	0.2	0.03	0.2

Table 2: Can model metrics explain why augmentation affects OOD robustness?

5 Discussion

Our work is only the beginning of a true understanding of what makes models effectively robust to distribution shifts. For example, Section 4.2 showed different types of model pruning can confer robustness to different distribution shifts, and while some of this robustness can be explained by improved behavior in response to amplitude and phase perturbations, much remains an open mystery.

As we saw in Section 4.3 (and has been noted by prior work [36, 15]), models can achieve robustness to some distribution shifts by training with data augmentations designed to imitate the expected OOD data. This approach can be a powerful way to mitigate expected test-time distribution shifts but leaves open the question of how to prepare for the unexpected.

We learned in Section 4.4 that one way in which CLIP models achieve OOD robustness is by reducing their sensitivity to perturbations of Fourier amplitude, perhaps making them more attuned to human perceptions of semantic meaning. This finding provides an exciting opportunity for future work to both

confirm our results across additional robust models and incorporate these Fourier interpolation metrics into training paradigms, for instance as regularizers, to explicitly encode robustness in future models.

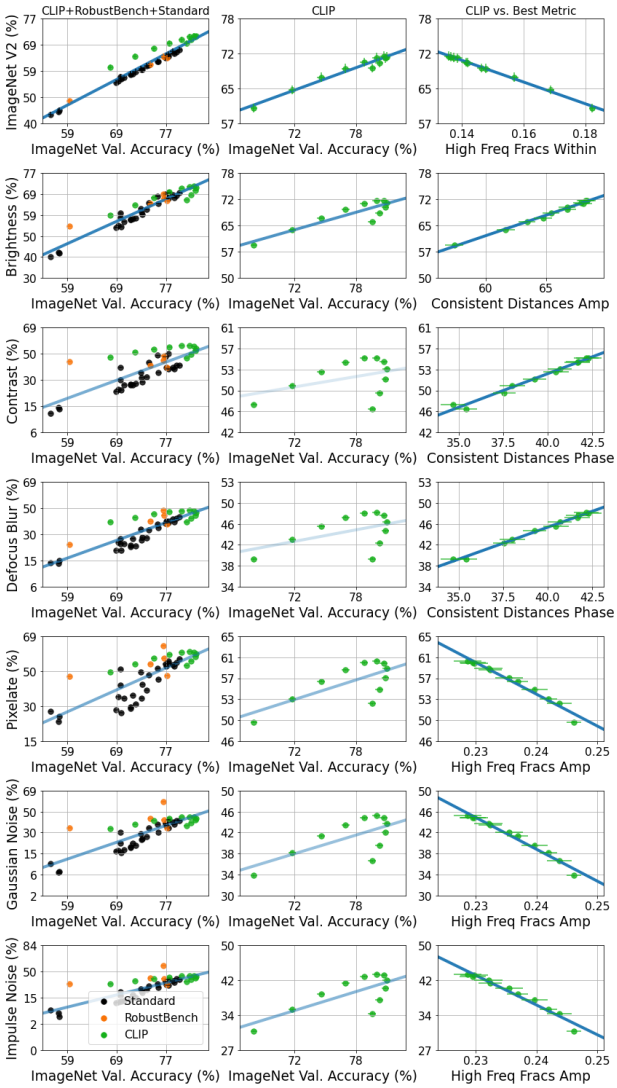

Figure 5: When and why does CLIP ensembling affect OOD robustness?

		All		CLIP		Non-CLIP	
		m	R^2	m	R^2	m	R^2
ImageNet V2	ID accuracy	0.98	**0.94**	0.64	**0.96**	0.93	**1.0**
	Jacobian	0.0	0.14	-0.0	**0.96**	0.0	0.0
	Pixel - within	-10.2	**0.88**	-6.21	**1.0**	-9.8	**0.92**
	Pixel - between	8.39	**0.74**	6.46	**0.78**	14.29	**0.73**
	Amp - HFF	-12.27	0.36	-11.56	**0.57**	-0.66	0.0
	Amp - CD	0.02	**0.77**	0.03	**0.9**	0.03	**0.67**
	Phase - HFF	-4.68	0.05	-5.05	0.08	19.65	0.42
	Phase - CD	0.02	**0.64**	0.02	0.42	0.04	**0.57**
Brightness	ID accuracy	1.06	**0.91**	0.66	**0.79**	1.12	**0.92**
	Jacobian	0.0	0.06	-0.0	**0.81**	-0.0	0.0
	Pixel - within	-10.9	**0.81**	-6.8	**0.91**	-11.4	**0.79**
	Pixel - between	8.36	**0.59**	6.03	**0.51**	18.99	**0.81**
	Amp - HFF	-11.2	0.25	-15.99	**0.83**	-8.87	0.01
	Amp - CD	0.02	**0.71**	0.03	**0.99**	0.04	**0.82**
	Phase - HFF	-2.09	0.01	-11.16	0.29	23.56	0.39
	Phase - CD	0.02	**0.51**	0.03	**0.7**	0.05	**0.65**
Contrast	ID accuracy	1.38	**0.6**	0.25	0.16	1.28	**0.6**
	Jacobian	0.0	0.2	-0.0	0.2	-0.0	0.01
	Pixel - within	-13.46	0.48	-3.2	0.3	-12.12	0.45
	Pixel - between	14.19	**0.67**	0.83	0.01	24.8	**0.7**
	Amp - HFF	-23.97	0.44	-13.62	**0.91**	-26.29	0.06
	Amp - CD	0.04	**0.86**	0.02	**0.64**	0.05	**0.85**
	Phase - HFF	-12.78	0.11	-15.9	**0.88**	25.25	0.22
	Phase - CD	0.04	**0.69**	0.03	**0.98**	0.07	**0.59**
Defocus Blur	ID accuracy	1.36	**0.75**	0.27	0.23	1.39	**0.81**
	Jacobian	0.0	0.15	-0.0	0.27	-0.0	0.01
	Pixel - within	-13.75	**0.66**	-3.38	0.38	-14.13	**0.7**
	Pixel - between	12.1	**0.64**	1.29	0.04	24.5	**0.78**
	Amp - HFF	-18.59	0.35	-13.03	**0.94**	-17.62	0.03
	Amp - CD	0.03	**0.79**	0.02	**0.71**	0.05	**0.82**
	Phase - HFF	-6.15	0.03	-14.51	**0.83**	32.63	0.43
	Phase - CD	0.03	**0.62**	0.03	**0.99**	0.07	**0.61**
Pixelate	ID accuracy	1.26	**0.6**	0.46	**0.5**	1.24	**0.55**
	Jacobian	0.0	0.08	-0.0	**0.55**	-0.0	0.06
	Pixel - within	-12.06	0.47	-5.12	**0.67**	-11.47	0.4
	Pixel - between	12.07	**0.59**	3.47	0.22	25.27	**0.71**
	Amp - HFF	-18.74	0.33	-15.18	**0.97**	-29.89	0.08
	Amp - CD	0.03	**0.76**	0.03	**0.93**	0.05	**0.83**
	Phase - HFF	-5.38	0.02	-14.03	**0.59**	32.24	0.36
	Phase - CD	0.03	**0.55**	0.03	**0.93**	0.07	**0.57**
Gaussian Noise	ID accuracy	1.6	**0.65**	0.44	0.46	1.73	**0.65**
	Jacobian	0.0	0.05	-0.0	0.5	-0.0	0.04
	Pixel - within	-15.68	**0.54**	-4.93	**0.62**	-16.75	**0.51**
	Pixel - between	13.82	**0.52**	3.18	0.19	33.06	**0.74**
	Amp - HFF	-20.7	0.27	-15.19	**0.98**	-50.62	0.14
	Amp - CD	0.04	**0.73**	0.03	**0.9**	0.07	**0.93**
	Phase - HFF	-4.93	0.01	-14.38	**0.62**	35.48	0.26
	Phase - CD	0.04	**0.51**	0.03	**0.94**	0.1	**0.72**
Impulse Noise	ID accuracy	1.86	**0.66**	0.54	**0.53**	1.94	**0.64**
	Jacobian	0.0	0.07	-0.0	**0.57**	-0.0	0.04
	Pixel - within	-18.45	**0.55**	-5.88	**0.69**	-19.1	**0.52**
	Pixel - between	16.68	**0.56**	4.15	0.25	37.04	**0.72**
	Amp - HFF	-25.88	0.31	-17.02	**0.96**	-58.31	0.14
	Amp - CD	0.05	**0.76**	0.03	**0.94**	0.08	**0.9**
	Phase - HFF	-7.67	0.02	-15.28	**0.55**	40.22	0.26
	Phase - CD	0.05	**0.55**	0.04	**0.91**	0.11	**0.69**

Table 3: Can model metrics explain why CLIP ensembling affects OOD robustness?

Our results provide a new perspective on the effective robustness problem by showing that no one single metric rules them all. This dictates the need for a multi-faceted analysis to understand OOD robustness, which is in sharp contrast to the in-distribution setting where a single property such as model flatness can reasonably explain generalization performance.

***RobustNets* benchmark.** To aid further research into understanding, modeling, and designing for OOD robustness, we will publish a database of our pretrained models on CIFAR-10 (with diverse effective robustness) with the full range of model architectures, data augmentations, and pruning types we study. The models we study on ImageNet are already available publically. We will also publish the ID and OOD accuracy values for each model, and our code used to evaluate Jacobian norm and high frequency fraction and consistent distance for Fourier amplitude and phase interpolation, so that future researchers can apply it to study new and more robust models as they become available.

References

[1] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning, 2021. URL https://arxiv.org/abs/2106.15831.

[2] Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. 2019. doi: 10.48550/ARXIV.1906.00425. URL https://arxiv.org/abs/1906.00425.

[3] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark, 2020. URL https://arxiv.org/abs/2010.09670.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] James Diffenderfer and Bhavya Kailkhura. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. *International Conference on Learning Representations*, 2021.

[7] James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/file/0607f4c705595b911a4f3e7a127b44e0-Paper.pdf.

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[9] Sara Fridovich-Keil, Raphael Gontijo Lopes, and Rebecca Roelofs. Spectral bias in practice: The role of function frequency in generalization. *CoRR*, abs/2110.02424, 2021. URL https://arxiv.org/abs/2110.02424.

[10] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions, 2021. URL https://arxiv.org/abs/2107.03315.

[11] Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

[14] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

[15] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2019.

[16] Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

[17] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

[18] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[19] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. *arXiv preprint arXiv:2103.02325*, 2021.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[21] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL https://doi.org/10.1145/1873951.1874254.

[22] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization, 2021. URL https://arxiv.org/abs/2107.04649.

[23] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[24] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries, 2020. URL `https://arxiv.org/abs/2002.06349`.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[26] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. 2018. doi: 10.48550/ARXIV.1806.08734. URL `https://arxiv.org/abs/1806.08734`.

[27] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11893–11902, 2020.

[28] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10?, 2018. URL `https://arxiv.org/abs/1806.00451`.

[29] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL `https://arxiv.org/abs/1902.10811`.

[30] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning, 2020.

[31] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time?, 2019. URL `https://arxiv.org/abs/1906.02168`.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[33] Jiachen Sun, Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, Dan Hendrycks, Jihun Hamm, and Z Morley Mao. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*, 2021.

[34] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2021. URL `https://arxiv.org/abs/2109.01903`.

[35] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain, 2018. URL `https://arxiv.org/abs/1807.01251`.

[36] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision, 2019. URL `https://arxiv.org/abs/1906.08988`.

[37] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are included in the discussion section.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code will be made public upon acceptance. We do also provide sufficient details in the text that anyone could reproduce our code.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Most of these are provided in the methods section, and some are in the supplement.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We use 95% confidence intervals in all of our figures.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Our experiments did require substantial compute to evaluate all of our models on all of the interpolation paths, but we use pretrained models so we do not require any compute for training.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of all the models and datasets we use.

(b) Did you mention the license of the assets? [No] Licenses for individual assets can be found by visiting the websites of the corresponding creators.

(c) Did you include any new assets either in the supplemental material or as a URL? [No] We do not include new assets in the submission; however, we will release our *RobustNets* database of pretrained models evaluated in this paper upon acceptance.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]