Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori
UC Berkeley

Achal Dave
CMU

Vaishaal Shankar
UC Berkeley

Nicholas Carlini
Google Brain

Benjamin Recht
UC Berkeley

Ludwig Schmidt
UC Berkeley

Abstract

We study how robust current ImageNet models are to distribution shifts arising from natural variations in datasets. Most research on robustness focuses on synthetic image perturbations (noise, simulated weather artifacts, adversarial examples, etc.), which leaves open how robustness on synthetic distribution shift relates to distribution shift arising in real data. Informed by an evaluation of 204 ImageNet models in 213 different test conditions, we find that there is often little to no transfer of robustness from current synthetic to natural distribution shift. Moreover, most current techniques provide no robustness to the natural distribution shifts in our testbed. The main exception is training on larger and more diverse datasets, which in multiple cases increases robustness, but is still far from closing the performance gaps. Our results indicate that distribution shifts arising in real data are currently an open research problem. We provide our testbed and data as a resource for future work at https://modestyachts.github.io/imagenet-testbed/.

1 Introduction

Reliable classification under distribution shift is still out of reach for current machine learning [65, 68, 91]. As a result, the research community has proposed a wide range of evaluation protocols that go beyond a single, static test set. Common examples include noise corruptions [33, 38], spatial transformations [28, 29], and adversarial examples [5, 84]. Encouragingly, the past few years have seen substantial progress in robustness to these distribution shifts, e.g., see [13, 28, 34, 55, 57, 66, 93, 96, 105, 114, 115] among many others. However, this progress comes with an important limitation: all of the aforementioned distribution shifts are *synthetic*: the test examples are derived from well-characterized image modifications at the pixel level.

Synthetic distribution shifts are a good starting point for experiments since they are precisely defined and easy to apply to arbitrary images. However, classifiers ultimately must be robust to distribution shifts arising naturally in the real world. These distribution shifts may include subtle changes in scene compositions, object types, lighting conditions, and many others. Importantly, these variations are *not* precisely defined because they have not been created artificially. The hope is that an ideal robust classifier is still robust to such natural distribution shifts.

In this paper, we investigate how robust current machine learning techniques are to distribution shift arising naturally from real image data without synthetic modifications. To this end, we conduct a comprehensive experimental study in the context of ImageNet [18, 70]. ImageNet is a natural starting point since it has been the focus of intense research efforts over the past decade and a large number of pre-trained classification models, some with robustness interventions, are available for this task. The core of our experimental study is a testbed of 204 pre-trained ImageNet models that we evaluate in 213 different settings, covering both the most popular models and distribution shifts. Our testbed consists of 10^9 model predictions and is 100 times larger than prior work [27, 33, 47, 68]. This allows us to draw several new conclusions about current robustness interventions:

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Robustness measurements should control for accuracy. Existing work typically argues that an intervention improves robustness by showing that the accuracy on a robustness test set has improved (e.g., see [34, 40, 63, 102, 115]). We find that in many cases, this improved robustness can be explained by the model performing better on the standard, unperturbed test set. For instance, using different model architectures does not substantially improve the robustness of a model beyond what would be expected from having a higher standard accuracy. While training more accurate models is clearly useful, it is important to separate accuracy improvements from robustness improvements when interpreting the results.

Current synthetic robustness measures do not imply natural robustness. Prior work often evaluates on synthetic distribution shifts to measure robustness [9, 32, 38]. We find that current robustness measures for synthetic distribution shift are at most weakly predictive for robustness on the natural distribution shifts presently available. While there are good reasons to study synthetic forms of robustness – for instance, adversarial examples are interesting from a security perspective – synthetic distribution shifts alone do not provide a comprehensive measure of robustness at this time. Moreover, as the right plot in Figure 1 exemplifies, current robustness interventions are often (but not always) ineffective on the natural distribution shifts in our testbed.

Training on more diverse data improves robustness. Across all of our experiments, the only intervention that improves robustness to multiple natural distribution shifts is training with a more diverse dataset. This overarching trend has not previously been identified and stands out only through our large testbed. Quantifying when and why training with more data helps is an interesting open question: while more data is generally helpful, we find some models that are trained on 100 times more data than the standard ImageNet training set but do not provide any robustness.

The goal of our paper is specifically *not* to introduce a new classification method or image dataset. Instead, our paper is a meta-study of current robustness research to identify overarching trends that span multiple evaluation settings. This is particularly important if the ultimate goal of a research direction is to produce models that function reliably in a wide variety of contexts. Our findings highlight robustness on real data as a clear challenge for future work. Due to the diminishing returns of larger training datasets, addressing this robustness challenge will likely require new algorithmic ideas and more evaluations on natural distribution shifts.

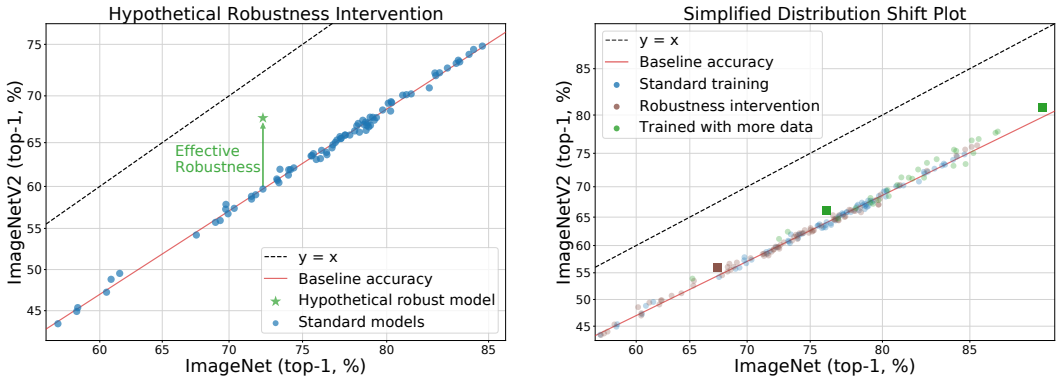

Figure 1: (Left) We plot 78 standard models trained on ImageNet without any robustness interventions, showing both their accuracy on the standard test set (ImageNet, x-axis) and on a test set with distribution shift (ImageNetV2, y-axis). All models lie below the $y = x$ line: their accuracy under this distribution shift is lower than on the standard test set. Nevertheless, improvements in accuracy on the standard test set almost perfectly predict a consistent improvement under distribution shift, as shown by the linear fit (red line). A hypothetical robustness intervention, shown in green, should provide *effective robustness*, i.e., the intervention should improve the accuracy under distribution shift beyond what is predicted by the linear fit.
(Right) We plot most of the 204 models in our testbed, highlighting those with the highest effective robustness using square markers. These models are still far from closing the accuracy gap induced by the distribution shift (ideally a robust model would fall on the $y = x$ line). Figure 2 shows a more detailed version of this plot with error bars for all points.

2 Measuring robustness

We first discuss how to measure robustness as a quantity distinct from accuracy. In our experiments, we always have two evaluation settings: the "standard" test set, and the test set with distribution shift. For a model f, we denote the two accuracies with $\text{acc}_1(f)$ and $\text{acc}_2(f)$, respectively.

When comparing the robustness of two models f_a and f_b, one approach would be to rank the models by their accuracy under distribution shift. However, this approach does not disentangle the robustness of a model from its accuracy on the standard test set. As an example, consider a pair of models with accuracy $\text{acc}_1(f_a) = 0.8$, $\text{acc}_2(f_a) = 0.75$ (i.e., a 5% drop in accuracy from the distribution shift), and $\text{acc}_1(f_b) = 0.9$, $\text{acc}_2(f_b) = 0.76$ (a 14% drop). Model f_b has higher accuracy on the second test set, but overall sees a drop of 14% from the standard to the shifted test set. In contrast, the first model sees only a 5% drop. Hence we would like to refer to the first model as more robust, even though it achieves lower accuracy on the shifted test set.

Effective robustness. The core issue in the preceding example is that standard accuracy (acc_1) acts as a confounder. Instead of directly comparing accuracies under distribution shift, we would like to understand if a model f_b offers higher accuracy on the shifted test set *beyond what is expected from having higher accuracy on the original test set*. We call this notion of robustness beyond a baseline *effective robustness*. Graphically, effective robustness corresponds to a model being above the linear trend (red line) given by our testbed of standard models in Figure 1 (left).

To precisely define effective robustness, we introduce $\beta(x)$, the baseline accuracy on the shifted test set for a given accuracy x on the standard test set. On the distribution shifts in our testbed, we instantiate β by computing the parameters of a log-linear fit for the models without a robustness intervention (the red line in Figure 1). Empirically, this approach yields a good fit to the data. For other distribution shifts, the baseline accuracy may follow different trends and may also depend on properties beyond the standard accuracy, e.g., model architecture. Appendix J.1 contains detailed information on how to compute β.

Given the accuracy baseline β, we define the effective robustness of a model as

$$\rho(f) = \text{acc}_2(f) - \beta(\text{acc}_1(f)) \ .$$

A model without special robustness properties falls on the linear fit and hence has $\rho(f) = 0$. The main goal of a robustness intervention is to increase ρ. Models with large ρ offer robustness beyond what we can currently achieve with standard models.

Relative robustness. Effective robustness alone does not imply that a robustness intervention is useful. In particular, a robustness intervention could increase ρ for a model it is applied to, but at the same time *decrease* both acc_1 and acc_2. Such a robustness intervention would offer no benefits. So to complement effective robustness, we also introduce *relative robustness* to directly quantify the effect of an intervention on the accuracy under distribution shift. For a model f' with robustness intervention, derived from a model f without the intervention, the relative robustness is $\tau(f') = \text{acc}_2(f') - \text{acc}_2(f)$. We graphically illustrate this notion of robustness in Appendix C.1.

Overall, a useful robustness intervention should obtain *both* positive effective and relative robustness. As we will see, only few classification models currently achieve this goal, and no models achieve both large effective and relative robustness.

3 Experimental setup

We now describe our experimental setup. A model f is first trained on a fixed training set. We then evaluate this model on two test sets: the "standard" test set (denoted S_1) and the test set with a distribution shift (denoted S_2).

A crucial question in this setup is what accuracy the model f can possibly achieve on the test set with distribution shift. In order to ensure that the accuracy on the two test sets are comparable, we focus on natural distribution shifts where humans have thoroughly reviewed the test sets to include only correctly labeled images [2, 18, 39, 68, 76].[1] This implies that an ideal robust classifier does not have a substantial accuracy gap between the two test sets. Indeed, recent work experimentally

[1] For ObjectNet [2], Borji [7] has pointed out potential label quality issues, but also found that a substantial accuracy drop remains when taking these issues into account.

confirms that humans achieve similar classification accuracy on the original ImageNet test set and the ImageNetV2 replication study (one of the distribution shifts in our testbed) [77].

3.1 Types of distribution shifts

At a high level, we distinguish between two main types of distribution shift. We use the term *natural* distribution shift for datasets that rely only on unmodified images. In contrast, we refer to distribution shifts as *synthetic* if they involve modifications of existing images specifically to test robustness. To be concrete, we next provide an overview of the distribution shifts in our robustness evaluation, with further details in Appendix F and visual overviews in Appendices A and K.

3.1.1 Natural distribution shifts

We evaluate on seven natural distribution shifts that we classify into three categories.

Consistency shifts. To evaluate a notion of robustness similar to ℓ_p-adversarial examples but without synthetic perturbations, we measure robustness to small changes across video frames as introduced by Gu et al. [35] and Shankar et al. [76]. The authors assembled sets of contiguous video frames that appear perceptually similar to humans, but produce inconsistent predictions for classifiers. We define S_1 to be the set of "anchor" frames in each video, and evaluate the accuracy under distribution shift by choosing the worst frame from each frame set for a classifier. This is the "`pm-k`" metric introduced by Shankar et al. [76].

Dataset shifts. Next, we consider datasets S_2 that are collected in a different manner from S_1 but still evaluate a classification task with a compatible set of classes. These distribution shifts test to what extent current robustness interventions help with natural variations between datasets that are hard to model explicitly. We consider four datasets of this variety: (i) ImageNetV2, a reproduction of the ImageNet test set collected by Recht et al. [68]; (ii) ObjectNet, a test set of objects in a variety of scenes with 113 classes that overlap with ImageNet [2]; and, (iii) ImageNetVid-Robust-anchor and YTBB-Robust-anchor [76], which are the datasets constructed from only the anchor frames in the consistency datasets described above. These two datasets contain 30 and 24 super-classes of the ImageNet class hierarchy, respectively. For each of these distribution shifts, we define S_1 to be a subset of the ImageNet test set with the same label set as S_2 so that the accuracies are comparable.

Adversarially filtered shifts. Finally, we consider an adversarially collected dataset, ImageNet-A [39]. Hendrycks et al. [39] assembled the dataset by downloading a large number of labeled images from Flickr, DuckDuckGo, iNaturalist, and other sites, and then selected the subset that was misclassified by a ResNet-50 model. We include ImageNet-A in our testbed to investigate whether the adversarial filtering process leads to qualitatively different results. Since ImageNet-A contains only 200 classes, the standard test set S_1 here is again a subset of the ImageNet test set that has the same 200 classes as ImageNet-A.

3.1.2 Synthetic distribution shifts

The research community has developed a wide range of synthetic robustness notions for image classification over the past five years. In our study, we consider the following classes of synthetic distribution shifts, which cover the most common types of image perturbations.

Image corruptions. We include all corruptions from [38], as well as some corruptions from [33]. These include common examples of image noise (Gaussian, shot noise), various blurs (Gaussian, motion), simulated weather conditions (fog, snow), and "digital" corruptions such as various JPEG compression levels. We refer the reader to Appendix F.2 for a full list of the 38 corruptions.

Style transfer. We use a stylized version of the ImageNet test set [34, 44].

Adversarial examples. We include untargeted adversarial perturbations bounded in ℓ_∞- or ℓ_2-norm by running projected gradient descent as described in [55]. We use $\varepsilon = \{\frac{0.5}{255}, \frac{2}{255}\}$ for ℓ_∞ and $\varepsilon = \{0.1, 0.5\}$ for ℓ_2 (further details in Appendix F.3).

3.2 Classification models

Our model testbed includes 204 ImageNet models covering a variety of different architectures and training methods. The models can be divided into the following three categories (see Appendix G for a full list of all models and their categories).

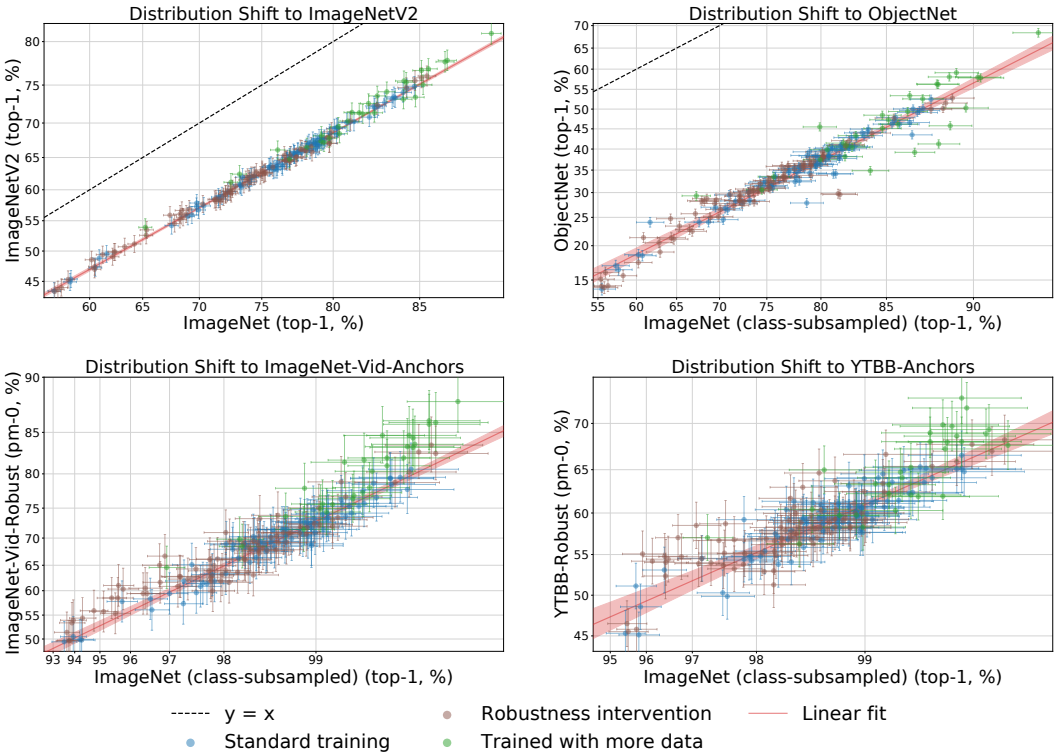

Figure 2: Model accuracies on the four natural dataset shifts: ImageNetV2 (top left), ObjectNet (top right), ImageNet-Vid-Robust-anchor (bottom left), and YTBB-Robust-anchor (bottom right). These plots demonstrate that the standard test accuracy (x-axis) is a reliable predictor for the test accuracy under distribution shift (y-axis), especially for models trained without a robustness intervention. The notable outliers to this trend are some models trained on substantially more data. For ObjectNet, ImageNet-Vid-Robust-anchor, and YTBB-Robust-anchor, we show the accuracy on a subset of the ImageNet classes on the x-axis to match the label space of the target task (y-axis). Each data point corresponds to one model in our testbed and is shown with 99.5% Clopper-Pearson confidence intervals. The axes were adjusted using logit scaling and the linear fit was computed in the scaled space on only the standard models. The red shaded region is a 95% confidence region for the linear fit from 1,000 bootstrap samples.

Standard models. We refer to models trained on the ILSVRC 2012 training set without a specific robustness focus as *standard* models. This category includes 78 models with architectures ranging from AlexNet to EfficietNet, e.g., [37, 50, 78, 85, 88].

Robust models. This category includes 86 models with an explicit robustness intervention such as adversarially robust models [13, 27, 72, 74, 101], models with special data augmentation [20, 28, 34, 41, 100, 108, 113], and models with architecture modifications [115].

Models trained on more data. Finally, our testbed contains 30 models that utilize substantially more training data than the standard ImageNet training set. This subset includes models trained on (i) Facebook's collection of 1 billion Instagram images [56, 104], (ii) the YFCC 100 million dataset [104], (iii) Google's JFT 300 million dataset [82, 102], (iv) a subset of OpenImages [98], or (v) a subset of the full ImageNet dataset of 21,841 classes [11, 49, 99].

4 Main results

We now present our main experiments. First, we measure how much effective and relative robustness models achieve on the natural distribution shifts in our testbed. Then we investigate to what extent robustness on synthetic distribution shift is predictive of robustness on natural distribution shift.

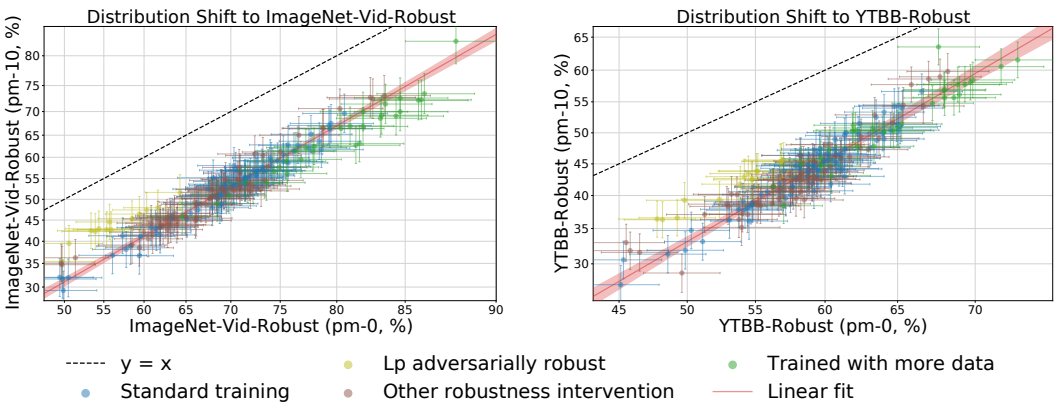

Figure 3: Model accuracies on the two consistency shifts: ImageNet-Vid-Robust (left), and YTBB-Robust (right). Both plots are shown with evaluation on pm-0 (anchor frames) on the x-axis and pm-10 (worst case prediction in a 20-frame neighborhood) on the y-axis. This plot shows that most current robustness interventions do not provide robustness to consistency distribution shifts. The notable outliers to this trend are ℓ_p-adversarially robust models and EfficientNet-L2 (NoisyStudent). We color the adversarially robust models separately in this figure to illustrate this phenomenon. Confidence intervals, axis scaling, and the linear fit are computed similarly to Figure 2.

4.1 Results on natural distribution shifts

Following the categorization in Section 3, we measure the robustness of classification models on three types of natural distribution shift. Appendix J.2 contains variations of the figures referenced in this section. For further detail, we have made interactive plots available at http://robustness.imagenetv2.org/.

Dataset shifts. Figure 2 shows the effective robustness of models on the four dataset shifts in our testbed. In each case, we find that the standard test accuracy (x-axis) is a good predictor for the test accuracy under distribution shift (y-axis). The linear fit is best for ImageNetV2, ObjectNet, and ImageNet-Vid-Robust with respective r^2 scores of 1.00, 0.95, and 0.95, but is more noisy for YTBB-Robust ($r^2 = 0.83$). The noisy fit on YTBB-Robust is likely due to the fact that the categories in YTBB-Robust are not well aligned with those of ImageNet, where the models were trained [76]. Another potential reason is that the video test sets are significantly smaller (2,530 images in YTBB and 1,109 images in ImageNet-Vid-Robust).

In the high accuracy regime (above the 76% achieved by a ResNet-50), the main outliers in terms of positive effective robustness are models trained on substantially more data than the standard ImageNet training set. This includes a ResNet152 model trained on 11,000 ImageNet classes ($\rho = 2.1\%$) [99], several ResNeXt models trained on 1 billion images from Instagram ($\rho = 1.5\%$) [56], and the EfficientNet-L2 (NoisyStudent) model trained on a Google-internal JFT-300M dataset of 300 million images ($\rho = 1.1\%$) [102]. However, not all models trained on more data display positive effective robustness. For instance, a ResNet101 trained on the same JFT-300M dataset has an effective robustness of $\rho = -0.23\%$ [82]. We conduct additional experiments to investigate the effect of training data in Appendix B. Appendix H contains a full list of models with their effective robustness numbers. On YTBB-Robust, a few data augmentation strategies and ℓ_p-robust models display positive effective robustness; we investigate this further in Appendix C.2.

Consistency shifts. We plot the effective robustness of models on consistency shifts in Figure 3. Interestingly, we observe that ℓ_p-adversarially robust models display substantial effective robustness to ImageNet-Vid-Robust (average $\rho = 6.7\%$) and YTBB-Robust (average $\rho = 4.9\%$). This suggests that these models are not only more robust to synthetic perturbations, but also offer robustness for the perceptually small variations between consecutive video frames.

However, these gains in effective robustness do not necessarily lead to relative robustness. On average, relative robustness on both datasets is negative (average $\tau = -8.5\%$ on ImageNet-Vid-Robust and average $\tau = -0.7\%$ on YTBB-Robust for ResNet50 models). See Appendix C.2 (Figure 10) for a visual comparison. Among the models trained on more data, only one achieves both high accuracy and substantial effective robustness: EfficientNet-L2 (NoisyStudent) [102] has $\rho = 2.4\%$ and $\rho = 7.4\%$ on ImageNet-Vid-Robust and YTBB-Robust, respectively.

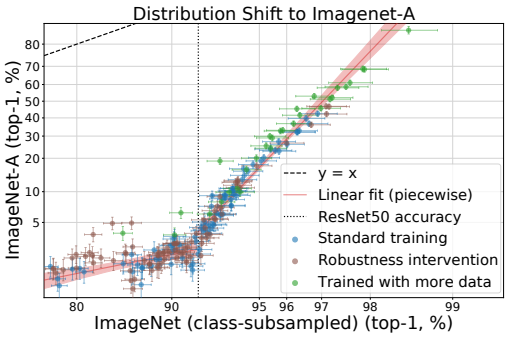

Figure 4: Model accuracies on ImageNet-A, a dataset adversarially filtered to contain only images incorrectly classified by a ResNet50 trained on ImageNet. This filtering results in a 'knee' curve: models with lower ImageNet accuracy than ResNet-50 have near-chance performance on ImageNet-A, while models with higher ImageNet accuracy improve drastically on ImageNet-A. The linear fit is computed piecewise around the ResNet50 model accuracy.

Adversarially filtered shifts. ImageNet-A [39] was created by classifying a set of images with a ResNet50 and only keeping the misclassified images. Interestingly, this approach creates a "knee" in the resulting scatter plot (see Figure 4): models below a ResNet50's standard accuracy have close to chance performance on ImageNet-A,[2] and models above a ResNet50's standard accuracy quickly close the accuracy gap. In the high accuracy regime, every percentage point improvement on ImageNet brings at least an 8% improvement on ImageNet-A. This is in contrast to datasets that are not constructed adversarially, where the initial accuracy drops are smaller, but later models make slow progress on closing the gap. These results demonstrate that adversarial filtering does not necessarily lead to harder distribution shifts.

4.2 Results on synthetic distribution shifts

Given the difficulty of collecting real world data to measure a model's robustness to natural distribution shifts, an important question is whether there are synthetic proxies. We now study to what extent robustness to the above synthetic distribution shifts predicts robustness on these natural distribution shifts.

In Figure 5, we analyze the predictiveness of one commonly studied synthetic robustness metric: average accuracy on image corruptions [38]. We compare this metric with effective robustness on ImageNetV2. While effective robustness is only one aspect (c.f. Section 2), it is a necessary prerequisite for a model to have helpful robustness properties.

The plots show that robustness under this synthetic distribution shift does not imply that the corresponding model has effective robustness on ImageNetV2 (the Pearson correlation coefficient is $r = 0.24$). In Appendix D.1, we repeat the above experiment for accuracy drop under PGD adversarial attacks [55] and also find a weak correlation ($r = -0.05$). Appendix D.2 further extends the experiment by comparing both synthetic distribution shift measures with the remaining natural distribution shifts in our testbed and reaches similiar conclusions.

Our analysis of the aggregate measures proposed in prior work does not preclude that specific synthetic distribution shifts do predict behavior on natural distribution shifts. Instead, our results show that averaging a large number of synthetic corruptions does not yield a comprehensive robustness measure that also predicts robustness on natural distribution shift.

To extend on this analysis, in Appendix I we find that no individual synthetic measure in our testbed is a consistent predictor of natural distribution shift, but some synthetic shifts are substantially more predictive than others. For instance, ℓ_p-robustness has the highest correlation with consistency shifts, and some image corruptions such as brightness or Gaussian blur have higher correlation with dataset shifts. However, our testbed indicates that these synthetic measures are not necessarily causal, i.e., models trained with brightness or Gaussian blur do not have substantial effective robustness on dataset shifts. Further analyzing relationships between individual synthetic and natural distribution shifts is an interesting avenue for future work.

4.3 Takeaways and discussion

To recap our results, we now discuss two of the central questions in our paper: Do current robustness interventions help on real data? And is synthetic robustness correlated with natural robustness?

[2]Chance performance is 0.5% as ImageNet-A contains 200 classes.

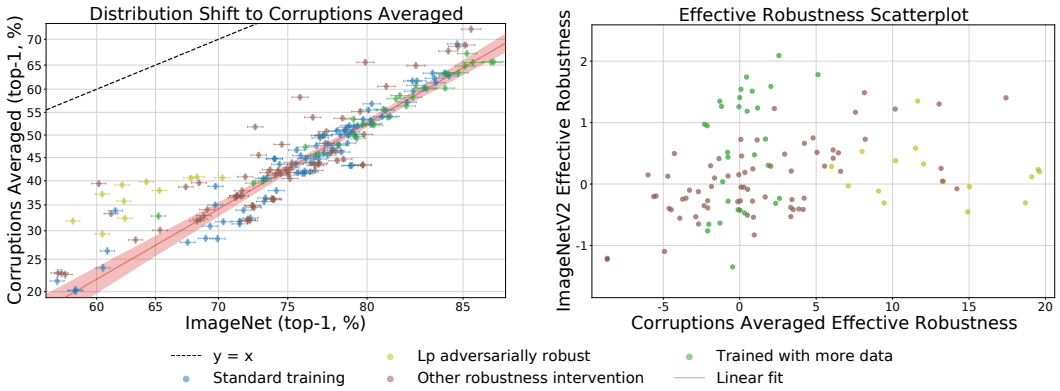

Figure 5: Model accuracies under image corruptions. Similar to Figure 2, the left plot shows the effective robustness for this synthetic distribution shift. Multiple non-standard models achieve substantial effective robustness, corroborating recent research progress on creating models robust to synthetic shifts. The right plot shows the correlation between the effective robustness for image corruptions and the ImageNetV2 distribution shift (top left in Figure 2) for the non-standard models. Image corruptions are only weakly predictive of effective robustness on ImageNetV2: there are several models that achieve high effective robustness under image corruptions but little to no effective robustness on ImageNetV2.

Across our study, current robustness interventions offer little to no improvement on the natural distribution shifts presently available.

For dataset shifts, we find that models trained with substantially more data yield a small improvement. However, the amount of extra data needed is orders of magnitude larger than the standard ImageNet training set, and the models show only small gains (in the best case improving the accuracy drop from 8.6% to 7.5% on ImageNetV2 for EfficientNet-L2 NoisyStudent). These results suggest that current robustness interventions methods do not provide benefits on the dataset shifts in our study.

For consistency shifts, adversarially trained models generally have effective robustness, but usually little or no relative robustness. On ImageNet-Vid-Robust, the baseline models without adversarial training still achieve higher accuracy under distribution shift. A notable outlier is EfficietNet-L2 (NoisyStudent) [102], which utilizes self-training and exhibits high effective robustness in the high accuracy regime. Self-training has recently been shown to help adversarial robustness as well [10, 61, 94]. Investigating the effect of self-training on robustness is an interesting direction for future work.

Moreover, we find that current aggregate metrics for synthetic robustness are at most weakly correlated with natural robustness. Effective robustness under non-adversarial image corruptions or ℓ_p-attacks does not imply effective robustness to natural distribution shifts. While much progress has been made on creating models robust to synthetic distribution shift, new methods may be needed to handle natural shifts.

5 Related work

Our work is best seen as a unification of two independent lines of research—synthetic and natural distribution shift—not previously studied together. Synthetic distribution shifts have been studied extensively in the literature [28, 33, 38, 48, 57, 93]. We incorporate as many prior synthetic measures of robustness as possible. Our dataset largely confirms the high-level results from these papers (see Appendix E for additional discussion). For example, Ford et al. [31] provide evidence for the relationship between adversarial robustness and robustness to Gaussian noise. The study of natural distribution shifts has been an equally extensive research direction [2, 68, 76, 91]. When examining each natural distribution shift individually, we confirm the findings of earlier work that there is a consistent drop with a linear trend going from ImageNet to each of the other test sets [2, 68, 76].

We study the relationship between these two previously independent lines of work. By creating a testbed $100\times$ larger than prior work [27, 33, 47, 68], we are able to make several new observations. For instance, we show that robustness to synthetic distribution shift often behaves differently from robustness to natural distribution shift. We argue that it is important to control for accuracy when

measuring the efficacy of a robustness intervention. Viewed in this light, most interventions do not provide effective robustness. The main exception is training with more data, which improves robustness across natural distribution shifts. In some situations, ℓ_p-adversarial robustness helps with natural distribution shift that asks for consistency across similar looking images.

Appendix L contains additional discussion of related work in more detail. Appendix L.1 broadly discusses the relationship of this work to other areas in machine learning. Appendix L.3 specifically revisits consistency shifts and explains why, in contrast to previous work [35], we find consistency robustness is only weakly correlated with color corruption robustness.

Concurrent and subsequent work. An early version of this paper with results on ImageNetV2 and ImageNet-Vid-Robust appeared on OpenReview in late 2019 [90]. Since then, two closely related papers have been published concurrently with the updated version of this paper.

Djolonga et al. [21] evaluate 40 models on the same natural distribution shifts as our paper. Our testbed is larger and contains 200 models with more robustness interventions. Overall both papers reach similar conclusions. Their focus is more on the connections to transfer learning while we focus more on comparisons between synthetic and natural distribution shifts. Djolonga et al. [21] also explore the performance of various models with a synthetic image dataset.

Hendrycks et al. [40] also study the connections between synthetic robustness and robustness to natural distribution shifts. This paper introduces a new dataset, ImageNet-R, that contains various renditions (sculptures, paintings, etc.) of 200 ImageNet classes as a new example of natural distribution shift. The paper then introduces DeepAugment, a new data augmentation technique based on synthetic image transformations, and find that this robustness intervention is effective on ImageNet-R. In Appendix L.2, we analyze the ImageNet-R test set and DeepAugment models, as well as the closely related ImageNet-Sketch test set [95], in more detail.

At a high-level, ImageNet-R and ImageNet-Sketch follow the trends of the other dataset shifts in our testbed, with models trained on extra data providing the most robustness (up to $\rho = 29.1\%$ on ImageNet-R, though the effect is not uniform, similar to the other dataset shifts). After the models trained on more data, we find that DeepAugment (in combination with AugMix [41]) achieves substantial effective robustness ($\rho = 11.2\%$). Interestingly, adversarial robustness also leads to effective robustness on ImageNet-R. An AdvProp model [100] achieves the highest absolute accuracy on ImageNet-R for a model trained without extra data (57.8%) and has effective robustness $\rho = 7.5\%$. A model with feature denoising and trained with PGD-style robust optimization [55, 101] achieves the highest effective robustness on ImageNet-R ($\rho = 22.7\%$) and also positive relative robustness ($\tau = 5.7\%$).

6 Conclusion

The goal of robust machine learning is to develop methods that function reliably in a wide variety of settings. So far, this research direction has focused mainly on synthetic perturbations of existing test sets, highlighting important failure cases and initiating progress towards more robust models. Ultimately, the hope is that the resulting techniques also provide benefits on real data. Our paper takes a step in this direction and complements the current synthetic robustness tests with comprehensive experiments on distribution shifts arising from real data.

We find that current image classification models still suffer from substantial accuracy drops on natural distribution shifts. Moreover, current robustness interventions – while effective against synthetic perturbations – yield little to no consistent improvements on real data. The only approach providing broad benefits is training on larger datasets, but the gains are small and inconsistent.

Overall, our results show a clear challenge for future research. Even training on 1,000 times more data is far from closing the accuracy gaps, so robustness on real data will likely require new algorithmic ideas and better understanding of how training data affects robustness. Our results indicate two immediate steps for work in this area: robustness metrics should control for baseline accuracy, and robust models should additionally be evaluated on natural distribution shifts. We hope that our comprehensive testbed with nuanced robustness metrics and multiple types of distribution shift will provide a clear indicator of progress on the path towards reliable machine learning on real data.

Broader Impact

Robustness is one of the key problems that prevents deploying machine learning in the real world and harnessing the associated benefits. A canonical example is image classification for medical diagnosis. As was found when researchers attempted to deploy a neural network to detect diabetes from retina images, "an accuracy assessment from a lab goes only so far. It says nothing of how the AI will perform in the chaos of a real-world environment" [3]. Similarly, researcher also found that current methods for chest X-ray classification are brittle even in the absence of recognized confounders [110]. If models were robust, then this transfer to the real world would be straightforward. Unfortunately, achieving robustness on real data is still a substantial challenge for machine learning.

Our work studies how robust current image classification methods are to distribution shifts arising in real data. We hope that our paper will have a positive effect on the study of distribution shifts and allow researchers to more accurately evaluate to what extent a proposed technique increases the robustness to particular forms of distribution shift. This will allow researchers to better understand how a deployed system will work in practice, without actually having to deploy it first and users potentially suffering negative consequences.

However, there are several potential ways in which our study could cause unintended harm. It is possible that our paper might be used as an argument to stop performing research on some synthetic forms of robustness, e.g., adversarial examples or common corruptions. This is not our intention. These forms of corruption are interesting independent of any correlation to existing natural distribution shift (e.g., adversarial examples are a genuine security problem).

We only capture a small number of natural distribution shifts among all the possible distribution shifts. We selected these shifts because they have been used extensively in the literature and are concrete examples of the types of distribution shift we would like models to be robust to. It is likely that there are shifts that we do not capture, and so even if the shifts we define were to be completely solved, other shifts would remain a concern.

One significant form of distribution shift we do not evaluate is dataset bias in representing different demographic groups. For example, the Inclusive Images dataset [75] attempts to correct for the geographical bias introduced in the Open Images dataset [51] by including a more balanced representation of images from Africa, Asia, and South America. Neglecting such implicit biases in the data distribution can harm underrepresented demographic groups. Ultimately, evaluating on fixed datasets may not be enough, and validating the fairness and safety of deployable machine learning requires careful analysis in the application domain.

Finally, more reliable machine learning can also enable negative uses cases, e.g., widespread surveillance or autonomous weapon systems. As with many technologies, these risks require careful regulation and awareness of unintended consequences arising from technological advances.

Acknowledgments and Disclosure of Funding

We would like to thank Logan Engstrom, Justin Gilmer, Moritz Hardt, Daniel Kang, Jerry Li, Percy Liang, Nelson Liu, John Miller, Preetum Nakkiran, Rebecca Roelofs, Aman Sinha, Jacob Steinhardt, and Dimitris Tsipras for helpful conversations while working on this paper.

This research was generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, a Siemens Futuremakers Fellowship, an Amazon AWS AI Research Award.

References

[1] Atwood, J., Baljekar, P., Barnes, P., Batra, A., Breck, E., Chi, P., Doshi, T., Elliott, J., Kour, G., Gaur, A., Halpern, Y., Jicha, H., Long, M., Saxena, J., Singh, R., and Sculley., D. The Inclusive Images competition, 2018. =https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html.

[2] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. http://papers.nips.cc/paper/9142-objectnet-a-large-scale-bias-controlled-dataset-for-pushing-the-limits-of-object-recognition-models.

[3] Beede, E., Baylor, E., Hersh, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L. M. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *CHI Conference on Human Factors in Computing Systems*, 2020. https://dl.acm.org/doi/abs/10.1145/3313831.3376718.

[4] Belinkov, Y. and Bisk, Y. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1711.02173.

[5] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD)*, 2013. https://arxiv.org/abs/1708.06131.

[6] Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, 2011. https://papers.nips.cc/paper/4312-generalizing-from-several-related-classification-tasks-to-a-new-unlabeled-sample.

[7] Borji, A. Objectnet dataset: Reanalysis and correction, 2020. https://arxiv.org/abs/2004.02042.

[8] Bras, R. L., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. Adversarial filters of dataset biases. In *International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/2002.04108.

[9] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness, 2019. https://arxiv.org/abs/1902.06705.

[10] Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. C. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13736.

[11] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. Dual path networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. https://arxiv.org/abs/1707.01629.

[12] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1610.02357.

[13] Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.02918.

[14] Crawford, K. and Paglen, T. Excavating AI: The politics of training sets for machine learning, 2019. https://www.excavating.ai/.

[15] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1805.09501.

[16] Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space, 2019. https://arxiv.org/abs/1909.13719.

[17] Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. https://ieeexplore.ieee.org/document/1467360.

[18] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.

[19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, 2019. https://www.aclweb.org/anthology/N19-1423/.

[20] DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout, 2017. https://arxiv.org/abs/1708.04552.

[21] Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., Gelly, S., Houlsby, N., Zhai, X., and Lucic, M. On robustness and transferability of convolutional neural networks, 2020. https://arxiv.org/abs/2007.08558.

[22] Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *To appear in the Annals of Statistics*, 2018. https://arxiv.org/abs/1810.08750.

[23] Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. *Journal of Machine Learning Research (JMLR)*, 2019. https://arxiv.org/abs/1610.02581.

[24] Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures, 2019. https://arxiv.org/abs/2007.13982.

[25] Dulhanty, C. and Wong, A. Auditing ImageNet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets, 2019. https://arxiv.org/abs/1905.01347.

[26] Dunn, M., Sagun, L., Higgins, M., Guney, V. U., Cirik, V., and Cho, K. Searchqa: A new q&a dataset augmented with context from a search engine, 2017. https://arxiv.org/abs/1704.05179.

[27] Engstrom, L., Ilyas, A., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

[28] Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Madry, A. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1712.02779.

[29] Fawzi, A. and Frossard, P. Manitest: Are classifiers really invariant? In *BMVC*, 2015. https://arxiv.org/abs/1507.06535.

[30] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. https://ieeexplore.ieee.org/document/5255236.

[31] Ford, N., Gilmer, J., Carlini, N., and Cubuk, E. D. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning (ICML)*, 2019. http://arxiv.org/abs/1901.10513.

[32] Galloway, A., Tanay, T., and Taylor, G. W. Adversarial training versus weight decay, 2018. https://arxiv.org/abs/1804.03308.

[33] Geirhos, R., Temme, C. R. M., 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 (NeurIPS)*, 2018. https://papers.nips.cc/paper/7982-generalisation-in-humans-and-deep-neural-networks.

[34] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. https://arxiv.org/abs/1811.12231.

[35] Gu, K., Yang, B., Ngiam, J., Le, Q., and Shlens, J. Using videos to evaluate image model robustness. In *SafeML workshop International Conference on Learning Representations (ICLR)*, 2019. https://arxiv.org/abs/1904.10076.

[36] Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization, 2020. https://arxiv.org/abs/2007.01434.

[37] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.03385.

[38] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. https://arxiv.org/abs/1903.12261.

[39] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples, 2019. https://arxiv.org/abs/1907.07174.

[40] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2020. https://arxiv.org/abs/2006.16241.

[41] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/1912.02781.

[42] Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. Squeeze-and-excitation networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1709.01507.

[43] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1608.06993.

[44] Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. *International Conference on Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1703.06868.

[45] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, 2016. https://arxiv.org/abs/1602.07360.

[46] Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. https://arxiv.org/abs/1502.03167.

[47] Kang, D., Sun, Y., Brown, T., Hendrycks, D., and Steinhardt, J. Transfer of adversarial robustness between perturbation types, 2019. https://arxiv.org/abs/1905.01034.

[48] Kang, D., Sun, Y., Hendrycks, D., Brown, T., and Steinhardt, J. Testing robustness against unforeseen adversaries, 2019. https://arxiv.org/abs/1908.08016.

[49] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning, 2019. https://arxiv.org/abs/1912.11370.

[50] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[51] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., and et al. The open images dataset v4. *International Journal of Computer Vision (IJCV)*, 2020. https://arxiv.org/abs/1811.00982.

[52] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. https://arxiv.org/abs/1712.00559.

[53] Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., and Murphy, K. Progressive neural architecture search. In *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1712.00559.

[54] Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1807.11164.

[55] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1706.06083.

[56] Mahajan, D. K., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1805.00932.

[57] Maini, P., Wong, E., and Kolter, J. Z. Adversarial robustness against the union of multiple perturbation models, 2019. https://arxiv.org/abs/1909.04068.

[58] Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/2004.14444.

[59] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Conference on Fairness, Accountability, and Transparency (FAT)*, 2019. https://arxiv.org/abs/1810.03993.

[60] Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning (ICML)*, 2013. https://arxiv.org/abs/1301.2115.

[61] Najafi, A., Maeda, S.-i., Koyama, M., and Miyato, T. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13021.

[62] Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. https://arxiv.org/abs/1909.02060.

[63] Orhan, A. E. Robustness properties of facebook's resnext wsl models, 2019. https://arxiv.org/abs/1907.07640.

[64] Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010. https://ieeexplore.ieee.org/document/5288526.

[65] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[66] Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1801.09344.

[67] Real, E., Shlens, J., Mazzocchi, S., Pan, X., and Vanhoucke, V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[68] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.10811.

[69] Rowley, H. A., Baluja, S., and Kanade, T. Human face detection in visual scenes. In *Advances in Neural Information Processing Systems (NIPS)*. 1996. https://papers.nips.cc/paper/1168-human-face-detection-in-visual-scenes.

[70] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. https://arxiv.org/abs/1409.0575.

[71] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. https://arxiv.org/abs/1911.08731.

[72] Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I., and Bubeck, S. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.04584.

[73] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1801.04381.

[74] Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1904.12843.

[75] Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017. https://arxiv.org/abs/1711.08536.

[76] Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time?, 2019. https://arxiv.org/abs/1906.02168.

[77] Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*, 2020.

[78] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. https://arxiv.org/abs/1409.1556.

[79] Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018. https://arxiv.org/abs/1710.10571.

[80] Sperber, M., Niehues, J., and Waibel, A. Toward robust neural machine translation for noisy input sequences. In *International Workshop on Spoken Language Translation (IWSLT)*, 2017. http://workshop2017.iwslt.org/downloads/P04-Paper.pdf.

[81] Srivastava, M., Hashimoto, T., and Liang, P. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning (ICML)*, 2020. https://arxiv.org/abs/2007.06661.

[82] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1707.02968.

[83] Sung, K. . and Poggio, T. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. https://ieeexplore.ieee.org/document/655648.

[84] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. https://arxiv.org/abs/1312.6199.

[85] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. https://arxiv.org/abs/1409.4842v1.

[86] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.00567.

[87] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. https://arxiv.org/abs/1602.07261.

[88] Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1905.11946.

[89] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. Mnasnet: Platform-aware neural architecture search for mobile. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1807.11626.

[90] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on imagenet, 2019. https://openreview.net/pdf?id=HyxPIyrFvH.

[91] Torralba, A., Efros, A. A., et al. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. https://ieeexplore.ieee.org/document/5995347.

[92] Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.06423.

[93] Tramer, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1904.13000.

[94] Uesato, J., Alayrac, J.-B., Huang, P.-S., Stanforth, R., Fawzi, A., and Kohli, P. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13725.

[95] Wang, H., Ge, S., Xing, E. P., and Lipton, Z. C. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.13549.

[96] Wong, E., Schmidt, F. R., and Kolter, J. Z. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1902.07906.

[97] Wu, B., Chen, W., Fan, Y., Zhang, Y., Hou, J., Liu, J., and Zhang, T. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7:172683–172693, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2956775. URL http://dx.doi.org/10.1109/ACCESS.2019.2956775.

[98] Wu, B., Chen, W., Fan, Y., Zhang, Y., Hou, J., Liu, J., and Zhang, T. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *IEEE Access*, 7, 2019.

[99] Wu, W. Classifying images into 11k classes with pretrained model, 2016. https://github.com/tornadomeet/ResNet and https://github.com/awslabs/deeplearning-benchmark/blob/master/image_classification/common/modelzoo.py#L41.

[100] Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., and Le, Q. V. Adversarial examples improve image recognition, 2019. https://arxiv.org/abs/1911.09665.

[101] Xie, C., Wu, Y., van der Maaten, L., Yuille, A., and He, K. Feature denoising for improving adversarial robustness. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1812.03411.

[102] Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1911.04252.

[103] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1611.05431.

[104] Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. Billion-scale semi-supervised learning for image classification, 2019. https://arxiv.org/abs/1905.00546.

[105] Yang, F., Wang, Z., and Heinze-Deml, C. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.11235.

[106] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. https://arxiv.org/abs/1809.09600.

[107] Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1906.08988.

[108] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. https://arxiv.org/abs/1905.04899.

[109] Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. https://arxiv.org/abs/1605.07146.

[110] Zech, J., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219764.

[111] Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. https://www.aclweb.org/anthology/D18-1009/.

[112] Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., and Houlsby, N. The visual task adaptation benchmark, 2019. https://arxiv.org/abs/1910.04867.

[113] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2017. https://arxiv.org/abs/1710.09412.

[114] Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. http://proceedings.mlr.press/v97/zhang19p.html.

[115] Zhang, R. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1904.11486.

[116] Zhang, X., Li, Z., Loy, C. C., and Lin, D. Polynet: A pursuit of structural diversity in very deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1611.05725.

[117] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1707.07012.