Benchmarks for Neural Network Robustness to Common Distortions and Styles

Dan Hendrycks*
University of Chicago
dan@ttic.edu

Thomas G. Dietterich
Oregon State University
tgd@oregonstate.edu

Abstract

In this paper we establish a new benchmark for image classifier robustness which standardizes and expands this topic. Unlike recent robustness research, our benchmark IMAGENET-D *evaluates performance on commonplace not worst-case adversarial distortions. We observe that there are negligible changes in relative robustness from AlexNet to ResNet classifiers. To improve robustness, we experiment with numerous techniques and find evidence that stability training, 10-crop classification, image restoration, and model compression do not increase robustness. In contrast, we demonstrate that image histogram equalization, Multigrid architectures, and enormous models improve distortion robustness. At the end, we introduce a new dataset to open research on a new kind of robustness, style robustness.*

1. Introduction

The use of deep neural networks in pattern recognition systems has led to large advances on numerous benchmark tasks. However, vast gains on the test set belie an underlying fragility of these models: they are not robust to common input distortions that we expect them to face in real-world usage. When tasked with a visual classification problem, humans tend to be robust to numerous natural and unnatural distortions, such as snow, blur, and pixelation. Moreover, the introduction of novel distortions and variations on existing distortions tends not to change this fact. Endowing neural network-based systems with this level of robustness is hence an important direction for bringing the performance of deep learning systems closer to that of humans, as well as for improving the safety and security of these systems for deployment in real-world environments like foggy roads.

In the general setting, input distortions can be arbitrarily severe, and even adversarial in nature. To date, the latter case has received much attention. Most investigations of adversarial [46] noise consider minute perturbations of the input, resulting in examples that are close in Euclidean distance. We contend that less pathological distortions, which may be further away in Euclidean distance, are also important targets for generalization. In Figure 1, we note that adversarial images are a type of "worst-case" distortion, while we benchmark more common distortions.

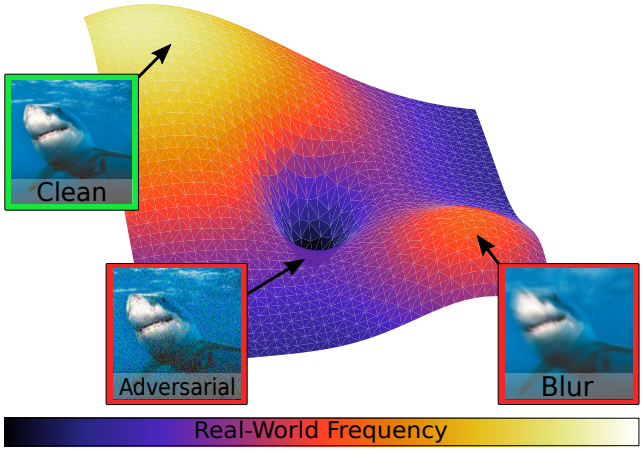

Figure 1: Adversarial and blurred images are often incorrectly classified, yet blurry images are far more frequent than crafted adversarial images. Rather than studying model generalization to images that are closest in a Euclidean sense, we benchmark model generalization to images with common distortions.

To this end, we propose a set of fifteen common visual distortions as a benchmark of the general robustness of deep neural network image classifiers. With this benchmark called IMAGENET-D, we attempt to standardize distortion research to prevent incomparable evaluations, moving goalposts, and experiment cherry-picking. At the same time, IMAGENET-D enables further classification research since there is wide room for improvement on this challenging benchmark. In extensive experiments, we demonstrate the shortcomings of current systems on this benchmark, and find methods and architectures which make tangible progress on differentially improving robustness while retaining accuracy. Finally, we consider a robustness goal

*Work done while at OSU.

parallel to distortion robustness which has not yet received proper treatment. We lay groundwork for research in image style robustness by creating a novel dataset, and we show current architectures lack style robustness.

2. Related Work

Adversarial Examples. An adversarial image is a clean image perturbed with a small, carefully crafted distortion so as to confuse a classifier [46]. These deceptive distortions can occasionally fool black-box classifiers [32], and they approximate the smallest image modification in RGB space necessary for classifier confusion [5]. Thus adversarial distortions serve as type of worst-case analysis for network robustness, and the eminence of such worst-case analysis has led "adversarial robustness" to become interchangeable with "robustness" [3, 42]. In efforts to ground adversarial robustness research, several competing robustness measures can be found scattered across the literature [3, 7, 5, 37, 28], but we find no indication of widespread adoption of one measure. Worse, defenses [36, 40, 38, 18] quickly succumb to new attacks [13, 8, 6], making absolute progress elusive. Worst, it is possible that this problem is unduly emphasized in the literature. For example, there is a lively tug-of-war between researchers creating stop-sign detectors and attacks attempting to fool said detectors. Yet it appears unjustified to worry about adversarial stop signs. This is because it is unclear why terrorists or meddlers would meticulously fabricate adversarial stop signs to fool autonomous vehicles when simply covering or removing stop signs fools both autonomous vehicles and humans. Additionally, terrorists could remove stop signs to fool humans *today*, yet the opportunity has not mobilized legions of terrorists to date. Defending against adversaries and the worst-case can be useful, but generalizing to plain distortions is already challenging.

Robustness in Speech. Speech recognition research approaches robustness differently [33, 39]. Common audio distortions (e.g., street noise, background chatter, wind) receive greater focus than adversarial audio because common distortions are ever-present and unsolved. In fact, there are several popular datasets containing noisy test audio [20, 19]. Robustness in noisy environments requires robust architectures, and some research finds convolutional networks more robust than fully connected networks [1]. Additional robustness stems from pre-processing techniques like standardizing the input's statistics [34, 39, 2, 16, 30].

ConvNet Fragility Studies. Several studies demonstrate the fragility of convolutional networks on simple distortions. For example, [21] use impulse noise to break Google's Cloud Vision API. Using Gaussian noise and blur, [11] demonstrate the superior robustness of human vision to convolutional networks, *even when networks are fine-tuned* on Gaussian noise or blur. [15] also compare networks to humans with noisy, contrast-reduced, and elastically deformed images. They determine that fine-tuning on specific distortions does not generalize, and classification error patterns made by networks and humans are not similar.

Robustness Enhancements. To reduce fragility, [47] describe how to fine-tune on blurred images to improve their recognition. They find it is not enough to fine-tune on one type of blur to obtain robustness to other blurs, and fine-tuning on several blurs can marginally decrease performance. [50] also find that fine-tuning on noisy images can cause underfitting, so they propose minimizing the cross-entropy from the softmax distribution of the noisy image to the softmax of the clean image. [10] address underfitting differently. They fine-tune each network on one type of distortion and classify with an mixture of these distortion-specific experts; generalization to images with combinations of known distortions is not assessed.

3. Distortion Robustness Benchmark

3.1. IMAGENET-D Dataset

IMAGENET-D Design. Our benchmark consists of 15 diverse distortion types exemplified in Figure 2. The benchmark covers noise, blur, weather, and digital categories. Research that improves performance on this benchmark should be a strong indication of general robustness, as the distortions are varied and great in number. These 15 distortion types each have five different levels of severity, as distortions like brightness are at different intensities in the real world. More, real-world distortions also have variation even at a fixed intensity. To that end, we introduce variation for each distortion when possible. For example, for each image the position of snowflakes and their descent angle are unique. These algorithmically generated distortions are applied to ImageNet [9] validation images, so we call the benchmark IMAGENET-D. Data can be downloaded or re-created by visiting this site. To enable further experimentation, we also designed an extra distortion for each noise, blur, weather, or digital category. Extra distortions are also available at the aforementioned URL, and they are depicted and explicated in Section 6. Even more experimentation is possible by applying these distortions to other images, like those from Places365. Overall, the dataset IMAGENET-D consists of 15 distortion types, each with five different severities, all applied to ImageNet validation images. These 15 distortion types are described below.

Common Distortions. The first distortion is *Gaussian*

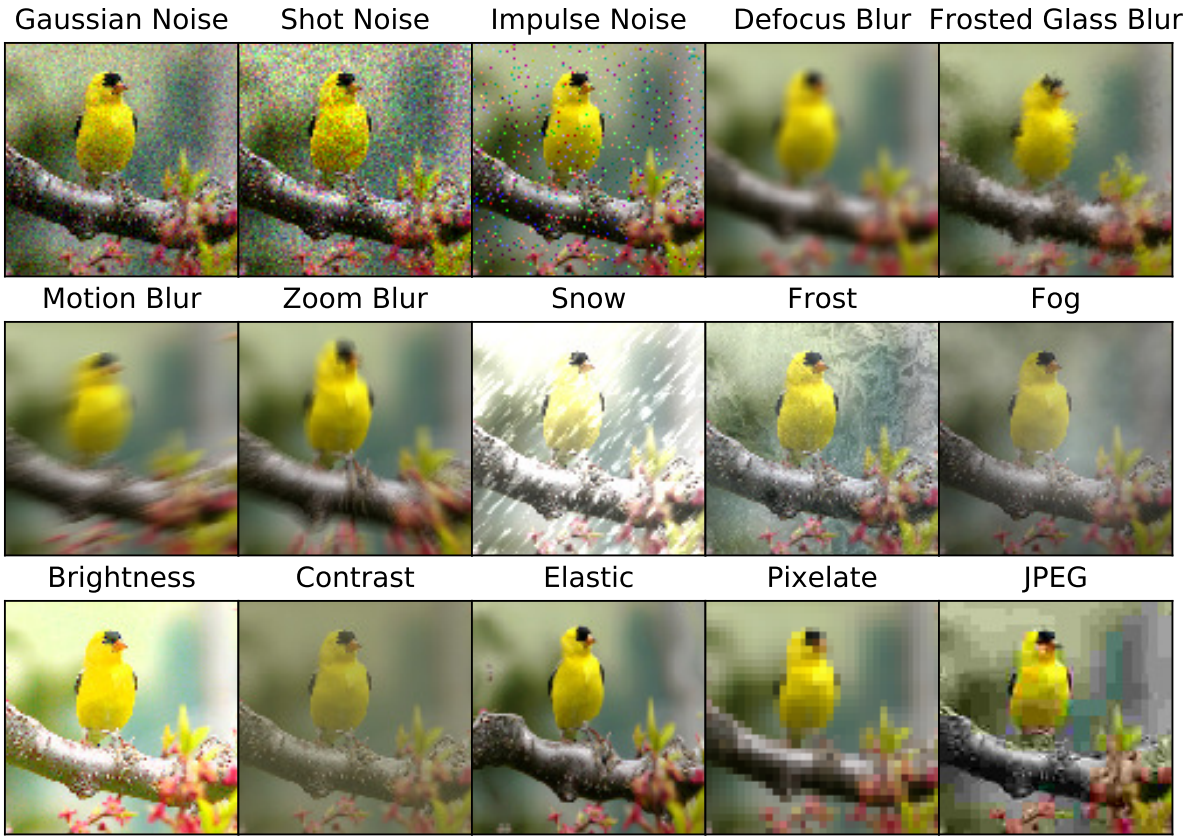

Figure 2: Our IMAGENET-D dataset consists of 15 types algorithmically generated distortions coming from noise, blur, weather, and digital categories. Each type of distortion has five levels of severity, resulting in 75 distinct distortions. When possible, each distortion has variation. For example, each fog cloud is unique to each image. We make this new dataset publicly available here.

noise, and this distortion can appear in low-lighting conditions. *Shot noise*, also called Poisson noise, is electronic noise caused by the discrete nature of light itself. *Impulse noise* is a color analogue of salt-and-pepper noise and can be caused by bit errors. *Defocus blur* occurs when an image is out of focus. *Frosted Glass Blur* appears with "frosted glass" windows or panels. *Motion blur* can occur when a camera is moving quickly; camera movement angles are between $[-45°, 45°]$. *Zoom blur* can occur when a camera moves toward an object quickly. *Snow* is a common form of precipitation, and when we render snow, it falls from $-135°$ to $-45°$. *Frost* can occur when lenses or windows are coated with ice crystals. *Fog* can reduce visibility and is generated with the diamond-square algorithm. *Brightness* can change due to sunlight intensity; we increase brightness by increasing the Value in HSV color space. *Contrast* can be high or low depending on lighting conditions and the photographed object's color. *Elastic* transformations stretch or contract image regions. *Pixelation* occurs when upsampling a low-resolution image. *JPEG* is a lossy image compression format that increases image pixelation and introduces artifacts. This broad range of distortions allow us to test model robustness with breadth, and with each distortion's five severity levels, our benchmark secures its depth.

3.2. Metric

AUDE. Gaussian noise can be invisible or destructive. For each distortion, we measure the classifier's performance across several distortion severity levels since real distortions appear with different intensities. First, we compute the clean dataset error rate. Then we record the error rate for the dataset distorted with five different severities. With these values, we compute the area under the distortion error curve (AUDE). The AUDE is our benchmark's metric. Figure 3 shows the how distortion severity increases error. At severity level 0, no Gaussian noise is applied so images are clean. At severity level 1, the MSSIM between the clean and noised image is roughly 0.9, on average. Likewise, a severity level of 5 corresponds to about a 0.5 MSSIM, on average. Our AUDE metric measures the classifier robust-

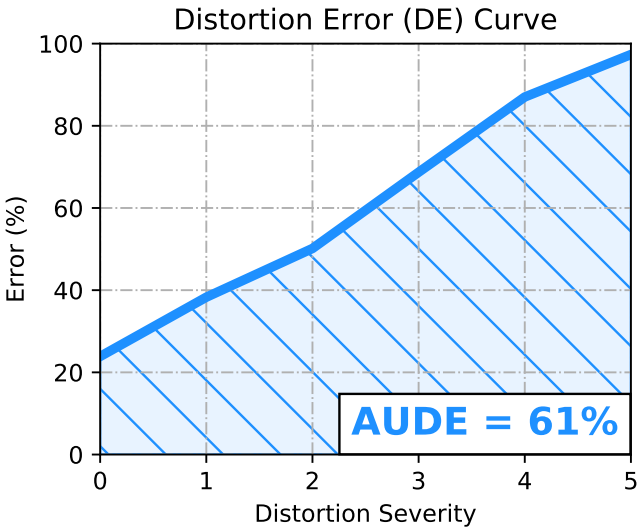

Figure 3: The Distortion Error Curve of a ResNet-50 for Gaussian noise applied to ImageNet images. Severity level 0 is the ResNet-50 performance on clean images. The Area Under the Distortion Error Curve (AUDE) is 61%.

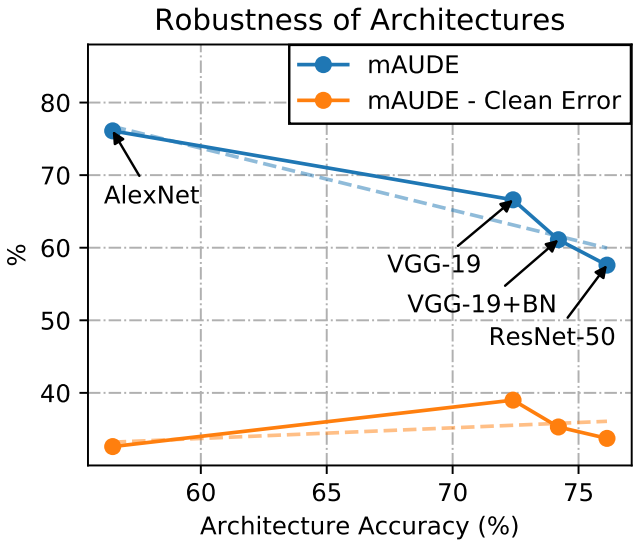

Figure 4: Robustness (mAUDE) and relative robustness (mAUDE − Clean Error) of various architectures on IMAGENET-D. mAUDE is the area under the distortion error curve averaged across all distortions. mAUDE − Clean Error shows that robustness gains are largely due to classification performance improvements. Here "BN" abbreviates Batch Normalization [26].

ness to a distortion by testing the classifier on the distortion at several severities. To summarize model robustness across several distortions, we average AUDE values, resulting in the *mean AUDE* or *mAUDE* for short.

Metric Validity. Preserving the usefulness of the metric requires that researchers not directly train on any of these 15 distortions. To reduce implicitly overfitting these distortions, we provide extra distortions with which to validate models. See Section 6. We discourage directly fitting the test distortions because we aim to benchmark how robustly a system generalizes, and fine-tuning a model on each distortion is not in the spirit of generalization to new settings. Note that demanding generalization to a novel distortion is a reasonable since, for example, humans can generalize to new Instagram filters with ease. Additionally, there is a bevy of other distortions for fine-tuning like uniform noise, fisheye lens distortion, and style transfer, and exponentially many combinations of other distortions. What is more is that fine-tuning on specific distortions tends not to provide generalization to new distortions [47, 15]. Even when a network learns to cope with a distortion like Gaussian noise, performance remains less robust than human performance for that distortion [11]. For those reasons, test distortions should remain unseen until test time.

3.3. Architecture Robustness

Have architectures become more robust since AlexNet [31]? In Table 1, as architectures improve, so too do their area under the Distortion Error Curve (AUDE). By this measure, architectures have become progressively more successful at generalizing to distorted distributions. Note that models of similar accuracy have similar AUDEs across different distortions. As such, we do not find that one architecture is exceptionally suited for motion blur and another for noise, nor do we find any relatively large shifts in a distortion's AUDE. Consequently, it would seem that architectures have slowly improved their representations over time.

However, it appears that robustness gains are mostly explained by classification accuracy gains. In Figure 4, we plot mAUDE − Error on Clean Data to visualize how error increases (as distortion severity increases) relative to clean data error. An implication is that the AUDE decreases mostly because the Distortion Error Curve shifts down with classification error reduction, not because the Error vs Distortion Severity graph is flatter. Consequently, from AlexNet to ResNet [17], robustness has barely improved. Then robustness has remained beneath human-level robustness, revealing our "superhuman" classifiers to be decidedly subhuman.

4. Increasing Distortion Robustness

4.1. Failed Attempts

Stability Training. Stability training a technique to improve the robustness of deep networks [50]. The method's

			Noise			Blur				Weather				Digital			
Network	Clean Error	**mAUDE**	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
AlexNet	43.5	76.1	81	82	85	81	78	75	79	81	73	81	68	79	60	75	69
SqueezeNet	41.8	79.5	89	89	92	80	81	76	79	82	74	77	66	76	65	80	87
VGG-19	27.6	66.6	70	73	80	73	77	65	70	68	60	57	47	54	56	79	68
VGG-19+BN	25.8	61.1	65	67	76	69	74	60	67	61	56	52	41	48	55	66	61
ResNet-18	30.2	63.9	68	70	75	69	72	64	69	68	60	63	47	55	53	62	63
ResNet-50	23.9	57.6	61	64	66	62	69	57	63	61	53	52	40	47	50	60	60

Table 1: Areas under the Distortion Error Curves across different distortions and architectures on IMAGENET-D. An mAUDE value is the mean AUDE, or the mean of the values in Noise, Blur, Weather, and Digital rows. All models are trained on clean ImageNet images. Here "BN" abbreviates Batch Normalization.

creators found that training on images distorted with Gaussian noise can lead to underfitting, so they instead propose minimizing the cross-entropy from the softmax distribution of the noisy image to the softmax of the clean image. The authors evaluated its performance on images with subtle differences and suggested that the method provides additional robustness to JPEG distortions. As this technique attempts to improve robustness, we benchmark this technique. Therefore, we fine-tune a ResNet-50 with stability training for five epochs. We distort images with uniform noise where the maximum and minimum of the uniform noise is tuned over $\{0.01, 0.05, 0.1\}$ and the stability weight is tuned over $\{0.01, 0.05, 0.1\}$. Across all noise strengths and stability weight combinations, the resulting model's mAUDE is greater than the baseline ResNet-50's mAUDE. Even on unseen noises, stability training did not increase robustness. An upshot of this failure is that benchmarking robustness-enhancing techniques requires a diverse test-set, lest the method only improve robustness on seen distortions.

Image Restoration. Modifying the model representations with stability training failed, so now we turn to modifying model inputs. An immediate idea is to restore the model inputs. Yet *general* image restoration techniques are pre-mature, but denoising restoration techniques are not. Then we can to restore an image with a denoising technique like non-local means [4]. However, we are not given the amount of noise within an image. To meet this challenge, we estimate the amount of noise with [12]. Combining noise estimation with a denoising technique, we restore images with non-local means, and the amount of denoising is informed by a noise estimation technique. Thus clean images receive nearly no modifications from the restoration method, while noisy images should undergo considerable restoration. But for all that effort, this image restoration technique increased the AUDE from 33.7% to 34.1%. A plausible explanation is that the non-local means algorithm slightly smoothed images even when images lacked noise, despite having the non-local means algorithm governed by the noise estimate. Therefore, the gains in noise robustness were wiped out by subtle blurs to images with other types of distortions. Consequently, a first pass at image restoration proved harmful for robustness.

10-Crop Classification. Viewing an object at several different locations may give way to a more stable prediction. Having this goal in mind, we perform 10-crop classification. 10-crop classification is executed by way of cropping all four corners and cropping the center of an image. These crops and their horizontal mirror are processed through a network. Given these ten probability distributions from the network, we average the distributions and finally issue our prediction. Of course, a prediction informed by 10-crops rather than a single central crop is more accurate. Ideally, this revised prediction should be more robust too. However, the gains in mAUDE do not outpace the gains in accuracy on a ResNet-50. In all, 10-crop classification is a computationally expensive option which contributes to classification accuracy but not noticeably to robustness.

Smaller Models. All else equal, "simpler" models are often expected to generalize better, and "simplicity" frequently translates to model size. Accordingly, smaller models may provide superior robustness. It is for this reason that we test SqueezeNet1.1 [25] and CondenseNet [23]. First, SqueezeNet obtains is orders of magnitude smaller in file size than the likes of VGGNet [44] and AlexNet. As it happens, SqueezeNet and AlexNet have similar error rates: 41.8% and 43.5% respectively. But mere gains in accuracy do not necessarily yield gains in robustness. In this instance, the reverse occurred. Squeezenet lost robustness and has an mAUDE of 79.5%, up from AlexNet's 76.1%. Furthermore, on some distortions like JPEG compression, the AUDE for SqueezeNet is approximately 18% worse than AlexNet.

At a higher tier of accuracy than SqueezeNet, CondenseNet maintains its small size by virtue of sparse convolutions and pruned filter weights. An off-the-shelf CondenseNet ($C = G = 4$) obtains a 26.3% error rate and a

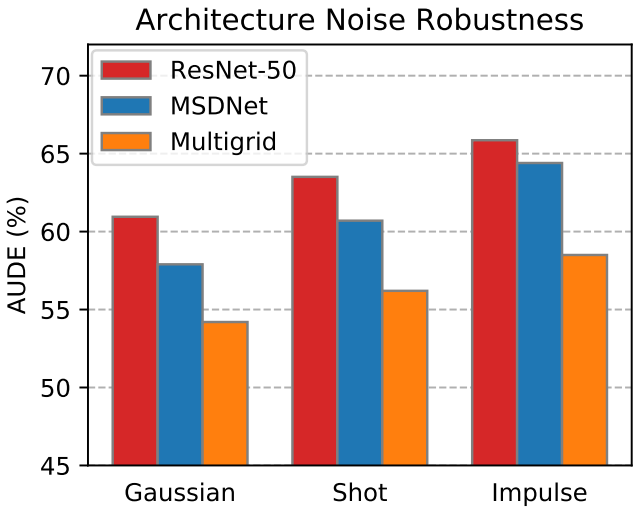

Figure 5: Architectures like Multigrid networks operate on representations across different scales and can more effectively resist noise distortions.

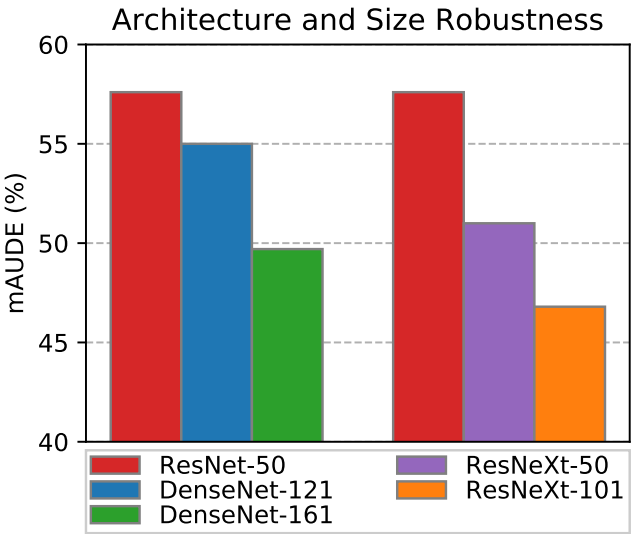

Figure 6: Larger networks can have robustness gains that substantially outpace their accuracy gains.

61.5% mAUDE. On the whole, this CondenseNet is slightly less robust than larger models on the same accuracy rung. Even more pruning and sparsification yields a CondenseNet ($C = G = 8$) with both deteriorated performance (28.9% error rate) and robustness (64.8% mAUDE). Here again robustness is worse than larger models. In consequence, models fashioned for mobile devices will be smaller and in some sense simpler, but this is not enough for robustness gains let alone its preservation.

4.2. Successful Robustness Enhancements

Multigrid Architectures. Having put forward then cast aside several plausible but flawed proposals, we finally find an architecture with greater robustness than ResNets. Multigrid convolutional neural networks [29] possess greater robustness by amalgamating information across scale at each layer rather than slowly gaining a global representation of the input like in typical convolutional neural networks. A pyramid of grids in each layer enables the subsequent layer to operate across scales. Along similar lines, Multi-Scale Dense Networks (MSDNets) [22] use information across scales in a different manner. Distinctly, MSDNets bind network layers with DenseNet-like [24] skip connections, while Multigrid networks do without dense connections and may propagate features with residual connections.

These two different multiscale networks both enhance robustness. Before comparing robustness, we first must note the Multigrid network has 24.6% top-1 error, as does the MSDNet, while the ResNet has 23.9% top-1 error. With those similar error rates specified, we find that on noisy inputs Multigrid networks surpass ResNets and MSDNets, as depicted in Figure 5. Since these multiscale and Multigrid architectures have high-level representations processed in tandem with fine details, the architectures appear better equipped to suppress otherwise distracting pixel noise. When all distortions are considered, we compute that the ResNet has an mAUDE of 57.6%, the MSDNet has an mAUDE of 56.1%, and the Multigrid network has an mAUDE of 55.1%. Consequently, original architectures are an avenue to more robust models.

Histogram Equalization. So far attempts at input cleaning do sanitize the image of a few distortions, but the attempts do not help with other distortions. In fact, our previous cleaning attempts exacerbate other distortions and degrade clean inputs, overwhelming the robustness gains plucked from erasing specific distortions. We then come to a different technique that standardizes images corrupted by arbitrary distortions. These images are standardized by histogram equalization which spreads image intensity. Histogram equalization is used to standardize speech data for robust speech recognition [2, 16], and we make this input cleaning technique suitable for images by using Contrast Limited Adaptive Histogram Equalization [41] to the input before feeding the input to the network. CLAHE reduces the effect of some distortions while not hindering performance on most others. To see this, note that CLAHE improves the mAUDE, unlike techniques such as median blur which can remove impulse noise but magnify other distortions and increase the mAUDE. To demonstrate CLAHE's net improvement, we take a pretrained ResNet-

Figure 7: Image samples are from 12 of 50 ICONS-50 classes. These images showcase the dataset's high image quality, interclass and intraclass diversity, and its many styles. We make this new dataset publicly available at this URL.

50 and fine-tune it on inputs with CLAHE for five epochs. The ResNet-50 has a 23.87% error rate, but ResNet50 with CLAHE has an error rate of 23.55%. On some distortions, CLAHE truncates the distortion. For example, on impulse noise the AUDE decreases from 66% to 57% by applying CLAHE. Overall, a the ResNet-50 without CLAHE has an mAUDE of 57.6%, while with CLAHE the ResNet-50's mAUDE is trimmed to 55.4%.

Increased Feature Aggregation and Model Size. Several recent architectures fine-tune the ResNet architecture by increasing feature aggregation. Of these, DenseNets [24] and ResNeXts [48] are most prominent. Each purports to have stronger representations than ResNets, and the evidence is largely a hard-won ImageNet accuracy uptick. However, we find that the representations are stronger than the ImageNet accuracy gains would suggest. Whereas ImageNet accuracy measures are nearly saturated, our robustness measure is not. So, our robustness benchmark clearly indicates that DenseNets and ResNeXts have superior representations. Accordingly, a switch from a ResNet-50 (23.9% top-1 error) to a DenseNet-121 (25.6% top-1) decreases the IMAGENET-D mAUDE from 57.6% to 55.0%. More starkly, switching from a ResNet-50 to a ResNeXt-50 (22.9% top-1) drops the mAUDE from 57.6% to 51.0%. The top-1 ImageNet accuracies gains underrated but IMAGENET-D appropriately corroborated the higher caliber of DenseNet and ResNeXt representations.

Some of the greatest and simplest robustness gains sometimes emerge from making recent models more monolithic. Apparently more layers, more connections, and more capacity allow these massive models to operate more stably on distorted inputs. In point of fact, the replacement of a DenseNet-121 (25.6% top-1) with a DenseNet-161 (22.9% top-1) sinks the mAUDE from 55.0% to 49.7%. In a similar fashion, a ResNeXt-50 (22.9% top-1) is less robust than the a giant ResNeXt-101 (21.0% top-1) since the mAUDEs are 51.0% and 46.8% respectively. Both model size and feature aggregation results are summarized in Figure 6. Consequently, future models which make use of more depth, width, and feature aggregation may advance future robustness improvements.

5. Testing Style Robustness with ICONS-50

Originally, it will be recalled, distortion robustness and IMAGENET-D were introduced to benchmark average-case distortions rather than adversarial worst-case distortions. What was tested were 15 distortions which introduced artifacts, ruffled pixels, cluttered textures, and yet preserved the underlying image structure—the noisy bird image still has the yellow bird, upright and perched. We wish to distort the higher-level image structure and have the model be robust to these changes too. At a first pass, we might think to render the bird with spread wings, but this part of the clean

image distribution. A better idea is to immerse the yellow bird in the swirling scenery of *The Starry Night* by applying style transfer [14] to the image. Separately, the bird could be rendered as a cartoon. Third, consider data in a different domain like audio. One could test of style robustness by playing music pieces on different instruments, thus creating different styles, and classifiers would be expected to generalize to pieces played on different instruments. Each would distort the higher-level structure, and classifiers would do well to be robust to these style changes. Since humans can generalize to different stylizations of real objects, a goal parallel to average-case and worst-case distortion robustness is style robustness.

5.1. The ICONS-50 Dataset

Given that classifiers should be capable of generalizing to new styles, we need a way to test and improve style robustness. And seeing an absence of a style robustness dataset, we create one called ICONS-50. The new ICONS-50 dataset covers icons of animals, people, food, activity, places, objects, and symbols, for a total of 50 diverse classes. A full list of ICONS-50 classes are in Appendix A. Now, a subset of ICONS-50 classes is visible in Figure 7. As evident in the figure, each class consists icons with different styles. For example, icons with the thick, black outlines (like the bottom right "Drink" icon) are stylized by Microsoft for their operating systems. Other styles in the ICONS-50 dataset are from Apple, Samsung, Google, Facebook, and other platforms. With these various icon styles at hand, we can test style robustness by simply holding out, say, Microsoft-styled icons and training on all remaining icons. ICONS-50 consists in these icons, and this style robustness dataset is publicly available to the research community at this URL.

A given dataset should be dense, in that they have numerous examples per class. We note that ICONS-50 has more density than related datasets. The most comparable existing datasets are logo datasets. The dataset of [43] has approximately 70 logos per class, with 32 total classes. Likewise, the dataset of [27] has approximately 53 logos per class, with 37 total classes. Contrariwise, ICONS-50 has approximately 164 icons per class, with 50 total classes. This density is not artificial and is balanced. Duplicates icons were carefully removed, and classes with inordinately many icons were downsized. Relative to logo datasets, ICONS-50 has over twice or thrice the density.

Beyond dataset density is diversity. The datasets of [43] and [27] include the same logo under different lighting and backgrounds, while ICONS-50 has several different stylings and schemas for each class. For example, a plain class like "Cloud" has icons in which clouds cooccur with the sun, lightning, rain, and snow. And each of these different schemas include different styles (e.g., Apple's realistic style, Microsoft's flat vector graphics style, etc.) and they include different renditions using that style (icon files changed across operating system versions). In summary ICONS-50 has relatively greater dataset density and diversity.

5.2. Lack of Style Robustness

In what follows, we train a network on ICONS-50, hold out Microsoft icons, and test on the held-out icons. Holding out Microsoft icons allows us to test a network's style robustness because the network has trained on icons fashioned with several styles, and then it must generalize to a new style. To accomplish this robustness experiment, we use a 16-4 Wide Residual Network [49]. Trained for 50 epochs with the cosine learning rate schedule [35], this network learned with a dropout [45] rate of 30%. Like others, we use image cropping and flipping data augmentation, as well as color jittering and Random Erasure data augmentation [51] for further regularization. What we find is that the network lacks extensive style robustness—the network only obtains 55% accuracy on the held-out Microsoft-styled icons. This result is not symptomatic of a training data shortage; if we hold out Apple icons instead of Microsoft icons and test on Apple icons, the accuracy is over 90%. With only 55% accuracy on Microsoft icons, the network demonstrates a clear lacking in style robustness.

6. Conclusion

In this paper, we introduced IMAGENET-D: a complete robustness benchmark that standardizes and expands research in this space. The dataset showed that many years of architectural advancements gave rise to robustness gains that were mostly explained by accuracy increases, and that there were minuscule robustness gains in itself during those years. Still with wide room for improvement, this benchmark enables architectures with better representations to demonstrate their superiority, rather than have their gains shrouded in the noise by a nearly saturated metric like top-1 ImageNet error. To give a sense of both promising and also fruitless avenues to improved robustness, we first considered four reasonable but ultimately disappointing methods. Thereafter, we investigated obscure architectures like Multigrid architectures and found robustness gains, especially on noisy inputs. Then, taking a queue from years of robustness research in automatic speech recognition, we found that image histogram equalization causes greater distortion robustness. Afterward we established that robustness sharply improves simply by making recent architectures more massive. Shifting attention to a twin goal of distortion robustness, we created a new dataset called ICONS-50 that opens research in style robustness and found modern networks to be lacking in style robustness. These datasets allow for effective study of model robustness, a necessity as models are unleashed into the real world.

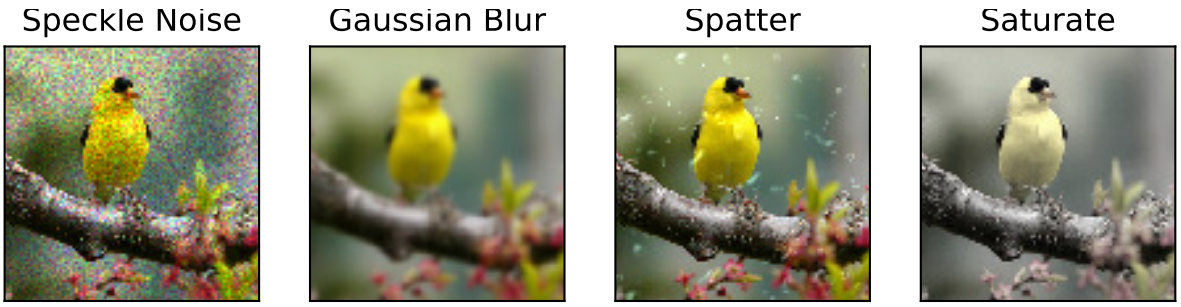

Figure 8: Extra IMAGENET-D distortion examples are available for model validation and sounder experimentation.

A. Extra IMAGENET-D Distortions

Directly fitting the types of IMAGENET-D distortions is worth avoiding, as it would cause researchers to overestimate a model's robustness. Therefore, it is incumbent on us to simplify model validation. For this reason, we provide extra distortions that are available for download here. There is one distortion type for each Noise, Blur, Weather, and Digital category. The first distortion type is *speckle noise*, an additive noise where the noise added to a pixel tends to grow as the pixel intensity increases. *Gaussian blur* is a low-pass filter where a blurred pixel is a result of a weighted average of its neighbors, and farther pixels have less weight. *Spatter* can occlude a lens in the form of rain or mud. Finally, *saturate* is common in edited images where images are made more or less colorful. See Figure 8 for instances of each distortion type.

B. Classes in ICONS-50

The 50 classes of ICONS-50 are as follows: Airplane, Drink, Arrow Directions, Automobile, Ball, Biking, Boat, Books, Bunny Ears, Cartwheeling, Cat, Chick, Clock, Cloud, Disk, Emotion Face, Envelope, Factory Worker, Family, Fast Train, Flag, Flower, Golfing, Hand, Heart, Holding Hands, House, Japanese Ideograph, Judge, Kiss, Lock, Mailbox, Mechanic, Medal, Money, Monkey, Moon, Mountain, Numbers, Phone, Prohibit Sign, Rowing, Scientist, Shoe, Surfing, Tree, Umbrella, Water Polo, Wrestling, Writing Utensil.

References

[1] O. Abdel-Hamid, A. rahman Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *ICASSP*, pages 4277–4280. IEEE. 2

[2] Ángel de la Torre, A. Peinado, J. Segura, J. Pérez-Córdoba, M. C. Benítez, and A. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Signal Processing Society*, 2005. 2, 6

[3] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi. Measuring neural net robustness with constraints. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2613–2621. Curran Associates, Inc., 2016. 2

[4] A. Buades and B. Coll. A non-local algorithm for image denoising. In *CVPR 2005*, 2005. 5

[5] N. Carlini, G. Katz, C. Barrett, and D. L. Dill. Ground-truth adversarial examples, 2017. 2

[6] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples, 2016. 2

[7] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks, 2016. 2

[8] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods, 2017. 2

[9] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR 2009*, 2009. 2

[10] S. Dodge and L. Karam. Quality resilient deep neural networks, 2017. 2

[11] S. Dodge and L. Karam. A study and comparison of human and deep learning recognition performance under visual distortions, 2017. 2, 4

[12] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 1993. 5

[13] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on deep learning models, 2017. 2

[14] L. Gatys, A. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. *CVPR 2016*, 2016. 8

[15] R. Geirhos, D. H. J. Janssen, H. H. Schtt, J. Rauber, M. Bethge, and F. A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker, 2017. 2, 4

[16] M. Harvilla and R. Stern. Histogram-based subband powerwarping and spectral averaging for robust speech recognition under matched and multistyle training, 2012. 2, 6

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR 2016*, 2015. 4

[18] D. Hendrycks and K. Gimpel. Early methods for detecting adversarial images, 2016. 2

[19] H.-G. Hirsch. Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments, 2007. 2

[20] H.-G. Hirsch and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ISCA ITRW ASR2000*, 2000. 2

[21] H. Hosseini, B. Xiao, and R. Poovendran. Google's cloud vision api is not robust to noise, 2017. 2

[22] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger. Multi-scale dense networks for resource efficient image classification. *ICLR 2018*, 2017. 6

[23] G. Huang, S. Liu, L. van der Maaten, and K. Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions. *arXiv preprint*, 2017. 5

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6, 7

[25] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. 2016. 5

[26] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *JMLR 2015*, 2015. 4

[27] A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. *ACM International Conference on Multimedia Retrieval 2009*, 2009. 8

[28] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Towards proving the adversarial robustness of deep neural networks. 2017. 2

[29] T.-W. Ke, M. Maire, and S. X. Yu. Multigrid neural architectures, 2016. 6

[30] C. Kim and R. M. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(7):1315–1329, July 2016. 2

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS 2012*. 4

[32] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ICLR 2017*, 2017. 2

[33] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. An overview of noise-robust automatic speech recognition. volume 22, pages 745 – 777. IEEE Institute of Electrical and Electronics Engineers, April 2014. 2

[34] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *Proc. of DARPA Speech and Natural Language Workshop*, pages 69–74, March 1993. 2

[35] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 8

[36] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. Standard detectors aren't (currently) fooled by physical adversarial stop signs, 2017. 2

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2017. 2

[38] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations, 2017. 2

[39] V. Mitra, H. Franco, R. Stern, J. V. Hout, L. Ferrer, M. Graciarena, W. Wang, D. Vergyri, A. Alwan, and J. H. Hansen. Robust features in deep learning based speech recognition, 2017. 2

[40] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks, 2017. 2

[41] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. T. H. Romeny, and J. B. Zimmerman. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 1987. 6

[42] J. Rauber, W. Brendel, and M. Bethge. Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models, 2017. 2

[43] S. Romberg, L. Pueyo, R. Lienhart, and R. van Zwol. Scalable logo recognition in real-world images. *ACM International Conference on Multimedia Retrieval 2011*, 2011. 8

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR 2015*, 2015. 5

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR 2014*, 2014. 8

[46] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2013. 1, 2

[47] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich. Examining the impact of blur on recognition by convolutional networks, 2016. 2, 4

[48] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CVPR 2017*, 2016. 7

[49] S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, 2016. 8

[50] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training, 2016. 2, 4

[51] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 8