Adversarial Examples Are a Natural Consequence of Test Error in Noise

Nicolas Ford [*] [1] [2] **Justin Gilmer** [*] [1] **Nicholas Carlini** [1] **Ekin D. Cubuk** [1]

Abstract

Over the last few years, the phenomenon of *adversarial examples* — maliciously constructed inputs that fool trained machine learning models — has captured the attention of the research community, especially when the adversary is restricted to small modifications of a correctly handled input. Less surprisingly, image classifiers also lack human-level performance on randomly corrupted images, such as images with additive Gaussian noise. In this paper we provide both empirical and theoretical evidence that these are two manifestations of the same underlying phenomenon. We establish close connections between the adversarial robustness and corruption robustness research programs, with the strongest connection in the case of additive Gaussian noise. This suggests that improving adversarial robustness should go hand in hand with improving performance in the presence of more general and realistic image corruptions. Based on our results we recommend that future adversarial defenses consider evaluating the robustness of their methods to distributional shift with benchmarks such as ImageNet-C.

1. Introduction

State-of-the-art computer vision models can achieve impressive performance on many image classification tasks. Despite this, these same models still lack the robustness of the human visual system to various forms of image corruptions. For example, they are distinctly subhuman when classifying images distorted with additive Gaussian noise (Dodge & Karam, 2017), they lack robustness to different types of blur, pixelation, and changes in brightness (Hendrycks & Dietterich, 2019), lack robustness to random translations of the input (Azulay & Weiss, 2018), and even make errors when foreign objects are inserted into the field of view (Rosenfeld et al., 2018). At the same time, they are also sensitive to small, worst-case perturbations of the input, so-called "adversarial examples" (Szegedy et al., 2014). This latter phenomenon has struck many in the machine learning community as surprising and has attracted a great deal of research interest, while the former has received less attention.

The machine learning community has researchers working on each of these two types of errors: adversarial example researchers seek to measure and improve robustness to small-worst case perturbations of the input while corruption robustness researchers seek to measure and improve model robustness to distributional shift. In this work we analyze the connection between these two research directions, and we see that adversarial robustness is closely related to robustness to certain kinds of distributional shift. In other words, the existence of adversarial examples follows naturally from the fact that our models have nonzero test error in certain corrupted image distributions.

We make this connection in several ways. First, in Section 4, we provide a novel analysis of the error set of an image classifier. We see that, given the error rates we observe in Gaussian noise, the small adversarial perturbations we observe in practice appear at roughly the distances we would expect from a *linear* model, and that therefore there is no need to invoke any strange properties of the decision boundary to explain them. This relationship was also explored in Fawzi et al. (2018b; 2016).

In Section 5, we show that improving an alternate notion of adversarial robustness *requires* that error rates under large additive noise be reduced to essentially zero.

Finally, this suggests that methods which increase the distance to the decision boundary should also improve robustness to Gaussian noise, and vice versa. In Section 6 we confirm that this is true by examining both adversarially trained models and models trained with additive Gaussian noise. We also show that measuring corruption robustness can effectively distinguish successful adversarial defense methods from ones that merely cause vanishing gradients.

We hope that this work will encourage both the adversarial and corruption robustness communities to work more closely together, since their goals seem to be so closely related. In particular, it is not common for adversarial defense

[*] Equal contribution [1] Google Brain [2] This work was completed as part of the Google AI Residency. Correspondence to: Nicolas Ford <nicf@google.com>, Justin Gilmer <gilmer@google.com>.

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

methods to measure corruption robustness. Given that successful adversarial defense methods should also improve some types of corruption robustness we recommend that future researchers consider evaluating corruption robustness in addition to adversarial robustness.

2. Related Work

Adversarial machine learning studies general ways in which an adversary may interact with an ML system, and dates back to 2004 (Dalvi et al., 2004; Biggio & Roli, 2018). Since Szegedy et al. (2014), a subfield has focused specifically on small adversarial perturbations of the input, or "adversarial examples." Many algorithms have been developed to find the smallest perturbation in input space which fool a classifier (Carlini & Wagner, 2017; Madry et al., 2017). Defenses have been proposed for increasing the robustness of classifiers to small adversarial perturbations, however many have later been shown ineffective (Carlini & Wagner, 2017). To our knowledge the only method which has been confirmed by a third party to increase l_p-robustness (for certain values of ϵ) is adversarial training (Madry et al., 2017). However, this method remains sensitive to slightly larger perturbations (Sharma & Chen, 2017).

Several recent papers (Gilmer et al., 2018b; Mahloujifar et al., 2018; Dohmatob, 2018; Fawzi et al., 2018a) use *concentation of measure* to prove rigorous upper bounds on adversarial robustness for certain distributions in terms of test error, suggesting non-zero test error may imply the existence of adversarial perturbations. This may seem in contradiction with empirical observations that increasing small perturbation robustness tends to reduce model accuracy (Tsipras et al., 2019). We note that these two conclusions do not necessarily contradict each other. It could be the case that hard bounds on adversarial robustness in terms of test error exist, but current classifiers have yet to approach these hard bounds.

Because we establish a connection between adversarial robustness and model accuracy in *corrupted* image distributions, our results do not contradict reports that adversarial training reduces accuracy in the *clean* distribution (Tsipras et al., 2019). In fact, we find that improving adversarial robustness also improves corruption robustness.

3. Adversarial and Corruption Robustness

Both adversarial robustness and corruption robustness can be thought of as functions of the **error set** of a statistical classifier. This set, which we will denote E, is the set of points in the input space on which the classifier makes an incorrect prediction. In this paper we will only consider perturbed versions of training or test points, and we will always assume the input is corrupted such that the "correct" label for the corrupted point is the same as for the clean point. This assumption is commonly made in works which study model robustness to random corruptions of the input (Hendrycks & Dietterich, 2019; Dodge & Karam, 2017).

Because we are interested in how our models perform on both clean images and corrupted ones, we introduce some notation for both distributions. We will write p for the *natural* image distribution, that is, the distribution from which the training data was sampled. We will use q to denote whichever *corrupted* image distribution we are working with. A sample from q will always look like a sample from p with a random corruption applied to it, like some amount of Gaussian noise. Some examples of noisy images can be found in Figure 10 in the appendix.

We will be interested in two quantities. The first, **corruption robustness** under a given corrupted image distribution q, is $\mathbb{P}_{x\sim q}[x \notin E]$, the probability that a random sample from the q is not an error. The second is called **adversarial robustness**. For a clean input x and a metric on the input space d, let $d(x, E)$ denote the distance from x to the nearest point in E. The adversarial robustness of the model is then $\mathbb{P}_{x\sim p}[d(x, E) > \epsilon]$, the probability that a random sample from p is not within distance ϵ of some point in the error set. When we refer to "adversarial examples" in this paper, we will always mean these nearby errors.

In this work we will investigate several different models trained on the CIFAR-10 and ImageNet datasets. For CIFAR-10 we look at the naturally trained and adversarially trained models which have been open-sourced by Madry et al. (2017). We also trained the same model on CIFAR-10 with Gaussian data augmentation. For ImageNet, we investigate an Inception v3 (Szegedy et al., 2016) trained with Gaussian data augmentation. In all cases, Gaussian data augmentation was performed by first sampling a σ uniformly between 0 and some specified upper bound and then adding random Gaussian noise at that scale. Additional training details can be found in Appendix A. We were unable to study the effects of adversarial training on ImageNet because no robust open sourced model exists. (The models released in Tramèr et al. (2017) only minimally improve robustness to the white box PGD adversaries we consider here.)

4. Errors in Gaussian Noise Suggest Adversarial Examples

We will start by examining the relationship between adversarial and corruption robustness in the case where q consists of images with additive Gaussian noise.

The Linear Case. For linear models, the error rate in Gaussian noise exactly determines the distance to the decision boundary. This observation was also made in Fawzi et al. (2016; 2018b).

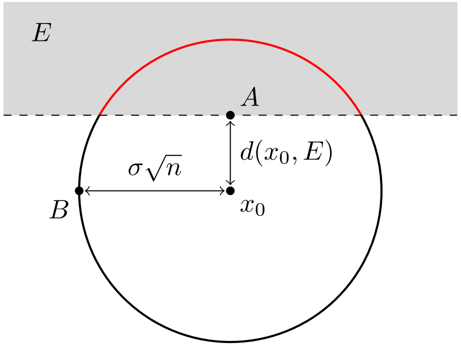

Figure 1. When the input dimension, n, is large and the model is linear, even a small error rate in additive noise implies the existence of small adversarial perturbations. For a point x_0 in image space, most samples from $\mathcal{N}(x_0; \sigma^2 I)$ (point B) lie close to a sphere of radius $\sigma\sqrt{n}$ around x_0, drawn here as a circle. For a linear model the error set E is a half-space, and the error rate μ is approximately equal to the fraction of the sphere lying in this half-space. The distance $d(x_0, E)$ from x_0 to its nearest error (point A) is also drawn. Note the relationship between σ, μ, and $d(x_0, E)$ does not depend on the dimension. However, because the typical distance to a sample from the Gaussian is $\sigma\sqrt{n}$ the ratio between the distance from x_0 to A and the distance from x_0 to B shrinks as the dimension increases.

It will be useful to keep the following intuitive picture in mind. In high dimensions, most samples from the Gaussian distribution $\mathcal{N}(x_0; \sigma^2 I)$ lie close to the surface of a sphere of radius σ centered at x_0. The decision boundary of a linear model is a plane, and since we are assuming that the "correct" label for each noisy point is the same as the label for x_0, our error set is simply the half-space on the far side of this plane.

The relationship between adversarial and corruption robustness corresponds to a simple geometric picture. If we slice a sphere with a plane, as in Figure 1, the distance to the nearest error is equal to the *distance* from the plane to the center of the sphere, and the corruption robustness is the fraction of the *surface area* cut off by the plane. This relationship changes drastically as the dimension increases: most of the surface area of a high-dimensional sphere lies very close to the equator, which means that cutting off even, say, 1% of the surface area requires a plane which is very close to the center. Thus, for a linear model, even a relatively small error rate on Gaussian noise implies the existence of errors very close to the clean image (i.e., an adversarial example).

To formalize this relationship, pick some clean image x_0 and consider the Gaussian distribution $\mathcal{N}(x_0; \sigma^2 I)$. For a fixed μ, let $\sigma(x_0, \mu)$ be the σ for which the error rate is μ, that is, for which

$$\mathbb{P}_{x \sim \mathcal{N}(x_0; \sigma^2 I)}[x \in E] = \mu.$$

Then, letting d denote l_2 distance, we have

$$d(x_0, E) = -\sigma(x_0, \mu)\Phi^{-1}(\mu), \tag{1}$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-x^2/2)dx$$

is the cdf of the univariate standard normal distribution. (Note that $\Phi^{-1}(\mu)$ is negative when $\mu < \frac{1}{2}$.)

This expression depends only on the error rate μ and the standard deviation σ of a single component, and not directly on the dimension, but the dimension appears if we consider the distance from x_0 to a typical sample from $\mathcal{N}(x_0; \sigma^2 I)$, which is $\sigma\sqrt{n}$. When the dimension is large the distance to the decision boundary will be significantly smaller than the distance to a noisy image.

For example, this formula says that a linear model with an error rate of 0.01 in noise with $\sigma = 0.1$ will have an error at distance about 0.23. In three dimensions, a typical sample from this noise distribution will be at a distance of around $0.1\sqrt{3} \approx 0.17$. However when $n = 150528$, the dimension of the ImageNet image space, these samples lie at a distance of about 38.8. So, in the latter case, a 1% error rate on random perturbations of size 38.8 implies an error at distance 0.23, more than 150 times closer. Detailed curves showing this relationship can be found in Appendix F.

Comparing Neural Networks to the Linear Case. The decision boundary of a neural network is, of course, not linear. However, by comparing the ratio between $d(x_0, E)$ and $\sigma(x_0, \mu)$ for neural networks to what it would be for a linear model, we can investigate the relationship between adversarial and corruption robustness. We ran experiments on several neural network image classifiers and found results that closely resemble Equation 1. Adversarial examples therefore are not "surprisingly" close to x_0 given the performance of each model in Gaussian noise.

Concretely, we examine this relationship when $\mu = 0.01$. For each test point, we compare $\sigma(x_0, 0.01)$ to an estimate of $d(x_0, E)$. Because it is not feasible to compute $d(x_0, E)$ exactly, we instead search for an error using PGD (Madry et al., 2017) and report the nearest error we can find.

Figure 2 shows the results for several CIFAR-10 and ImageNet models, including ordinarily trained models, models trained with Gaussian data augmentation with $\sigma = 0.4$, and an adversarially trained CIFAR-10 model. We also included a line representing how these quantities would be related for a linear model, as in Equation 1. Because most test points lie close to the predicted relationship for a linear model, we see that the half-space model shown in Figure 1 accurately predicts the existence of small perturbation adversarial examples.

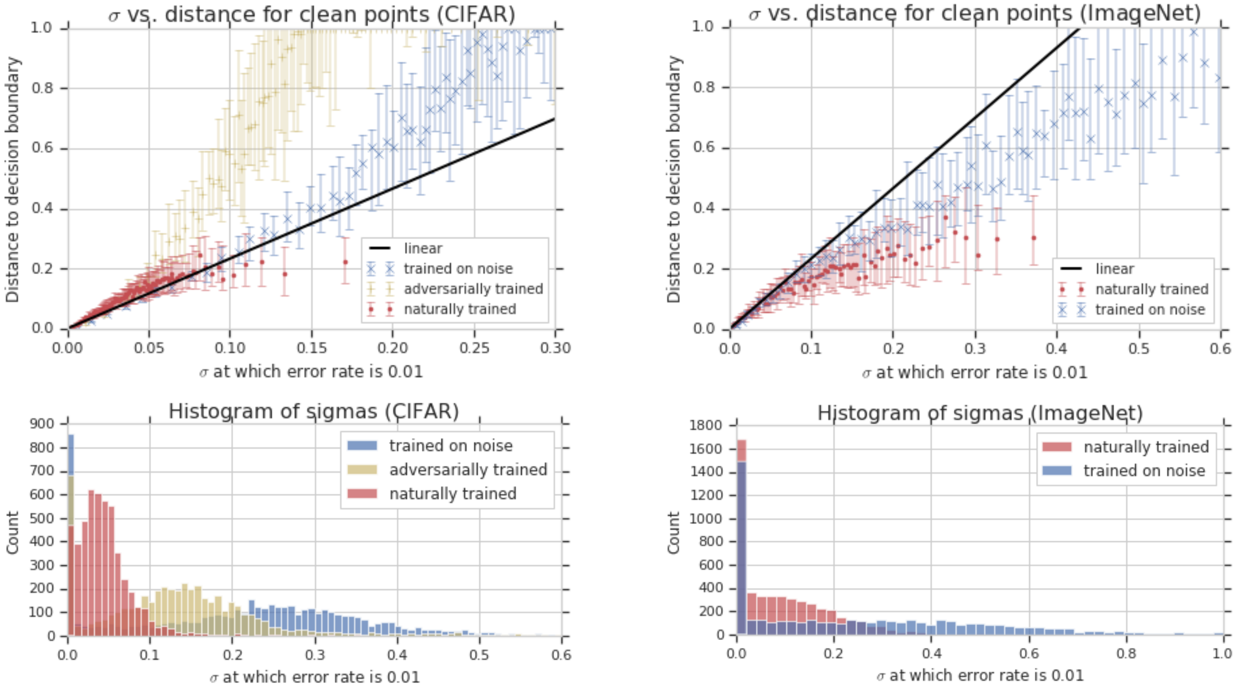

Figure 2. (Top) Comparing the l_2 distance to the decision boundary with the σ for which the error rate in Gaussian noise is 1%. Each point represents 50 images from the test set, and the median values for each coordinate are shown. The error bars cover the 25th to 75th percentiles. The PGD attack was run with $\epsilon = 1$, so the distances to the decision boundary reported here are cut off at 1. *(Bottom)* Histograms of the x coordinates from the above plots. A misclassified point is assigned $\sigma = 0$.

It is interesting to observe how each training procedure affected the two quantities we measured. First, adversarial training and Gaussian data augmentation increased *both* $\sigma(x_0, 0.01)$ and $d(x_0, E)$ on average. The adversarially trained model deviates from the linear case the most, but it does so in the direction of *greater* distances to the decision boundary. While both augmentation methods do improve both quantities, Gaussian data augmentation had a greater effect on σ (as seen in the histograms) while adversarial training had a greater effect on d. We explore this further in Section 6.

Visual Confirmation of the Half-space Model In Figure 3 we draw two-dimensional slices in image space through three points. (Similar visualizations have appeared in Fawzi et al. (2018b), and are called "church window plots.")

This visualized decision boundary closely matches the half-space model in Figure 1. We see that an error found in Gaussian noise lies in the same connected component of the error set as an error found using PGD, and that at this scale that component visually resembles a half-space. This figure also illustrates the connection between adversarial example research and corruption robustness research. To measure adversarial robustness is to ask whether or not there are any errors in the l_∞ ball — the small diamond-shaped region in the center of the image — and to measure corruption robustness is to measure the volume of the error set in the defined noise distribution. At least in this slice, nothing distinguishes the PGD error from any other point in the error set apart from its proximity to the clean image.

We give many more church window plots in Appendix G.

5. Concentration of Measure for Noisy Images

There is an existing research program (Gilmer et al., 2018b; Mahloujifar et al., 2018; Dohmatob, 2018) which proves hard upper bounds on adversarial robustness in terms of the error rate of a model. This phenomenon is sometimes called *concentration of measure*. Because proving a theorem like this requires understanding the distribution in question precisely, these results typically deal with simple "toy" distributions rather than those corresponding to real data. In this section we take a first step toward bridging this gap. By comparing our models to a classical concentration of measure bound for the Gaussian distribution, we gain another perspective on our motivating question.

The Gaussian Isoperimetric Inequality. As in Section 4, let x_0 be a correctly classified image and consider the distribution $q = \mathcal{N}(x_0; \sigma^2 I)$. Note q is the distribution of random Gaussian perturbations of x_0. The previous section discussed the distance from x_0 to its nearest error. In this

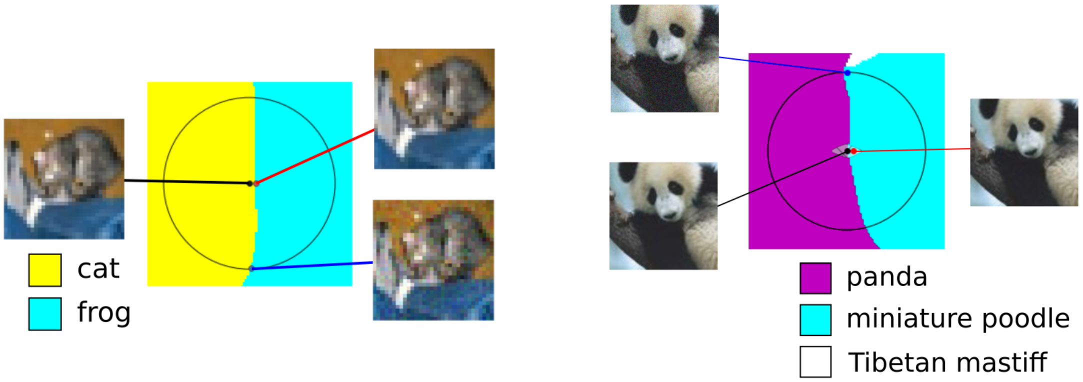

Figure 3. Two-dimensional slices of image space together with the classes assigned by trained models. Each slice goes through three points, a clean image from the test set (black), an error found by randomly perturbing the center image with Gaussian noise (blue), and an error found using a targeted PGD attack (red). The black circles have radius $\sigma\sqrt{n}$, indicating the typical size of the Gaussian perturbation used. The diamond-shaped region in the center of the right image shows the l_∞ ball of radius $8/255$. In both slices, the decision boundary resembles a half-space as predicted in Figure 1, demonstrating how non-zero error rate in noise predicts the existence of small adversarial perturbations. The CIFAR-10 model on the left was evaluated with $\sigma = 0.04$ (black circle has radius 2.22), where 0.21% of Gaussian perturbations are classified as "frog" (cyan region). The adversarial error was found at distance 0.159 while the half-space model predicts errors at distance 0.081. The ImageNet model on the right was evaluated at $\sigma = 0.08$ (black circle has radius 31.4) where 0.1% of Gaussian perturbations were misclassified as "miniture poodle" (cyan). The adversarial error has distance 0.189 while the half-space model predicts errors at distance 0.246. For the panda picture on the right we also found closer errors than what is shown by using an untargeted attack (an image was assigned class "indri" at distance 0.024). Slices showing more complicated behavior can be found in Appendix G.

section we will instead discuss the distance from a typical sample from q (e.g. point B in Figure 1) to its nearest error.

For random samples from q, there is a precise sense in which small adversarial perturbations exist only because test error is nonzero. That is, given the error rates we actually observe on noisy images, most noisy images *must* be close to the error set. This result holds completely independently of any assumptions about the model and follows from a fundamental geometric property of the Gaussian distribution, which we will now make precise.

Let $\epsilon_q^*(E)$ be the median distance from one of these noisy images to the nearest error. (In other words, it is the ϵ for which $\mathbb{P}_{x\sim q}[d(x, E) \leq \epsilon] = \frac{1}{2}$.) As before, let $\mathbb{P}_{x\sim q}[x \in E]$ be the probability that a random Gaussian perturbation of x_0 lies in E. It is possible to deduce a bound relating these two quantities from the *Gaussian isoperimetric inequality* (Borell, 1975). The form we will use is:

Theorem (Gaussian Isoperimetric Inequality)**.** *Let $q = \mathcal{N}(0; \sigma^2 I)$ be the Gaussian distribution on $\mathbb{R}^n$ with variance $\sigma^2 I$, and, for some set $E \subseteq \mathbb{R}^n$, let $\mu = \mathbb{P}_{x\sim q}[x \in E]$.*

As before, write Φ for the cdf of the univariate standard normal distribution. If $\mu \geq \frac{1}{2}$, then $\epsilon_q^*(E) = 0$. Otherwise, $\epsilon_q^*(E) \leq -\sigma\Phi^{-1}(\mu)$, with equality when E is a half space.

In particular, for any machine learning model for which the error rate in the distribution q is at least μ, the median distance to the nearest error is at most $-\sigma\Phi^{-1}(\mu)$. Because each coordinate of a multivariate normal is a univariate normal, $-\sigma\Phi^{-1}(\mu)$ is the distance to a half space for which the error rate is μ. In other words, the right hand side of the inequality is the same expression that appears in Equation 1.

So, among models with some fixed error rate $\mathbb{P}_{x\sim q}[x \in E]$, the most robust are the ones whose error set is a half space (as shown in Figure 1). In Appendix E we will give a more common statement of the Gaussian isoperimetric inequality along with a proof of the version presented here.

Comparing Neural Networks to the Isoperimetric Bound. We evaluated these quantities for several models on the CIFAR-10 and ImageNet test sets.

As in Section 4, we report an estimate of ϵ_q^*. For each test image, we took 1,000 samples from the corresponding Gaussian and estimated ϵ_q^* using PGD with 200 steps on each sample and reported the median.

We find that for the five models we considered, the relationship between our estimate of $\epsilon_q^*(E)$ and $\mathbb{P}_{x\sim q}[x \in E]$ is already close to optimal. This is visualized in Figure 4. For CIFAR-10, adversarial training improves robustness to small perturbations, but the gains are primarily because error rates in Gaussian noise were improved. In particular, it is clear from the graph on the bottom left that adversarial training increases the σ at which the error rate is 1% on average. This shows that improved adversarial robustness results in

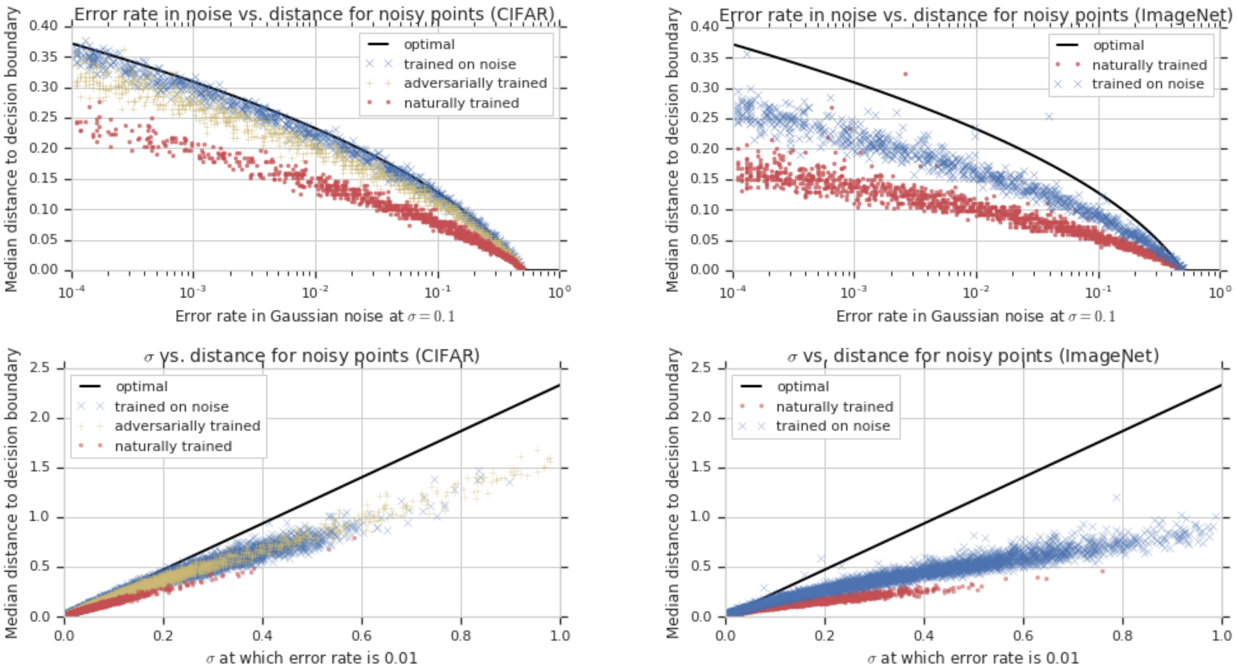

Figure 4. These plots give two ways to visualize the relationship between the error rate in noise and the distance from noisy points to the decision boundary (found using PGD). Each point on each plot represents one image from the test set. On the top row, we compare the error rate of the model with Gaussian perturbations at $\sigma = 0.1$ to the distance from the median *noisy* point to its nearest error. On the bottom row, we compare the σ at which the error rate is 0.01 to this same median distance. (These are therefore similar to the plots in Figure 2.) The thick black line at the top of each plot is the upper bound provided by the Gaussian isoperimetric inequality. We include data from a model trained on clean images, an adversarially trained model, and a model trained on Gaussian noise ($\sigma = 0.4$.)

improved robustness to large random perturbations, as the isoperimetric inequality says it *must*.

6. Evaluating Corruption Robustness

The previous two sections show a relationship between adversarial robustness and one type of corruption robustness. This suggests that methods designed to improve adversarial robustness ought to also improve corruption robustness, and vice versa. In this section we investigate this relationship.

We analyzed the performance of our models on the corruption robustness benchmark described in Hendrycks & Dietterich (2019). There are 15 different corruptions in this benchmark, each of which is tested at five different levels of severity. The results are summarized in Figure 5, where we have aggregated the corruption types based on whether the ordinarily trained model did better or worse than the augmented models. We found a significant difference in performance on this benchmark when the model is evaluated on the compressed images provided with the benchmark rather than applying the corruptions in memory. (In this section we report performance on corruptions applied in-memory.) Figure 8 in the appendix shows an example for the Gaussian-5 corruption, where performance degraded from 57% accuracy (in memory) to 10% accuracy (compressed images). Detailed results on both versions of this benchmark are presented in Appendix B.

Gaussian data augmentation and adversarial training both improve the overall benchmark[1], which requires averaging the performance across all corruptions, and the results were quite close. Adversarial training helped more with blurring corruptions and Gaussian data augmentation helped more with noise corruptions. Interestingly, both methods performed much worse than the clean model on the fog and contrast corruptions. For example, the adversarially trained model was 55% accurate on the most severe contrast corruption compared to 85% for the clean model. Note that Hendrycks & Dietterich (2019) also observed that adversarial training improves robustness on this benchmark on Tiny ImageNet.

The fact that adversarial training is so successful against the noise corruptions further supports the connection we have been describing. For other corruptions, the relationship is more complicated, and it would be interesting to explore this in future work.

[1] In reporting overall performance on this benchmark, we omit the Gaussian noise corruption.

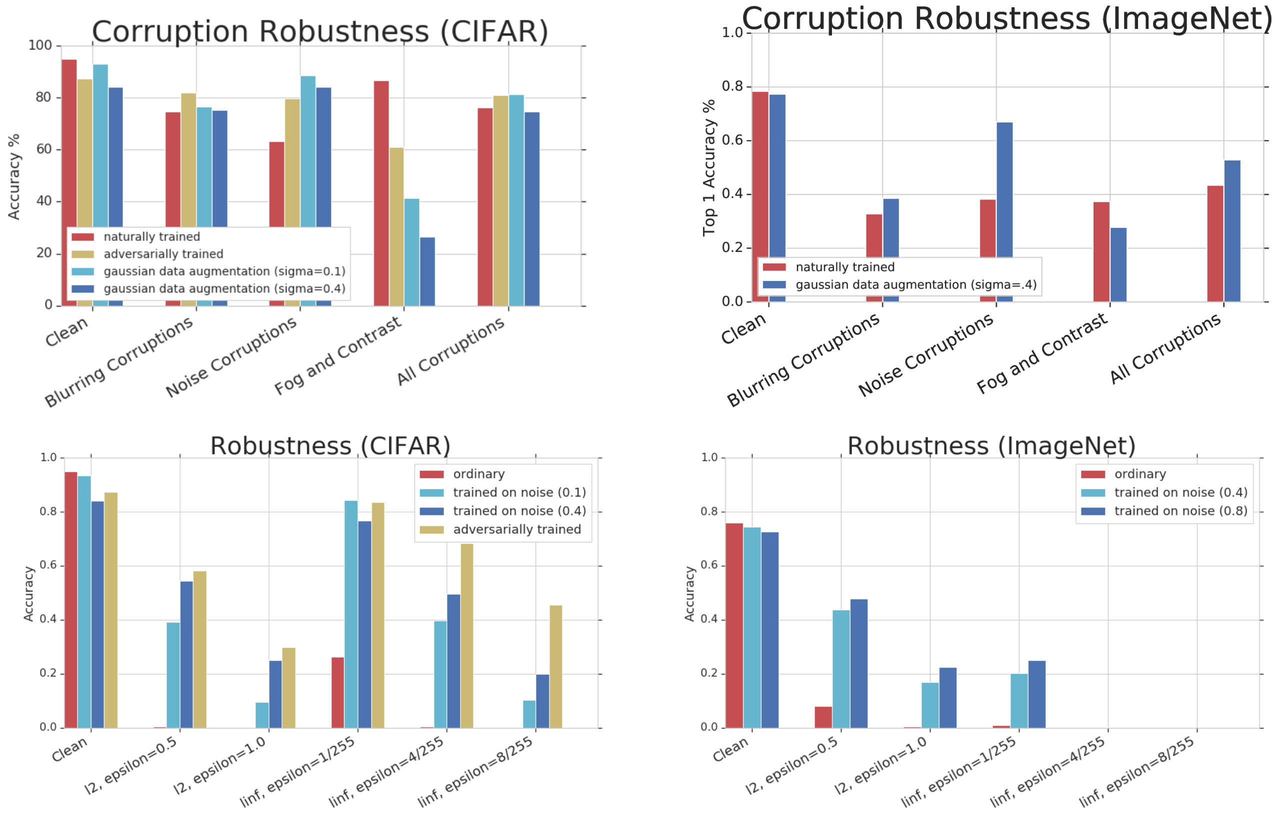

Figure 5. The performance of the models we considered on the corruption robustness benchmark, together with our measurements of those models' robustness to small l_p perturbations. For all the robustness tests we used PGD with 100 steps and a step size of $\epsilon/25$. The adversarially trained CIFAR-10 model is the open sourced model from Madry et al. (2017).

We also evaluated these two augmentation methods on standard measures of l_p robustness. We see a similar story there: while adversarial training performs better, Gaussian data augmentation improves adversarial robustness as well. Gaussian data augmenation has been proposed as an adversarial defense in prior work (Zantedeschi et al., 2017). Here we evaluate this method not to propose it as a novel defense but to provide further evidence of the connection between adversarial and corruption robustness.

We also considered the MNIST adversarially trained model from Madry et al. (2017), and found it to be a special case where robustness to small perturbations was increased while generalization in noise was not improved (see Appendix D). This is because this model violates the linearity assumption discussed in Section 4.

Corruption Robustness as a Sanity Check for Defenses. We also analyzed the performance several previously published adversarial defense strategies in Gaussian noise. These methods have already been shown to result in vanishing gradients, which causes standard optimization procedures to fail to find errors, rather than actually improving adversarial robustness (Athalye et al., 2018). We find that these methods also show no improvement in Gaussian noise. The results are shown in Figure 6. Had these prior defenses performed an analysis like this, they would have been able to determine that their methods relied on vanishing gradients and fail to improve robustness.

Obtaining Zero Test Error in Noise is Nontrivial. It is important to note that applying Gaussian data augmentation does not reduce error rates in Gaussian noise to zero. For example, we performed Gaussian data augmentation on CIFAR-10 at $\sigma = .15$ and obtained 99.9% training accuracy but 77.5% test accuracy in the same noise distribution. (For comparison, the naturally trained obtains 95% clean test accuracy.) Previous work (Dodge & Karam, 2017) has also observed that obtaining perfect generalization in large Gaussian noise is nontrivial. This mirrors Schmidt et al. (2018), which found that adversarial robustness did not generalize to the test set, providing yet another similarity between adversarial and corruption robustness. This is perhaps not surprising given that error rates on the *clean* test set are also non-zero. Although the model is in some sense "superhuman" with respect to clean test accuracy, it still makes many mistakes on the clean test set that a human would never

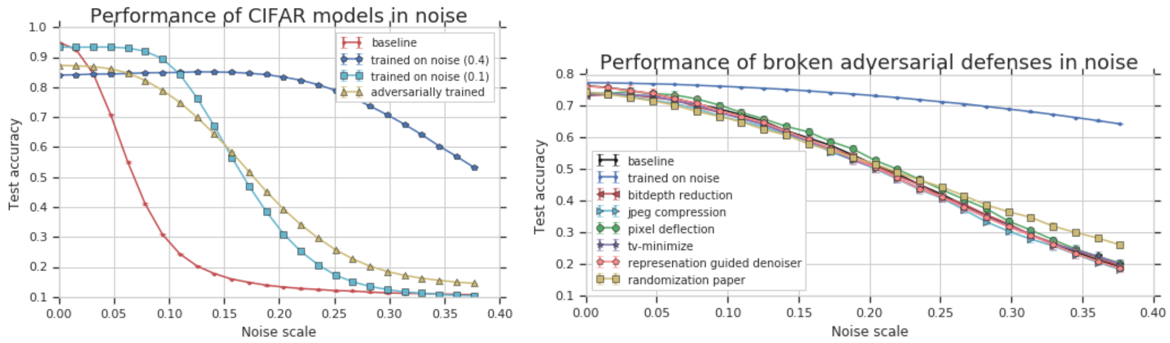

Figure 6. (Left) The performance in Gaussian noise of the CIFAR models described in this paper. *(Right)* The performance in Gaussian noise of several previously published defenses for ImageNet, along with an ImageNet model trained on Gaussian noise at $\sigma = 0.4$ for comparison. For each point we ran ten trials; the error bars show one standard deviation. All of these defenses are now known not to improve adversarial robustness (Athalye et al., 2018). The defense strategies include bitdepth reduction (Guo et al., 2017), JPEG compression (Guo et al., 2017; Dziugaite et al., 2016; Liu et al., 2018; Aydemir et al., 2018; Das et al., 2018; 2017), Pixel Deflection (Prakash et al., 2018), total variance minimization (Guo et al., 2017), respresentation-guided denoising (Liao et al., 2018), and random resizing and random padding of the input image (Xie et al., 2017).

make. We collected some examples in Appendix I. More detailed results on training and testing in noise can be found in Appendices C and H.

7. Conclusion

This paper investigates whether we should be surprised to find adversarial examples as close as we do, given the error rates we observe in corrupted image distributions. After running several experiments, we argue that the answer to this question is no. Specifically:

1. The nearby errors we can find show up at the same distance scales we would expect from a linear model with the same corruption robustness.
2. Concentration of measure shows that a non-zero error rate in Gaussian noise *logically implies* the existence of small adversarial perturbations of noisy images.
3. Finally, training procedures designed to improve adversarial robustness also improve many types of corruption robustness, and training on Gaussian noise moderately improves adversarial robustness.

In light of this, we believe it would be beneficial for the adversarial defense literature to start reporting generalization to distributional shift, such as the common corruption benchmark introduced in Hendrycks & Dietterich (2019), in addition to empirical estimates of adversarial robustness. There are several reasons for this recommendation.

First, a varied suite of corruptions can expose failure modes of a model that we might otherwise miss. For example, we found that adversarial training significantly degraded performance on the fog and contrast corruptions despite improving small perturbation robustness. In particular, performance on constrast-5 dropped to 55.3% accuracy vs 85.7% for the vanilla model (see Appendix B for more details).

Second, measuring corruption robustness is significantly easier than measuring adversarial robustness — computing adversarial robustness perfectly requires solving an NP-hard problem for every point in the test set (Katz et al., 2017). Since Szegedy et al. (2014), hundreds of adversarial defense papers have been published. To our knowledge, only one (Madry et al., 2017) has reported robustness numbers which were confirmed by a third party. We believe the difficulty of measuring robustness under the usual definition has contributed to this unproductive situation.

Third, all of the failed defense strategies we examined also failed to improve performance in Gaussian noise. For this reason, we should be highly skeptical of defense strategies that only claim improved l_p robustness but are unable to demonstrate robustness to distributional shift.

Finally, if the goal is improving the security of our models in adversarial settings, errors on corrupted images already imply that our models are not secure. Until our models are perfectly robust in the presence of average-case corruptions, they will not be robust in worst-case settings.

The communities of researchers studying adversarial and corruption robustness seem to be attacking essentially the same problem in two different ways. We believe that the corruption robustness problem is also interesting independently of its connection to adversarial examples, and we hope that the results presented here will encourage more collaboration between these two communities.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Ayse Elvan Aydemir, Alptekin Temizel, and Tugba Taskaya Temizel. The effects of jpeg and jpeg2000 compression on attacks using adversarial examples. *arXiv preprint arXiv:1803.10418*, 2018.

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

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108. ACM, 2004.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. *arXiv preprint arXiv:1802.06816*, 2018.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *Computer Communication and Networks (ICCCN), 2017 26th International Conference on*, pp. 1–7. IEEE, 2017.

Elvis Dohmatob. Limitations of adversarial robustness: strong no free lunch theorem. *arXiv preprint arXiv:1810.04065*, 2018.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pp. 1632–1640, 2016.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018a.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *IEEE CVPR*, number CONF, 2018b.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018a.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018b.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *International Conference on Learning Representations*, 2019.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu, and Xiaolin Hu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Zihao Liu, Qi Liu, Tao Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. *arXiv preprint arXiv:1803.05787*, 2018.

Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial examples. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *arXiv preprint arXiv:1809.03063*, 2018.

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8571–8580, 2018.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Yash Sharma and Pin-Yu Chen. Breaking the madry defense model with l1-based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 39–49. ACM, 2017.