On the Sensitivity of Adversarial Robustness to Input Data Distributions

Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, Ruitong Huang

Borealis AI, Canada

Abstract

Neural networks are vulnerable to small adversarial perturbations. While existing literature largely focused on the vulnerability of learned models, *we demonstrate an intriguing phenomenon that adversarial robustness, unlike clean accuracy, is sensitive to the* input data distribution. *Even a semantics-preserving transformations on the input data distribution can cause a significantly different robustness for the adversarially trained model that is both trained and evaluated on the new distribution. We show this by constructing semantically-identical variants for MNIST and CIFAR10 respectively, and show that standardly trained models achieve similar clean accuracies on them, but adversarially trained models achieve significantly different robustness accuracies. This counter-intuitive phenomenon indicates that input data distribution alone can affect the adversarial robustness of trained neural networks, not necessarily the tasks themselves. The full paper (ICLR 2019) can be found at* https://openreview.net/forum?id=S1xNEhR9KX.

1. Introduction

We study the relationship between adversarial robustness and the input data distribution. We focus on the adversarial training method [1], arguably the most popular defense method so far due to its simplicity, effectiveness and scalability. Our main contribution is the finding that adversarial robustness is highly sensitive to the input data distribution:

> *A semantically-lossless shift on the data distribution could result in a drastically different robustness for adversarially trained models.*

Note that this is different from the transferability of a fixed model that is trained on one data distribution but tested on another distribution. Even retraining the model on the new data distribution may give us a completely different adversarial robustness on the same new distribution. This is also in sharp contrast to the clean accuracy of standard training, which, as we show in later sections, is insensitive to such shifts. To our best knowledge, our paper is the first work in the literature that demonstrates such sensitivity.

Such sensitivity raises the question of how to properly evaluate adversarial robustness. In particular, the sensitivity of adversarial robustness suggests that certain datasets may not be sufficiently representative when benchmarking different robust learning algorithms. It also raises serious concerns about the deployment of believed-to-be-robust training algorithm in a real product. In a standard development procedure, various models would be prototyped and measured on the existing data. However, the sensitivity of adversarial robustness makes the truthfulness of the performance estimations questionable, as one would expect future data to be slightly shifted. We illustrate the practical implications in Section 3: the robust accuracy of PGD trained model is sensitive to gamma values of gamma-corrected CIFAR10 images. This indicates that image datasets collected under different lighting conditions may have different robustness properties.

Finally, our finding opens up a new angle and provides novel insights to the adversarial vulnerability problem, complementing several recent works on the issue of data distributions' influences on robustness. [4] hypothesizes that there is an intrinsic tradeoff between clean accuracy and adversarial robustness. Our studies complement this result, showing that there are different levels of tradeoffs depending on the characteristics of input data distribution, under the same learning settings (training algorithm, model and training set size). [2] shows that different data distributions could have drastically different properties of adversarially robust generalization, theoretically on Bernoulli vs mixtures of Gaussians, and empirically on standard benchmark datasets. From the sensitivity perspective, we demonstrate that being from completely different distributions (e.g. binary vs Gaussian or MNIST vs CIFAR10) may not be the essential reason for having large robustness difference. Gradual semantics-preserving transformations of data distribution can also cause large changes to datasets' achievable robustness.

2. Robustness on Datasets Variants with Different Input Distributions

In this section we carefully design a series of datasets and experiments to further study its influence. One important property of our new datasets is that they have different input data distributions $\mathbb{P}(x)$'s while keeping the true classification $\mathbb{P}(y|x)$ reasonably fixed, thus these datasets are only different in a "semantic-lossless" shift. Our experiments reveal an unexpected phenomenon that while standard learning methods manage to achieve stable clean accuracies across different data distributions under "semantic-lossless" shifts, however, adversarial training, arguably the most popular method to achieve robust models, loses this desirable property, in that its robust accuracy becomes unstable even under a "semantic-lossless" shift on the data distribution. We emphasize that different from preprocessing steps or transfer learning, here we treat the shifted data distribution as a new underlying distribution. We both train the models and test the robust accuracies on the same new distribution.

2.1. Smoothing and Saturation

In general, MNIST has a more binary distribution of pixels, while CIFAR10 has a more continuous spectrum of pixel values. We apply different levels of "smoothing" on MNIST to create more CIFAR-like datasets, and different levels of "saturation" on CIFAR10 to create more "binary" ones, as shown in Figure 1a and 1b. Note that we would like to maintain the semantic information of the original data, which means that such operations should be semantics-lossless.

Smoothing is applied on MNIST images, to make images "less binary". Given an image x_i, its smoothed version $\tilde{x_i}^{(s)}$ is generated by first applying average filter of kernel size s to x_i to generate an intermediate smooth image, and then take pixel-wise maximum between x_i and the intermediate smooth image.

Saturation of the image x is denoted by $\widehat{x}^{(p)}$, and the procedure is defined as: $\widehat{x}^{(p)} = \mathrm{sign}(2x-1)\frac{|2x-1|^{\frac{2}{p}}}{2} + \frac{1}{2}$, where all the operations are pixel-wise and each element of $\widehat{x}^{(p)}$ is guaranteed to be in $[0, 1]$. Saturation is used to generate variants of the CIFAR10 dataset with less centered pixel values. For different saturation level p's, one can see from Figure 1b that $\widehat{x}^{(p)}$ is still semantically similar to x in the same classification task.

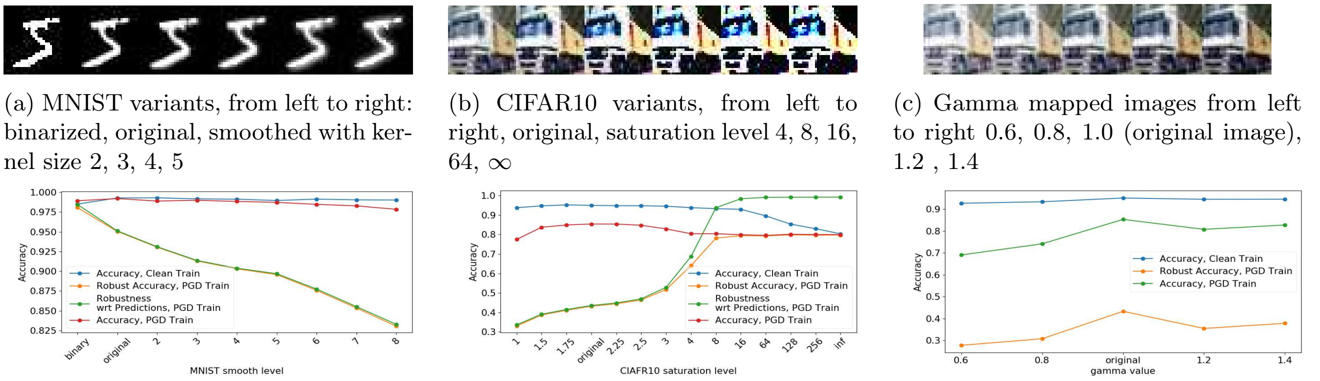
2.2. Experimental Setups

We use the smoothing and saturation to manipulate the data distributions of MNIST and CIFAR10, and show empirical results on how data distributions affects robust accuracies of neural networks trained on them. To measure the difficulty of the classification task, we perform standard neural network training and test *accuracies* on clean data. To measure the difficulty to achieve robustness, we perform ℓ_∞ projected gradient descent (PGD) based adversarial training [1] and test *robust accuracies* on adversarially perturbed data. To understand whether low robust accuracy is due to low clean accuracy or vulnerability of model, we also report *robustness w.r.t. predictions*, where the attack is used to perturb against the model's clean prediction, instead of the true label. We use LeNet5 on all the MNIST variants, and use wide residual networks [5] with widen factor 4 and depth 28 for all the CIFAR10 variants. Unless otherwise specified, PGD training on MNIST variants and CIFAR10 variants all follows the settings in [1]. PGD attacks on MNIST variants run with $\epsilon = 0.3$, step size of 0.01 and 40 iterations, and runs with $\epsilon = 8/255$, step size of 2/255 and 10 iterations on CIFAR10 variants , same as in [1].

2.3. Sensitivity of Robust Accuracy to Data Transformations

Results on MNIST variants are presented in Figure 1d. The clean accuracy of standard training is very stable across different MNIST variants. This indicates that their classification tasks have similar difficulties, if the training has no robust considerations. When performing PGD adversarial training, clean accuracy drops only slightly. However, both robust accuracy and robustness w.r.t. predictions drop significantly. This indicates that as smooth level goes up, it is significantly harder to achieve robustness. Note that for binarized MNIST with adversarial training, the clean accuracy and the robust accuracy are almost the same. Indicating that getting high robust accuracy on binarized MNIST does not conflict with achieving high clean accuracy.

CIFAR10 result tell a similar story, as reported in Figure 1e. For standard training, the clean accuracy maintains almost at the original level until saturation level 16, despite that it is already perceptually very saturated. In contrast, PGD training has a different trend. Before level 16, the robust accuracy significantly increases from 43.2% until 79.7%, while the clean test accuracy drops only in a comparatively small range, from 85.4% to 80.0%. After level 16, PGD training has almost the same clean accuracy and robust accuracy. However, robustness w.r.t. predictions still keeps increasing, which again indicates the instability of the



(a) MNIST variants, from left to right: binarized, original, smoothed with kernel size 2, 3, 4, 5

(b) CIFAR10 variants, from left to right, original, saturation level 4, 8, 16, 64, ∞

(c) Gamma mapped images from left to right 0.6, 0.8, 1.0 (original image), 1.2 , 1.4

(d) MNIST results under different smooth levels

(e) CIFAR10 results under different saturation levels

(f) Robustness results on gamma mapped CIFAR10 variant

Figure 1: Variants of MNIST and CIFAR10 datasets (a, b, c), and Accuracy, Robust Accuracy and Robustness w.r.t. Predictions on different data variants (c, d, e).

robustness. On the other hand, if the saturation level is smaller than 2, we get worse robust accuracy after PGD training, e.g. at saturation level 1 the robust accuracy is 33.0%. Simultaneously, the clean accuracy maintains almost the same.

Note that after saturation level 64 the standard training accuracies starts to drop significantly. This is likely due to that high degree of saturation has caused "information loss". Models trained on highly saturated CIFAR10 are quite robust and the gap between robust accuracy and robustness w.r.t. predictions is due to lower clean accuracy. In contrast, In MNIST variants, the robustness w.r.t. predictions is always almost the same as robust accuracy, indicating that drops in robust accuracy is due to adversarial vulnerability.

From these results, we can conclude that robust accuracy under PGD training is much more sensitive than clean accuracy under standard training to the differences in input data distribution. More importantly, a semantically-lossless shift on the data transformation, while not introducing any unexpected risk for the clean accuracy of standard training, can lead to large variations in robust accuracy. Such previously unnoticed sensitivity raised serious concerns in practice, as discussed in the next section.

3. Sensitivity to Image Acquisition Condition and Preprocessing

The natural images are acquired under different lighting conditions, with different cameras and different camera settings. They are usually preprocessed in different ways. All these factors could lead to mild shifts on the input distribution. Therefore, we might get very different performance measures when performing adversarial training on images taken under different conditions. In this section, we demonstrate this phenomenon on variants of CIFAR10 images under different gamma mappings. These variants are then used to represent image dataset acquired under different conditions. Gamma mapping is a simple element-wise operation that takes the original image x, and output the gamma mapped image $\tilde{x}^{(\gamma)}$ by performing $\tilde{x}^{(\gamma)} = x^{\gamma}$. Gamma mapping is commonly used to adjust the exposure of an images. We refer the readers to [3] on more details about gamma mappings. Figure 1c shows variants of the same image processed with different gamma values. Lower gamma value leads to brighter images and higher gamma values gives darker images, since pixel values range from 0 to 1. Despite the changes in brightness, the semantic information is preserved.

We perform the same experiments as in the saturated CIFAR10 variants experiment in Section 2, with results displayed in Figure 1f. Clean accuracies almost remain the same across different gamma values. However, under PGD training, both accuracy and robust accuracy varies largely under different gamma values.

These results should raise practitioners' attention on how to interpret robustness benchmark "values". For the same adversarial training setting, the robustness measure might change drastically between image datasets with different "exposures". In other words, if a training algorithm achieves good robustness on one image dataset, it doesn't necessarily achieve similar robustness on another semantically-identical but slightly varied datasets. Therefore, the actual robustness could be underestimated or overestimated significantly. This raises the questions on whether we are evaluating image classifier robustness in a reliable way, and how we choose benchmark settings that can match the real robustness requirements in practice. We defer this important open question to future research.

References

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

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

[3] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[4] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

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