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

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.

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. 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.

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.

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. 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.

2.1. Sensitivity to Data Transformations

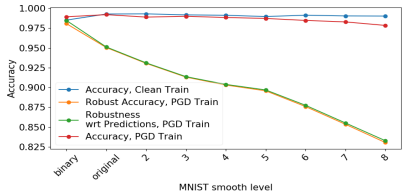
Results on MNIST variants are presented in Figure 1d. The clean accuracy of standard training is very stable across different MNIST variants. This indicates that their classification tasks have similar difficulties, if the training has no robust considerations. When performing PGD adversarial training, 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

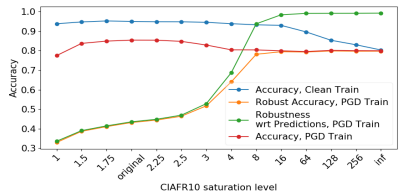
(a) MNIST variants, from left to right: binarized, original, smoothed with kernel size 2, 3, 4, 5

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

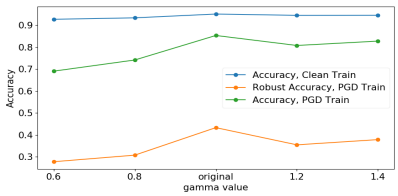
(c) Gamma mapped images from left to right 0.6, 0.8, 1.0 (original image), 1.2 , 1.4



(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).

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 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.

3. Sensitivity to the Gamma Mapping

Different factors could lead to distributional shifts on image datasets, such as acquired under different lighting conditions or preprocessed differently. Do they lead to different levels of robustness? We answer this question by testing on variants of CIFAR10 images under different gamma mappings. 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 $\tilde{x}^{(\gamma)} = x^{\gamma}$. It is commonly used to adjust the exposure of an images. Figure 1c shows variants of the same image processed with different gamma values.

We perform similar experiments as 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.