Adversarially Robust Generalization Requires More Data

Zachary Barlow; zachary.barlow@mail.utoronto.ca; (1003088026)
Anthony Inthavong; anthony.inthavaong@mail.utoronto.ca; (1003158646)

Due: April 20^{th}, 2020

1 Introduction

This paper shows the difference between standard generalization and robust generalization and proves that the sample complexities can be considerably different from one another. This describes and proves that the accuracy of the training depends on the size of the data set that is provided to the model. With adversarially training there needs to be more data than of standard training to have the same outcome. Which brings up the following question that the researchers propose:

> "How does the sample complexity of standard generalization compare to that of adversarially robust generalization?"[10]

They answer this question by looking at the distributions of Bernoulli and Gaussian and by analyzing the robust generalization for them both. This is looked at by subsampling the dataset at various rater and study the impact of each sample size on their adversarial robustness. For this review, we will only be looking at the MNIST dataset rather than MNIST, SVHN and CIFAR10 all together, and the goal is to learn a classifier that achieves good test accuracy even under ℓ_∞-bounded perturbations.

Going further, most of their analysis was to create a lower bound for the two distributions in order to show the number of samples required for robust generalization. This lower bound is important and provides a hardness that shows that any similar distributions will follow the same hardness. A good standard error can be achieved from a single sample whereas the robust generalization approach requires a significant more samples to provide the same error which is difficult to find the right number of samples for each dataset. Thus, no algorithm can produce a robust classifier without many samples.

2 Main Result:

We will begin by defining some important terms that help with understanding the topic of this paper further, such as adversarial examples, robust learning and specifically what a perturbation is. Which are defined as the following:

- *Adversarial examples:* machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution[8]
- *Robust learning:* is learning that either or both better senses making through deep conceptual understanding and fast and accurate procedural fluency[3]
- *Perturbation:* method for solving a problem by comparing it with a similar one for which the solution is known

With these definitions, we can get into the theoretical aspects of this paper that it revolves around. Firstly, the Gaussian and Bernoulli models are the focuses of this paper and studying the differences between their standard classification and robust classification errors. For the norms in this paper, the authors focused on the ℓ_∞-bounded adversaries defining the following perturbation set that is used further, $\mathbf{B}^{\epsilon}_{\infty}(x) = \{x' \in \mathbb{R}^d | \, \|x' - x\|_\infty \leq \epsilon\}$ [10].

Lets first look at the Gaussian model. Looking at the generalization for robustness, which is more of a property for an algorithm. It is merely defined as "if a testing sample is close to a training sample, then the testing error is also close."[13]

Definition 2.1. Algorithm $\mathcal{A}$ is $(\mathcal{K}, \epsilon(s))$-robust if $\mathcal{Z}$ can be partitioned into $\mathcal{K}$ disjoint sets, denoted as $\{C_i\}_{i=1}^{K}$, such that $\forall s \in \mathbf{s}$,

$$s, z \in C_i, \Rightarrow |\ell(\mathcal{A}_s, s) - \ell(\mathcal{A}_s, z)| \leq \epsilon(s)$$

[13]

where $\ell(\mathcal{A}_s, s)$ & $\ell(\mathcal{A}_s, z)$ are loss functions. With this, looking at the Gaussian model, we can clearly distinguish betweeen standard generalization and robust generalization. From their paper, Corollary 19 proves that with high probability, tuning model parameters properly can achieve a small error probability from a sample size of one for standard generalization. Whereas, to achieve robust generalization, ϵ needs to be bounded by a small constant and have a large number of samples (n) to learn the ℓ^{ϵ}_{∞}-robust classifier according to Theorem 5 [10]. This lower bound, compared to Theorem 6, provides that,

$$\frac{c}{logd} \leq \frac{n}{\epsilon^2 \sqrt{d}} \leq c'$$

Giving transferable adversarial examples as well as using only a single adversarial perturbation per class [10] that is polynomial by $\sqrt{d}$ in dimension more than standard generalization.

The ultimate goal is to decrease the lower bound so that the sample complexity does not need to be a large to provide a small error of 1% with high probability. This is ultimately increasing the robustness of the neural nets and strengthening their defenses against adversarial examples trying to attack and disrupt the system [7]. Going further, the tradeoff between sample complexity and ℓ_∞-robustness of a classifier is very tight, as seen in Theorems 5 & 6 [10], which gives very little leeway in the number of samples that are needed to provide a good classifier for the (θ, σ)-Gaussian model. Their proof shows that a single "perturbation is transferable across examples as well as across architectures and learning procedures" [10].

When looking at the Bernoulli model, it is very similar to the Gaussian model in terms of the distribution but is quite significantly different with respect to the robust sample complexity. In regards to achieving standard generalization from a single sample, Theorem 8 provides the same bound as the (θ, σ)-Gaussian distribution but with τ as the parameter of interest instead of σ. Where τ when small, makes the samples less correlated with their respective class vectors. The lower bound for the Bernoulli model is,

$$n \geq c \frac{\epsilon^2 d}{log d}$$

and from Theorem 9, the expected ℓ_∞^ϵ-robust classification error is at least $\frac{1}{2} - \gamma$ where $0 < \gamma < \frac{1}{2}$ and $n \leq c_2 \frac{\epsilon^2 \gamma^2 d}{log \frac{d}{\gamma}}$. Linear classifiers, the same as in the Gaussian model require a significant amount of samples to provide the ℓ_∞-robustness whereas in this model, "non-linear classifiers can achieve a significantly improved robustness"[10] when there is a thresholding operator applied to the function to undo the ℓ_∞-bounded adversaries.

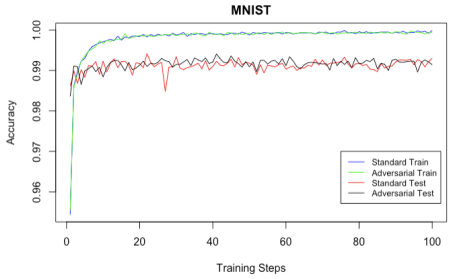
Since the MNIST dataset is binary, it follows the (θ^*, τ)-Bernoulli distribution, and being familiar with working with the MNIST dataset in Assignment 3, we use the procedure in following experiments section.

3 Experiment:

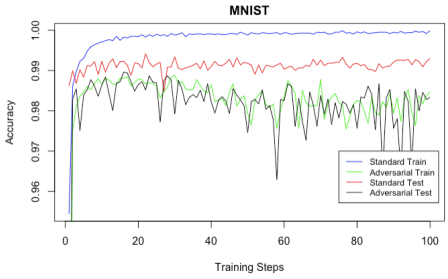
For our experiments, we performed two different methods of adding perturbations and using adversarial examples for the MNIST and CIFAR-10 datasets.

The first method was a simple linear addition of size $\epsilon = 0.05$. We can see for the MNIST dataset there was very little difference in the generalization of the test and training trends. For the CIFAR-10 dataset there is a large difference between standard and adversarial accuracies.

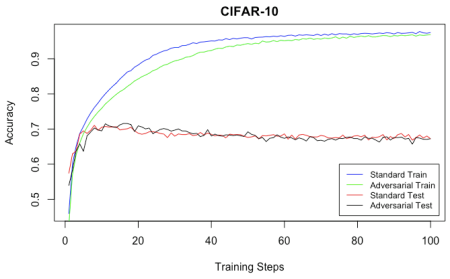
The second method was using project gradient descent as our adversary. This method maximized the loss for a given epsilon (we took the gradient in terms of ℓ_∞ norm). For both MINST and CIFAR-10, it can be seem there is a rather significant difference between the standard and adversarial training accuracies. This method, which may not be exactly the same as the one that the researchers have, shows the same outcome. The model needs



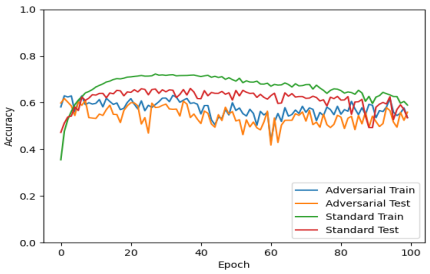
(a) Linear Transformation



(b) Gradient Descent



(c) Linear Transformation

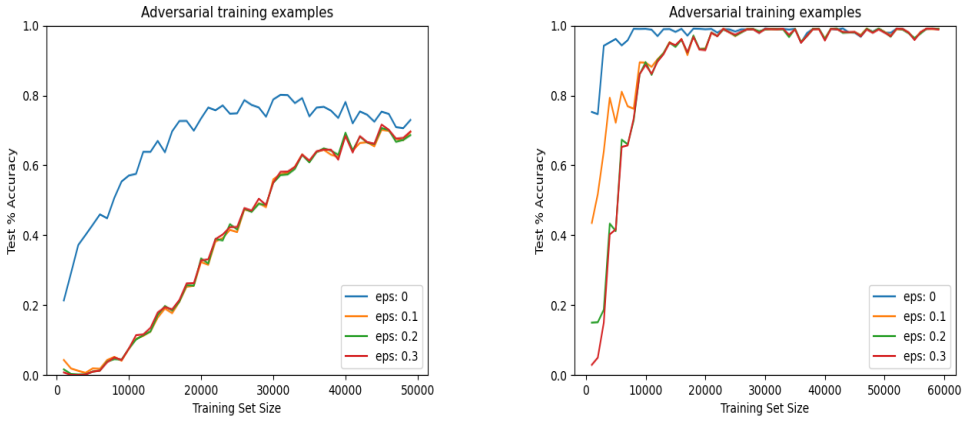


(d) Gradient Descent (CIFAR-10)

Figure 1: Follows that the adversarial examples, for increasing training samples the training follows the testing and the standard train and test vary. This is similar to Figure 1 in [10].

significantly more samples to provide a good accuracy. From the change in epsilon in 2 we used standard CNN as did the researchers and we used attacks with 10000 and 20000 adversarial examples and looked at their testing accuracy of the defence of the mode. We then used this same model and ran it on the original model and adversarial tested model and noticed a significant increase in accuracy of our epochs which reached 95-99% consistently.

MNIST and CIFAR-10. We ran 100 iterations of projected gradient descent as our adversary, with perturbations of size 0.05. Similarly, we performed the same method but instead with 100 iterations of linear difference of 0.05. [9]



(a) CIFAR-10 (b) MNIST

Figure 2: These graphs, mostly the MNIST dataset was designed to follow Figure 2 & 3 in [10]. Starting at 1000 training set size for 10000 adversarial examples and increasing by 1000 each time on different values of epsilon. Our model is similar in that the higher the epsilon for the pertubations, the less it was able to defend the examples. Whereas, all examples for CIFAR-10 provided similar outcomes.

4 Limitations:

As with the majority of studies, the design of the current study is subject to limitations. Robustness comes at a cost and the tight bounds on the sample complexity only sure a guarantee on accuracy to a certain degree. This behaves properly when there is a limit on the amount of samples that can be trained on the model but when there is an infinite supply of data, it is very difficult to learn a robust classifier and be accurate. The paper focuses on ℓ_∞ norm. However due to this, a adversarial model train specifically for ℓ_∞ norm can be tricked by other restrictions. We can see from our results, when we applied a linear transformation there was a significant gap in complexity, especially for more complex datasets such as CIFAR-10. Performing a more complex perturbation on data that is already complex leads to a larger sample complexity, and thus a larger generalization gap between standard and adversarial training and testing accuracy.

5 Future Directions:

As the main contribution of this paper was their lower bounds on the sample complexity for ℓ_∞-robustness, future courses of study could be:

- **Even stronger lower bounds:**

  The lower bound that Schmidt et al. is a tight bound on the sample complexity. This bound still causes misclassification on adversarially

or regular examples. To fix this issue, there needs to a greater restriction to increase the accuracy the model trains with high probability.

- **Does Robustness cause a decrease in accuracy:**

  The robustness of an algorithm causes an increase in sample complexity due to the influx of adversarial examples trying to perturb the model. Due to this increase, a thought would be, does the robustness cause the accuracy to change for standard generalization, and if so, by how much?

- **Generalization of Adversarial training**

  Though the focus of the paper was on adversarial sample complexity on ℓ_∞ norm, future directions can be focused on generalizing adversarial training. For example, although we can train a model on a ℓ_∞ norm adversary, a more complex adversary can still mislead this model.

References

[1] Convoluted neural net. URL: https://rpubs.com/juanhklopper/example_of_a_CNN.

[2] Module: tf.keras.losses : Tensorflow core v2.1.0. URL: https://www.tensorflow.org/api_docs/python/tf/keras/losses.

[3] Robust learning, 2020. URL: https://learnlab.org/research/wiki/Robust_learning.

[4] J.J. Allaire. Tensorflow for r: Cnn. URL: https://tensorflow.rstudio.com/tutorials/advanced/images/cnn/.

[5] J.J. Allaire. Tensorflow for r: Gpu. URL: https://tensorflow.rstudio.com/installation/gpu/local_gpu/.

[6] J.J. Allaire. Tensorflow for r: Keras. URL: https://tensorflow.rstudio.com/guide/keras/.

[7] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *CoRR*, abs/1802.00420, 2018. URL: http://arxiv.org/abs/1802.00420, arXiv:1802.00420.

[8] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. 2017. URL: https://arxiv.org/abs/1611.01236.

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

[10] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *CoRR*, abs/1804.11285, 2018. URL: http://arxiv.org/abs/1804.11285, arXiv:1804.11285.

[11] Sebastian Theiler. Implementing adversarial attacks and defenses in keras & tensorflow 2.0, Dec 2019. URL: https://medium.com/analytics-vidhya/implementing-adversarial-attacks-and-defenses-in-keras-tensorflow-2-0-cab6120c5715

[12] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2018. arXiv:1805.12152.

[13] Huan Xu and Shie Mannor. Robustness and generalization. *CoRR*, abs/1005.2243, 2010. URL: http://arxiv.org/abs/1005.2243, arXiv:1005.2243.