HOW TRAINING DATA AFFECT THE ACCURACY AND ROBUSTNESS OF NEURAL NETWORKS FOR IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has demonstrated the lack of robustness of well-trained deep neural networks (DNNs) to adversarial examples. For example, visually indistinguishable perturbations, when mixed with an original image, can easily lead deep learning models to misclassifications. In light of a recent study on the mutual influence between robustness and accuracy over 18 different ImageNet models, this paper investigates how training data affect the accuracy and robustness of deep neural networks. We conduct extensive experiments on four different datasets, including CIFAR-10, MNIST, STL-10, and Tiny ImageNet, with several representative neural networks. Our results reveal previously unknown phenomena that exist between the size of training data and characteristics of the resulting models. In particular, besides confirming that the model accuracy improves as the amount of training data increases, we also observe that the model robustness improves initially, but there exists a turning point after which robustness starts to decrease. How and when such turning points occur vary for different neural networks and different datasets.

1 INTRODUCTION

Deep neural models have achieved ground-breaking results for a growing list of tasks such as image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012), and the game of Go (Silver et al., 2016). Despite these accomplishments, research has discovered that existing deep neural networks are easily susceptible to various attacks. Szegedy et al. (2013) are the first to show the existence of *adversarial examples* in the image classification. Specifically, they demonstrate how to cause a network to misclassify an image by applying certain visually imperceptible perturbations. Clearly, this finding hinders the adoption of deep neural networks in practice, especially in safety-critical scenarios. Indeed, Evtimov et al. (2017) reveal that slight alterations to road signs can cause classifiers to predict a "STOP" sign as a "Speed Limit" sign.

Apart from image classification (Carlini & Wagner, 2016; Kurakin et al., 2016; Chen et al., 2017b), existing work has also investigated the robustness of deep learning models in other application domains, such as natural language processing (Jia & Liang, 2017; Cheng et al., 2018), image captioning (Chen et al., 2017a; Xu et al., 2018), and speech recognition (Carlini & Wagner, 2018). In a recent study, Su et al. (2018) empirically analyze how model accuracy and robustness interact on 18 deep image classification models, and provide several findings, the most interesting of which is that the sole pursuit of accuracy sacrifices robustness.

In this paper, we study how accuracy and robustness interact from a different perspective: the influence of training data on deep learning models. In particular, we aim at demystifying the relationship between the size of training data and important properties of the trained neural network, namely its accuracy and robustness. Our experiments are conducted with the CIFAR-10, MNIST, STL-10, and Tiny ImageNet datasets; and neural networks are chosen from a simple 2-layer perceptron to state-of-the-art DenseNet models. For each dataset, we split the training data into sub-datasets with strictly inclusion relationship (*i.e.*, having the current sub-dataset a *strict subset* of the next sub-dataset) and ensure all the sub-datasets are balanced (*i.e.*, having the same number of images for each label in the original dataset). We then train all targeted neural networks on these sub-datasets and

investigate how their robustness and accuracy vary *w.r.t.* the size of the underlying training dataset. Our study reveals several insights, and we summarize the main contributions below:

- We present the first comprehensive study to show how model robustness changes with increased training data for five representative neural networks, including a 2-layer perceptron, multi-layer CNNs (with a similar structure as those of Alexnet and VGG), ResNet, and DenseNet, on four different datasets (CIFAR-10, MNIST, STL-10, and Tiny ImageNet).
- We demonstrate that the robustness of a simple linear regression model can decrease as the amount of training data increases via a closed-form calculation.
- We find that model accuracy continues to improve with increased training data. Similarly, model robustness also improves, but may start to deteriorate when training data continue to increase. The occurrence of turning points depends on the deep neural network as well as the dataset on which it is trained.

2 MOTIVATION

Increasing training data helps to increase the accuracy of models, however, the general trend in model robustness is relatively understudied in the literature. This section presents a simple example to illustrate that, with more training data, a model can be more accurate but less robust. To allow a closed-form calculation, and simple and clear presentation, we consider the linear regression model.

First, we generate a set of data points in the form of (x, y), $y = a * x + \mu$, where a is a constant drawn from $(-10, 10)$, μ randomly sampled from $(-1, 1)$, x the input, and y the output. Then, we split the generated dataset into training data and testing data randomly.

Given the generated dataset, we apply the closed-form formula to compute the optimal value for the coefficients of a linear regression model. The accuracy of the model $\mathcal{M}$ is obtained by calculating the mean squared error (MSE) from the testing set. Regarding the robustness of the model, we search for the minimum $|\delta|$ such that $|(\mathcal{M}(x_t) - y_t)/y_t| < \theta$ and $|(\mathcal{M}(x_t + \delta) - y_t)/y_t| > \theta$ holds, where (x_t, y_t) denotes a data pair in the testing set. Essentially, we define an estimation to be correct if its relative error is within a bound, denoted by θ. In other words, we look for the largest amount of distortion to which the model $\mathcal{M}$ is immune. Here, we set θ 2.5% and enumerate δ from 0 by a step of 0.0001.

Let us consider a concrete example with training sets S_1 and S_2, and testing set T:

$$\begin{aligned} S_1 &= \{(1.35, 9.52), (2.42, 16.7), (4.02, 28.03)\} \\ S_2 &= S_1 \cup \{(8.59, 60.22), (3.85, 25.74), (6.71, 47.2)\} \\ T &= \{(4.78, 33.24), (9.71, 67.78)\} \end{aligned}$$

We apply closed-form formula to compute two models $\mathcal{M}_1$ and $\mathcal{M}_2$ on S_1 and S_2 respectively. On testing set T, the MSE of $\mathcal{M}_1$ is 0.0450 and that of $\mathcal{M}_2$ is 0.0396, meaning that $\mathcal{M}_2$ is more accurate. Averaging the value of δ over the entire testing set T, we get 0.2194 for $\mathcal{M}_1$ and 0.2125 for $\mathcal{M}_2$, which indicates that $\mathcal{M}_1$ is more robust than $\mathcal{M}_2$.

To summarize, we show, in a linear regression model, as the volume of training data increases, the model can become more accurate but less robust. For the rest of this paper, we will study if this phenomenon is transferrable to deep neural networks in image classification.

3 BACKGROUND

This section introduces the components of our study, including the evaluated neural networks, the robustness property, and the datasets.

3.1 NEURAL NETWORKS AND NOTATION

Our study covers the following neural networks:

– **2-Layer Perceptron.** This is the simplest model with one hidden layer and a softmax output layer.

- **Simple CNN.** We use a 7-layer Alex-like (Krizhevsky et al., 2012) model for MNIST and CIFAR (with the same structure in Carlini & Wagner (2016)). It consists of four convolutional layers and three fully connected layers followed by the softmax output layer. The kernel sizes are all 3×3 for the convolutional layers. For STL-10 and Tiny ImageNet, we also use a deeper CNN (20 layers) with a similar structure as VGG (Krizhevsky et al., 2012).
- **ResNet.** ResNet was proposed to alleviate the vanishing gradient problem for training very deep neural networks (He et al., 2016). ResNet introduces a novel structure where each layer learns the residual functions with reference to the input by adding skip-layer paths, or "identity shortcut connections." In this paper, we use ResNet-32 for CIFAR-10 and ResNet-20 for MNIST, STL-10, Tiny ImageNet, respectively.
- **DenseNet.** This model was proposed by Huang et al. (2017) to further utilize "identity shortcut connections" across different layers. It connects all layers with each other within a dense block. In this paper, we study DenseNet with a depth of 40 for CIFAR-10, STL-10 and Tiny ImageNet.

3.2 Robustness of Neural Networks

In this paper, we evaluate the robustness of models using state-of-the-art adversarial attacks. There are two kinds of adversarial attacks: *targeted* and *untargeted*. Given a valid input with label t, an untargeted attack searches for an input x' such that the prediction on x' is different from t, and x' is very close to x. As for an targeted attack, if the target label is t' ($t' \neq t$), it searches for an input x' such that the prediction on x' is t' and x' is close to x. To measure the distance between x' and the original x, we use two widely adopted metrics — ℓ_2 and ℓ_∞ — for adversarial perturbations (Carlini & Wagner, 2016; 2017; Chen et al., 2017b).

When evaluating the robustness of a model, the key is to measure the amount of distortions, when added to an image, cause the model's prediction to be incorrect regardless of predicted labels. Thus, in this paper, we focus on untargeted attacks of the following attack methods:

- **Iterative Fast Gradient Sign Mehtod (I-FGSM).** Kurakin *et al.* (Kurakin et al., 2016) proposed the I-FGSM attack to address the low success rate of the FGSM attack (Goodfellow et al., 2014) by applying FGSM multiple times with a finer distortion. One step of I-FGSM is:

$$x_{i+1} \leftarrow \text{clip}[x_i + \epsilon \ \text{sgn}(\nabla J(x_i, t))] \tag{1}$$

  where $\text{sgn}(\nabla J(x_0, t))$ is the sign of the gradient of the training loss *w.r.t.* x_i, and $\text{clip}(x)$ is used to guarantee that the generated image x is valid within the pixel range. This operation maximizes the training loss J by updating x. The I-FGSM repeats this step multiple times, if the number of iterations is set to T, the per-iteration perturbation is set to $\frac{\epsilon}{T}\text{sgn}(\nabla J(x_0, t))$. It usually finds adversarial examples with small ℓ_∞ distortions. In this paper, we use the average ℓ_∞ distortions of adversarial images constructed by the I-FGSM attack as a measurement of the robustness of models.

- **Carlini and Wagner's attack (CW).** Carlini & Wagner (2016) formulates the problem of generating untargeted adversarial examples as the following optimization problem:

$$\begin{array}{ll} \min\limits_{x} & cf(x,t) + ||x - x_0||_2^2 \\ \text{s.t.} & x \in [0,1]^n \end{array} \tag{2}$$

  where n is the dimension of the images, and $f(x, t)$ is a loss function to determine if an attack succeeds for input x with an original label t. In this work, the following loss function is used:

$$f(x,t) = \max\left\{\text{Logit}(x)_t - \max_{i \neq t}\left[\text{Logit}(x)_i\right], -\kappa\right\} \tag{3}$$

  where $\text{Logit}(x)$ denotes the vector representation of x at the logit layer, and κ denotes the confidence level. CW is by far one of the strongest attacks that construct adversarial images with small ℓ_2 perturbations. In this paper, we use the average ℓ_2 distortions of adversarial images constructed by this attack as another measurement of the robustness of models.

Table 1: Architecture and size of neural networks under robustness evaluation for each dataset

Dataset	Name	Layers	Parameters
CIFAR-10	2-Layer Perceptron	2	3,157,002
	Simple CNN	7	1,147,978
	ResNet	32	467,946
	DenseNet	40	1,059,298
MNIST	2-Layer Perceptron	2	814,090
	Simple CNN	7	312,202
	ResNet	20	272,778
STL-10	Simple CNN	20	7,257,322
	ResNet	20	278,186
	DenseNet	40	1,059,298
Tiny ImageNet	Simple CNN	20	6,044,072
	ResNet	20	323,816

3.3 DATASETS

We choose four representative datasets for our experiments: MNIST, CIFAR-10, STL-10[1] and Tiny ImageNet[2]. MNIST (LeCun et al., 1998) consists of 60,000 training data and 10,000 testing data for handwritten digit recognition. CIAFR-10 (Krizhevsky & Hinton, 2009) consists of 60,000 color images with 10 classes. STL-10 contains images with higher resolutions, which are acquired from labeled examples on ImageNet (Coates et al., 2011). It has 13,000 labeled data in total for 10 classes. In this paper, we divide the labeled data into 10,000 training data (1,000 per class) and 3,000 testing data (300 per class). Tiny ImageNet is a scaled-down version of ImageNet with 200 classes: each class has 500 training images, 50 validation images, and 50 test images. We use the validation images for accuracy and robustness evaluation since the labels of test images are not publicly available.

4 EXPERIMENTS

This section presents the details of our experiments, and summarizes and discusses our findings. Recall our goal is to reveal how training data affect the accuracy and robustness of neural networks.

4.1 SETUP

For each dataset in Section 3.3, we partition its training set into n sub-datasets with a strict inclusion relationship: $S_1 \subset S_2 \subset ... \subset S_n$. Each sub-dataset contains the same number of images for each training label. For each neural network architecture $\mathcal{M}$ described in Section 3.1, we train a model on each of these sub-datasets to obtain $\mathcal{M}_1, ..., \mathcal{M}_n$. Finally, we record the accuracy and robustness of each $\mathcal{M}_i$. We use the same set of sub-datasets to train respective neural networks for each dataset.

For each dataset, we study several different neural networks shown in Table 1. On one hand, the neural networks are chosen based on their performance on the datasets. For example, 2-layer perceptron and 7-layer Simple CNN are used on CIFAR-10 and MNIST, but not STL-10 and Tiny ImageNet since they are generally considered to be beyond simple model architectures[3]. On the other hand, our choices include fully-connected network (2-layer perceptron), vanilla CNN (Simple CNN), CNN with residual connections (ResNet) and CNN with densely connected blocks (DenseNet) to ensure the sufficient diversity in neural network architectures.

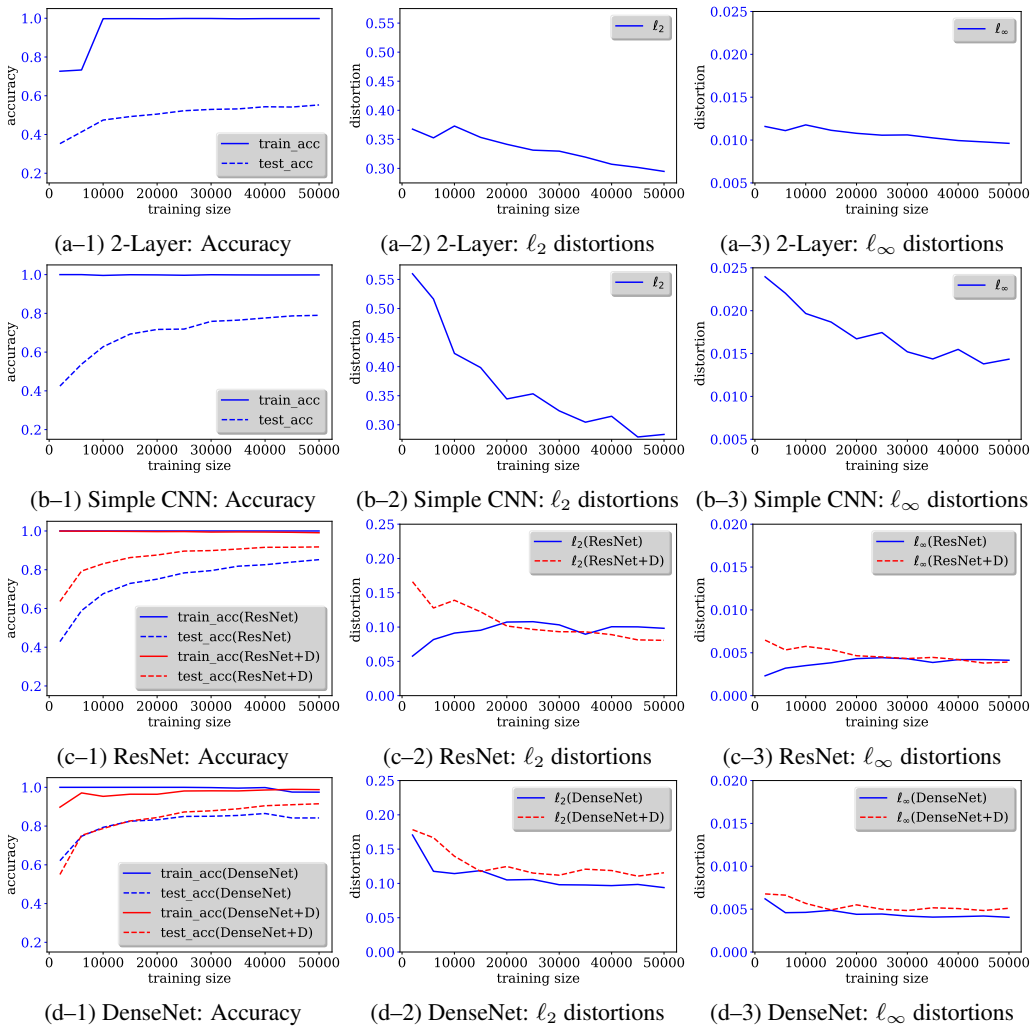
[1] https://cs.stanford.edu/~acoates/stl10/

[2] https://tiny-imagenet.herokuapp.com/

[3] We do not apply DenseNet on MNIST due to its simplicity.

To evaluate the robustness of each model, we conduct the I-FGSM attack to obtain adversarial images in ℓ_∞ distortions and the CW attack to obtain those in ℓ_2 distortions. We only consider original images that are correctly classified to avoid trial attacks when generating adversarial images. Given an attack on $\mathcal{M}_i$ with a valid input image x of label t, we define the attack to be successful if $\mathcal{M}_i(x') \neq t$, where x' is the adversarial image generated by the attack. The *attack success rate* is defined by the percentage of successful attacks. Both the CW and I-FGSM attacks can achieve 100% success rate on all models while the average distortions vary with different models and datasets. Equation (4) computes the robustness score of a model $\mathcal{M}_i$ where X is the set of images that all $\mathcal{M}_i$ $(1 \leq i \leq n)$ predict correctly.

$$Robustness(\mathcal{M}_i) = \frac{\sum_{x \in X} ||x - x'||_p}{|X|} \tag{4}$$



(a–1) 2-Layer: Accuracy (a–2) 2-Layer: ℓ_2 distortions (a–3) 2-Layer: ℓ_∞ distortions

(b–1) Simple CNN: Accuracy (b–2) Simple CNN: ℓ_2 distortions (b–3) Simple CNN: ℓ_∞ distortions

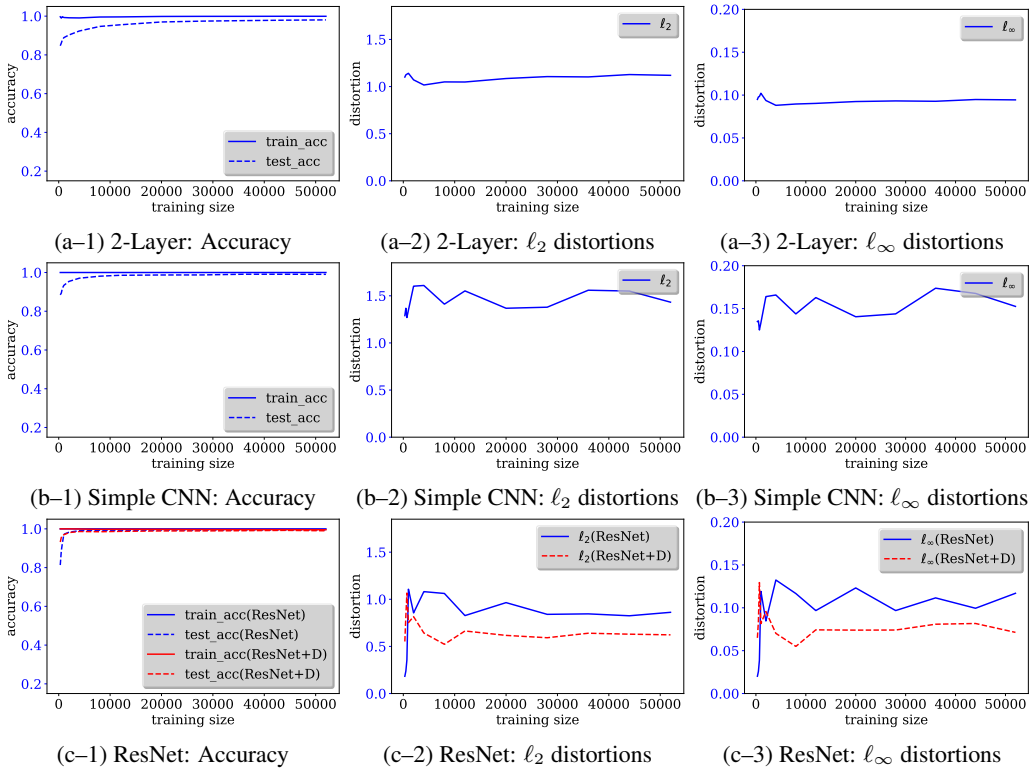
(c–1) ResNet: Accuracy (c–2) ResNet: ℓ_2 distortions (c–3) ResNet: ℓ_∞ distortions

(d–1) DenseNet: Accuracy (d–2) DenseNet: ℓ_2 distortions (d–3) DenseNet: ℓ_∞ distortions

Figure 1: Results of accuracy and robustness of different neural networks on increasing training data on CIFAR-10 dataset. (ResNet+D represents ResNet models trained with data augmentation; DenseNet+D represents DenseNet models trained with data augmentation.)

4.2 Relationship between size of training data, model accuracy and robustness

Results on CIFAR-10, MNIST, STL-10 and Tiny ImageNet are shown in Figs 1–4, respectively. The x-axis of each figure shows the size of sub-datasets, and the y-axis shows the accuracy or robustness

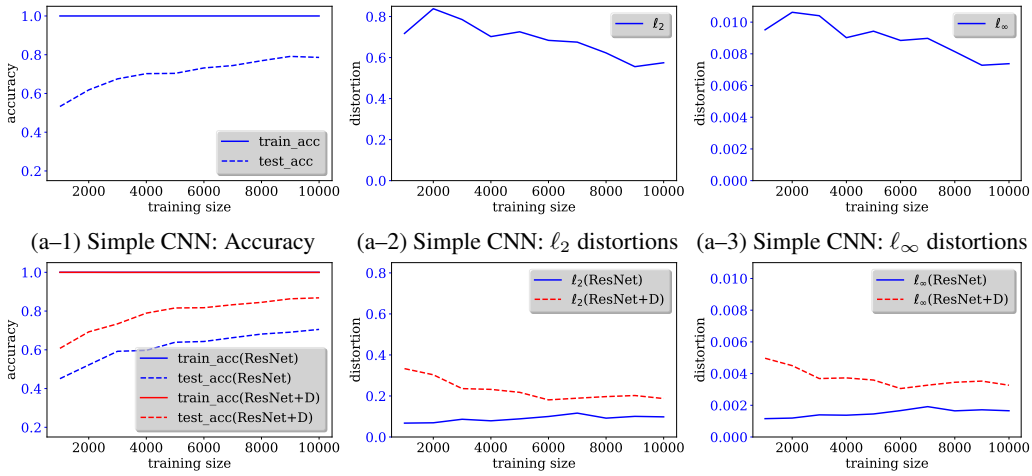


(a–1) 2-Layer: Accuracy (a–2) 2-Layer: ℓ_2 distortions (a–3) 2-Layer: ℓ_∞ distortions

(b–1) Simple CNN: Accuracy (b–2) Simple CNN: ℓ_2 distortions (b–3) Simple CNN: ℓ_∞ distortions

(c–1) ResNet: Accuracy (c–2) ResNet: ℓ_2 distortions (c–3) ResNet: ℓ_∞ distortions

Figure 2: Results of accuracy and robustness of different neural networks on increasing training data on MNIST dataset.(ResNet+D represents ResNet models trained with data augmentation.)

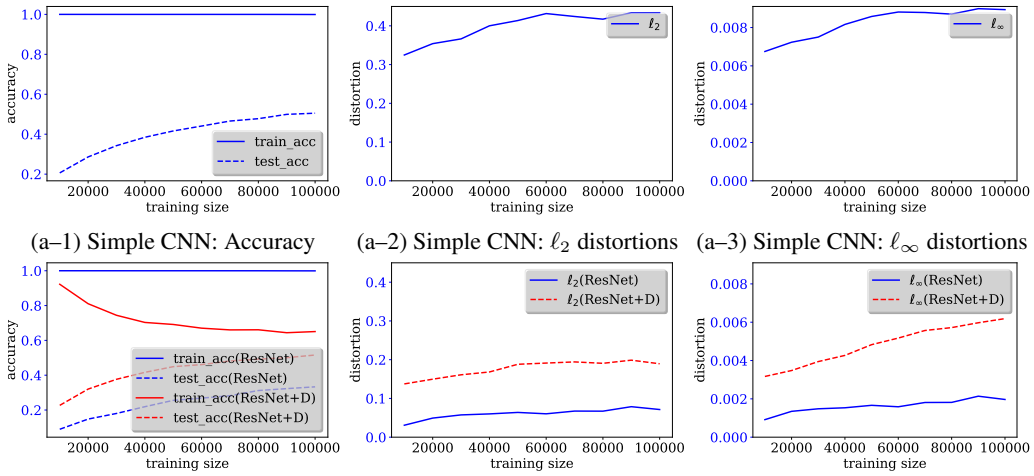


(a–1) Simple CNN: Accuracy (a–2) Simple CNN: ℓ_2 distortions (a–3) Simple CNN: ℓ_∞ distortions

(b–1) ResNet: Accuracy (b–2) ResNet: ℓ_2 distortions (b–3) ResNet: ℓ_∞ distortions

Figure 3: Results of accuracy and robustness of different neural networks on increasing training data on STL-10 dataset. (ResNet+D represents ResNet models trained with data augmentation.)

of the models trained on each sub-dataset. The first column of each figure shows how the accuracy of different neural networks changes with increased size of the training data. When training more complex models like ResNet and DenseNet, we consider both with and without data augmentation. Regarding data augmentation, we apply random shifting and flipping for ResNet; and random shifting, flipping and rotation for DenseNet.



(a–1) Simple CNN: Accuracy (a–2) Simple CNN: ℓ_2 distortions (a–3) Simple CNN: ℓ_∞ distortions

(b–1) ResNet: Accuracy (b–2) ResNet: ℓ_2 distortions (b–3) ResNet: ℓ_∞ distortions

Figure 4: Results of accuracy and robustness of different neural networks on increasing training data on Tiny ImageNet dataset. (ResNet+D represents ResNet models trained with data augmentation.)

The second and third column of each figure show how the robustness of models trained on increasingly larger training datasets changes measured by the ℓ_2 distortions and ℓ_∞ distortions. For example, in Fig (a–2), the x-axis shows the size of the sub-datasets, and the y-axis shows the robustness of all 2-layer perceptron models (11 in total) trained with each of the sub-datasets. Their robustness is measured by the average ℓ_2 distortions of the successful CW attacks on all 11 models. Similarly, the third column of Fig 1 shows how the robustness changes measured by the ℓ_∞ distortions of I-FGSM. On each dataset, we use the same scale to draw the robustness of different neural networks for an easy comparison across different neural networks. Note that the robustness of different models might reside in different ranges. For example, on CIFAR-10, 7-layer Simple CNN model is overall more robust than DenseNet (ℓ_2 distortion in the range of 0.30 to 0.55, versus 0 to 0.25).

Summarizing results on all four datasets and different network architectures, we observe the following:

- For CIFAR-10 and STL-10 in Figs 1 and 3, we observe that, when the accuracy has reached a certain level, increasing the size of training dataset will continue to improve the accuracy but sacrifice the robustness. ResNet (without data augmentation) is the only exception, where the robustness increases along with the accuracy.
- In the case of Tiny ImageNet in Figure 4, both accuracy and robustness increase when a larger training dataset is used. It is worth noting that Tiny ImageNet only contains a small subset of ImageNet and is insufficient for obtaining good test accuracy. The model is still in the region of starving for training data. In this case, increasing the number of training data will benefit both the accuracy and robustness.
- For the MNIST dataset, test accuracy quickly increases to a very high level and then saturates. The robustness of the model does not quite change when increasing the dataset size, and the accuracy neither changes. Since MNIST is a simple dataset, newly added training examples are likely redundant, and does not increase accuracy nor robustness.
- Generally, a more accurate model has worse robustness. For example, for CIFAR-10, ResNet and DenseNet show much worse robustness than simple models like 2-layer perceptron and 7-layer Simple CNN. This observation is consistent with the findings of Su et al. (2018) on the ImageNet dataset.

Overall, when the models are supplied with small amount of data, increasing training data helps them to find a better and clearer decision boundary, resulted in better accuracy and robustness. However, while more training data may lead to higher accuracy, the decision boundaries can become complicated and delicate, causing adversarial examples to forge. Thus, there is usually a "turning point" where the robustness starts to decrease while the accuracy keeps increasing. The turning point

is model and dataset dependent. In our experiments, the turning points occurred in CIFAR-10 and STL-10 due to the abundant date they contain *w.r.t.* Simple CNN, ResNet with data augmentation and DenseNet. For ResNet without data augmentation, the training accuracy has not improved sufficiently due to the lack of data, and therefore the turning point has not reached. Data augmentation essentially increases the size of dataset; consequently models can achieve higher accuracy leading to the emergence of turning points, as a result, decreasing robustness. Similarly the turning point did not occur on Tiny ImageNet due to the data shortage.

5 Related Work

Robustness Evaluation. One way to evaluate the robustness of a DNN is to find the minimal adversarial distortion (in terms of a particular form of ℓ_p norm). A neural network is said to be more robust if it tolerates larger amount of adversarial perturbations. Although finding the exact minimal adversarial distortion is NP-hard (Katz et al., 2017; Sinha et al., 2017; Ehlers, 2017), an adversarial attack can be deemed as an upper bound of the minimal adversarial distortion and is widely adopted for robustness evaluation. Existing work (Goodfellow et al., 2014; Carlini & Wagner, 2016; Xu et al., 2018; Metzen et al., 2017; Xiao et al., 2018b; Sun et al., 2018; Cheng et al., 2018; Carlini & Wagner, 2018; Xiao et al., 2018a; Chen et al., 2018) adopts this technique to evaluate the robustness of DNNs for many different tasks.

On the other hand, several approaches have been proposed to find the lower bounds of minimal adversarial distortion. Szegedy et al. (2013) offers a very loose result by bounding the global Lipschitz constants; Hein & Andriushchenko (2017) provides instance-specific lower bounds by analytically deriving the local Cross-Lipschitz constant for 2-layer networks. Weng et al. (2018b) empirically estimates the local Cross-Lipschitz constant for larger networks without statistical guarantees. Gehr et al. (2018); Weng et al. (2018a) give certified lower bounds of the minimum distortion by exploiting the special structure of ReLU networks. Unfortunately, these approaches can only be applied to relatively small networks due to the restriction of methodology or limited computational resources, therefore providing guaranteed lower bounds for arbitrary networks is still out of reach.

Accuracy and Robustness Tradeoffs. Su et al. (2018) study the relationship between accuracy and robustness. The authors investigate 18 well-trained neural networks on ImageNet and compare their accuracy and the robustness. They observe that more accurate models can exhibit worse robustness. For example, according to this study, ResNet is more accurate than AlexNet while being significantly less robust. Schmidt et al. (2018) conclude that robust generalization (*i.e.*, good accuracy and robustness) is significantly harder than standard generalization (*i.e.*, good accuracy); and robust generalization may need significantly more data even under a simple (but unrealistic) assumption that data were drawn from a Gaussian distribution.

Our paper sheds light on how training data affect both accuracy and robustness. Specifically, for a given model architecture, we study how accuracy and robustness change with increased amount of training data.

Data Poisoning Attacks. Data poisoning attacks generally inject false training data with the aim of corrupting the learned model by malicious users (Steinhardt et al., 2017; Biggio et al., 2012; 2014; Xiao et al., 2015; Rubinstein et al., 2009; Mei & Zhu, 2015; Yang et al., 2017), *etc.*. Under this setting, the accuracy of the learned model drops due to the injected false training data. In our study, we add more natural training data to train a model, the accuracy of the model increases while its robustness may drop.

6 Conclusion

This paper studies how robustness changes with increased training data for several representative neural networks on different datasets. Our experimental results show that with increased training data, both accuracy and robustness improve initially, however, there exists a turning point after which accuracy keeps increasing while robustness starts to decrease. Such turning points are different for different datasets and neural networks. This study also sheds light on several possible directions for future work such as how to select training data to make a model more robust without sacrificing its

accuracy and how to devise new attacks on robustness by flooding a model with excessive amount of natural training data.

REFERENCES

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv:1206.6389*, 2012.

Battista Biggio, Samuel Rota Bulò, Ignazio Pillai, Michele Mura, Eyasu Zemene Mequanint, Marcello Pelillo, and Fabio Roli. Poisoning complete-linkage hierarchical clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 42–52, 2014.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv:1608.04644*, 2016.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv:1801.01944*, 2018.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Show-and-fool: Crafting adversarial examples for neural image captioning. *arXiv:1712.02051*, 2017a.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 2587–2597, 2018.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. *arXiv:1709.04114*, 2017b.

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *arXiv:1803.01128*, 2018.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.

Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pp. 269–286, 2017.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv:1707.08945*, 1, 2017.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin T. Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy*, pp. 948–963, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pp. 2266–2276, 2017.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, 2017.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv:1707.07328*, 2017.

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, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv:1611.01236*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Shike Mei and Xiaojin Zhu. The security of latent dirichlet allocation. In *Artificial Intelligence and Statistics*, pp. 681–689, 2015.

Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *The IEEE International Conference on Computer Vision*, 2017.

Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and JD Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *ACM SIGCOMM Conference on Internet Measurement*, pp. 1–14, 2009.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv:1710.10571*, 2017.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, pp. 3517–3529, 2017.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

Mengying Sun, Fengyi Tang, Jinfeng Yi, Fei Wang, and Jiayu Zhou. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. *arXiv:1802.04822*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for ReLU networks. *arXiv:1804.09699*, 2018a.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv:1801.10578*, 2018b.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv:1801.02610*, 2018a.

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

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, pp. 1689–1698, 2015.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4951–4961, 2018.

Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. Generative poisoning attack method against neural networks. *arXiv:1703.01340*, 2017.