DATA QUALITY MATTERS FOR ADVERSARIAL TRAINING: AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple intriguing problems are hovering in adversarial training, including robust overfitting, robustness overestimation, and robustness-accuracy trade-off. These problems pose great challenges to both reliable evaluation and practical deployment. Here, we empirically show that these problems share one common cause—low-quality samples in the dataset. Specifically, we first propose a strategy to measure the data quality based on the learning behaviors of the data during adversarial training and find that low-quality data may not be useful and even detrimental to the adversarial robustness. We then design controlled experiments to investigate the interconnections between data quality and problems in adversarial training. We find that when low-quality data is removed, robust overfitting and robustness overestimation can be largely alleviated; and robustness-accuracy trade-off becomes less significant. These observations not only verify our intuition about data quality but may also open new opportunities to advance adversarial training.

1 INTRODUCTION

Adversarial training (Goodfellow et al., 2015; Huang et al., 2015; Kurakin et al., 2017; Madry et al., 2018) is arguably the most effective way to establish the robustness of deep neural networks against adversarial perturbations. Nevertheless, several intriguing problems and phenomena hover in adversarial training practice, including (1) robust overfitting (Rice et al., 2020), (2) robustness overestimation[1] (Uesato et al., 2018; Mosbach et al., 2018; Croce & Hein, 2020b; Chen & Gu, 2020), and (3) robustness-accuracy trade-off (Papernot et al., 2016; Su et al., 2018; Tsipras et al., 2019; Zhang et al., 2019).

In this work, we focus on the data in adversarial training and show that the three aforementioned problems are all interconnected with the quality of the data employed in adversarial training. Specifically, we measure the data quality based on their stability of being learned through training and show that unstably learned examples are of low quality to adversarial training. We show that low-quality data can be one common cause of the three aforementioned problems. To the best of our knowledge, this is the first time the effect of the data quality being systematically studied in adversarially robust learning. This is also the first time that all these problems are investigated in a holistic manner.

As a demonstration, we partition the training set of CIFAR-10 (Krizhevsky, 2009) into two exclusive subsets with equal size and balanced classes, while ensuring all the examples in one subset have higher quality than all the examples in the other one in a class-wise manner. We then conduct both standard training and adversarial training with Projected Gradient Descent (PGD) (Madry et al., 2018). The model is fixed as pre-activation ResNet-18 (He et al., 2016). As shown in Figure 1, adversarial training on the high-quality half yields *high* robustness, *almost no* robust overfitting, *genuine* robustness against the popular PGD attack, and *minor* robustness-accuracy trade-off that is only noticeable in the late stage of training, which is in sharp contrast to that on the low-quality half.

We further conduct controlled experiments on three real-world datasets, CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and Tiny-ImageNet (Le & Yang, 2015), to explore the interconnection between data quality and those three problems in adversarial training. Our experiments cover various

[1] Robustness overestimation refers to the problem that the adversarial robustness may be spuriously high against certain types of adversaries.

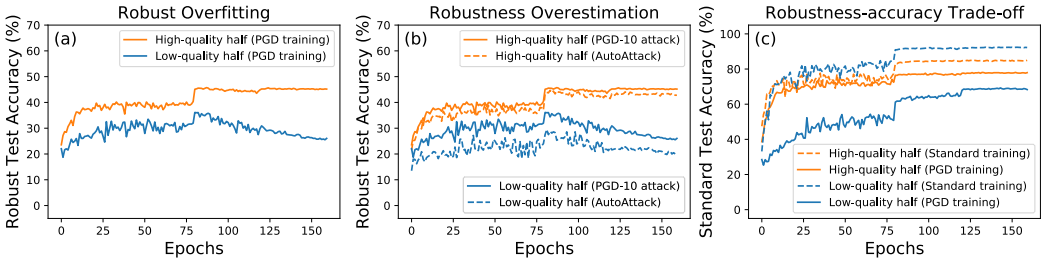


Figure 1: Despite being equal-size and class-balanced, the high-quality half and low-quality half of the CIFAR-10 training set have drastically different behaviors in adversarial training.

sample sizes, training methods (e.g., PGD and TRADES (Zhang et al., 2019)), and neural architectures (e.g., pre-activation ResNet and Wide ResNet (WRN) (Zagoruyko & Komodakis, 2016)).

Our major findings are summarized as follows.

- Low-quality data may not be useful or even detrimental to adversarial training.
- Low-quality data causes robust overfitting in adversarial training.
- Adversarial training with low-quality data can lead to overestimated robustness against certain adversaries such as the PGD attack.
- Robustness-accuracy trade-off is only prominent when the dataset contains low-quality data. When using high-quality data only, the standard accuracy achieved by adversarial training and standard training is comparable.

The remainder of this paper is organized as follows. In Section 2, we briefly review the existing works that are most related to our data-centric investigation of adversarial training. In Section 3, we propose a strategy to estimate the data quality based on learning stability, and show that unstably-learned data are of low quality to adversarial training. Sections 4.1, 4.2 and 4.3 investigate the interconnection between data quality and major problems in adversarial training including robust overfitting, robustness overestimation, and robustness-accuracy trade-off respectively. Conclusions and further implications are discussed in Section 5.

2 Related Work

Increasing attention has recently been paid to analyzing and improving the quality of the (training) data. Compared with standard learning, only a limited number of all learnable features would be beneficial to the adversarial robustness (Ilyas et al., 2019), thus making data quality more important in adversarially robust learning. Existing works have been focusing on the effect of sample size and data distribution in adversarial training. Schmidt et al. (2018) show that adversarially robust learning requires a much higher sample complexity than standard learning. They also show that such sample complexity requirement is highly sensitive to the data distribution. Ding et al. (2019) further observe that adversarial robustness achieved by adversarial training is sensitive to semantically-lossless distribution shift. Shafahi et al. (2019) show that the adversarial robustness achieved by any classifier is fundamentally limited by the properties of the data distribution. Our study is distinguished from previous works by investigating the datasets with the same size and within the same distribution. We also systematically study the other important problems in adversarial training beyond the performance.

Towards advancing adversarial robustness, substantially more works have been devoted to customizing the inner maximization or outer minimization in adversarial training based on the properties and behaviors of training data. Representative works including Wang et al. (2020), which shows that misclassified examples have greater impacts on adversarial robustness, and Zhang et al. (2020b), which observes that some training data are more robust to adversarial attack and will significantly improve the robustness against certain types of attacks if they are emphasized in adversarial training. We review more works along this line in Appendix D. Our study is distinguished from these works by both methodology and objective. In terms of the methodology, we identify the low-quality examples based on their behaviors of being learned throughout the adversarial training. Our approach to measure data quality yields more fine-grained results than misclassification to distinguish the individual examples, and more consistent results than the resistance to perturbation among different

training methods and neural architectures (see Appendix D). In terms of the objective, we are not aiming to compete with the state-of-the-art methods, but rather to reveal the important influence of data quality on the problems in adversarial training through rigorously controlled experiments and reliable robustness evaluation. Such comparative and systematic analyses are lacking in the current adversarially robust learning research.

Due to the space constraint, we will discuss more works specifically related to the benefit of data in adversarially robust learning, robust overfitting, robustness overestimation, and robustness-accuracy trade-off when it comes to Sections 3.2, 4.1, 4.2 and 4.3, respectively. More broadly related work regarding data quality in standard and robust learning will be discussed in Appendix A.

3 Data Quality in Adversarial Training

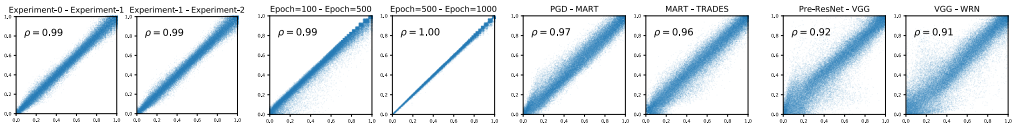
3.1 Data Quality Measure: Learning Stability as an Example

We propose to measure the quality of the data in adversarial training based on the frequency of events that an example is predicted correctly under adversarial perturbation throughout the training, which we will refer as the *learning stability* in the rest of the paper. Specifically, given an example x, the model f and the optimizer ω, the learning stability $s(x; f, \omega)$ can be formulated as

$$s(x; f, \omega) = \frac{1}{T} \left| \{ t | f(t, x + \delta) = y(x),\ t \in \{1, \cdots, T\} \} \right|, \tag{1}$$

where δ is a ℓ_∞ norm-bounded adversarial perturbation, $f(t, \cdot)$ denotes the classifier at epoch t, $y(x)$ denotes the true label of example x, and T is the total number of epochs the model will be trained. A similar measure in standard training has been utilized by Moon et al. (2020) to regularize the confidence estimates of deep neural networks, and is shown to be roughly proportional to the probability that an example will be correctly predicted by a model.

We find that, although the absolute value of the learning stability can vary a lot (see Appendix C.2), the ranking of examples based on the learning stability is consistent across different experimental setups, such as random initializations, training epochs, training methods, and neural architectures as shown in Figure 2. This suggests that the learning stability can reflect the *intrinsic* "hardness" of an example in a *relative manner*.



(a) Random initializations (b) Training epochs (c) Different methods (d) Different models

Figure 2: Correlation between the rankings of the examples (scaled by the total number of examples) based on learning stability estimated by different training settings. Spearman's rank coefficient (ρ) is annotated at the upper left of each subfigure. Here we consider adversarial training methods including PGD, TRADES and MART (Wang et al., 2020), and neural architectures including pre-activation ResNet-18, VGG-19 (Simonyan & Zisserman, 2015) and WRN-16-10.

Towards a more accurate estimation of such hardness, we calculate the ranking of the learning stability based on the ensemble of multiple experiments (see Appendix C.1). Figure 3a and Figure 3b show samples of the examples with the lowest learning stability and the examples with the highest learning stability, respectively, in the CIFAR-10 training set. One can find that unstably-learned examples appear more ambiguous and may not align well with their given labels. Under adversarial perturbation with relatively small size, they cannot easily be identified even by humans. This aligns with our intuition that unstably-learned examples may be intrinsically hard.

In standard learning, hard examples are often essential to the model performance, which is manifested in real-world applications such as hard example mining (Kumar et al., 2010; Shrivastava et al., 2016) and ablation studies on the effect of removing hard examples (Toneva et al., 2019). Empirical evidence can also be found in our Figure 1(a) where standard training on the unstably-learned half yields a higher standard accuracy than that on the stably-learned half.

(a) Examples with the lowest learning stability (b) Examples with the highest learning stability

Figure 3: Examples selected from CIFAR-10 training set by learning stability. The top row shows the original images annotated with their true labels and the bottom row shows the corresponding adversarially-perturbed images annotated with their labels predicted by a model.

In adversarial training, rather surprisingly, we observe an opposite case. Unstably-learned examples are often not useful and may even hurt the robustness achieved by adversarial training (see Section 3.2). Therefore, in the rest of the paper, we refer those unstably-learned examples as *low-quality examples*, and measure the data quality relatively using the learning stability. It is worth mentioning that, in Appendix D, we show that it is possible to estimate the data quality based on other behaviors and properties of the data (e.g., predictive probability, minimum perturbation, and learning order) and all these estimations will yield consistent findings.

3.2 Unstably-learned data is of low-quality for adversarial training

In this section, we conduct controlled experiments on three real-world datsets and show that low-quality examples are not useful and may even be detrimental to adversarial training.

Related work. In adversarially robust learning, more data will not always yield better performance and sometimes even be detrimental. Ding et al. (2019) observes that the robustness achieved by adversarial training plateaus after training size is sufficiently large on MNIST (LeCun et al., 1998). Introducing additional data unrelated to the original dataset can hurt the adversarial robustness (Uesato et al., 2019; Gowal et al., 2020). Yang et al. (2020a) shows that a proper pruning of the dataset will benefit the adversarial robustness for non-parametric classifiers. A recent study also shows that label noise in the dataset may be the origin of adversarial vulnerability (Sanyal et al., 2020). However, we note that low-quality data is not simply label noise after manually inspecting the samples as those in Figure 3a. Moreover, in-depth studies show that the fractions of label noise in benchmark datasets are only marginal (Northcutt et al., 2021).

Experimental setup of controlled experiments. We conduct controlled experiments to dissect the effect of training data on adversarial training. Specifically, we gradually remove more examples from the training set and record the robustness achieved by adversarial training from scratch on an increasingly smaller training subset, while all other experiment settings remain the same. The training examples are removed either randomly, or in an ascending order of data quality, i.e., low-quality data will be removed first. By removing the same number of random data and low-quality data, it should be able to eliminate the effect of sample size in the comparison and reveal the effect of data quality.

We consider the robustness against ℓ_∞ norm-bounded adversarial attack with perturbation radius 8/255 throughout the paper. We evaluate the adversarial robustness against AutoAttack (AA) (Croce & Hein, 2020b), one of the adversarial attacks known to be relatively reliable. We employ popular adversarial training methods including PGD and TRADES combined with early stopping (Rice et al., 2020), which have been shown to attain the improvements of almost all other adversarial training variants (Croce & Hein, 2020b; Chen & Gu, 2020; Pang et al., 2021). We conduct experiments with pre-activation ResNet-18 in the main paper and Wide ResNet in the appendix. We present results on three datasets including CIFAR-10, CIFAR-100 and Tiny-ImageNet. Other experiment details can be found in Appendix F. The same experimental setup will be applied to Sections 4, 5 and 6 as well.

Results & Discussions. Figure 4 shows that removing random examples from the training set degrades the robustness achieved by adversarial training rapidly. In contrast, when removing low-quality examples first, the achieved robustness maintains or even increases. For instance, by PGD training, up to 40%, 60%, and 50% low-quality training examples can be removed without reducing

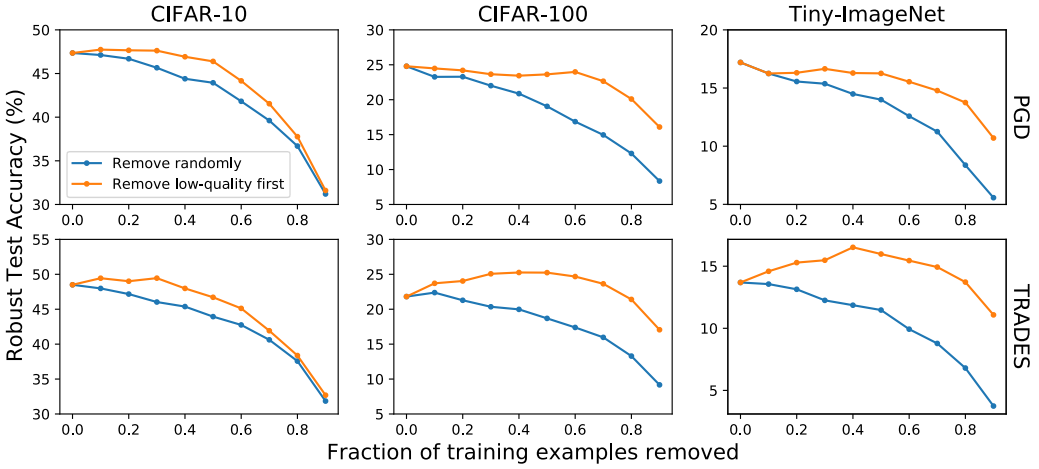


Figure 4: Best robust test accuracy obtained when removing different fractions of training examples either randomly or in an ascending order of data quality.

the robustness significantly ($< 1\%$) on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. More surprisingly, removing examples with the lowest data quality can achieve robustness even higher than that achieved on the entire training set for both PGD and TRADES on CIFAR-10, and TRADES for other datasets, although the training set is actually becoming smaller. We also validate the above observations with different perturbation sizes and neural achitectures in Appendix B.1.

4 Data Quality and Problems in Adversarial Training

In this section, we investigate the interconnection between data quality and problems in adversarial training by conducting extensive controlled experiments.

4.1 Robust Overfitting

The problem and related work. *Robust overfitting* is a prevalent phenomenon in adversarial training (Rice et al., 2020). Specifically, the robust test accuracy will constantly decrease after a certain point in adversarial training, resulting in an inferior final performance. Such overfitting occurs consistently across different datasets, training settings, adversary settings, and neural architectures, and cannot be completely eliminated other than using early stopping (Rice et al., 2020).

Low-quality data causes robust overfitting. Our controlled experiments suggest that the robust overfitting results from the low-quality examples. Figure 1 already shows that adversarial training on the high-quality half induces almost no robust overfitting. This suggests that robust overfitting is against our conventional understanding of overfitting in the sense that it can be mitigated by a smaller training sample size. Figure 5 further shows that when a particular fraction of low-quality examples is removed from the training set (e.g., on CIFAR-10, about 30% for PGD and 50% for TRADES), the robust overfitting would vanish. The smaller number of low-quality examples are removed from the training set, the more severe the robust overfitting will be. In contrast, when the examples are randomly removed from the training set, the robust overfitting is often prominent consistently across different training set sizes. This suggests that low-quality examples cause the robust overfitting, and more low-quality examples will make it more severe.

Motivated from the feature learning perspective, we conjecture that low-quality examples cause robust overfitting because robust features might be forgotten when low-quality examples are adversarially learned. More detailed explanation can be found in Appendix G.2.

Mitigate robust overfitting. Based on the above understanding, to mitigate robust overfitting, one may need to intentionally prevent the model from learning low-quality examples. Removing low-quality examples from the training set and the early stopping (Rice et al., 2020) are the most straightforward methods to prevent such learning. Recent works to mitigate robust overfitting

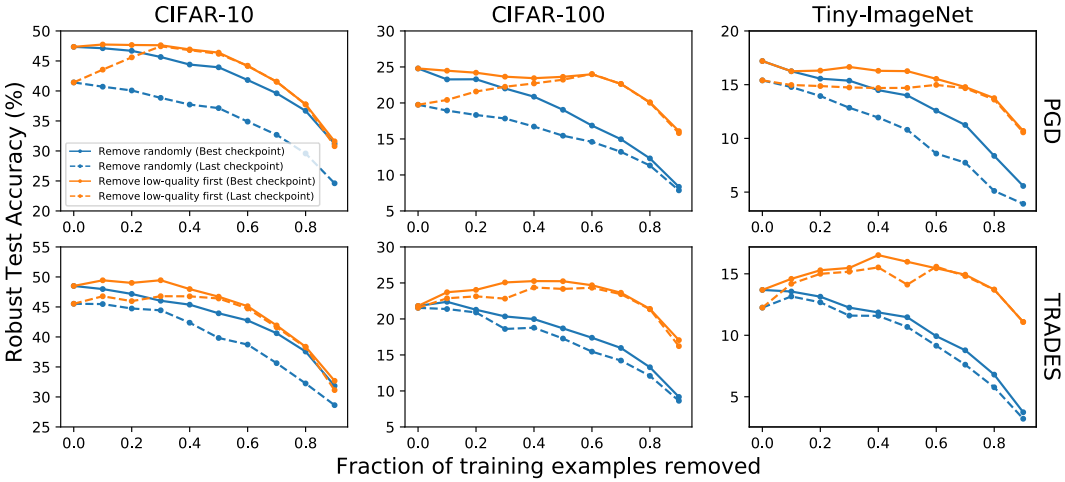


Figure 5: Robust test accuracy obtained at the best and last checkpoints by adversarial training when removing training examples either randomly or in an ascending order of data quality. The robust overfitting can be captured by the visual gap between the dashed and solid curves of the same color.

including regularizations of the flatness of the weight loss landscape (Wu et al., 2020; Stutz et al., 2021), introducing low-curvature activation functions (Singla et al., 2021) and adopting stochastic weight averaging (Izmailov et al., 2018) and knowledge distillation (Hinton et al., 2015) (Chen et al., 2021). These methods are likely to be consistent with our data-centric understanding as various regularization techniques may suppress the learning of low-quality examples.

4.2 Robustness Overestimation

The problem and related work. We refer *robustness overestimation* as the problem that the robustness of an adversarial training method may appear spuriously high against certain types of adversaries, but be significantly diminished by stronger adversaries (Chen & Gu, 2020). Such problem might result from a significantly more difficult inner maximization induced by highly non-linear loss landscape (Qin et al., 2019), thus is more subtle than the conventional *gradient masking* problem where the gradients of the model are completely not useful (Papernot et al., 2017; Tramèr et al., 2018; Athalye et al., 2018; Uesato et al., 2018; Engstrom et al., 2018).

We focus on the PGD attack as it is the most popular method to evaluate adversarial robustness. Previous works have already shown that PGD attack with insufficient iterations (Mosbach et al., 2018) and fixed step size (Croce & Hein, 2020b) may yield misleading robustness. Existing adversarial training methods may suffer from unreliable evaluation against PGD attack as mentioned in recent comprehensive studies (Croce & Hein, 2020b; Chen & Gu, 2020; Pang et al., 2021).

Low-quality data can cause robustness overestimation. For adversarial training methods, we show that low-quality data can cause overestimated robustness if evaluated against PGD attack by our controlled experiments. We report the **best robust test accuracy** throughout the training to rule out the effect of robust overfitting. We employ AutoAttack as a stronger adversary to evaluate the adversarial robustness. Although AutoAttack not necessarily reflects the true robustness (Tramèr et al., 2020), it is more reliable than PGD attack and is sufficient to reveal the significance of overestimation.

We denote the *overestimation gap* as the difference between the robust test accuracy evaluated against PGD-10 attack and AutoAttack. Figure 6 shows that the overestimation gap shrinks significantly when low-quality data is removed from the training set. For PGD training on all datasets, the robustness obtained on the entire training set is often higher than that obtained on a smaller training set with low-quality data removed if evaluated against PGD attack, but will instead be close to or even lower than the latter if evaluated against AutoAttack. In contrast, when the data is removed randomly from the training set, the overestimation gap largely maintains and is also consistently larger than that produced by removing an equal amount of low-quality data. These evidences show that low-quality data in the training set is an important source of the robustness overestimation.

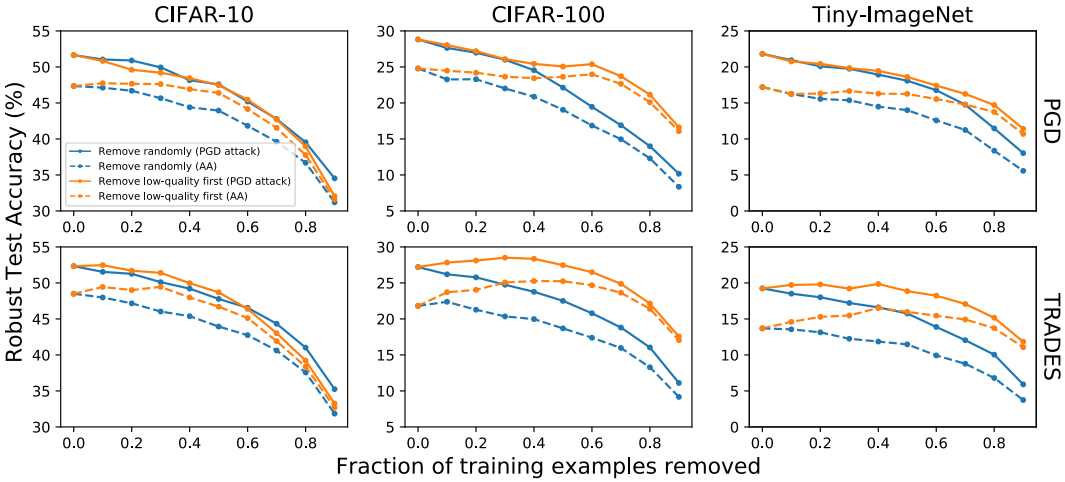


Figure 6: Best robust test accuracy obtained and evaluated against PGD attack and AA when removing training examples either randomly or in an ascending order of data quality. The overestimation gap can be captured by the visual gap between the dashed and solid curves of the same color.

In Appendix B.2, we also evaluate the robustness against a variety of black-box attacks including Square Attack (Andriushchenko et al., 2020), RayS (Chen & Gu, 2020) and transfer attack from a separately trained surrogate model, and show the robustness overestimation caused by low-quality data is consistent.

From a geometrical perspective, we suspect low-quality data causes robustness overestimation as they lie close to the decision boundary (see Appendix D.2). The perturbation balls of low-quality data from different classes may overlap, thus twisting the local loss landscape and complicating the maximization problem in gradient-based attacks (See more discussion in Appendix G.3).

Table 1: Robust test accuracy achieved by various training methods on the entire training set, evaluated by various attack methods. "*" indicates the transfer attack from a surrogate model.

	White-box Attack			Black-box Attack		
Dataset	PGD-10 (%)	PGD-1000 (%)	AA (%)	Square (%)	RayS (%)	PGD-1000* (%)
PGD	51.74	50.85	48.12	55.41	56.19	61.99
GAIRAT	59.51	59.33	32.34	41.98	44.17	58.06
Difference	+7.77	+8.48	−15.78	−13.43	−12.02	−3.93
TRADES	52.56	51.75	48.58	55.20	55.63	63.19
MART	53.81	53.39	46.57	52.76	53.45	60.73
Difference	+1.25	+1.64	−2.01	−2.44	−2.18	−2.46

Implications. Since low-quality data can cause robustness overestimation, it can be intentionally exploited to create spuriously high robustness against weak adversaries such as PGD attack. We find that this problem is ubiquitous among sophisticated adversarial training variants. We select two recently published methods GAIRAT (Zhang et al., 2020b) and MART and show that they suffer from robust overestimation due to the emphases on low-quality data in their training objectives (See Appendix A.1). As shown in Table 1, these methods trained with recommended settings (see Appendix F) can achieve significant improvements compared to their corresponding baselines when evaluated against white-box PGD attack, even with a large number of attack iterations. However, when evaluating against a stronger adversary such as AutoAttack, or black-box attacks, the performance of these methods are instead inferior. Similar observations have been made for these methods in recent works (Gowal et al., 2020; Hitaj et al., 2021). These results imply that emphasizing on the low-quality data in adversarial training may not be appropriate and it is important to validate the robustness against a wide variety of evaluation metrics as more methods are focusing on customizing the training for individual examples.

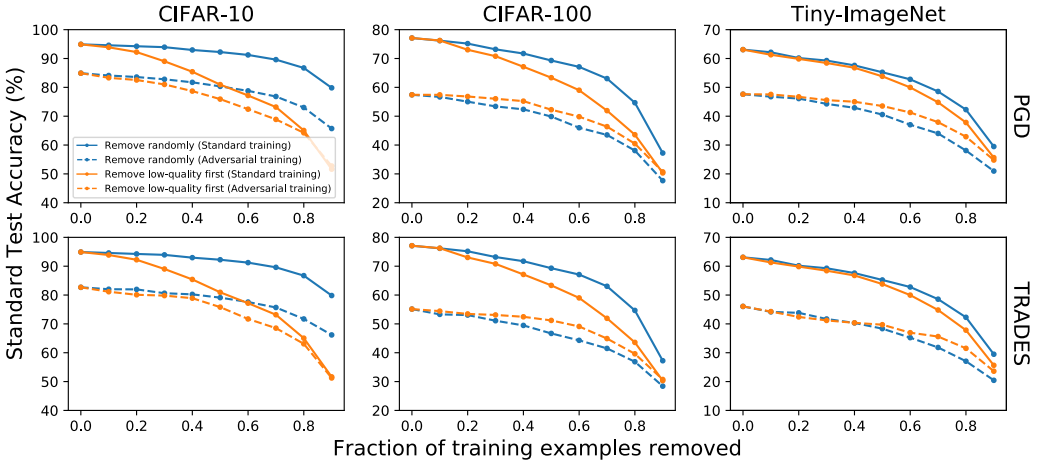


Figure 7: Standard test accuracy obtained by standard training and adversarial training when removing different fractions of training examples either randomly or in ascending order of data quality. The cross-generalization gap can be captured by the visual gap between the dashed and solid curves of the same color.

4.3 Robustness-accuracy Trade-off

The problem and related work. It is widely observed that adversarially training the model comes at the cost of standard accuracy, which is typically referred as the *robustness-accuracy trade-off* (Papernot et al., 2016; Su et al., 2018; Tsipras et al., 2019; Zhang et al., 2019). For certain learning problems, this trade-off might be inevitable either because no optimal classifier exists (Tsipras et al., 2019; Zhang et al., 2019) or the hypothesis classifier is not expressive enough (Nakkiran, 2019). Meanwhile, this trade-off is data-sensitive. Specifically, Ding et al. (2019) showed that the level of trade-off depends on the data distribution, and Tsipras et al. (2019) observed that the adversarial training is beneficial to standard accuracy when the training data is insufficient to train the model.

Robustness-accuracy trade-off is correlated with the data quality. We show that the robustness-accuracy trade-off is sensitive to the data quality within the same dataset. On high-quality data, this trade-off is not significant; however, on low-quality data, it is prominent.

We present the correlation between data quality and *cross-generalization gap* by controlled experiments. Following previous work (Chen et al., 2020), we define the cross-generalization gap as the standard test accuracy difference between standard training and adversarial training on the same dataset, which reflects the degree of robustness-accuracy trade-off induced by adversarial training on a specific dataset. Note here we report the standard test accuracy **at the last checkpoint** for all settings in case the selected checkpoint varies due to robust overfitting. As in Figure 7, the cross-generalization gap maintains the same or even enlarges as more randomly selected examples are removed from the training set. In contrast, the cross-generalization gap gradually shrinks as more low-quality examples are removed from the training set. For example, on all three datasets, adversarial training (either PGD or TRADES) and standard training yield comparable standard accuracy, when using the 10% highest quality examples only (i.e., 90% lowest quality ones removed).

Implications. Our identified correlation between data quality and robustness-accuracy trade-off complements the existing observations of the impact of data distribution on robustness. Besides the training set size (Tsipras et al., 2019) and distribution shift (Ding et al., 2019), the quality of the training data can also influence this trade-off. In Figure 7, the cross-generalization gap induced by PGD training is negative when using the 10% most high-quality examples on CIFAR-10 and CIFAR-100, which means that adversarial training is actually beneficial to the standard accuracy. Similar phenomena were previously only observed on MNIST or extremely small training sets (~ 100 examples) (Tsipras et al., 2019). Meanwhile, low-quality examples dominate the trade-off potentially because they are intrinsically ambiguous as shown in Figure 3a, where robust features may be difficult to learn. This is aligned with the understanding of trade-off from the feature learning

perspective (Tsipras et al., 2019; Ilyas et al., 2019). More detailed explanations motivated by this perspective will be discussed in Appendix G.1.

Mitigate robustness-accuracy trade-off. Based on the above understanding, to mitigate the robustness-accuracy trade-off one may resort to properly deal with low-quality data in adversarial training. Multiple existing works are likely to implicitly perform this by customizing either the inner maximization (Balaji et al., 2019; Ding et al., 2020; Cheng et al., 2020; Zhang et al., 2020a) or outer minimization (Huang et al., 2020; Wang et al., 2020; Zhang et al., 2020b) for individual examples (see more in Appendix D). Although these methods can mitigate the trade-off by increasing either robustness or accuracy while maintaining the other, the effectiveness may be unclear under state-of-the-art robust evaluation methods (Croce & Hein, 2020b) as shown in Section 4.2.

5 Conclusion and Discussions

In this work we systematically study the difficulties in adversarial training from a data quality perspective. We show that the data quality is an important but previously neglected aspect for adversarial training. It is strongly correlated with the current problems in adversarial training including robust overfitting, robustness overestimation and robustness-accuracy trade-off.

Towards understanding the difficulty of robust learning, our analyses of the data quality complements the existing analyses on the impact of data properties such as distribution and sample size. Robust learning through adversarial training can exhibit diverse levels of difficulty on individual examples in the same benchmark dataset. From human perspective, low-quality data is often ambiguous, which implies that adversarial training might be susceptible to complex and indistinct image conditions, thus challenging the feasibility of adversarial training in a realistic scenario where ambiguous data may prevail. One may note that achieving robustness through adversarial training on ImageNet (Russakovsky et al., 2015) is already shown to be much harder (Rozsa et al., 2016; Kurakin et al., 2017).

In practice, our analyses of the data quality provides a new methodology to analyze the effectiveness of adversarial training methods. The learning behavior of a method on low-quality data reveals the potential reason why it may work or not work. For example, TRADES employs an alternative adversarial loss that penalizes the difference between the predictive probability from an adversarial example and its clean counterpart, instead of the true label, which will impose weaker supervision on low-quality examples since low-quality examples typically have low predictive probabilities (see Appendix D.1). As a result, TRADES can underfit low-quality data thus potentially gaining robustness compared to PGD training under the same perturbation setting, which resembles the effect of directly removing low-quality data.

Furthermore, our analyses encourage a new perspective to utilize the benchmark dataset for adversarial training. The fact that removing low-quality data maintains or even improves robustness implies that adversarial training cannot achieve the optimal robustness on the entire dataset, especially when the model capacity is relatively low. A promising direction to boost the robustness is to design methods able to learn low-quality data properly such that it can also contribute to the robustness. Nevertheless, We empirically find that no existing adversarial training method, without resort to larger model capacity, can achieve reliable robustness improvement through proper learning of low-quality data. Whether or not such a method exists remains an interesting problem.

Lastly, as the introduction of additional labeled or unlabeled data becomes increasingly popular (Hendrycks et al., 2019a; Uesato et al., 2019; Carmon et al., 2019; Najafi et al., 2019), extra effort needs to be paid to picking high-quality data, which echos the practical suggestions made in previous works (Uesato et al., 2019; Gowal et al., 2020).

Limitations. We emphasize that removing low-quality examples only serves as a demonstration of our intuition of data quality in adversarial training, and should not be advocated as a competing method. Although removing low-quality examples can potentially advance the best robustness, it will inevitably cost standard accuracy (see Appendix E), since those hard examples are of high quality to standard training.

REPRODUCIBLITY STATEMENT

We conduct experiments on public benchmark. We will release implementations for all methods and scripts for all experiments on GitHub, under the Apache-2.0 license.

ETHIC STATEMENT

In this paper, we conduct a series of controlled experiments to investigate the interconnections between data quality and problems in adversarial training. We experiment on three benchmark datasets that are publicly available. We have shown that low-quality examples are helpful for standard training but may hurt the adversarial robustness. We emphasize that removing low-quality examples only serves as a probing method to demonstrate our intuition of data quality in adversarial training, and should not be advocated as a competing method. Therefore, we believe our work is ethically on the right side of spectrum.

REFERENCES

Maksym Andriushchenko, F. Croce, Nicolas Flammarion, and M. Hein. Square attack: a query-efficient black-box adversarial attack via random search. *ArXiv*, abs/1912.00049, 2020.

D. Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and S. Lacoste-Julien. A closer look at memorization in deep networks. *ArXiv*, abs/1706.05394, 2017.

Anish Athalye, Nicholas Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ArXiv*, abs/1802.00420, 2018.

D. Ayala. The theory and practice of item response theory. 2008.

Y. Balaji, T. Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *ArXiv*, abs/1910.08051, 2019.

Yoshua Bengio, J. Louradour, Ronan Collobert, and J. Weston. Curriculum learning. In *ICML '09*, 2009.

Nicholas Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Y. Carmon, Aditi Raghunathan, L. Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness. *ArXiv*, abs/1905.13736, 2019.

Haw-Shiuan Chang, E. Learned-Miller, and A. McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NIPS*, 2017.

J. Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *ICML*, 2020.

P. Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and C. Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *ArXiv*, abs/1709.04114, 2018.

Tianlong Chen, Zhenyu (Allen) Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*, 2021.

Minhao Cheng, Qi Lei, Pin-Yu Chen, I. Dhillon, and C. Hsieh. Cat: Customized adversarial training for improved robustness. *ArXiv*, abs/2002.06789, 2020.

F. Croce and M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020a.

F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020b.

Jiequan Cui, Shu Liu, L. Wang, and J. Jia. Learnable boundary guided adversarial training. *ArXiv*, abs/2011.11164, 2020.

G. W. Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. *ArXiv*, abs/1902.08336, 2019.

G. W. Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training. *ArXiv*, abs/1812.02637, 2020.

Elvis Dohmatob. Limitations of adversarial robustness: strong no free lunch theorem. *ArXiv*, abs/1810.04065, 2018.

S. Embretson and S. Reise. Item response theory for psychologists. 2000.

L. Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *ArXiv*, abs/1807.10272, 2018.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT*, 1995.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and P. Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *ArXiv*, abs/2010.03593, 2020.

Guy Hacohen, Leshem Choshen, and D. Weinshall. Let's agree to agree: Neural networks share classification order on real datasets. *arXiv: Learning*, 2019.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *ArXiv*, abs/1603.05027, 2016.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019a.

Dan Hendrycks, K. Zhao, Steven Basart, J. Steinhardt, and D. Song. Natural adversarial examples. *ArXiv*, abs/1907.07174, 2019b.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Dorjan Hitaj, Giulio Pagnotta, Iacopo Masi, and L. Mancini. Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *ArXiv*, abs/2103.01914, 2021.

Lang Huang, C. Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *ArXiv*, abs/2002.10319, 2020.

Ruitong Huang, B. Xu, Dale Schuurmans, and Csaba Szepesvari. Learning with a strong adversary. *ArXiv*, abs/1511.03034, 2015.

Andrew Ilyas, Shibani Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

Pavel Izmailov, Dmitrii Podoprikhin, T. Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *ArXiv*, abs/1803.05407, 2018.

A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *COLT*, 2020.

Harini Kannan, A. Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *ArXiv*, abs/1803.06373, 2018.

G. Katz, C. Barrett, D. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. *ArXiv*, abs/1702.01135, 2017.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

M. Kumar, Ben Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

A. Kurakin, Ian J. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015.

Y. LeCun, L. Bottou, Yoshua Bengio, and P. Haffner. Gradient-based learning applied to document recognition. 1998.

Mingchen Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *ArXiv*, abs/1903.11680, 2020.

A. Madry, Aleksandar Makelov, L. Schmidt, D. Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

Takeru Miyato, S. Maeda, Masanori Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, 2016.

Marius Mosbach, Maksym Andriushchenko, T. A. Trost, M. Hein, and D. Klakow. Logit pairing methods can fool gradient-based attacks. *ArXiv*, abs/1810.12042, 2018.

A. Najafi, S. Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. *ArXiv*, abs/1905.13021, 2019.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *ArXiv*, abs/1901.00532, 2019.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018. URL https://arxiv.org/pdf/1807.01069.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *ArXiv*, abs/2103.14749, 2021.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Xb8xvrtB8Ce.

Nicolas Papernot, P. McDaniel, Arunesh Sinha, and Michael P. Wellman. Towards the science of security and privacy in machine learning. *ArXiv*, abs/1611.03814, 2016.

Nicolas Papernot, P. McDaniel, Ian J. Goodfellow, S. Jha, Z. Y. Celik, and A. Swami. Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017.

Camilo Pestana, N. Akhtar, W. Liu, David Glance, and A. Mian. Adversarial perturbations prevail in the y-channel of the ycbcr color space. *ArXiv*, abs/2003.00883, 2020a.

Camilo Pestana, Wei Liu, David Glance, and A. Mian. Defense-friendly images in adversarial attacks: Dataset and metrics for perturbation difficulty. *ArXiv*, abs/2011.02675, 2020b.

R. Prudêncio, J. Hernández-Orallo, and A. Martinez-Usó. Analysis of instance hardness in machine learning using item response theory. 2015.

Chongli Qin, J. Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and P. Kohli. Adversarial robustness through local linearization. In *NeurIPS*, 2019.

Aditi Raghunathan, Sang Michael Xie, F. Yang, John C. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *ICML*, 2020.

Leslie Rice, Eric Wong, and J. Z. Kolter. Overfitting in adversarially robust deep learning. *ArXiv*, abs/2002.11569, 2020.

Kevin Roth, Yannic Kilcher, and T. Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *arXiv: Learning*, 2020.

Andras Rozsa, M. Günther, and T. Boult. Are accuracy and robustness correlated. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 227–232, 2016.

Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Amartya Sanyal, P. Dokania, Varun Kanade, and P. Torr. How benign is benign overfitting? *ArXiv*, abs/2007.04028, 2020.

L. Schmidt, Shibani Santurkar, D. Tsipras, Kunal Talwar, and A. Madry. Adversarially robust generalization requires more data. In *NeurIPS*, 2018.

A. Shafahi, W. Huang, Christoph Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? *ArXiv*, abs/1809.02104, 2019.

Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, 2016.

K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

Vasu Singla, Sahil Singla, David Jacobs, and Soheil Feizi. Low curvature activations reduce overfitting in adversarial training. *ArXiv*, abs/2102.07861, 2021.

M. Smith and T. Martinez. Improving classification accuracy by identifying and removing instances that should be misclassified. *The 2011 International Joint Conference on Neural Networks*, pp. 2690–2697, 2011.

M. Smith, T. Martinez, and C. Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95:225–256, 2013.

Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R. Venkatesh Babu. Guided adversarial attack for evaluating and enhancing adversarial defenses. *ArXiv*, abs/2011.14969, 2020.

David Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6969–6980, 2019.

David Stutz, Matthias Hein, and B. Schiele. Relating adversarially robust generalization to flat minima. *ArXiv*, abs/2104.04448, 2021.

D. Su, Huan Zhang, H. Chen, Jinfeng Yi, P. Chen, and Yupeng Gao. Is robustness the cost of accuracy? - a comprehensive study on the robustness of 18 deep image classification models. In *ECCV*, 2018.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and G. Gordon. An empirical study of example forgetting during deep neural network learning. *ArXiv*, abs/1812.05159, 2019.

Florian Tramèr, A. Kurakin, Nicolas Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *ArXiv*, abs/1705.07204, 2018.

Florian Tramèr, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *ArXiv*, abs/2002.08347, 2020.

D. Tsipras, Shibani Santurkar, L. Engstrom, A. Turner, and A. Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *ArXiv*, abs/1805.12152, 2018.

D. Tsipras, Shibani Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv: Machine Learning*, 2019.

Jonathan Uesato, Brendan O'Donoghue, A. Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *ArXiv*, abs/1802.05666, 2018.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? *ArXiv*, abs/1905.13725, 2019.

Yisen Wang, Difan Zou, Jinfeng Yi, J. Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. *ArXiv*, abs/2011.07478, 2020.

Tsui-Wei Weng, Huan Zhang, P. Chen, Jinfeng Yi, D. Su, Yupeng Gao, C. Hsieh, and L. Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *ArXiv*, abs/1801.10578, 2018.

Dongxian Wu, Shutao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *arXiv: Learning*, 2020.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, A. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 816–825, 2020.

Y. Yang, Cyrus Rashtchian, Yizhen Wang, and K. Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. In *AISTATS*, 2020a.

Y. Yang, Cyrus Rashtchian, Hongyang Zhang, R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. *arXiv: Learning*, 2020b.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.

Hongyang Zhang, Yaodong Yu, J. Jiao, E. Xing, L. Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *ArXiv*, abs/1901.08573, 2019.

Jingfeng Zhang, Xilie Xu, B. Han, Gang Niu, Li zhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. *ArXiv*, abs/2002.11242, 2020a.

Jingfeng Zhang, Jianing Zhu, Gang Niu, B. Han, M. Sugiyama, and M. Kankanhalli. Geometry-aware instance-reweighted adversarial training. *ArXiv*, abs/2010.01736, 2020b.

Tianyi Zhou, S. Wang, and J. Bilmes. Curriculum learning by dynamic instance hardness. In *NeurIPS*, 2020.