Improved Adversarial Robustness via Data Augmentation

Wenqian Ye, Xu Cao, Yunsheng Ma
Courant Institue of Mathematical Science
New York University
{wy2029, xc2057, ym2382}@nyu.edu

1 Introduction

Deep Neural networks are being used in a wide range of applications, such as self-driving cars [1] and medical diagnosis [2]. It's become more and more crucial to make sure that models are robust and generalize to a variety of input perturbations. Unfortunately, the addition of imperceptible adversarial perturbations can paradoxically lead to incorrect predictions of neural networks with high confidence [3]. There has been a lot of research towards understanding and producing adversarial perturbations [4], as well as developing defenses that are resistant to them [5]. While resilience and invariance to input perturbations are important for the deployment of machine learning models in a variety of applications, they can also have larger negative consequences for society, such as compromising privacy or increasing bias.

Madry et al. [6] suggested an adversarial training approach that feeds adversarially disturbed instances back into the training data. It is commonly considered to be one of the most effective methods for training robust deep neural networks. Figure 4 illustrates the intuition of this stratedy.

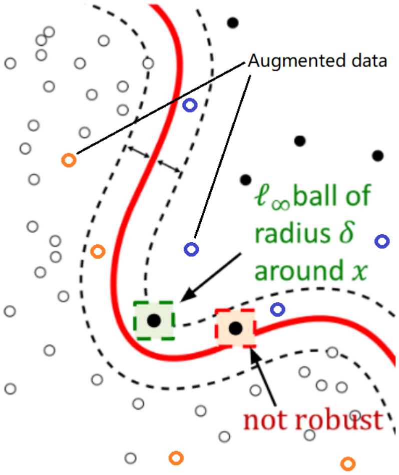


Figure 1: Illustration of the intuition of the data augmentation strategy. In this figure, the thick red line is the decision boundary. We consider ℓ_∞ norm perturbations. A point will be misclassified with ℓ_∞ perturbation if we can shift it by ϵ in the infinity norm and move it to the other side of the decision boundary. In other words, if we draw an ϵ ball around a data point and intersects the decision boundary. Then we are going to misclassify that point when added adversarial perturbation. This means that the red box is going to be misclassified, but the point in the green box will be classified correctly. The augmented data, in this example are the blue and orange points can help refine the decision boundary to increase the robust accuracy.

Preprint. Under review.

Various modifications to their orginal implementation have been offered [7]. However, in the last two years, the rate of increase in robust accuracy brought on by these newer methodologies has dropped substantially. When using additional data on CIFAR-10 against ℓ_∞ perturbations of size $\epsilon = 8/255$, the best known model[8] achieves a robust accuracy of 65.87%. Without this data, the same model achieves a robust accuracy of 57.14%. Focusing on approaches that don't employ any additional data, Rebuffi et al. [9, 10] propose to integrate data taken from generative models[11] with data augmentation approaches such as Cutout [12], based on the observation that model weight averaging [13] promotes robust generalization to a greater extent when robust overfitting is minimized. They train models with 64.20% robust accuracy on CIFAR-10 against ℓ_∞ norm-bounded perturbations of size $\epsilon = 8/255$, which achieved the state-of-the-art performance. Specifically, except for Gowal et al. [8]'s work, they outperformed all strategies that used additional data.

Code is available at `https://github.com/wenqian-ye/FML-robust`.

2 Related Works

Vision data augmentation is proven useful for many machine learning tasks. Such modification to the input image increases the overall size of the training set, therefore improving the model's robustness. Data augmentation can be divided into two sub-strategies: Heuristics-driven data augmentation and Data-driven data augmentation. Heuristics-driven data augmentation includes random crop, random flip, random rotation, and some sophisticated heuristic modification such as Cutout [12], CutMix [14], MixUp [15]. Mainstream data-driven data augmentation applies generative modeling. Generative modeling refers to the process of creating synthesis instances from the original dataset such that they retain similar characteristics [16]. The most famous generative modeling for data augmentation is GAN. Bowles et al. [17] describe GANs as a way to "unlock" additional information from the dataset. In recent years, diffusion probabilistic models are also widely used as generative modeling, such as DDPM [11].

For adversarial robustness against ℓ_p-norm attacks, Data augmentation has shown a major breakthrough in 2021. Gowal and Rebuffi et al. [9, 10] proposed to combine heuristics-driven and data-driven data augmentation together into adversarial training. Prior to their research, seldom researcher used data augmentation especially data generation to improve adversarial robustness [18, 19].

3 Method

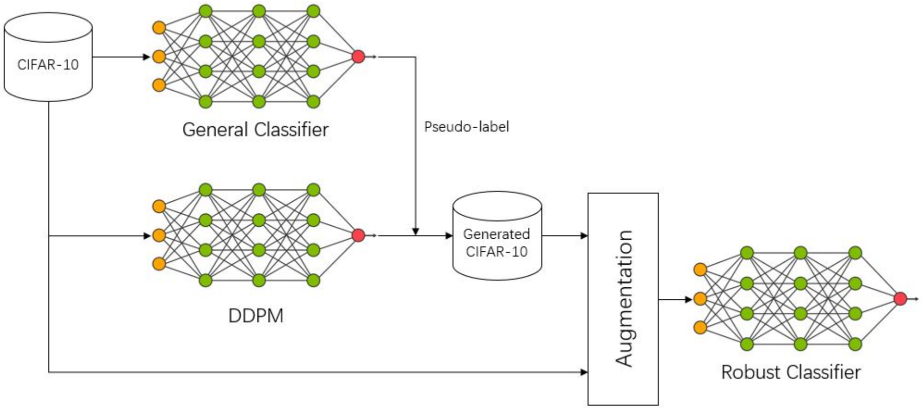


Figure 2

Our method initially trains a generative model (here we select DDPM suggested by [10]) and a non-robust classifier.

The non-robust classifier is used to provide pseudo-labels of the original data (CIFAR 10). The generative takes the original data (CIFAR 10) alongside with the pseudo-labels to produce complementary new augmented data.

After using heuristic data augmentation, such as random crop, random flip and random rotation, generated and original training data are combined to train a robust classifier. Stochastic Model Averaging [13] is applied in the robust classifier to ensure good generalization properties.

4 Theoretical Analysis

To show data augmentation can provably improve robustness, we analyze the problem through the lens of margin in the simplified binary non-linear classification setup. We provide the lower number of augmented data and the bounds on the margin.

Definition of the Margin. Given $S \subseteq \mathbb{R}^d \times \{\pm 1\}$, $f : \mathbb{R}^d \to \{\pm 1\}$ separates S, if $f(x) = y$ for all $(x, y) \in S$. Let $\mathcal{R}(S)$ denote the collection of separators of S. If $\mathcal{R}(S)$ is non-empty, we say that S is separable. Define the margin on S as :

$$\gamma_f(S) := \min_{(x,y)\in S} \operatorname{dist}(x, f^{-1}(-y))$$

We define $\gamma_f(S) = -\infty$ if $f \notin \mathcal{R}(S)$.

Augmented Margin. Let S' be the augmented set of the data. The margin $\gamma_f\left(S, S'\right)$ of f with respect to S, S' is defined by

$$\gamma_f\left(S, S'\right) = \gamma_f(S)$$

if $f \in \mathcal{R}\left(S^{\text{aug}}\right)$ and $-\infty$ otherwise.

Minimum number of augmented data. Suppose S is separable. If $\left|X'_+\right| \leq d$ or $\left|X'_-\right| \leq d$, then for any $\epsilon \in [0, \infty]$, either $\mathcal{R}_\epsilon\left(S^{\text{aug}}\right) = \emptyset$, or $\exists f \in \mathcal{R}_\epsilon\left(S^{\text{aug}}\right)$ such that $\gamma_f\left(S, S'\right) = 0$.

For all $n \geq 1$ and $\epsilon, r \in (0, \infty)$, there is some S of size n such that if $|S'| \leq |S|(d+1)$, then $\exists f \in \mathcal{R}_\epsilon\left(S^{\text{aug}}\right)$ such that $\gamma_f\left(S, S'\right) = 0$.

Therefore, for $S' \subseteq S_r$, to ensure that any $f \in \mathcal{R}_\epsilon\left(S^{\text{aug}}\right)$ has positive margin, we need $r \leq \epsilon, |S'| \geq 2d+2$, and $|S'| \geq |S|(d+1)$. These three conditions are sufficient to ensure positive margin.

Lower bounds on margin. If $r \leq \epsilon$, then there is a universal constant C such that if $N \geq Cd$, then with probability at least $1 - ne^{-d}$, $\forall f \in \mathcal{R}_\epsilon\left(S^{\text{aug}}\right)$,

$$\gamma_f\left(S, S'\right) \geq \frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}r.$$

Furthermore, if $\epsilon < \frac{\operatorname{dist}(X_+, X_-)}{4}$ then $\mathcal{R}_\epsilon\left(S^{\text{aug}}\right) \neq \emptyset$

Proof. Suppose $\epsilon \in (0, \infty]$ and $r \in (0, \infty)$ satisfies $r \leq \epsilon$.

Recall that $S' = \left\{x_i + z_i^{(j)}, y\right\}_{i\in[n], j\in[N]}$ where each $z_i^{(j)}$ is drawn uniformly at random from the sphere of radius r. Define $A_i = \left\{x_i + z_i^{(j)}\right\}_{j\in[N]}$, and $K_i = \operatorname{conv}\left(A_i\right)$.

Lemma. [20] Let $z_1, \ldots, z_N$ be drawn uniformly at random on $r\mathcal{S}^{d-1}$. Let $K = \operatorname{conv}\left(z_1, \ldots, z_N\right)$. Then there exists a constant $C > 0$ such that if $N \geq Cd$, then

$$\mathbb{P}\left(\frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}\mathcal{B}_r(0) \not\subset K\right) \leq e^{-d}$$

By this lemma, we know that with probability at least $1 - e^{-d}$, $\mathcal{B}_\rho\left(x_i\right) \subseteq K_i$ where

$$\rho = \frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}r.$$

Since $\left\|z_i^{(j)}\right\|_2 = r \leq \epsilon$, we have $A_i \subseteq \mathcal{B}_\epsilon\left(x_i\right)$. Hence, $R\left(A_i\right) \leq \epsilon$. By ϵ-respectfulness, we know that if $f \in \mathcal{R}_\epsilon\left(S \cup S'\right)$, then for all $x \in K_i, f(x) = y_i$. Also, $f(x) = y_i$ for all $x \in \mathcal{B}_\rho\left(x_i\right)$. This implies $d\left(x_i, f^{-1}\left(-y_i\right)\right) \geq \rho$.

Taking a union bound over $\forall i \in [n]$, with probability at least $1 - ne^{-d}$, if $f \in \mathcal{R}_\epsilon\left(S, S'\right)$, we can get $\gamma_f\left(S, S'\right) \geq \rho$. Thus proved. ■

Upper bounds on margin. Fix $\epsilon \in [0, \infty]$ and $r > 0$. There are absolute constants C_1, C_2 such that if $N > d$ and $\mathcal{R}_\epsilon\left(S^{\text{aug}}\right) \neq \emptyset$, then with probability at least $1 - 2e^{-C_2 d \log(N/d)}$, $\exists f \in \mathcal{R}_\epsilon\left(S^{\text{aug}}\right)$ such that

$$\gamma_f\left(S, S'\right) \leq \sqrt{C_1 \frac{\log(2N/d)}{d}} r.$$

Proof. We may assume $X_+ = \{0\}$ and $X'_+ = \{z_i\}_{i \in [N]}$. Let $K = \operatorname{conv}\left(\{0\} \cup X_+\right) = \operatorname{conv}\left(\{0\} \cup \{z_i\}_{i \in [N]}\right)$ and let $K' = \operatorname{conv}\left(\{0\} \cup X_+^{\text{aug}}\right) = \operatorname{conv}\left(\{0\} \cup \{\pm z_i\}_{i \in [N]}\right)$. Note that $K \subseteq K'$. Applying Proposition 3 in [20] to K' and Jensen's inequality, exist some constants C_1, C_2 such that

$$\mathbb{P}\left(\forall i \in [l], \frac{1}{\operatorname{vol}\left(F_i\right)} \int_{F_i} |x| dx \leq \sqrt{C_1 \frac{\log(2N/d)}{d}} r\right) \geq 1 - 2e^{-C_2 d \log(N/d)}$$

Here, $\{F_i\}_{i \in [l]}$ are the facets of K'.
Define

$$\delta = \sqrt{C_1 \frac{\log(2N/d)}{d}} r.$$

The term $\operatorname{vol}\left(F_i\right)^{-1} \int_{F_i} |x| dx$ is the average distance of the points on the facet F_i to the origin. If this average is bounded above by δ, then there is at least on point on the facet that is at a distance less than or equal to δ to the origin. Therefore, with probability at least $1 - 2\exp\left(-C_2 d \log(N/d)\right)$, there is a point on each F_i of distance at most δ to the origin. Thus, $d\left(0, \partial K'\right) \leq \delta$. Since $K \subseteq K'$, this implies that with the same probability, $d(0, \partial K) \leq \delta$.

Now, define a function $f : \mathbb{R}^d \to \mathbb{R}$ to be $+1$ on K and -1 elsewhere. Since $R\left(X'_+\right), R\left(X'_-\right) \leq r \leq \epsilon$, Proposition 2 implies that $\operatorname{conv}\left(X'_+\right) \cap \operatorname{conv}\left(X'_-\right) = \emptyset$. Therefore f is well-defined and $f \in \mathcal{F}_\epsilon\left(S^{\text{aug}}\right)$. We then have by construction of f,

$$\gamma_f\left(S, S'\right) \leq d\left(0, f^{-1}(-1)\right) = d(0, \partial K)$$

Thus proved. ■

5 Experiments

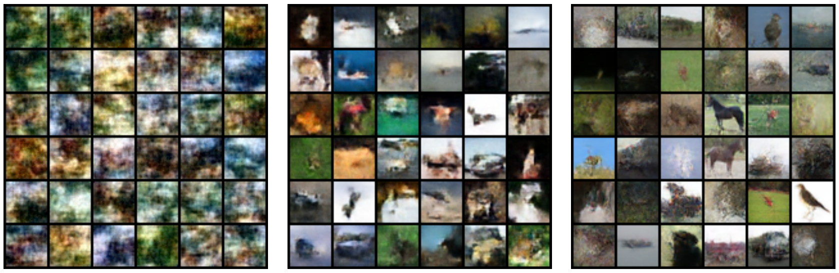
In this section, we present the detailed experimental setup and comparison results.

5.1 Configuration Details

We use RoBERTa [21] as the backbone in our main experiment. WRN-28-10 is a small-sized deep residual networks which outperforms in accuracy and efficiency previous deep residual networks. During the training, we set 0.04 as learning rate and use 256 as the batch size. We also apply cosine decay learning rate schedule and weight decay of 0.0005. All our experiments are conducted on NVIDIA RTX 8000 GPU clusters.

5.2 Analysis of DDPM Generation Data

Denoising Diffusion Probabilistic Models (DDPM) [11] are a type of artificial intelligence application that generates data. The data generated by DDPM is not always perfect, but they do have the ability to generate large amounts of training data. In our framework, we trained the DDPM on the CIFAR-10 training set (50,000 images). Figure. 3 shows the samples generated by DDPM in different training checkpoints. After training DDPM 320,000 steps, the synthesis CIFAR-10 images are more clear and bound to the 10 classes.



(a) (b) (c)

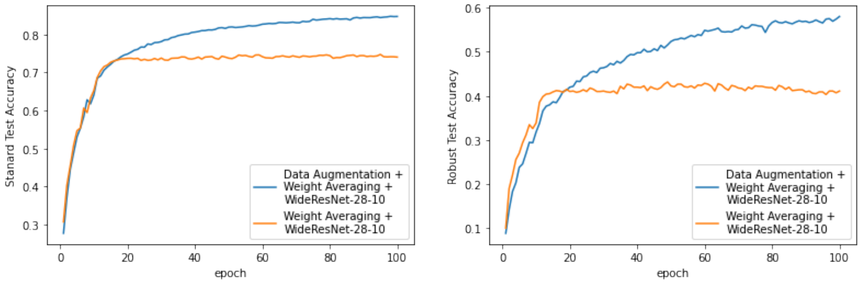
Figure 3: (a) Training DDPM 4000 steps; (b) Training DDPM 40000 steps; (c) Training DDPM 320000 steps

Augmentation	Architecture	Standard Accuracy	Auto-attack robust Accuracy
✗	WideResNet-28-10	73.75%	41.68%
✓	WideResNet-28-10	**85.75%**	**57.51%**

Table 1: Comparisons to non-augmentation baseline on CIFAR-10, $\ell_\infty = 8/255$, untargeted attack

5.3 Comparison Results

Table.1 shows the performance of models obtained by combining random crop, flip, rotation with samples generated by the DDPM. We also trained a non-augmentation baseline for comparison. The results show that the auto-attack robust accuracy of using both heuristic-driven and data-driven augmentation can achieve 57.5%. We believe that if we apply a large architecture, the final result would be much better. Besides, we also observed that without using data augmentation, the robust test accuracy will converge very quickly and stop increasing after 20 epochs (see Figure. 4). However, using data augmentation will help the robust test accuracy continually increase. The experiments prove that data augmentation is a useful tool for improving Adversarial Robustness.



(a) (b)

Figure 4

6 Conclusion

In this work, we proved that data augmentation can improve the adversarial robustness, especially the complementary generated data based on the original dataset. Also, with a simplified setup, we provide the minimum number of augmented data and the bounds of the margin of the decision. These theoretical analysis demonstrates the performance of data augmentation can be meaningful for future works.

7 Future directions

There are several interesting open problems that shows potentials of applying data augmentation in the field of adversarial robustness.

1. Theoretical analysis on certain practical data augmentation techniques (e.g. random crop, flip, rotation) and data generation techniques (e.g. Cycle-GAN, DDPM) could be interesting to provide precise learning bound.
2. Adversarial robustness can benefit from data augmentation while sometimes sacrificing standard accuracy. It would be meaningful to provide an exact bound of the number of augmented data to preserve the original gain of the network.
3. Due to the limitations of computation resources, we only tested our method on WRN-28-10. We believe that deeper and larger networks (e.g. WRN-70-16) can improve the adversarial accuracy further.

References

[1] Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *ArXiv*, abs/2006.06091, 2020.

[2] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

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

[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *ArXiv*, abs/1803.06373, 2018.

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

[7] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020.

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

[9] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[13] 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.

[14] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[15] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[16] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[17] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[18] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019.

[19] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

[20] Shashank Rajput, Zhili Feng, Zachary Charles, Po-Ling Loh, and Dimitris Papailiopoulos. Does data augmentation lead to positive margin? In *International Conference on Machine Learning*, pages 5321–5330. PMLR, 2019.

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