ON IMPROVING ADVERSARIAL ROBUSTNESS USING PROXY DISTRIBUTIONS

Vikash Sehwag[1], **Saeed Mahloujifar**[1], **Tinashe Handina**[1], **Sihui Dai**[1], **Chong Xiang**[1], **Mung Chiang**[2], **Prateek Mittal**[1]
[1]Princeton University, [2]Purdue University

ABSTRACT

We focus on the use of *proxy* distributions, approximations of the underlying distribution of the training dataset, in improving robust generalization of adversarial training. Adversarially trained networks, when trained on a limited number of samples available in the training set, suffer from a large generalization gap in the robust accuracy. Using proxy distribution, from which we can sample an unlimited number of data points, can enable us to 1) investigate the effect of the number of training samples 2) reduce the robust generalization gap in adversarial training. Earlier Min et al. (2020) argued that more training data can both help or hurt generalization based on the strength of the adversary. Here, using state-of-the-art attacks in adversarial training and training with up to 2M images, we find that more data continues to help generalization in deep neural networks. Next, we ask when incorporating additional samples from the proxy distribution will help? Here we prove that the difference of the robustness of a classifier on proxy and training dataset distribution is upper bounded by the conditional Wasserstein distance between them. It confirms the intuition that samples from a proxy distribution closely approximating training dataset distribution should be able to boost performance. Motivated by this, we leverage samples from state-of-the-art generative models, which can closely approximate training distribution, to improve robustness. In particular, we improve robust accuracy up to 6.5% and 5.0% in l_∞ and l_2 threat model, respectively, on the CIFAR-10 dataset.

1 INTRODUCTION

To instill robustness against adversarial examples in deep neural networks, adversarial training remains the most effective technique (Madry et al., 2018; Zhang et al., 2019; Pang et al., 2021). However, adversarially trained networks, when trained on a limited number of images available in curated datasets such as CIFAR-10 (Krizhevsky et al., 2009), suffers from a large generalization gap in robust accuracy. In this work, we approach adversarial training on these datasets in conjunction with a *proxy* distribution. We refer to approximations of the underlying distribution of these curated datasets as proxy distributions. Our use of proxy distributions allows us to target the following two objectives. 1) Since we can sample an unlimited number of samples from the proxy distribution, it allows us to investigate the effect of the number of training samples on the robust generalization of adversarial training 2) Using samples from proxy distribution to reduce the robust generalization gap in adversarial training.

We first investigate the effect of the number of training samples in adversarial training. Earlier works (Carmon et al., 2019; Uesato et al., 2019) demonstrated improved generalization of adversarial training by expanding the training dataset size using an additional set of curated images. Recently Min et al. (2020) argued that more training data can both help or hurt generalization based on the strength of the adversary in adversarial training. However, Min et al. (2020) works with simple problems such as Gaussian mixture classification or a two-dimensional classification. It remains unclear how adversarial training behaves with the number of samples on the scale of computer vision tasks and deep neural networks. We answer this question by training a deep neural network on an increasing number of images (ranging from 1K to 2M). We observe that both clean and robust test accuracy continues to improve with the number of samples.

Next, we ask when incorporating additional samples from the proxy distribution will help in improving robustness on datasets such as CIFAR-10? Here we provide a theoretical analysis on how robustness of a classifier trained on one distribution transfers to the other. In particular, we prove that the difference in robustness of a classifier on the two distributions is upper bounded by the Wasserstein distance between them. It confirms the intuition that samples from a proxy distribution that closely approximates training dataset distribution should be able to boost robustness.

Motivated by this intuition, we aim to leverage proxy distributions, which closely approximate the underlying distribution of training data, to improve performance. Here we propose to simultaneously train on the original training set and a set of additional images sampled from the proxy distribution. We use state-of-the-art generative models as a model for proxy distribution since they can closely approximate the training distribution from only a limited number of training samples (Karras et al., 2020; Ho et al., 2020; Gui et al., 2020). Our experimental results demonstrate that the use of synthetic images improves robust accuracy up to 6.4% and 5.0% in l_∞ and l_2 threat model, respectively, on the CIFAR-10 dataset.

Contributions. We make following three key contributions. 1) We investigate the effect of increasing number of training samples (from 1K to 2M) on the performance of adversarial training with deep neural networks, 2) we provide theoretical insights on how robustness of a classifier on one distribution transfer to another distribution. In particular, we provide a tight upper bound on the difference of the robustness of a classifier between the distributions, and 3) by leveraging additional images sampled from the proxy distributions, we improve robust accuracy up to 6.5% and 5.0% in l_∞ and l_2 threat model, respectively, on the CIFAR-10 dataset.

2 Related work

Adversarial training (Madry et al., 2018; Zhang et al., 2019; Gowal et al., 2020; Wu et al., 2020) still remains the most effective defense against adversarial examples (Szegedy et al., 2013). However, earlier works have also shown that adversarial training suffers from a large robust generalization gap (Schmidt et al., 2018; Raghunathan et al., 2019). We refer the reader to RobustBench (Croce et al., 2020) for a comparison and overview of state-of-the-art methods.

State-of-the-art generative models are capable of modeling the distribution of current large scale image dataset. In particular, generative adversarial networks (GANs) have excelled at this task (Goodfellow et al., 2014; Karras et al., 2020; Gui et al., 2020). Though GANs generate images with high fidelity, they lack high diversity (Ravuri & Vinyals, 2019). However, samples from recently proposed diffusion process based models achieve both highly diversity and fidelity (Ho et al., 2020).

Recent works have also explored the use of synthetic samples in training deep neural networks (Ravuri & Vinyals, 2019; Shmelkov et al., 2018). While these earlier works focuses on benign training our focus is on adversarial training. Another concurrent work (Rebuffi et al., 2021) also uses samples from generative models to improve adversarial robustness. While Rebuffi et al. (2021) broadly focuses on the effect of data augmentation, our goal is understand the effect of different proxy distributions. However, similar benefits of using generative model from two independent works further ascertain the importance of this direction.

3 Integrating proxy distribution with adversarial training

Formulation of adversarial training. The key objective in adversarial training is to minimize the training loss on adversarial examples obtained with iterative adversarial attacks, such as projected gradient descent (PGD) (Madry et al., 2018) based attacks, under the following formulation.

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim D} L_{adv}(\theta, x, y, \Omega), \qquad L_{adv}(\theta, x, y, \Omega) = L(\theta, PGD(x, \Omega), y) \tag{1}$$

where D is the training data distribution, Ω is the threat model, and θ represents network parameters.

Understanding generalization of adversarial robustness. We assume access to two distributions, namely D and $\tilde{D}$ supported on $X \times Y$. Our goal is to understand how robustness achieved on one distribution generalizes to the other. We first define the average robustness of a classifier on a distribution followed by the definition of conditional Wasserstein distance, a measure of the distance

between two labeled distributions. We prove that the difference in average robustness of a classifier on the two distributions is upper bounded by the conditional Wasserstein distance between them.

Definition 1 (Average Robustness). *We define average robustness for a classifier h on a distribution D according to a distance metric d as follows:*

$$\mathsf{Rob}_d(h, D) = \mathop{\mathbb{E}}_{(x,y)\leftarrow D}[\inf_{h(x')\neq y} d(x', x)]$$

Definition 2 (Conditional Wasserstein distance). *For two labeled distributions D and $\tilde{D}$ supported on $X \times Y$, we define conditional wasserstein distance according to a distance metric d as follows:*

$$\mathsf{CWD}_d(D, \tilde{D}) = \mathop{\mathbb{E}}_{(\cdot,y)\leftarrow D}\left[\inf_{J\in\mathcal{J}(D|y,\tilde{D}|y)} \mathop{\mathbb{E}}_{(x,x')\leftarrow J}[d(x, x')]\right]$$

where $\mathcal{J}(D, \tilde{D})$ is the set of joint distributions whose marginals are identical to D and $\tilde{D}$.

Theorem 1. *Let D and $\tilde{D}$ be two labeled distributions supported on $X \times Y$ with identical label distributions, i.e. $\forall y^* \in Y, \Pr_{(x,y)\leftarrow D}[y = y^*] = \Pr_{(x,y)\leftarrow \tilde{D}}[y = y^*]$. Then for any classifier $h : X \to Y$*

$$|\mathsf{Rob}_d(h, \tilde{D}) - \mathsf{Rob}_d(h, D)| \leq \mathsf{CWD}_d(D, \tilde{D}).$$

Theorem 2 (Tightness of Theorem 1). *For any distribution D supported on $X \times Y$, any classifier h, any homogeneous distance d and any $\epsilon \leq \mathsf{Rob}_d(h, D)$, there is a labeled distribution $\tilde{D}$ such that*

$$\mathsf{Rob}_d(h, D) - \mathsf{Rob}_d(h, \tilde{D}) = \mathsf{CWD}(D, \tilde{D}) = \epsilon.$$

Due to space constraints, we provide the proof of Theorem 1, 2 in supplementary material.

Using proxy distribution to close the generalization gap. Now we focus on improving robustness on training distribution (D) with access to only a limited number of training samples. As Theorem 1 suggests, robust training on a close proxy distribution ($\tilde{D}$) also generalize to training distribution (D). Therefore, to improve robustness on D, we proposed to augment original training set with samples from $\tilde{D}$. In particular, we use the following adversarial training formulation.

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim D} a * L_{adv}(\theta, x, y, \Omega) + \mathop{\mathbb{E}}_{(x,y)\sim \tilde{D}} b * L_{adv}(\theta, x, y, \Omega) \tag{2}$$
$$\text{where,} \quad a + b = 1; L_{adv}(\theta, x, y, \Omega) = L(\theta, PGD(x, \Omega), y)$$

4 Experimental results

We experiment with both l_∞ and l_2 threat model on the CIFAR-10 dataset. We use projected gradient descent based adversarial training. Our detailed training setup is provided in the supplementary material. We work with two state-of-the-art generative models, namely StyleGAN-C (Karras et al., 2020) and DDPM (Ho et al., 2020). While the former is a generative adversarial network (GAN), latter is based on the diffusion process. We sample 10M labeled images from the conditional StyleGAN-C and another set of 6M labeled images from the DDPM model[1]. Both models generate high fidelity images (example images are provided in supplementary material) but we observe significantly higher performance when using images from DDPM model in our experiments.

4.1 Adversarial robustness with an increasing number of training samples

Now we investigate the effect of an increasing number of training samples in adversarial training on both training distributions and further generalization to other distributions.

Setup. We robustly train a ResNet-18 network on 1k to 2M synthetic images from StyleGAN-C model trained on CIFAR-10 dataset. To downscale our setup to a manageable computational cost, we solve the binary classification problem between class-1 (automobile/car) and class-9 (truck). Additionally, we use the single-step FGSM attack in the adversarial training. We test each network on a fixed set of 100k images from the StyleGAN-C and the 10k images from the CIFAR-10 test set.

[1]It is a pre-sampled set of images made available by Nakkiran et al. (2021).

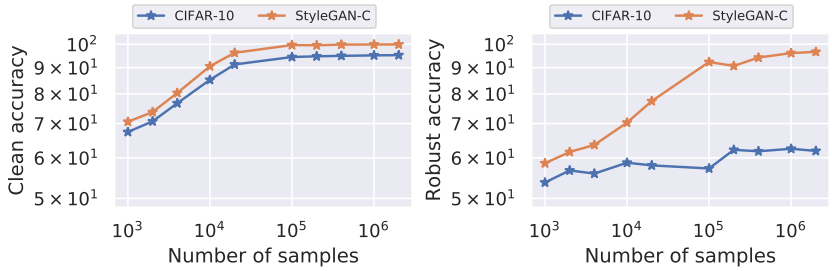

Figure 1: Generalization of clean and robust accuracy when training on *only* StyleGAN-C images but testing on both StyleGAN-C and CIFAR-10 test set. While both clean and robust accuracy improve consistently on the validation set from StyleGAN-C model, generalization of robust accuracy to CIFAR-10 is limited, even with increasing number of training samples to the order of millions.

Results. We present our results in Figure 1. We find that the robust accuracy of samples from the generative model keeps improving consistently. We also observe non-trivial generalization between the two distributions, as demonstrate by non-trivial clean and robust accuracy achieved on the CIFAR-10 dataset. In particular, clean accuracy crosses the 90% threshold with only 100k images. However, generalization of robustness remains harder than accuracy, where it increases much slowly with number of samples. It motivates us to use a large number of samples from the proxy to benefit the most in the generalization of robustness.

4.2 Achieving state-of-the-art robustness

Now we demonstrate state-of-the-art performance by following the improved adversarial training formulation from Equation 2. We present our results in Table 1.

State of the art robust accuracy. We observe that incorporating samples from the DDPM model improves robust accuracy significantly. In l_∞ threat model, it improves it to 59.5%, an improvement up to 6.4% over previous work. Similarly, we observe improvement up to 5.0% for l_2 attacks. Note that clean accuracy also improves simultaneously.

Proxy distribution offsets increase in network parameters. Note that gains from the generative model are equivalent to ones obtained by scaling network size by an order. For example, a ResNet-18 network with synthetic data achieves higher robust accuracy (l_∞) than a WRN-34-20 trained without it, while having $16\times$ fewer parameters than the latter. Similarly trend holds for WRN-34-10 networks, when compared with a much larger WRN-70-16 network.

Method	Architecture	Parameters (M)	*Clean*	*Auto*
Zhang et al. (2019)	ResNet-18	11.2	82.0	48.7
Madry et al. (2018)	ResNet-50	23.5	87.0	49.0
Zhang et al. (2019)	WRN-34-10	46.2	84.9	53.1
Rice et al. (2020)	WRN-34-20	184.5	85.3	53.4
Gowal et al. (2020)	WRN-70-16	266.8	85.3	57.2
Ours	ResNet-18	11.2	84.0	**54.3**
Ours	WRN-34-10	46.2	86.1	**59.5**

(a) ℓ_∞

Method	Architecture	Parameters (M)	*Clean*	*Auto*
Rice et al. (2020)	ResNet-18	11.2	88.7	67.7
Madry et al. (2018)	ResNet-50	23.5	90.8	69.2
Wu et al. (2020)	WRN-34-10	46.2	88.5	73.7
Gowal et al. (2020)	WRN-70-16	266.8	90.9	74.5
Ours	ResNet-18	11.2	89.2	**72.7**
Ours	WRN-34-10	46.2	90.0	**76.0**

(b) ℓ_2

Table 1: Experimental results on the CIFAR-10 dataset where our adversarial training formulation brings a large gain in adversarial robustness.

5 Conclusion

In this work, we focus on the use of proxy distributions in adversarial training. We model the proxy distribution using state-of-the-art generative models. Using samples from these models, we show that with an increasing number of training samples, adversarial training can continue to improve the robustness of deep neural networks. Next, we prove the relationship between the robustness achieved on the proxy distribution and its transfer to the underlying distribution of the training set. Finally, we use these insights to improve the performance of adversarial training significantly.

REFERENCES

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11190–11201, 2019.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The bootstrap framework: Generalization through the lens of online optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=guetrIHLFGI.

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.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/fcf55a303b71b84d326fb1d06e332a26-Paper.pdf.

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.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2018.

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

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/file/bea6cfd50b4f5e3c735a972cf0eb8450-Paper.pdf.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019.