GREAT Score: Evaluating Global Adversarial Robustness using Generative Models

ZAITANG LI
The Chinese University of Hong Kong
Shatin, Hong Kong, China
ztli@link.cuhk.edu.hk

Pin-Yu Chen
IBM Research
Yorktown Heights, NY 10598, USA
pin-yu.chen@ibm.com

Tsung-Yi Ho
The Chinese University of Hong Kong
Shatin, Hong Kong, China
tyho@cs.nthu.edu.tw

Abstract

Current studies on adversarial robustness mainly focus on aggregating local robustness results from a set of data samples to evaluate and rank different models. However, the local statistics may not well represent the true global robustness of the underlying unknown data distribution. To address this challenge, this paper makes the first attempt to present a new framework, called GREAT Score, for global robustness evaluation of adversarial perturbation using generative models. Formally, GREAT Score carries the physical meaning of a global statistic capturing a mean certified attack-proof perturbation level over all samples drawn from a generative model. For finite-sample evaluation, we also derive a probabilistic guarantee on the sample complexity and the difference between the sample mean and the true mean. GREAT Score has several advantages: (1) Robustness evaluations using GREAT Score are efficient and scalable to large models, by sparing the need of running adversarial attacks. In particular, we show high correlation and significantly reduced computation cost of GREAT Score when compared to the attack-based model ranking on RobustBench [4]. (2) The use of generative models facilitates the approximation of the unknown data distribution. In our ablation study with different generative adversarial networks (GANs), we observe consistency between global robustness evaluation and the quality of GANs.

1. Introduction

Adversarial robustness is the study of model performance in the worst-case scenario, which is a key element in trustworthy machine learning. Without further remediation, state-of-the-art machine learning models, especially neural networks, are known to be overly sensitive to small human-imperceptible perturbations to data inputs [11]. Such a property of over sensitivity could be exploited by bad actors to craft adversarial perturbations leading to prediction-evasive adversarial examples.

The methodology for adversarial robustness evaluation can be divided into two categories: *attack-dependent* and *attack-independent*. Attack-dependent approaches aim to devise the strongest possible attack and use it for performance assessment. On the other hand, attack-independent approaches aim to develop a certified or estimated score for adversarial robustness, reflecting a quantifiable level of attack-proof certificate.

To address the challenges including (i) lack of proper global adversarial robustness evaluation, (ii) limitation to white-box settings, requiring detailed knowledge about the target model and (iii) computational inefficiency, in this paper we present a novel attack-independent evaluation framework called *GREAT Score*, which is short for global robustness evaluation of adversarial perturbation using generative models. We tackle challenge (i) by using a generative adversarial network (GAN) [9, 10] as a proxy of the true unknown data distribution. Formally, GREAT score is defined as the mean of a certified lower bound on minimal adversarial perturbation over the data sampling distribution of a GAN. For challenge (ii), our derivation of GREAT score leads to a neat closed-form solution that only requires data forward-passing and accessing the model outputs. Finally, for challenge (iii), the computation of GREAT score is lightweight since each data sample only requires one forward pass through the model to obtain the final predictions.

We highlight the main contributions of this paper as fol-

lows:

- We present GREAT Score as a novel framework for deriving a global statistic representative of the distribution-wise robustness to adversarial perturbation
- Theoretically, we show that GREAT Score corresponds to a mean certified attack-proof level of $\mathcal{L}_2$-norm bounded input perturbation over the sampling distribution of GANs (Theorem 2.2). We further develop a formal probabilistic guarantee on the quality of GREAT Score. (Theorem 1).
- We show that the model ranking of GREAT score is highly aligned with that of the original ranking on RobustBench using Auto-Attack [5], while GREAT Score significantly reduces the computation time.

Notations and Backgrounds All the main notations used in the paper are summarized in Appendix A.1. We also provide related works and background introduction in Appendix B

2. GREAT Score: Methodology and Algorithms

In this section, we start by defining the true global robustness and its certified estimate in Section 2.1. Then, we propose using GANs to obtain a certified estimate for the true global robustness in Section 2.2 and develop a probabilistic guarantee on its effectiveness in finite-sample settings in Section 2.3.

2.1. True Global Robustness and Certified Estimate

Let $f = [f_1, \dots, f_K] : \mathbb{R}^d \to \mathbb{R}^K$ denote a fixed K-way classifier with flattened data input of dimension d, (x, y) denote a pair of data sample x and its corresponding groundtruth label $y \in \{1, \dots, K\}$, P denote the true data distribution which in practice is unknown, and $\Delta_{\min}(x)$ denote the minimal perturbation of a sample-label pair $(x, y) \sim P$ causing the change of the top-1 class prediction such that $\arg\max_{k\in\{1,\dots,K\}} f_k(x + \Delta_{\min}(x)) \neq \arg\max_{k\in\{1,\dots,K\}} f_k(x)$.

Definition 1 (True global robustness)**.** *The true global robustness of a classifier f with respect to a data distribution P is defined as:*

$$\Omega(f) = \mathbb{E}_{x\sim P}[\Delta_{min}(x)] = \int_{x\sim P} \Delta_{\min}(x)p(x)dx \quad (1)$$

Extending Definition 1, let $g(x)$ be a local robustness statistic. Then the corresponding global robustness estimate is defined as

$$\widehat{\Omega}(f) = \mathbb{E}_{x\sim P}[g(x)] = \int_{x\sim P} g(x)p(x)dx \quad (2)$$

2.2. Using GANs to Evaluate Global Robustness

Recall that a GAN takes a random vector $z \sim \mathcal{N}(0, I)$ sampled from a zero-mean isotropic Gaussian distribution as input to generate a data sample $G(z)$. We further denote c as the groundtruth class of x.

We now formally define a local robustness score function as

$$g\left(G(z)\right) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_c(G(z)) - \max_{k\in 1,\dots,K, k\neq c} f_k(G(z)), 0\} \quad (3)$$

We further offer several insights into understanding the physical meaning of the considered local robustness score in Appendix A.2.

Next, we use the local robustness score g defined in (3) to formally state our theorem on establishing a certified lower bound on the true global robustness, followed by a proof sketch.The complete proof is given in Appendix A.3.

Theorem 1 (certified global robustness estimate)**.** *Let $f : [0, 1]^d \mapsto \mathbb{R}^K$ be a K-way classifier and let $f_k(\cdot)$ be the predicted likelihood of class k, with c denoting the groundtruth class. Given a generator G such that it generates a sample $G(z)$ with $z \sim \mathcal{N}(0, I)$. Define $g\left(G(z)\right) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_c(G(z)) - \max_{k\in 1,\dots,K, k\neq c} f_k(G(z)), 0\}$. Then the global robustness estimate of f evaluated with $\mathcal{L}_2$-norm bounded perturbations, defined as $\widehat{\Omega}(f) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g(G(z))]$, is a certified lower bound of the true global robustness $\Omega(f)$ with respect to G.*

2.3. Probabilistic Guarantee on Sample Mean

As defined in Theorem 1, the global robustness estimate $\widehat{\Omega}(f) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g(G(z))]$ is the mean of the local robustness score function introduced in (3) evaluated through a generator G and its sampling distribution. In practice, one can use a finite number of samples $\{G(z_i|y_i)\}_{i=1}^n$ generated from a conditional generator $G(\cdot|y)$ to estimate $\widehat{\Omega}(f)$. The simplest estimator of $\widehat{\Omega}(f)$ is the sample mean, defined as

$$\widehat{\Omega}_S(f) = \frac{1}{n}\sum_{i=1}^{n} g(G(z_i|y_i)) \quad (4)$$

In what follows, we deliver a probabilistic guarantee on the sample complexity to achieve ϵ difference between the sample mean $\widehat{\Omega}_S(f)$ and the true mean $\widehat{\Omega}(f)$.

Theorem 2 (probabilistic guarantee on sample mean)**.** *Let f be a K-way classifier with its outputs bounded by $[0, 1]^K$ and let e denote the natural base. For any $\epsilon, \delta > 0$, if the sample size $n \geq \frac{32e\cdot\log(2/\delta)}{\epsilon^2}$, then with probability at least $1 - \delta$, the sample mean $\widehat{\Omega}_S(f)$ is ϵ-close to the true mean $\widehat{\Omega}(f)$. That is, $|\widehat{\Omega}_S(f) - \widehat{\Omega}(f)| \leq \epsilon$.*

The complete proof is given in Appendix A.4.

3. Experimental Results

3.1. Experiment Setup

Experiment Setup. We conduct our experiment on CIFAR-10 [13] and ImageNet-1K [6] datasets(Result in Appendix). For neural network models, we use the available models on RobustBench [3], which includes 17/5 models on CIFAR-10/ImageNet, correspondingly. We also use several off-the-shelf GANs trained on CIFAR-10 and ImageNet for computing GREAT Score. All our experiments were run on a GTX 2080 Ti GPU with 12GB RAM.

Grouping of models on RobustBench. We select all non-trivial models (having non-zero RA) submitted to the CIFAR-10 benchmarks of RobustBench[1] and evaluated with $\mathcal{L}_2$-norm perturbation with a fixed perturbation level of 0.5 using Auto-Attack. To control the variations of the submitted models, on CIFAR-10 we divided the models into 5 groups. Their summary is also presented in Table 2.

GREAT Score implementation. The implementation follows Algorithm 1 with a sigmoid/softmax function applied to the logits of the CIFAR-10/ImageNet classifier. 500 samples drawn from a GAN were used for computing GREAT Score.

Comparative methods. We compare the effectiveness of GREAT Score in two objectives: robustness ranking (global robustness) and per-sample perturbation. For the former, we compare to the RA reported in RobustBench on test dataset (named RobustBench Accuracy) as well as the RA of Auto-Attack on the generated data samples (named AutoAttack Accuracy). For the latter, we run $\mathcal{L}_2$-norm based CW attack [2] (with learning 0.01 and 100 iterations) on each generated data sample to find minimal adversarial perturbation and use them to compare to our local robustness score in (3).

Evaluation metrics. For robustness ranking, we report the Spearman's rank correlation coefficient between two sets of model rankings.

3.2. Local and Global Robustness Analysis

Recall from Theorem 1 that the local robustness score proposed in 3 gives a certified perturbation level for generated samples from a GAN.

Figure 1 shows the perturbation level of local GREAT Score (equation 3) and that of corresponding CW attack per generated sample. We can see that the local GREAT Score is indeed a lower bound of CW attack. Figure 2 shows that by sweeping the $\mathcal{L}_2$ perturbation level from 0 to 1 with a 0.05 increment for Auto-Attack. The cumulative RA of GREAT Score at a perturbation level r means the fraction of samples having their local GREAT scores greater than r, which gives an attack-proof guarantee that no attacks can achieve

[1] https://robustbench.github.io/

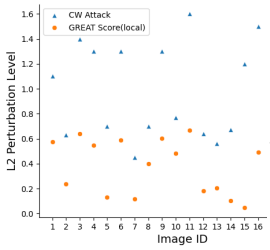

Figure 1. Comparison of local GREAT Score and CW attack in $\mathcal{L}_2$ perturbation on CIFAR-10 with F1 model. The x axis is the image id. The local GREAT Score is indeed a lower bound.

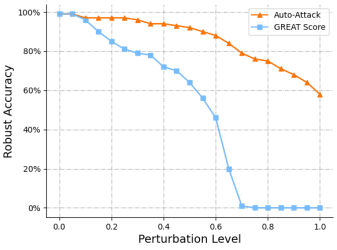

Figure 2. Cumulative robust accuracy (RA) with varying $\mathcal{L}_2$ perturbation level using 500 samples. Note that GREAT Score gives a certified RA for attack-proof robustness.

a lower RA at the same perturbation level. We see that the trend of attack-independent certified robustness (GREAT Score) is similar to that of empirical attacks (Auto-Attack). Accodring to Figure 2, the GREAT score seems saturate to zero after a certain perturbation size. The explanation is that AutoAttack, albeit being a powerful attack, does not guarantee there won't exist unfound adversarial examples when AutoAttack fails, which is a known common pitfall of empirical rubstness evaluation. On ther other hand, our GREAT Score provides a certified robustness guarantee that there won't exist any adversarial examples with perturbation levels within the certified range. Therefore, the gap between our certified curve versus the empirical curve of AutoAttack does not necessarily mean our method is not useful, it could mean that there exist undiscovered adversarial examples at higher perturbation radii. Unless the attacks used for evaluation are sound and complete, meaning that one can confirm no adversarial examples exist if these attacks fail, we cannot rule out the possibility of unfound adversarial examples in these high-radii regimes, as indicated by certified robustness analysis.

Table 1 compares the global robustness statistics of the 17 grouped CIFAR-10 models on RobustBench, in terms of the GREAT Score and the average distortion of CW attack, which again verifies GREAT Score is a certified lower bound on the true global robustness (see Section 2.1 its definition), while any attack with 100% attack success rate only gives on upper bound on the true global robustness.

3.3. Model Ranking on CIFAR-10 and ImageNet

Following the experiment setup in Section 3.1, Table 2 compares the group-level model ranking on CIFAR-10 using GREAT Score , RobustBench , and Auto-Attack. We find that the Spearman's rank correlation coefficient between GREAT Score and RobustBench is consistently higher or same as that between GREAT Score and Auto-Attack. Fur-

Table 1. Comparison of GREAT Score v.s. minimal distortion found by CW attack [2] on CIFAR-10. The results are averaged over 500 generated samples.

Group Name	Model Name	Synthetic Data	Extra Data	GREAT Score	CW Distortion
Fixing [15]	F1		Tiny ImageNet	0.507	0.65
	F2	DDPM		0.451	0.63
	F3	DDPM		0.424	0.62
	F4	DDPM		0.369	0.61
Uncover [12]	U1		Tiny ImageNet	0.534	0.62
	U2			0.124	0.6
RATIO [1]	R1		Tiny ImageNet	0.583	0.59
	R2		Tiny ImageNet	0.554	0.63
	R3		Tiny ImageNet	0.569	0.57
Proxy [18]	P1	DDPM		0.287	0.59
	P2	DDPM		0.236	0.6
Others	HAT [14]	DDPM		0.413	0.62
	AWP [19]			0.128	0.59
	LIBRARY [8]			0.160	0.61
	OVER [16]			0.152	0.61
	DDN [17]			0.275	0.57
	MMA [7]			0.112	0.55

Table 2. Group-wise robustness evaluation and Spearman's rank correlation on CIFAR-10 using GREAT Score, RobustBench (with test set), and Auto Attack (with generated samples).

Group Name	Model Name	RobustBench Accuracy(%)	AutoAttack Accuracy(%)	GREAT Score	GREAT Score v.s. RobustBench Correlation	GREAT Score v.s. AutoAttack Correlation
Fixing	F1	82.32	87.20	0.507	1	0.2
	F2	80.42	90.60	0.451		
	F3	78.80	90.00	0.424		
	F4	75.86	87.60	0.369		
Uncover	U1	80.53	85.60	0.534	1	1
	U2	74.50	86.40	0.124		
RATIO	R1	78.79	86.20	0.583	0.5	0.5
	R2	76.25	86.40	0.554		
	R3	72.91	85.20	0.569		
Proxy	P1	77.24	89.20	0.287	1	1
	P2	74.41	88.60	0.236		
Others	HAT	76.15	86.60	0.413	0.486	0.486
	AWP	73.66	84.60	0.128		
	LIBRARY	69.24	82.20	0.160		
	OVER	67.68	81.80	0.152		
	DDN	66.44	79.20	0.275		
	MMA	66.09	77.60	0.112		

thermore, in 3 out of 4 groups using similar training method (F,R,P), GREAT Score has exactly the same model ranking as RobustBench. The results suggest that GREAT Score can be a good alternative metric for robustness evaluation.

3.4. Run-time Analysis

As Figure 3 implies that comparing to Autoattack [5] with $\epsilon = 0.5$ on 500 generated samples. Our GREAT Score achieved a significantly smaller time cost. Hence, we can claim that the computational cost of GREAT Score method can be very small even applied to robustness deep neural network.

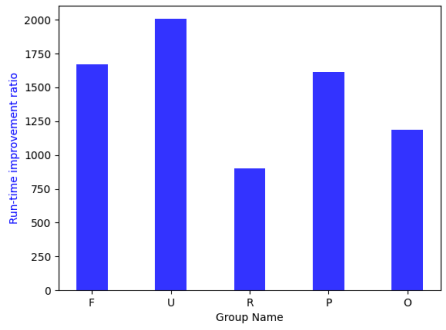

Figure 3. Run-time improvement (GREAT Score over Auto-Attack) on 500 generated CIFAR-10 images.

4. Conclusion

In this paper, we presented GREAT Score, a novel and computation-efficient attack-independent metric for global robustness evaluation against adversarial perturbations. GREAT Score uses an off-the-shelf generative model such as GANs for evaluation and enjoys theoretical guarantees on its estimation of the true global robustness. Its computation is lightweight and scalable because it only requires accessing the model predictions on the generated data samples. Our extensive experimental results on CIFAR-10 and ImageNet also verified high consistency between GREAT Score and the attack-based model ranking on RobustBench, demonstrating that GREAT Score can be used as an efficient alternative for robustness benchmarks.

Limitations and Societal Impacts. One limitation could be our framework of global adversarial robustness evaluation using generative models is centered on $\mathcal{L}_2$-norm based perturbations. This limitation could be addressed if the Stein's Lemma can be extended for other $\mathcal{L}_p$ norms. We do not see any ethical or negative impacts in our work.

References

[1] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pages 228–245. Springer, 2020.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

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

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer*

vision and pattern recognition, pages 248–255. Ieee, 2009.

[7] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

[8] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[12] 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.

[13] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[14] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[15] 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.

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

[17] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019.

[18] 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.

[19] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

A. Appendix

A.1. Notations

Table 3. Main notations used in this paper

Notation	Description
d	dimensionality of the input vector
K	number of output classes
$f : \mathbb{R}^d \to \mathbb{R}^K$	neural network classifier
$x \in \mathbb{R}^d$	data sample
y	groundtruth class label
$\delta \in \mathbb{R}^d$	input perturbation
$\|\delta\|_p$	$\mathcal{L}_p$ norm of perturbation, $p \geq 1$
$\Delta_{\min}$	minimum adversarial perturbation
G	(conditional) generative model
$z \sim \mathcal{N}(0, I)$	latent vector sampled from Gaussian distribution
g	robustness score function defined in (3)
$\Omega(f)/\widehat{\Omega}(f)$	true/estimated global robustness defined in Section 2.1

A.2. Physical meaning of local robustness score in (3)

We define the minimal perturbation for altering model prediction as $\Delta_{\min}(x) = 0$. The intuition is that an attacker does not need to take any action to make the sample x evades the correct prediction by f, and therefore the required minimal adversarial perturbation level is 0 (i.e., zero robustness).(i) The inner term $f_c(G(z)) - \max_{k\in 1,\ldots,K, k\neq c} f_k(G(z))$ represents the gap in the likelihood of model prediction between the correct class c and the second-best class. A positive and larger value of this gap reflects higher confidence of the correct prediction and thus better robustness. (ii) Following (i), a negative gap means the model is making an incorrect prediction, and thus the outer term $\max\{\text{gap}, 0\} = 0$, which corresponds to zero robustness.

A.3. Proof of Theorem 1

In this section, we will give a detailed prove for the certified global robustness estimate in Theorem 1. The proof contains 3 part: Derive the local robustness certificate, derive the closed-form global Lipschitz constant, and prove the proposed global robustness estimate is a lower bound on the true global robustness.

We provide a proof sketch below:

1. We use the local robustness certificate developed in [32], which shows an expression of a certified (attack-proof) $\mathcal{L}_p$-norm bounded perturbation for any $p \geq 1$. The certificate is a function of the gap between the best and second-best class predictions, as well as a local Lipschitz constant associated with the gap function.

2. We use the Stein's Lemma [30] which states that the mean of a measurable function integrated over a zero-mean isotropic Gaussian distribution has a closed-form global Lipschitz constant in the $\mathcal{L}_2$-norm. This result helps avoiding the computation of local Lipschitz constant in Step 1 for global robustness evaluation using GANs

3. We use the results from Steps 1 and 2 to prove that the proposed global robustness estimate $\widehat{\Omega}(f)$ is a lower bound on the true global robustness $\Omega(f)$ with respect to G.

A.3.1 Local robustness certificate In this part, we use the local robustness certificate in [32] to show an expression for local robustness certificate consisting of a gap function in model output and a local Lipschitz constant. The first lemma formally defines Lipschitz continuity and the second lemme introduces the the local robustness certificate in [32].

Lemma 1 (Lipschitz continuity in Gradient Form ([21]))**.** *Let $\boldsymbol{S} \subset \boldsymbol{R}^d$ be a convex bound closed set and let $f : \boldsymbol{S} \to \boldsymbol{R}$ be a continuously differentiable function on an open set containing $\boldsymbol{S}$. Then f is a Lipschitz continuous function if the following inequality holds for any $x, y \in \boldsymbol{S}$:*

$$|f(x) - f(y)| \leq L_q \|x - y\|_p \tag{5}$$

where $L_q = \max_{x\in \boldsymbol{S}} \|\nabla f(x)\|_q$: is the corresponding Lipschitz constant, and $\nabla f(x) = (\frac{\partial f}{\partial x_1}, \ldots \frac{\partial f}{\partial x_d})^\top$ is the gradient of the function f(x), and $1/q + 1/p = 1$, $p \geq 1, q \leq \infty$.

We say f is L_q-continuous in $\mathcal{L}_p$ norm if (5) is satisfied.

Lemma 2 (Formal guarantee on lower bound for untargeted attack of Theorem 3.2 in [32])**.** *Let $x_0 \in \boldsymbol{R}^d$ and $f : \boldsymbol{R}^d \to \boldsymbol{R}^K$ be a multi-class classifier, and f_i be the i-th output of f. For untargeted attack, to ensure that the adversarial examples can not be found for each class, for all $\delta \in \boldsymbol{R}^d$, the lower bound of minumum distortion can be expressed by:*

$$\|\delta\|_p \leq \min_{j\neq m} \frac{f_m(x_0) - f_i(x_0)}{L_q^i} \tag{6}$$

where $m = \arg\max_{1\leq i\leq K} f_i(x_0)$, $1/q + 1/p = 1$, $p \geq 1, q \leq \infty$, and L_q^i is the Lipschitz constant for the function $f_m(x) - f_i(x)$ in L_q norm.

A.3.2 Proof of closed-form global Lipschitz constant in the L2-norm over Gaussian distribution In this part, we present two lemmas towards developing the global Lipschitz constant of a function smoothed by a Gaussian distribution.

Lemma 3 (Stein's lemma([30]))**.** *Given a soft classifier $F : \boldsymbol{R}^d \to \boldsymbol{P}$, where $\boldsymbol{P}$ is the space of probability distributions over classes. The associated smooth classifier with parameter $\sigma \geq 0$ is defined as:*

$$\bar{F} := (F * \mathcal{N}(0, \sigma^2 I))(x) = \mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[F(x + \delta)] \tag{7}$$

Then, $\bar{F}$ is differentiable, and moreover,

$$\nabla \bar{F} = \frac{1}{\sigma^2}\mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}[\delta \cdot F(x+\delta)] \tag{8}$$

In a lecture note[2], Li used Stein's Lemma [30] to prove the following lemma:

Lemma 4 (Proof of global Lipschitz constant)**.** *Let $\sigma \geq 0$, let $h : \mathbb{R}^d \to [0,1]$ be measurable, and let $H = h * \mathcal{N}(0,\sigma^2 I)$. Then H is $\sqrt{\frac{2}{\pi\sigma^2}}$ – continuous in $\mathcal{L}_2$ norm*

A.3.3 Proof of the proposed global robustness estimate $\widehat{\Omega}(f)$ is a lower bound on the true global robustness $\Omega(f)$ with respect to G Recall that we assume a generative model $G(\cdot)$ generates a sample $G(z)$ with $z \sim \mathcal{N}(0, I)$. Following the form of Lemma 2 (but ignoring the local Lipschitz constant), let

$$g'\left(G(z)\right) = \max\{f_c(G(z)) - \max_{k\in 1,\ldots,K, k\neq c} f_k(G(z)), 0\} \tag{9}$$

denote the gap in the model likelihood of the correct class c and the second-best class of a given classifier f, where the gap is defined to be 0 if the model makes an incorrect top-1 class prediction on $G(z)$. Then, using Lemma 4 with g', we define

$$\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))] = g' * \mathcal{N}(0, I) \tag{10}$$

and thus $\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$ has a Lipschitz constant $\sqrt{\frac{2}{\pi}}$ in $\mathcal{L}_2$ norm. This implies that for any input perturbation δ,

$$|\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z)+\delta)] - \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]| \leq \tag{11}$$

$$\sqrt{\frac{2}{\pi}} \cdot \|\delta\|_2 \tag{12}$$

and therefore

$$\mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z)+\delta)] \geq \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))] - \tag{13}$$

$$\sqrt{\frac{2}{\pi}} \cdot \|\delta\|_2 \tag{14}$$

Note that if the right hand side of (13) is greater than zero, this will imply the classifier attains a nontrivial positive mean gap with respect to the generative model. This condition holds for any δ satisfying $\|\delta\|_2 < \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$.

[2]https://jerryzli.github.io/robust-ml-fall19/lec14.pdf

Note that by definition any minimum perturbation on $G(z)$ will be no smaller than $\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{z\sim\mathcal{N}(0,I)}[g'(G(z))]$ as it will make $g'(G(z)) = 0$ almost surely. Therefore, by defining $g = \sqrt{\frac{\pi}{2}} \cdot g'$, we conclude that the global robustness estimate $\widehat{\Omega}(f)$ in (2) using the proposed local robustness score g defined in (3) is a certified lower bound on the true global robustness $\Omega(f)$ with respect to G.

A.4. Proof of Theorem 2

To prove Theorem 2, we first define some notations as follows, with a slight abuse of the notation f as a generic function in this part. For a vector of independent random variables $X = (X_1 ..., X_n)$, define $X^{'} = (X_1^{'} ..., X_n^{'})$ is i.i.d. to X, $x = (x_1, ..., x_n) \in \mathbf{X}$, and the sub-exponential norms $\|\cdot\|_{\psi_2}$ for any random variable Z as

$$\|Z\|_{\psi_2} = \sup_{p\geq 1} \frac{\|Z\|_p}{\sqrt{p}} \tag{15}$$

Let $f : X^n \mapsto \mathbf{R}$. We further define the k-th centered conditional version of f as :

$$f_k(X) = f(X) - \mathbb{E}[f(X)|X_1, ..., X_{k-1}, X_{k+1}, ...X_n] \tag{16}$$

Lemma 5 (Concentration inequality from Theorem 3.1 in [19])**.** *Let $f : X^n \mapsto \boldsymbol{R}$ and $X = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in a space $\mathbb{X}$. Then for any $t > 0$ we have*

$$Pr(f(X) - E[f(X')] > t) \leq \exp\left(\frac{-t^2}{32e\left\|\sum_k \|f_k(X)\|_{\psi_2}^2\right\|_\infty}\right) \tag{17}$$

Recall that we aim to derive a probabilistic guarantee on the sample mean of the local robustness score in (3) from a K-way classifier with its outputs bounded by $[0,1]^K$. Following the definition of g (for simplicity, ignoring the constant $\sqrt{\pi/2}$), the sample mean f can be expressed as:

$$f(X) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \tag{18}$$

where $X_i \sim \mathcal{N}(0, I)$.

Following the definition of (16),

$$f_k(X) = f(X) - \mathbb{E}[f(X)|X_1, ..., X_{k-1}, X_{k+1}, ...X_n] \tag{19}$$

$$= \frac{1}{n}[g(X_k) - g(X_k^{'})] \leq \frac{1}{n} \tag{20}$$

This implies $f_k(X)$ is bounded by $\frac{1}{n}$, i.e., $\|f_k(X)\|_\infty \leq \frac{1}{n}$, and also $\|f_k(X)\|_{\psi_2} \leq \frac{1}{n}$.

Squaring over $\|f_k(X)\|_{\psi_2}$ gives

$$\|f_k(X)\|_{\psi_2}^2 \leq \frac{1}{n^2} \tag{21}$$

As a result,

$$\left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty \leq n \cdot \frac{1}{n^2} = \frac{1}{n} \tag{22}$$

Divide both side of (22) and multiply with $\frac{-t^2}{32e}$ gives:

$$\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \leq \frac{-t^2 n}{32e} \tag{23}$$

Take exponential function over both side of (23) gives

$$\exp\left(\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \leq \exp\left(\frac{-t^2 n}{32e} \right) \tag{24}$$

Recall Lemma5, since this bound holds on both sides of the central mean, we rewrite it as:

$$Pr(|f(X) - \mathbb{E}[f(X')]| > t) \leq 2 \exp \tag{25}$$

$$\left(\frac{-t^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \tag{26}$$

Hence to ensure that given a statistical tolerance $\epsilon > 0$ with δ as the maximum outage probability, i.e., $Pr(|f(X) - E[f(X')]| > \epsilon) \leq \delta$, we have

$$2 \cdot \exp\left(\frac{-\epsilon^2}{32e \left\| \sum_k \|f_k(X)\|_{\psi_2}^2 \right\|_\infty} \right) \leq 2 \exp\left(\frac{-\epsilon^2 n}{32e} \right) \leq \delta \tag{27}$$

Finally, (27) implies that: the sample complexity to reach the (ϵ, δ) condition is $n \geq \frac{32e \cdot \log(2/\delta)}{\epsilon^2}$.

The proof is built on a concentration inequality in [19]. It is worth noting that the bounded output assumption of the classifier f in Theorem 2 can be easily satisfied by applying a normalization layer at the final model output, such as the softmax function or the element-wise sigmoid function. In our implementation of the GREAT score sample mean estimator, we use the element-wise sigmoid function.

Table 4. Group-wise time efficiency evaluation on CIFAR-10 using GREAT Score and Auto Attack (with 500 generated samples).

Group Name	Model Name	GREAT Score(Per Sample)(s)	AutoAttack(Per Sample)(s)
Fixing	F1	0.034	60.872
	F2	0.03	61.3362
	F3	0.006	10.3828
	F4	0.004	4.4644
Uncover	U1	0.03	59.586
	U2	0.03	60.746
RATIO	R1	0.01	10.096
	R2	0.01	10.1056
	R3	0.01	6.9148
Proxy	P1	0.008	10.3662
	P2	0.002	3.8652
Others	HAT	0.002	4.4114
	AWP	0.008	10.9826
	LIBRARY	0.01	6.6462
	OVER	0.004	3.5776
	DDN	0.008	8.5834
	MMA	0.004	3.6194

A.5. Approximation Error and Sample Complexity

Figure 7 presents the sample complexity as analyzed in Theorem 2 with varying approximation error (ϵ) and three confidence parameters (δ) for quantifying the difference between the sample mean and the true mean for global robustness estimation. As expected, smaller δ or smaller ϵ will lead to higher sample complexity.

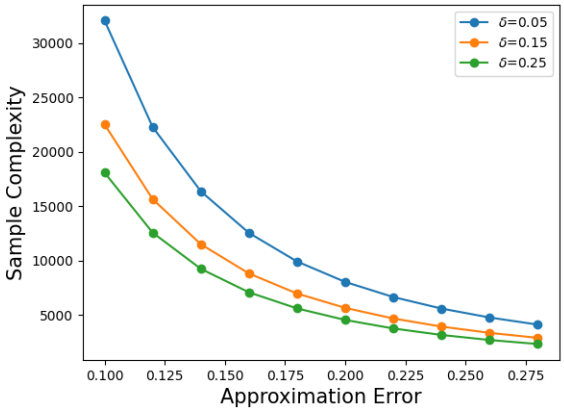

Figure 4. The relationship between the approximation error (ϵ) and sample complexity in Theorem 2, with three different confidence levels: $\delta = \{5, 15, 25\}\%$.

A.6. Sample Complexity and GREAT Score

Figure 5 and Figure 6 report the mean and variance of GREAT Score with a varying number of generated data samples using CIFAR-10 and the F1 model, ranging from 500 to 10000 with 500 increment. The results show that the statistics of GREAT Score are quite stable even with a small number of data samples.

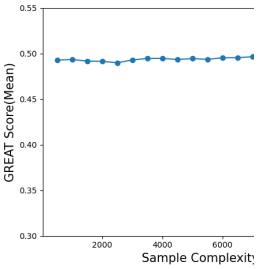

Figure 5. The relation of GREAT Score (mean) and sample complexity using CIFAR-10 and F1 model.

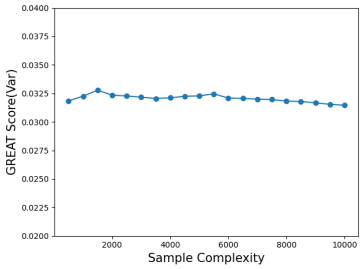

Figure 6. The relation of GREAT Score (variance) and sample complexity using CIFAR-10 and F1 model

Table 5. GREAT Score on CIFAR-10. The results are averaged over 500 original test samples.

Group Name	Model Name	Synthetic Data	Extra Data	GREAT Score
Fixing [24]	F1		Tiny ImageNet	0.465
	F2	DDPM		0.377
	F3	DDPM		0.344
	F4	DDPM		0.297
Uncover [10]	U1		Tiny ImageNet	0.481
	U2			0.109
RATIO [1]	R1		Tiny ImageNet	0.525
	R2		Tiny ImageNet	0.489
	R3		Tiny ImageNet	0.493
Proxy [29]	P1	DDPM		0.227
	P2	DDPM		0.177
Others	HAT [22]	DDPM		0.331
	AWP [34]			0.106
	LIBRARY [7]			0.127
	OVER [25]			0.120
	DDN [26]			0.221
	MMA [6]			0.08

A.7. GREAT Score evaluation on original test samples of CIFAR-10

Besides evaluating the GREAT Score on the generated samples from GAN, we also run the evaluation process on 500 test samples of CIFAR-10. We reported the evaluated Great score and the correlation coefficient between Robustbench and GREAT Score is same to generated samples.

B. Background and Related Works

Adversarial Attack and Defense. In classification tasks, adversarial attacks aim to generate adversarial examples that evade the correct prediction of a classifier. In principle, adversarial examples can be crafted by small perturbations to a native data sample, where the level of perturbation is measured by different $\mathcal{L}_p$ norms [3, 4, 31]. The procedure of finding adversarial perturbation within a perturbation level is often formulated as a constrained optimization problem, which can be solved by algorithms such as projected gradient descent (PGD) [17]. The state-of-the-art adversarial attack is the Auto-Attack [5], which uses an ensemble of white-box and black-box attacks. There are many methods (defenses) to improve adversarial robustness. A popular approach is adversarial training [17], which generates adversarial perturbation during model training for improved robustness. One common evaluation metric for adversarial robustness is robust accuracy, which is defined as the accuracy of correct classification under adversarial attacks, evaluated on a set of data samples. RobustBench [5] is the largest-scale standarized benchmark that ranks the submitted models using the robust accuracy against Auto-Attack on test sets from popular image classification datasets such as CIFAR-10 and ImageNet-1K.

Generative Adversarial Networks (GANs). Statistically speaking, let X denote the observable variable and let Y denote the corresponding label, the learning objective for a generative model is to model the conditional probability distribution $P(X \mid Y)$. Among all the generative models, GANs have gained a lot of attention in recent years due to its capability to generate realistic high-quality images [9]. The principle of training GANs is based on the formulation of a two-player zero-sum min-max game to learn the high-dimension data distribution. During training, GANs are composed of a generative model called the generator (G) and a discriminative model called Discriminator (D). Given a training dataset consisting of real-world data samples, the generator aims at capturing the true data distribution while the discriminator aims at discerning whether the data samples come from the generator or real data. The objective for the generator is to reduce the divergence between the discriminator's outputs based on the true versus generated samples. The objective for the discriminator is to correctly classify the true versus fake samples. The training parameters for G and D are iteratively updated till convergence. Eventually, these two players reach the Nash-equilibrium that D is unable to further discriminate real data versus generated samples. This adversarial learning methodology aids in obtaining high-quality generative models.

In practice, the generator $G(\cdot)$ takes a random vector z (i.e., a latent code) as input, which is generated from an zero-mean isotropic Gaussian distribution denoted as $z \sim \mathcal{N}(0, I)$, where I means an identity matrix. Conditional GANs refer to the conditional generator $G(\cdot|Y)$ given a class label Y. In our proposed framework, we use off-the-shelf conditional GANs that are publicly available as our generative models.

Formal Local Robustness Guarantee and Estimation. Given a data sample x, a formal local robustness guarantee refers to a certified range on its perturbation level such that within which the top-1 class prediction of a model will remain unchanged [14]. In $\mathcal{L}_p$-norm ($p \geq 1$) bounded perturbations centered at x, such a guarantee is often called a certified radius r such that any perturbation δ to x within this radius (i.e., $\|\delta\|_p \leq r$) will have the same top-1 class prediction as x. Therefore, the model is said to be provably

locally robust (i.e., attack-proof) to any perturbations within the certified radius r. By definition, the certified radius of x is also a lower bound on the minimal perturbation required to flip the model prediction.

Among all the related works on attack-independent local robustness evaluations, the CLEVER framework proposed in [32] is the closest to our study. weng2018evaluating derived a closed-form of certified local radius involving the maximum local Lipschitz constant of the model output with respect to the data input around a neighborhood of a data sample x. They then proposed to use extreme value theory to estimate such a constant and use it to obtain a local robustness score, which is not a certified local radius. Our proposed GREAT score has major differences from [32] in that our focus is on global robustness evaluation, and our GREAT score is the mean of a certified radius over the sampling distribution of a generative model. In addition, for every generated sample, our local estimate gives a certified radius.

Global Robustness Evaluation for Deep Neural Networks. There are some works studying "global robustness", while their contexts and scopes are different than ours. In [27], the global robustness is defined as the expectation of the maximal certified radius of $\mathcal{L}_0$-norm over a test dataset. Ours is not limited to a test set, and we take the novel perspective of the entire data distribution and use a generative model to define and evaluate global robustness. The other line of works consider to derive and compute the global Lipschitz constant of the classifier as a a global certificate of robustness guarantee, as it quantifies the maximal change of the classifier with respect to the entire input space [16]. The computation can be converted as a semidefinite program (SDP) [8]. However, the computation of SDP is expensive and hard to scale to larger neural networks. Our method does not require computing the global Lipschitz constant, and our computation is as simple as data forward pass for model inference.

B.1. GREAT Score Algorithm and Computational Complexity

To conclude this section, Algorithm 1 in Appendix summarizes the procedure of computing GREAT Score using the sample mean estimator. It can be seen that the computation complexity of GREAT Score is linear in the number of generated samples N_S. For each sample the computation of the statistic g defined in (3) only requires drawing a sample from the generator G and taking a forward pass to the classifier f to obtain the model predictions on each class. As a byproduct, GREAT Score applies to the setting when the classifier f is a black-box model, meaning only the model outputs are observable by an evaluator.

Algorithm 1: GREAT Score Computation using Sample Mean

Input: K-way classifier $f(\cdot)$, conditional generator $G(\cdot)$, local score function $g(\cdot)$ defined in (3), number of generated samples N_S
Output: GREAT Score $\widehat{\Omega}_S(f)$
for *i ← 1 to* N_S **do**
  Randomly select a class label $y \in \{1, 2, \ldots, K\}$
  Sample $z \sim \mathcal{N}(0, I)$ from a Gaussian distribution and generate a sample $G(z|y)$ with class y
  Pass $G(z|y)$ into the model f and get the prediction for each class $\{f_k(G(z|y))\}_{k=1}^K$
  Record the statistic $g^{(i)}(G(z|y)) = \sqrt{\frac{\pi}{2}} \cdot \max\{f_y(G(z|y)) - \max_{k\in\{1,\ldots,K\},\, k\neq y} f_k(G(z|y)), 0\}$
end
$\widehat{\Omega}_S(f) \leftarrow$ Evaluate the sample mean of $\{g^{(i)}(G(z|y))\}_{i=1}^{N_S}$

B.2. Ablation Study on GANs and Run-time Analysis

Ablation study on GANs. Using Group I (F) on CIFAR-10, Figure 8 compares the inception score (IS) and the Spearman's rank correlation coefficient between GREAT Score and RobustBench on five different GANs. One can observe that models with higher IS also attain better ranking consistency.

Run-time analysis. Figure **??** compares the group-level run-time efficiency of GREAT Score over Auto-Attack on the same 500 generated CIFAR-10 images. We show the ratio of their average per-sample run-time (wall clock time of GREAT Score/Auto-Attack in Appendix **??**) and observe around 800-2000 times improvement, validating the computational efficiency of GREAT Score.

Similarly, Table 6 presents the global robustness statistics of these three methods on ImageNet. We observe almost perfect ranking alignment between GREAT Score and RobustBench, with their Spearman's rank correlation coefficient being 0.9, which is higher than that of Auto-Attack and RobustBench (0.872).

B.3. Group information of models

Group I (F): Rebuffi et al. [24] proposed a fixing data augmentation method such as using CutMix [35] and GANs to prevent over-fitting. There are 4 models in Group I: F1 uses extra data from Tiny ImageNet in training, while F2 uses synthetic data from DDPM. F2/F3/F4 varies in the network architecture. They use WideResNet-70-16 [36]/WideResNet-28-10 [36]/PreActResNet-18 [13].
Group II (U): Gowal et al. [10] studied various training set-

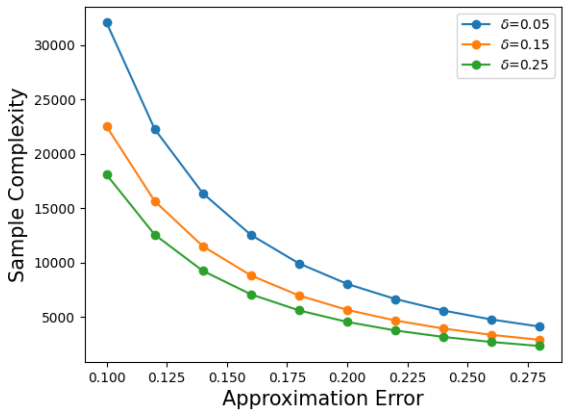

Figure 7. The relationship between the approximation error (ϵ) and sample complexity in Theorem 2, with three different confidence levels: $\delta = \{5, 15, 25\}\%$.

Table 6. Robustness evaluation on ImageNet using GREAT Score, RobustBench (with test set), and Auto Attack (with generated samples). The Spearman's rank correlation coefficient for GREAT score v.s. RobustBench and Auto-Attack v.s. RobustBench is 0.9 and 0.872, respectively.

Model Name	RobustBench Accuracy(%)	AutoAttack Accuracy(%)	GREAT Score
Trans1 [28]	38.14	30	0.483
Trans2 [28]	34.96	25	0.430
LIBRARY [7]	29.22	28	0.434
Fast [33]	26.24	19	0.271
Trans3 [28]	25.32	19	0.269

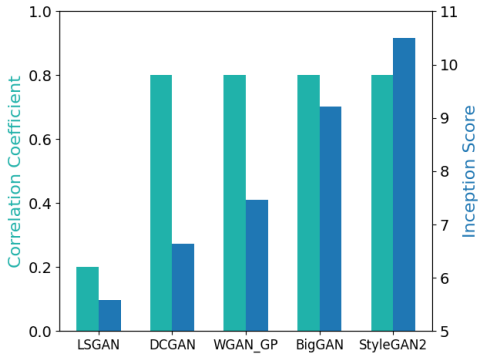

Figure 8. Comparison of Inception Score and Spearman's rank correlation to RobustBench using GREAT Score with different GANs.

tings such as training losses, model sizes, and model weight averaging. G1 differs from G2 in using extra data from Tiny ImageNet for training.

Group III (R): Augustin et al. [1] proposed RATIO, which trains with an out-Of-distribution dataset. R1 uses the out-of-distribution data samples for training while R2 does not.

Group IV (P): SehWag et al. [29] found that a proxy distribution containing extra data can help to improve the robust accuracy. P1/P2 uses WideResNet-34-10 [36]/ResNet-18 [12], respectively.

Group V (O): This group includes all other 6 standalone models. HAT [22] incorporates wrongly labeled data samples for training. AWP [34] regularizes weight loss landscape. LIBRARY [3] is a package used to train and evaluate the robustness of neural network.OVER [25] uses early stopping in reduce over-fitting during training. DDN [26] generates gradient-based attacks for robust training. MMA [6] enables adaptive selection of perturbation level during training.

For the 5 ImageNet models, Trans [28] incorporates transfer learning with adversarial training. Its model variants T1/T2/T3 use WideResNet-50-2 [36]/ResNet-50 [12]/ResNet-18 [12]. LIBRARY means using the package mentioned in Group V to train on ImageNet. Fast [33] means fast adversarial training. There is no $\mathcal{L}_2$-norm benchmark for ImageNet on RobustBench, so we use the $\mathcal{L}_\infty$-norm benchmark.

B.4. GAN

We used off-the-shelf GAN models provided by StudioGAN [20], a library containing released GAN models. We use the GAN model with the highest Inception Score (IS) as our default GAN for GREAT Score, which is StyleGAN2 [15] with IS = 10.477. For the ablation study of using different GANs in GREAT Score (Section B.2), we also use the following GAN models: LSGAN [18], DCGAN [23], WGAN-GP [11], BigGAN [2] and StyleGAN2 [15].

References

[1] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pages 228–245. Springer, 2020.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[4] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10–17, 2018.

[3] https://github.com/MadryLab/robustness

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[6] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

[7] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019.

[8] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] 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.

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[12] 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*, pages 770–778, 2016.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[14] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint arXiv:1705.08475*, 2017.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[16] Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning*, pages 6212–6222. PMLR, 2021.

[17] 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.

[18] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[19] Andreas Maurer and Massimiliano Pontil. Some hoeffding-and bernstein-type concentration inequalities. *arXiv preprint arXiv:2102.06304*, 2021.

[20] Minsu Cho Minguk Kang, Woohyeon Shim and Jaesik Park. Rebooting ACGAN: Auxiliary Classifier GANs with Stable Training. 2021.

[21] Remigijus Paulavičius and Julius Žilinskas. Analysis of different norms and corresponding lipschitz constants for global optimization. *Technological and Economic Development of Economy*, 12(4):301–306, 2006.

[22] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[23] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[24] 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.

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

[26] Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4322–4330, 2019.

[27] Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the l_0 norm. *arXiv preprint arXiv:1804.05805*, 2018.

[28] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neu-*

ral Information Processing Systems, 33:3533–3545, 2020.

[29] 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.

[30] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

[32] 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 preprint arXiv:1801.10578*, 2018.

[33] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[34] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[35] 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.

[36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.