Towards Million-Scale Adversarial Robustness Evaluation With Stronger Individual Attacks

Supplementary Material

1. Introduction

Due to the page limitation of the paper, we further illustrate our method in this supplementary material, which includes the following sections: 1)Visual illustrations of the attacked images. 2) A detailed analysis of the quantitative results for hyperparameters K' and n; 3) A comparison of experimental results between the PMA method and the AAA and ACG methods; 4) A comparison of experimental results between the traditional SGD+sign update strategy and optimizer-based strategies; 5) A detailed examination of the ablation results for P_{max} and P_y weights; 6) Detailed results of the million-scale adversarial robustness evaluation between the PMA method and other methods. 7) Supplementary experiments on CLIP.

2. isual illustrations of the attacked images.

Figure 1. Visual illustrations of the attacked images.

Fig. 1 visualizes adversarial examples generated by different attacks, along with their predicted labels. We evaluate their quality using PSNR, SSIM, and LPIPS, and report cross-entropy loss between adversarial predictions and ground-truth labels, demonstrating both visual and functional impact.

3. Detailed quantitative results of hyperparameter K' and n

Table 1. The models' robustness (%) evaluated on different K' values. The best results are boldfaced.

Dataset	Model	$K'=15$	$K'=20$	$K'=25$	$K'=30$	$K'=35$
CIFAR10	WRN-28-10[?]	67.79	67.76	**67.72**	67.77	67.77
CIFAR10	WRN-28-10[?]	67.35	67.34	**67.33**	67.36	67.35
CIFAR10	RWRN-70-16[?]	71.13	71.14	**71.1**	71.14	71.16
ImageNet	ViT-B+CS[?]	52.84	52.83	**52.82**	52.86	54.43
ImageNet	Swin-B[?]	54.41	54.41	**54.41**	54.41	54.43
ImageNet	ConvNeXt-S+CS[?]	49.74	49.74	**49.74**	49.75	49.8

Table 2. The models' robustness (%) evaluated on different n values. The best results are boldfaced.

Dataset	Model	$n=1$	$n=2$	$n=5$
CIFAR10	WRN-28-10[?]	**67.72**	67.8	68.18
CIFAR10	WRN-28-10[?]	**67.33**	67.35	67.65
CIFAR10	RWRN-70-16[?]	**71.1**	71.14	71.48
ImageNet	ViT-B+CS[?]	**52.82**	52.9	53.17
ImageNet	Swin-B[?]	**54.41**	54.45	54.68
ImageNet	ConvNeXt-S+CS[?]	**49.74**	49.75	49.97

In our study, we conducted a quantitative analysis of the hyperparameters K' and n for the PMA method. We tested three models from the CIFAR10 dataset and three defense models from ImageNet, with a total of 100 iterations set for the tests. The results are presented by evaluating the robust accuracy of different defense models under various attacks.

For the quantitative investigation of K', we compared five different values: 15, 20, 25, 30, and 35. The results, as shown in Table 1, indicate that the optimal performance is achieved when K is set to 25.

In the quantitative study of n, we set three different numbers of restarts: 1, 2, and 5, ensuring a total of 100 iterations. The results, as shown in Table 2, suggest that the best performance is obtained when n is set to 1.

4. Comparison of experimental results between PMA method and AAA, ACG methods

In this comparative analysis, we evaluated the AAA and ACG methods alongside our PMA method. Seven defense models from the CIFAR10 dataset were subjected to a constraint of 100 attack steps. The outcomes are detailed by assessing the robust accuracy of these models when confronted with diverse attack scenarios. As depicted in Table 3, our approach consistently outperformed the others, demonstrating superior effectiveness.

Table 3. The robustness (%) of different models on the CIFAR10 dataset, as evaluated by PMA, AAA, and ACG attacks.

Dataset	Model	PMA	AAA	ACG
CIFAR10	WRN-28-10[?]	**67.72**	68.85	68.63
CIFAR10	WRN-28-10[?]	**67.33**	68.49	68.26
CIFAR10	WRN-70-16[?]	**65.95**	71.27	69.39
CIFAR10	Mixing[?]	**68.43**	71.27	69.39
CIFAR10	WRN-70-16[?]	**66.80**	68.18	67.78
CIFAR10	WRN-106-16[?]	**64.69**	65.84	65.62
CIFAR10	WRN-70-16[?]	**70.67**	71.18	71.58

5. Comparison of experimental results between traditional SGD+sign update strategy and optimizer-based strategies

In our preliminary experiments, we adopted the SGD+sign update strategy, forgoing the integration of an optimizer. To extend our analysis, this section introduces comparative experiments with optimizer-based approaches, focusing on the widely recognized Adam optimizer.

We evaluated three models from the CIFAR10 dataset and three defense models from ImageNet, each subjected to a total of 100 iterations. For the optimizer configuration, we employed the tanh function to scale the noise within the in-

Table 4. The model's robustness(%) evaluated by individual attacks. The best results are boldfaced.

Dataset	Model	PGD_{ce}	PGD_{dlr}	PGD_{mg}	PGD_{pm}	PGD_{alt}	PGD_{mi}	MD	PMA
CIFAR10	WRN-28-10[**?**]	72.26/**70.65**	69.88/**68.65**	69.65/**68.62**	69.52/**68.47**	68.83/**67.88**	69.1/**67.95**	71.61/**67.79**	71.9/**67.72**
CIFAR10	WRN-28-10[**?**]	71.72/**70.31**	69.55/**68.31**	69.22/**68.22**	69.19/**68.10**	68.42/**67.46**	68.76/**67.51**	71.46/**67.42**	71.46/**67.33**
CIFAR10	RWRN-70-16[**?**]	75.14/**73.98**	73.43/**72.03**	73.09/**71.94**	73.03/**71.16**	72.25/**71.15**	72.49/**71.25**	74.9/**71.14**	74.9/**71.10**
ImageNet	ViT-B+CS[**?**]	57.39/**55.34**	56.84/**55.52**	55.93/**55.01**	54.72/**53.61**	54.21/**53.00**	55.7/**53.19**	55.48/**53.34**	55.10/**52.82**
ImageNet	Swin-B[**?**]	58.5/**57.26**	57.57/**56.69**	56.92/**56.28**	55.98/**54.93**	55.27/**54.55**	56.36/**54.57**	59.31/**54.48**	61.04/**54.57**
ImageNet	ConvNeXt-S+CS[**?**]	53.58/**52.69**	53.63/**52.72**	52.63/**51.94**	51.5/**50.38**	50.79/**49.93**	51.98/**49.98**	53.71/**50.12**	52.97/**49.74**

Table 5. The models's robustness results across various methods on CC1M, with the best performances in bold.

Dataset	Model	PGD_{ce}	PGD_{dlr}	PGD_{mg}	PGD_{pm}	PGD_{alt}	PGD_{mi}	MD	PMA	AA
ImageNet	Swin-L[**?**]	20.66	19.82	19.06	17.38	16.68	16.72	16.68	**16.54**	**16.3**
ImageNet	Mixing[**?**]	29.46	20.9	19.06	18.32	18.42	21.48	17.48	**17.4**	**16.64**
ImageNet	ConvNeXt-L[**?**]	20.26	20.9	20.3	18.5	17.4	17.3	17.3	**16.92**	**16.78**
ImageNet	ConvNeXt-L+CS[**?**]	22.76	21.09	21	18.88	18.2	18.14	19.02	**18**	**17.58**
ImageNet	Swin-B[**?**]	22.24	19.2	16.9	16.78	16.98	16.8	16.9	**16.64**	**16.54**

terval [-1, 1], multiplied this scaled noise by the perturbation magnitude, and subsequently added it to the original image. Clipping was applied to ensure pixel values remained within the permissible range. The initial learning rate was set to 0.05, with $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

The ensuing robustness outcomes, presented as a comparison between 'Adam' and 'SGD+sign' in Table 4, indicate that the Adam optimizer does not significantly enhance the efficacy of the attack. However, in this context, our pm loss, as implemented in the PMA method, consistently demonstrated superior performance. This underscores the critical importance of identifying more reliable optimization directions in the domain of adversarial attacks.

6. Detailed ablation results of P_{max} and P_y weights

Table 6. The robustness (%) of the models, evaluated using the **PGD_{pm}** attack with varying β values, on the CIFAR10 and ImageNet datasets.

Dataset	Model	$\beta = 0.5$	$\beta = 0.75$	$\beta = 1$	$\beta = 1.25$	$\beta = 1.5$
CIFAR10	WRN-28-10[**?**]	68.66	**68.47**	68.47	68.48	68.54
CIFAR10	WRN-28-10[**?**]	68.11	**68.09**	68.1	68.16	68.26
CIFAR10	RWRN-70-16[**?**]	71.95	**71.74**	71.76	71.79	71.87
ImageNet	ViT-B+CS[**?**]	53.48	**53.48**	53.61	53.8	54.04
ImageNet	Swin-B[**?**]	54.94	**54.83**	54.93	55.11	55.31
ImageNet	ConvNeXt-S+CS[**?**]	50.38	**50.26**	50.38	50.56	50.82

In this section, we present a quantitative analysis of the weights associated with P_{max} and P_y within the PMA and PGD_{pm} approaches. We utilized a weighted formulation of the PM loss, defined as $L_{pm} = \beta \cdot P_{max} - P_y$, to evaluate both the PGD and PMA methods. The evaluation encompassed three models from the CIFAR10 dataset and three defense models from ImageNet, each limited to a maximum of 100 iterations. The robust accuracy of these defense models under various attack scenarios was assessed to detail the outcomes.

For the parameter β, we investigated its influence across five distinct values: 0.5, 0.75, 1, 1.25, and 1.5. The results for the PGD_{pm} method are detailed in Table 6, while those for the PMA method are presented in Table 7. The findings reveal distinct performance trends: the PGD_{pm} method achieves marginally superior performance with $\beta = 0.75$, whereas the PMA method yields optimal results with $\beta = 1$.

Table 7. The robustness (%) of the models, evaluated using the **PMA** attack with varying β values, on the CIFAR10 and ImageNet datasets.

Dataset	Model	$\beta = 0.5$	$\beta = 0.75$	$\beta = 1$	$\beta = 1.25$	$\beta = 1.5$
CIFAR10	WRN-28-10[**?**]	67.98	67.78	**67.72**	67.79	67.78
CIFAR10	WRN-28-10[**?**]	67.56	67.39	**67.33**	67.43	67.37
CIFAR10	RWRN-70-16[**?**]	71.35	71.18	**71.1**	71.13	71.16
ImageNet	ViT-B+CS[**?**]	53.02	52.88	**52.82**	52.84	52.96
ImageNet	Swin-B[**?**]	54.72	54.45	**54.41**	54.39	54.42
ImageNet	ConvNeXt-S+CS[**?**]	50.01	49.77	**49.74**	49.74	49.89

7. Million-Scale adversarial robustness evaluation between the PMA method and other methods

In this supplementary section, we broaden our comparative analysis by incorporating the PMA and PGD_{ce} methods with other existing techniques. We assessed the same set of five ImageNet defense models discussed in the main body of the paper. Due to the extensive duration—estimated to span several months—to test AA on the CC1M dataset, we chose not to conduct this test. Instead, to enhance our evaluation, we randomly selected a subset of 10,000 images from CC1M to evaluate the comparative robustness of AA. For consistency, we allocated 100 steps for all methods, with the exception of AA, which includes four distinct attacks.

The results of these experiments are presented in Table 5, where we also document the computational time expended by one of the defense models when subjected to various attack methodologies. All experiments were conducted on an NVIDIA RTX 3090 GPU with a batch size of 32. As shown in Table 8, AA emerges as the superior approach; however, our PMA method closely matches AA in performance while requiring only 3% of AA's evaluation time.

Table 8. The efficiency results (in seconds) across various methods on CC1M, with the best performances in bold.

Dataset	Model	PGD_{ce}	PGD_{dlr}	PGD_{mg}	PGD_{pm}	PGD_{alt}	PGD_{mi}	MD	PMA	AA
ImageNet	Swin-B[**?**]	766	700	742	742	688	878	742	718	22328

8. Supplementary experiments on CLIP

To address the domain mismatch between Conceptual-Captions and ImageNet, we conducted supplementary experiments using CLIP on CC1M. We tested perturbation ranges of 1, 2, and 3, with a batch size of 32. All samples in a batch—except the text corresponding to the given image—were treated as negative samples. As shown in Table 9, our method consistently achieves the best performance.

ϵ	Clean	PGD_{ce}	PGD_{dlr}	PGD_{mg}	PGD_{pm}	PGD_{alt}	PGD_{mi}	MD	PMA	diff
1/255	73.74	26.22	22.32	21.92	21.9	21.4	21.32	21.32	**21.16**	**-0.16**
2/255	73.74	13.12	10.45	9.97	9.94	9.48	9.5	9.5	**9.4**	**-0.1**
3/255	73.74	7.42	6.2	5.87	5.89	5.5	5.55	5.57	**5.45**	**-0.05**

Table 9. Robustness (%) evaluated under varying ϵ.