Data-Driven Lipschitz Continuity: A Cost-Effective Approach to Improve Adversarial Robustness

Erh-Chung Chen [1] Pin-Yu Chen [2] I-Hsin Chung [2] Che-Rung Lee [1]
National Tsing Hua University [1] IBM Research [2]

A. Ablation Study

A.1. Hyper-parameter Selection

This experiment investigates how the choice of hyper-parameter c^r influences standard accuracy and robust accuracy. Since most models are represented in 16 bit format, and the widths of fraction bit for FP16 format defined by IEEE-754 standard and BFloat are 10 and 7 bits, respectively, truncated errors might easily occur when performing addition on two numbers with a magnitude difference of 2^8 or higher. On the other hand, when c^r is set to 2^{-5}, all models experience a significant drop in standard accuracy, and there is meaningless in evaluating robustness at this configuration. We suggest that the candidates of c^r are 2^{-8}, 2^{-7} and 2^{-6}.

The results on CIFAR10 and CIFAR100 are presented in Table 1 and Table 2, respectively. Moreover, the results of accuracy against CW attack on L_∞ norm for CIFAR10 and CIFAR100 datasets are presented in Tables 4a and 4b, respectively. As can be seen, when c^r is set to 2^{-8}, all models achieve better standard accuracy and robust accuracy. Additionally, the results for all models with $c^r = 2^{-7}$ are surpassed by those when c^r is set to 2^{-8}. Robust accuracy can be further enhanced by setting 2^{-6}, while standard accuracy might drop compared to the original. The results suggest that $c^r = 2^{-7}$ is a solution that balances standard accuracy and robustness. Nevertheless, when robustness is a major concern, $c^r = 2^{-6}$ is a better choice.

Intuitively, we expect that standard accuracy gradually decreases when the value of c^r increases. The phenomenon can be observed when c^r is 2^{-6} or higher but two counterexamples are reported in the ablation study when setting c^r to 2^{-7} and 2^{-8}. A possible explanation is that the optimizer becomes stuck in a saddle area, as ReLU is non-differentiable at the zero point. This might cause the gradient direction to become stuck in an oscillation when values are close to zero. By shifting those values to zero, antagonistic effects among different feature maps, filters, or channels are accidentally mitigated. However, further investigation and evidence are needed to support this conjecture.

We argue that any function that satisfies the conditions defined in (13) can shrink the largest eigenvalue. There might be another function that can perform better than the proposed one. Besides, the hyper-parameter is determined by choosing the maximum value appearing in the dataset.

A.2. Scaling Training Data for Hyperparameter Tuning

This experiment investigates how the number of training images used for hyperparameter selection affects both standard accuracy and robust accuracy. As shown in Table 3, robust accuracy decreases slightly as the amount of training data increases, but it consistently remains higher than that of the competing methods.

B. Full Experimental Results of Gradient Masking Verification

Table 6a and 6b present the robust accuracy against adversarial examples generated by the original models on CIFAR10 and CIFAR100 datasets, respectively. As observed, none of the models showed lower robust accuracy than the original model. It indicates that adversarial examples can be efficiently crafted by utilizing the gradients from the victim models.

Table 7 and 8 presents the robust accuracy against FGSM and PGD attacks among different radii of the ϵ-ball on the CIFAR10 and CIFAR100 datasets, respectively. As observed, the robust accuracy against FGSM, a one-step attack, is always higher than the robust accuracy against PGD, an iterative attack. This implies that the gradient is reliable, allowing the PGD attack to adjust the gradient direction multiple times to find adversarial examples. Additionally, we observe that the robust accuracy against PGD attacks for all models gradually decreases to zero as the radius of the ϵ-ball increases. This indicates that the quality of gradients is preserved, enabling PGD attacks to move the gradient toward examples not in the observed distribution.

Method	RobustBench		$c^r = 2^{-8}$		$c^r = 2^{-7}$		$c^r = 2^{-6}$	
	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}
RST-AWP	88.25	60.04	88.82	60.96	89.50	62.76	87.88	61.96
DefEAT	86.54	57.30	86.88	57.81	87.40	59.55	84.59	61.08
LTD	85.21	56.94	85.28	57.28	85.98	59.25	85.59	60.63
AWP	85.36	56.17	85.80	56.53	86.19	57.85	84.55	59.21
TRADES	85.34	52.86	85.57	52.97	85.78	53.80	85.49	55.37

Table 1. Ablation study on selecting optimal c^r for CIFAR10 dataset.

Method	RobustBench		$c^r = 2^{-8}$		$c^r = 2^{-7}$		$c^r = 2^{-6}$	
	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}
EffAug	68.75	31.85	68.81	32.00	69.14	32.57	68.44	33.64
DKLD	64.08	31.65	64.10	31.77	64.26	32.58	63.50	33.87
DefEAT	65.89	30.57	66.12	31.11	66.42	32.46	65.06	34.07
LTD	64.07	30.59	64.29	31.13	64.29	31.95	64.18	34.04
AWP	60.38	28.86	60.18	29.10	60.63	29.72	60.71	30.82

Table 2. Ablation study on selecting optimal c^r for CIFAR100 dataset.

#	Method	Training Data [#]	acc_{nat}	acc_{AA}
1	[2] + Ours	5,000	78.88	**60.04**
2	[2] + Ours	10,000	78.92	59.97
3	[2] + Ours	20,000	78.90	59.81
4	[2]	-	78.92	59.56
5	[1]	-	**81.48**	58.50

Table 3. Ablation study on the impact of scaling training data for ImageNet dataset.

Figure 1 illustrates the certified robustness achieved by random smoothing for various models on the CIFAR10 dataset, where *Original* refers to the certified robustness of the original model, while *Ours* denotes the robustness of the model combined with the proposed method. As can be seen, our method brings slight improvements in robustness, except for the AWP model. These results demonstrate that our algorithm does not suffer from the gradient masking issue. However, the empirical Lipschitz constant is derived from the observed data. As the input distribution drawn from random smoothing and the observed data might have discrepancies, this could result in fluctuations in robustness.

Alternatively, we can empirically assess the gradient masking using Lemma 2. The first step is the same as usual: obtaining the hyperparameter c^{th_i} for each layer. The second step involves counting the occurrences of specific elements in the input vectors that are remapped to 0. The corresponding columns of the weight matrix are then replaced with zero vectors. Table 5 presents the results, where *[2] + Ours + Weight* refers to the alternative implementation. As shown, the alternative implementation achieves better robust accuracy compared to the original model. However, the proposed method outperforms the alternative implementation. This is because the proposed method offers the flexibility to remap the input vectors on a sample-wise granularity, while the alternative implementation replaces the corresponding columns of the weight matrix, which has a broader

Method	Origin	c^r 2^{-8}	2^{-7}	2^{-6}
RST-AWP	58.98	61.84	68.24	80.92
DefEAT	56.92	58.02	61.06	65.56
LTD	58.12	58.56	60.50	64.86
AWP	56.84	57.34	60.58	66.50
TRADES	56.10	56.52	58.18	63.62

(a) CIFAR10 dataset

Method	Origin	c^r 2^{-8}	2^{-7}	2^{-6}
EffAug	37.40	37.70	38.70	43.00
DKLD	37.50	38.06	39.38	44.20
DefEAT	36.90	37.56	39.82	44.30
LTD	36.66	37.32	38.86	43.44
AWP	34.56	35.20	35.94	40.40

(b) CIFAR100 dataset

Table 4. The robust accuracy against CW attack on L_∞ norm.

Method	Architecutre	acc_{nat}	acc_{AA}
[2] + Ours	Swin-L	78.88	60.04
[2] + Ours + Weight	Swin-L	78.82	59.71
[2]	Swin-L	78.92	59.56

Table 5. Gradient masking verification by the weight adjustment.

Method	Origin	c^r 2^{-8}	2^{-7}	2^{-6}
RST-AWP	60.04	62.10	65.10	70.53
DefEAT	57.30	58.37	60.39	66.10
LTD	56.94	58.71	61.63	66.47
AWP	56.17	57.49	59.74	65.58
TRADES	52.86	55.55	55.09	58.68

(a) CIFAR10 dataset

Method	Origin	c^r 2^{-8}	2^{-7}	2^{-6}
EffAug	31.85	32.87	35.08	40.04
DKLD	31.65	32.91	35.04	40.58
DefEAT	30.57	31.82	33.94	40.67
LTD	30.59	32.05	34.07	39.11
AWP	28.86	29.88	32.18	36.67

(b) CIFAR100 dataset

Table 6. The robust accuracy against adversarial examples generated by the original models.

influence on all samples.

C. Quantitative Analysis of Empirical Lipschitz Constant

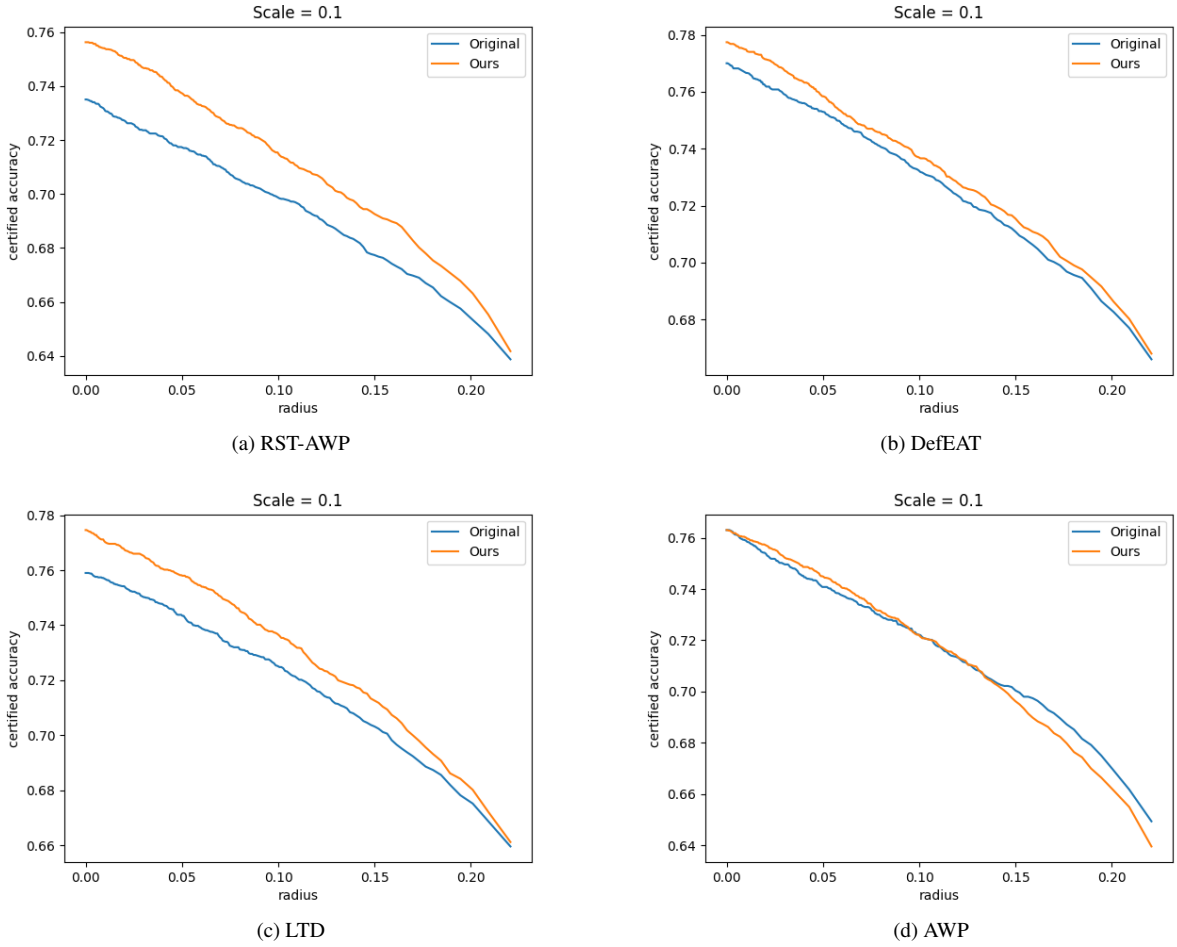
Our proposed approach, however, is data-driven, relying on observed data to automatically determine the appropriate parameters for constructing the forged function. The sparsity of the forged vectors is a crucial factor influencing the magnitude of the Lipschitz constant for the corresponding layers, although it is not the only factor. Figure 2 illustrates the average proportion of pruned activations, which varies depending on the location of each linear layer. FC1 and FC2 refer to the first and second fully connected layers in the MLP blocks, respectively. As shown, the proportion of pruned activations is approximately 20% for the FC1 layers, except for the last layer. This suggests that the forged function alone cannot significantly reduce the magnitude of the Lipschitz constant.

On the other hand, the pruned rates for the FC2 layer range from 30% to 95%, depending on the layer's location. However, we emphasize that a higher pruning rate does not necessarily lead to a substantial reduction in the empirical Lipschitz constant. This is because, although the Lipschitz constant estimated from most of the data may decline, the empirical Lipschitz constant is based on the worst-case scenario across the entire observed dataset. Experimental results show that, after applying our method, the eigenvalue of the worst-case scenario is approximately 95% of the original eigenvalue.

References

[1] Yatong Bai, Mo Zhou, Vishal M Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *arXiv preprint arXiv:2402.02263*, 2024. 2

[2] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *arXiv preprint arXiv:2302.14301*, 2023. 2, 3



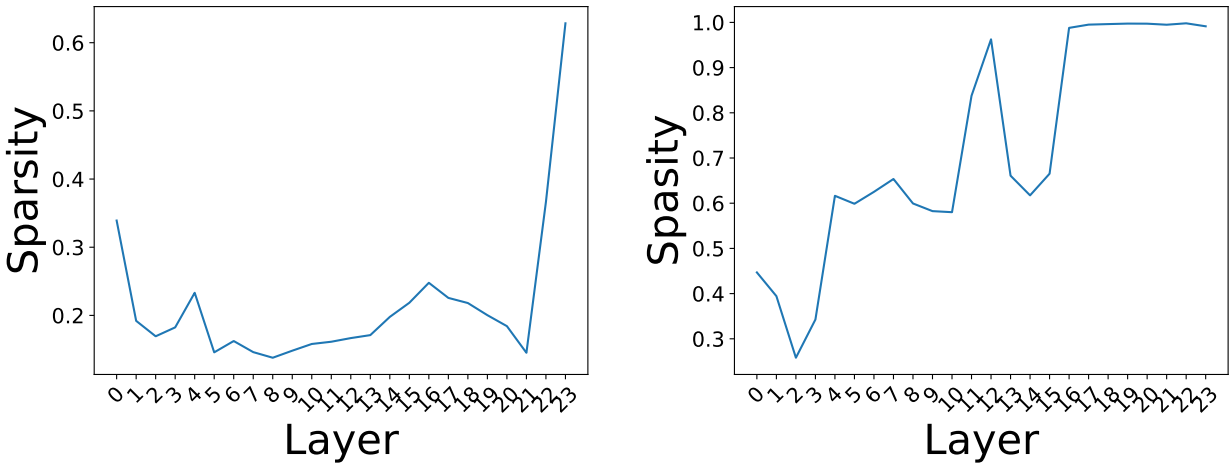
(a) RST-AWP

(b) DefEAT

(c) LTD

(d) AWP

Figure 1. Certified robustness that conducted by random smoothing.



(a) FC1 layers

(b) FC2 layers

Figure 2. The average proportion of pruned activations depends on the location of each linear layer. FC1 and FC2 refer to the first and second fully connected layers in MLP blocks, respectively.

Method	c^r	Attack	ϵ							
			$\frac{1}{255}$	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{32}{255}$	$\frac{64}{255}$	$\frac{96}{255}$
RST-AWP	2^{-8}	FGSM	88.28	86.94	83.80	75.12	57.23	34.04	18.80	19.03
		PGD	86.78	84.47	79.03	66.03	34.02	2.01	0.01	0.0
	2^{-7}	FGSM	89.46	88.28	85.64	77.62	60.60	35.91	19.39	20.68
		PGD	88.03	85.88	80.89	69.24	38.27	3.08	0.1	0.0
	2^{-6}	FGSM	87.70	86.63	84.38	77.91	60.93	33.48	16.12	18.81
		PGD	86.38	84.69	81.19	73.72	52.24	11.89	0.19	0.0
DefEAT	2^{-8}	FGSM	86.38	85.40	81.98	72.73	53.28	30.34	18.13	19.65
		PGD	84.52	82.07	76.51	63.71	33.87	1.76	0.0	0.0
	2^{-7}	FGSM	86.69	85.70	83.05	74.35	56.36	32.04	18.88	21.40
		PGD	85.11	82.87	78.00	66.54	38.70	3.12	0.0	0.0
	2^{-6}	FGSM	84.14	83.57	81.03	74.63	57.67	28.11	13.17	19.38
		PGD	82.96	81.35	77.37	69.52	50.70	10.05	0.2	0.0
LTD	2^{-8}	FGSM	84.94	83.87	81.15	72.80	55.45	33.06	18.44	17.48
		PGD	83.13	80.68	75.53	63.52	34.81	2.64	0.0	0.0
	2^{-7}	FGSM	85.48	84.67	82.24	74.10	57.41	35.53	17.57	17.50
		PGD	83.88	81.88	77.00	65.38	28.57	3.95	0.0	0.0
	2^{-6}	FGSM	85.06	84.28	82.21	75.77	60.79	34.39	14.34	15.42
		PGD	83.91	82.00	78.33	69.82	49.95	11.63	0.21	0.0
AWP	2^{-8}	FGSM	85.11	83.90	80.68	71.28	53.78	33.21	20.86	19.61
		PGD	83.34	80.34	75.08	61.53	30.50	1.89	0.03	0.0
	2^{-7}	FGSM	85.50	84.68	81.75	73.56	57.19	35.50	20.63	20.16
		PGD	83.94	81.62	76.34	65.57	34.16	2.89	0.02	0.0
	2^{-6}	FGSM	83.87	83.19	81.08	74.97	61.49	35.13	16.12	18.69
		PGD	83.00	81.42	78.13	71.50	54.95	14.78	0.41	0.0
TRADES	2^{-8}	FGSM	84.74	83.51	70.58	70.50	54.23	36.53	23.97	23.46
		PGD	82.62	79.72	72.81	57.30	24.21	1.21	0.01	0.0
	2^{-7}	FGSM	85.00	84.02	80.57	71.61	55.94	25.67	22.31	22.51
		PGD	82.96	80.31	73.75	58.91	26.22	1.31	0.02	0.0
	2^{-6}	FGSM	85.05	84.19	81.40	75.07	60.24	36.99	21.81	21.21
		PGD	83.23	81.33	75.81	64.56	36.76	3.37	0.02	0.0

Table 7. The robust accuracy against FGSM and PGD attacks among different radii of ϵ-ball on CIFAR10 dataset.

Method	c^r	Attack	ϵ							
			$\frac{1}{255}$	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{32}{255}$	$\frac{64}{255}$	$\frac{96}{255}$
EffAug	2^{-8}	FGSM	68.02	65.96	60.36	49.65	33.90	17.39	7.11	5.77
		PGD	64.92	61.04	52.82	39.37	17.53	1.81	0.0	0.0
	2^{-7}	FGSM	68.41	66.56	61.84	51.83	36.59	18.34	7.02	6.70
		PGD	65.66	62.02	54.58	41.55	19.70	2.26	0.0	0.0
	2^{-6}	FGSM	67.84	67.25	64.65	57.97	44.23	22.20	8.26	8.85
		PGD	66.38	64.20	59.63	50.89	33.98	7.27	0.17	0.0
DKLD	2^{-8}	FGSM	63.47	61.94	58.15	48.86	34.39	17.73	6.26	3.66
		PGD	60.71	57.14	50.44	38.14	17.38	1.97	0.0	0.0
	2^{-7}	FGSM	63.55	62.31	58.65	50.37	36.49	18.69	6.36	4.29
		PGD	61.06	57.84	51.51	39.99	19.71	2.37	0.0	0.0
	2^{-6}	FGSM	63.26	62.77	60.41	55.18	43.86	21.17	5.50	4.97
		PGD	61.67	59.62	55.83	48.06	33.51	7.55	0.17	0.0
DefEAT	2^{-8}	FGSM	65.57	64.36	59.88	49.38	32.48	15.38	5.50	3.30
		PGD	62.39	58.96	51.85	38.59	17.18	1.45	0.0	0.0
	2^{-7}	FGSM	65.97	64.96	60.67	51.44	34.84	16.36	5.33	3.96
		PGD	62.94	59.83	52.98	41.05	20.07	2.02	0.0	0.0
	2^{-6}	FGSM	64.69	63.58	61.31	54.47	40.06	16.91	4.64	4.46
		PGD	62.85	60.50	56.29	47.54	30.76	5.84	0.09	0.0
LTD	2^{-8}	FGSM	63.59	62.35	58.10	48.86	33.11	16.78	5.92	3.20
		PGD	61.06	57.65	50.70	38.21	18.21	1.98	0.0	0.0
	2^{-7}	FGSM	64.05	62.87	59.02	50.16	34.90	17.27	5.54	3.11
		PGD	61.51	58.32	51.84	39.89	20.32	2.26	0.0	0.0
	2^{-6}	FGSM	63.62	62.98	60.96	54.96	41.51	19.78	4.69	3.12
		PGD	61.90	59.68	55.23	46.49	29.13	5.74	0.05	0.0
AWP	2^{-8}	FGSM	59.77	58.06	54.21	45.54	30.98	16.69	6.49	3.97
		PGD	56.72	52.92	46.53	34.90	16.03	2.11	0.0	0.0
	2^{-7}	FGSM	60.00	58.52	54.93	46.65	32.66	17.42	6.04	3.60
		PGD	57.19	53.75	47.46	36.22	17.48	2.58	0.0	0.0
	2^{-6}	FGSM	60.20	59.71	57.47	52.05	39.78	21.10	5.70	3.90
		PGD	58.19	55.66	50.91	42.37	25.53	5.68	0.09	0.0

Table 8. The robust accuracy against FGSM and PGD attacks among different radii of ϵ-ball on CIFAR100 dataset.