A THE COMPUTATIONAL TIME IN THE WORST CASE

In this section, we discuss the computational time in the worst case. According to their attack mechanism as an ensemble of diverse attacks, AA and T-AA consider one attack first. If the attack succeeds, stop other attacks on the current example; else, continue to consider the next attack in the ensemble. According to the strategy of our STARS method, MM attack considers the false target with the largest predicted probability first, if the attack succeeds, stop attacks on other false targets; else, continue to consider the next target in the ranking of the predicted probability. The computational time of these methods is influenced by different datasets and models. Hence, in the worst case that all attacks inside fail to succeed, the computational time is the sum of the individual time of each attack. Hence, the computational cost of AA (or T-AA) is 109 times (or 440 times) more than PGD, and 34 times (or 139 times) more than MM3 in this case.

B THE REALIZATION OF ADVERSARIAL TRAINING OF MM ATTACK.

We summarize the adversarial training of MM Attack in Algorithm 2. We use MM3 attack to generate adversarial examples, and the computational time is about 2 times as much as PGD (Madry et al., 2018), which can be acceptable for most practitioners.

C POTENTIAL BENEFITS OF DIVERSE RESCALINGS.

We investigate the difference among different successful sets of seven rescaling methods mentioned above. In Table 2, the setting follows (Madry et al., 2018) (with 20 fixed steps). In Table 4, the setting follows (Croce & Hein, 2020) (with 100 adaptive steps). In Table 2 and Table 4, the non-empty difference sets $A \cup B_i - A$ and $A \cup B_i - B_i$ suggest that diverse rescaling methods can complement each other. Hence, when considerable computational resources are available, we recommend practitioners to consider diverse logits rescaling on a strong attack (e.g., our MM attack) rather than diverse weak attacks. Note that we do not argue that diverse weak attacks is unnecessary but rather that when a reliable enough attack exists, most relatively weak attacks have limited benefits other than increased computational cost.

D THE REPLACEMENT OF NATURAL DATA FOR THE RANKING IN STARS.

In our STARS method, we also investigate the difference of replacing the natural input x with adversarial examples. Table 5 shows that the replacement has limited improvements.

E DETAILED EXPERIMENTAL RESULTS.

To verify the rationality of minimum-margin, we conduct experiments on different step size, different step number and different $\mathcal{B}_\epsilon[x]$ in Table 6 and Table 7. We compare the reliability and the computational time between MM attacks and baselines. In Table 8 and Table 9, unless specified, the model structure is ResNet-18. The experiments verify that our MM attack achieves comparable performance but only incurs a very small amount of computational time.

F EXPERIMENTAL RESOURCES

We implement all methods on Python 3.7 (Pytorch 1.7.1) with an NVIDIA GeForce RTX 3090 GPU with AMD Ryzen Threadripper 3960X 24 Core Processor. The CIFAR-10 dataset, the SVHN and the CIFAR-100 dataset can be downloaded via Pytorch. Given the $50,000$ images from the CIFAR-10 and CIFAR-100 training set, $73,257$ digits from the SVHN training set, we conduct the adversarial training on ResNet-18 and Wide ResNet-34 for classification.

Algorithm 2 Adversarial Training of MM attack.

Input: network architecture f parametrized by θ, training dataset S, loss function l, learning rate η, number of epochs T, batch size n;
Output: Adversarial robust network f_θ;
for epoch = $1, 2, \ldots, T$ **do**
  for mini-batch = 1,2,...,N **do**
    Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^{n}$ from S;
    for $i = 1, 2, \ldots, n$ **do**
      Obtain adversarial data of MM attack x'_i of x_i by Algorithm 1;
    end for
    $\theta \leftarrow \theta - \eta \sum_{i=1}^{n} \nabla_\theta \ell \left(f_\theta \left(x'_i \right), y_i \right) / n$;
  end for
end for

Table 4: The successful set of different rescaling methods.

ID	Rescaling method	Formulation	Successful set	Ranking	diff.	$A \cup B_i - A$	$A \cup B_i - B_i$
A	Natural logits	$-(z_y - z_t)$	5379	1	0	0	0
B_1	Softmax	$-\frac{e^{z_y} - e^{z_t}}{\sum_{i=0}^{K} e^{z_i}}$	5377	2	-2	3	5
B_2	Max	$-\frac{z_y - z_t}{z_y}$	5374	=4	-5	3	8
B_3	Sum	$-\frac{z_y - z_t}{z_y + z_t}$	5374	=4	-5	3	8
B_4	Min-Max	$-\frac{z_y - z_t}{z_{\pi_1} - z_{\pi_{10}}}$	5376	3	-3	4	7
B_5	DLR	$-\frac{z_y - z_t}{z_{\pi_1} - \frac{1}{2} \cdot (z_{\pi_3} + z_{\pi_4})}$	5372	6	-7	2	9
B_6	Sigmoid	$-(\frac{e^{z_y}}{1+e^{z_y}} - \frac{e^{z_t}}{1+e^{z_t}})$	5311	7	-68	2	70

Table 5: Test accuracy (%): Replacing natural data with adversarial data in STARS method.

Dataset	Reference attack	Select-ϵ	MM3	Diff.	MM9	Diff.
CIFAR-10	None	8/255	48.23	-0.42	47.81	0.00
CIFAR-10	FGSM	8/255	48.05	-0.24	47.81	0.00
CIFAR-10	PGD-20	8/255	**47.92**	-0.11	47.81	0.00
CIFAR-10	PGD-20	6/255	47.98	-0.17	47.81	0.00
CIFAR-10	PGD-20	4/255	48.04	-0.23	47.81	0.00
SVHN	None	8/255	52.45	-0.61	51.84	0.00
SVHN	FGSM	8/255	52.07	-0.23	51.84	0.00
SVHN	PGD-20	8/255	**51.97**	-0.13	51.84	0.00
SVHN	PGD-20	6/255	52.00	-0.16	51.84	0.00
SVHN	PGD-20	4/255	52.07	-0.23	51.84	0.00
CIFAR-100	None	8/255	23.92	-0.41	23.51	0.00
CIFAR-100	FGSM	8/255	23.63	-0.12	23.51	0.00
CIFAR-100	PGD-20	8/255	**23.57**	-0.06	23.51	0.00
CIFAR-100	PGD-20	6/255	**23.57**	-0.06	23.51	0.00
CIFAR-100	PGD-20	4/255	23.63	-0.12	23.51	0.00

Table 6: Test accuracy (%): the rationality of MM under different step sizes and step numbers.

Step size	Step num	PGD-20	Diff.	CW	Diff.	MM3-F	Diff.	MM9-F	Diff.
CIFAR-10									
0.003	20	51.14	-3.33	49.95	-2.14	48.23	-0.42	**47.81**	0.00
1/255	40	50.16	-3.15	49.13	-2.12	47.46	-0.45	**47.01**	0.00
1/255	20	50.28	-3.22	49.19	-2.13	47.50	-0.44	**47.06**	0.00
1/255	40	49.30	-2.92	48.45	-2.07	46.88	-0.50	**46.38**	0.00
2/255	10	50.54	-3.26	49.38	-2.10	46.70	-0.42	**47.28**	0.00
2/255	20	49.36	-2.93	48.48	-2.05	46.92	-0.49	**46.43**	0.00
4/255	10	49.52	-2.97	48.60	-2.05	47.02	-0.47	**46.55**	0.00
SVHN									
0.003	20	57.68	-5.84	54.42	-2.58	52.45	-0.61	**51.84**	0.00
1/255	40	56.03	-5.78	52.90	-2.65	50.91	-0.66	**50.25**	0.00
1/255	20	56.81	-5.74	53.69	-2.62	51.72	-0.65	**51.07**	0.00
1/255	40	55.49	-5.50	52.59	-2.60	50.65	-0.66	**49.99**	0.00
2/255	10	57.30	-5.71	54.12	-2.53	52.19	-0.60	**51.59**	0.00
2/255	20	55.45	-5.32	52.70	-2.57	50.79	-0.66	**50.13**	0.00
4/255	10	56.16	-5.13	53.52	-2.49	51.62	-0.59	**51.03**	0.00

Table 7: Test accuracy (%): the rationality of MM under different $\mathcal{B}_\epsilon[x]$.

ϵ	PGD-20	Diff.	CW	Diff.	MM3-F	Diff.	MM9-F	Diff.
ResNet-18								
4	67.90	-0.70	68.06	-0.86	67.23	-0.03	**67.20**	0.00
8	51.14	-3.33	49.95	-2.14	48.23	-0.42	**47.81**	0.00
12	45.53	-4.62	43.85	-2.94	41.86	-0.95	**40.91**	0.00
WRN-34								
4	70.23	-0.30	70.55	-0.62	69.94	-0.01	**69.93**	0.00
8	53.69	-2.07	53.89	-2.27	51.95	-0.33	**51.62**	0.00
12	46.76	-3.68	46.24	-3.16	44.05	-0.97	**43.08**	0.00

Table 8: Evaluation: test accuracy (%) on different datasets and model structures.

Methods	CIFAR-10	Diff.	CIFAR-100	Diff.	SVHN	Diff.	[WRN34] CIFAR-10	Diff.
PGD	51.14	-5.03	26.45	-3.92	57.68	-10.39	53.70	-3.88
CW	49.95	-3.84	25.60	-3.07	54.50	-7.21	53.90	-4.08
A-CE	48.58	-2.47	24.71	-2.18	51.55	-4.26	51.00	-1.18
A-DLR	48.85	-2.74	24.85	-2.32	50.64	-3.35	52.24	-2.42
FAB	47.28	-1.17	23.16	-0.63	52.19	-4.90	51.04	-1.22
Square	54.46	-8.35	27.94	-5.41	53.80	-6.51	58.04	-8.22
AA	46.43	-0.32	23.07	-0.54	48.44	-1.15	50.21	-0.39
T-AA	46.12	-0.01	22.53	0.00	47.36	-0.07	49.82	0.00
MM3	46.69	-0.58	22.98	-0.45	49.15	-1.86	50.26	-0.44
MM5	46.34	-0.23	22.72	-0.19	48.69	-1.40	49.99	-0.17
MM+	46.11	0.00	22.53	0.00	47.29	0.00	49.82	0.00

Table 9: Evaluation: the computational time (s) on different datasets and model structures.

Methods	CIFAR-10	Diff.	CIFAR-100	Diff.	SVHN	Diff.	[WRN34] CIFAR-10	Diff.
PGD	60	0	60	0	166	-2	416	-10
CW	62	-2	64	-4	164	0	406	0
A-CE	289	-229	215	-155	777	-613	1910	-1504
A-DLR	305	-245	222	-162	871	-707	1901	-1495
FAB	2181	-2121	1980	-1920	6178	-6014	13809	-13403
Square	3768	-3708	2528	-2468	9506	-9342	22593	-22187
AA	3885	-3825	2187	-2127	11146	-10982	29637	-29231
T-AA	5970	-5910	2967	-2907	25116	-24952	40178	-39772
MM3	126	-66	91	-31	332	-168	796	-390
MM5	182	-122	137	-77	587	-423	1342	-936
MM+	1421	-1361	746	-686	4431	-4267	10773	-10367