Is FGSM Optimal or Necessary for L_∞ Adversarial Attack? Supplementary Material

Chaoning Zhang* chaoningzhang@kaist.ac.kr
Adil Karjauv* mikolez@gmail.com
Philipp Benz* pbenz@kaist.ac.kr
Soomin Ham smham@kaist.ac.kr
Gyusang Cho gyusang.cho@kaist.ac.kr
Chan-Hyun Youn chyoun@kaist.ac.kr
In So Kweon iskweon77@kaist.ac.kr
Korea Advanced Institute of Science and Technology (KAIST)

A. Appendix

A.1. Preliminary Knowledge

White-box Attack. Given a data distribution $\mathcal{D}$, consisting of sample, ground truth pairs (x, y_{gt}). We denote a classifier as $f_\theta(x) : \mathcal{X} \rightarrow \mathcal{Y}$ mapping a sample $x \in \mathcal{X}$ to a label $y \in \mathcal{Y}$. For simplicity, we omit θ in the following. The objective of an adversarial attack is to find a perturbation δ that fools the classifier, *i.e.* $f(x^{adv}) \neq y_{\text{gt}}$, where $x^{adv} = x + \delta$. In the more challenging targeted attack setting, the crafted adversarial example needs to fool the network to a predefined target class y_t, *i.e.* $f(x^{adv}) = y_t$. To ensure that the perturbation δ is (quasi)-imperceptible to the human eye, the perturbation is commonly constrained through an upper bound ϵ on the ℓ_∞-norm, *i.e.* $||\delta||_\infty \leq \epsilon$. To achieve the objective of an (untargeted) adversarial attack to fool a classifier it is a common choice to maximize the objective function $\mathcal{L}$, which was initially minimized for model training. A common choice for $\mathcal{L}$ is the cross-entropy function. Hence, the objectivecan be formalized as:

$$\arg\max_{\delta} \mathcal{L}(f(x+\delta, y), \quad \text{s.t. } ||\delta||_\infty \leq \epsilon, \tag{1}$$

To solve this objective Goodfellow *et al.* proposed a fast method, by calculating the sign of the input gradients, termed fast gradient sign method (FGSM):

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}(x, y)). \tag{2}$$

To improve the attack success rate, [3] proposed an iterative variant of FGSM, *i.e.* I-FGSM, which can be formulated as:

$$x^{adv}_{t=0} = x, \quad x^{adv}_{t+1} = x^{adv}_t + \alpha \cdot sign(\nabla_x \mathcal{L}(x^{adv}_t, y)), \tag{3}$$

where α indicates a step size and is commonly set to ϵ/N with N indicating the total number of iterations. The white-box attack is known as a well-resolved issue and a variety of works emerged discussing transfer-based black-box attacks, of which many are based on the I-FGSM attack.

Black-box attack. One intriguing property of adversarial examples is that they can transfer to an unseen model for performing a transfer-based black-box attack. Compared with FGSM, I-FGSM increases the white-box attack success rate but at the cost of low transferability. A new family of FGSM-based attack methods has been developed, such as MI-FGSM [1], DI-FGSM [4], and TI-FGSM [2] to improve the transferability. MI-FGSM [1] introduces a momentum term as:

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x \mathcal{L}(x^{adv}_t, y)}{||\nabla_x \mathcal{L}(x^{adv}_t, y)||_1}, \\ x^{adv}_{t+1} &= x^{adv}_t + \alpha \cdot sign(g_{t+1}), \end{aligned} \tag{4}$$

where g_t indicates the accumulated gradients at iteration t and μ, usually set to 1, indicates the momentum weight. DI-FGSM [4] improves transferability with input diversity by applying transformations to the input images and can be expressed as:

$$x^{adv}_{t+1} = x^{adv}_t + \alpha \cdot sign(\nabla_x \mathcal{L}(Tr(x^{adv}_t; p), y)), \tag{5}$$

where Tr indicates a transformation, which is applied with probability p. TI-FGSM [2] proposes a translation-invariant attack method byconvolving the gradient of the initial image with a pre-defined kernel W:

$$x^{adv}_{t+1} = x^{adv}_t + \alpha \cdot sign(W * \nabla_x \mathcal{L}(x^{adv}_t, y)). \tag{6}$$

A.2. On the Influence of APD.

I-FGSM is widely known to be stronger than FGSM but has a lower transfer rate [3]. The results in Table 1 show that this counter-intuitive phenomenon can be at least partly explained by the fact that FGSM has a perturbation with

*Equal contribution

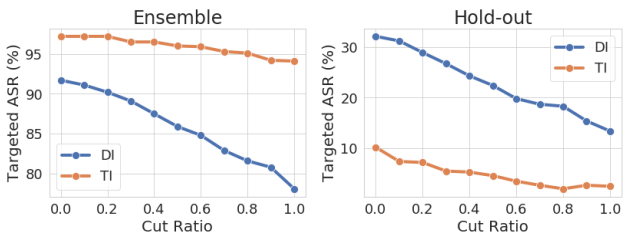

Figure 1. Ablation study of cut ratio parameter on the targeted attack success rate (%) for white-box and hold-out scenarios with DI and TI.

Table 1. The average success rates (%) of non-targeted attacks and APD under $\ell_\infty = 16$ norm constraint for images in the range $[0, 255]$.

Surrogate	Attack	Inc-V3	Inc-V4	IncRes-V2	Res-152	APD
Inc-V3	FGSM	81.0	37.6	33.9	32.9	15.4
	I-FGSM	100.0	29.7	20.9	19.9	3.9
	MI-FGSM	100	54.4	52.1	44.3	10.3
	DI-FGSM	99.8	53.9	42.7	33.1	4.0
	TI-FGSM	100	33.3	23.8	20.6	3.9
Res-152	FGSM	41.5	36.5	32.0	82.2	15.4
	I-FGSM	30.8	25.5	17.8	99.5	4.0
	MI-FGSM	55.7	50.9	48.0	99.4	10.2
	DI-FGSM	62.3	56.9	53.1	99.4	4.1
	TI-FGSM	31.7	26.5	20.1	99.5	4.1

higher APD even though they satisfy the same L_∞ constraint. Specifically, the APD of FGSM is 15.4 (not 16 because some perturbations are clipped for keeping the image in the range of [0, 255]), while that of I-FGSM is 3.9. The reason for low APD in I-FGSM can be attributed to the fact that gradient directions change between iterations, and the perturbation updates can offset each other when the sign changes. Intuitively, a lower APD tends to cause lower transferability. Different from DI-FGSM and TI-FGSM that improve the transferability without increasing the APD, MI-FGSM improves the transferability at the cost of much higher APD. A high APD increases its visibility to the human eye and thus constitutes a drawback of momentum-based iterative approach. The reason that MI-FGSM increases the APD is because MI-FGSM accumulates all historical gradient values, alleviating the offset effect, *i.e.* sign flipping between iterations, in the vanilla I-FGSM. As indicated [1], MI-FGSM *stabilizes* the update directions and escape from poor local maxima during iterations. Intuitively, the reason that the update directions need to be stabilized is that the sign operation of FGSM discards their actual values and maps all the positive/negative values to 1/-1. The sign directions of gradient values near zero are less reliable than those with large values. To illustrate this, we apply the gradient updates in FGSM to partial pixels, and the results are shown in Figure 2. We observe that only applying the perturbation update to pixels with large gradient values results in a significantly stronger and more transferable attack than only applying perturbation update to the pixels with small gradient values. For instance, updating only the top 20% of gradients shows better transferability across black-box models than updating the bottom 80%. In addition, updating only the top 10% of gradients already achieves white-box ASR of 100%, while it requires at least 50% of bottom gradients to get the same attack strength. A similar phenomenon is observed when MI-FGSM is only updated with small or large gradients. The results suggest that the large gradient values constitute a more reliable gradient direction update. Overall, the sign operation in FGSM treats all pixel values equally, causing unstable sign directions for gradient values close to zero, and momentum helps stabilize it, leading to a transferability boost. However, this stabilization effect comes at the cost of higher APD as well as being overly dependent on historical gradients for the perturbation at the current iteration. This indicates the momentum might not be beneficial for improving the performance if the sign instability issues can be directly addressed.

A.3. Ablation Study and Discussion

Cut ratio. The cut ratio indicates the percentage of pixels that get clipped. The higher cut ratio, the smaller t_{cut}. Setting a very small cut-ratio is equivalent to keeping the values of all gradients, *i.e.* close to an identity mapping, while setting a high cut-ratio is close to the original sign method. We explore the effect of the cut ratio of our Cut&Norm attack on the performance of targeted attack using an ensemble of networks of Inc-v4, Res-152, and IncRes-v2. Inc-v3 is used as a hold-out black-box model. Results are shown in Figure 1. We observe an overall trend that the targeted attack success rate decreases when the cut ratio is increased, suggesting the importance of keeping the values for more gradients. In the extreme case of setting the cut ratio to 0%, the &Norm approach is simplified into a linear mapping function. Due to its simplicity and effectiveness, we adopt this special case as our final approach.

Number of iterations (T). Given ϵ set to 16/255, we investigate the influence of T. Note that the step size $\alpha = \epsilon/T$. The results are shown in Figure 3. We observe that over a wide range of T, our approach consistently outperforms existing approaches by a large margin.

L_∞ budget (ϵ) or step size (α). Given the T fixed to 20, we further compare our approach with existing ones on a wide range of ϵ. Since we still follow $\alpha = \epsilon/T$, it is also equivalent to investigate the influence of step size α. The results in Figure 3 show that our approach consistently outperforms the existing approaches by a large margin.

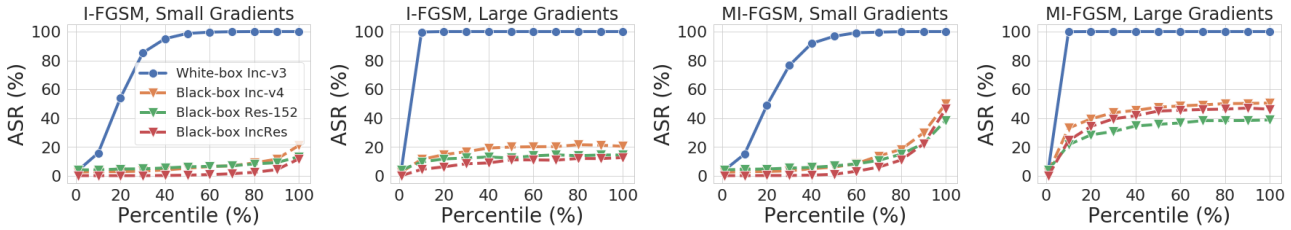

Figure 2. Training I-FGSM/MI-FGSM on a bottom (small gradients) or top (large gradients) percentile of pixels ranked by absolute value of their gradient.

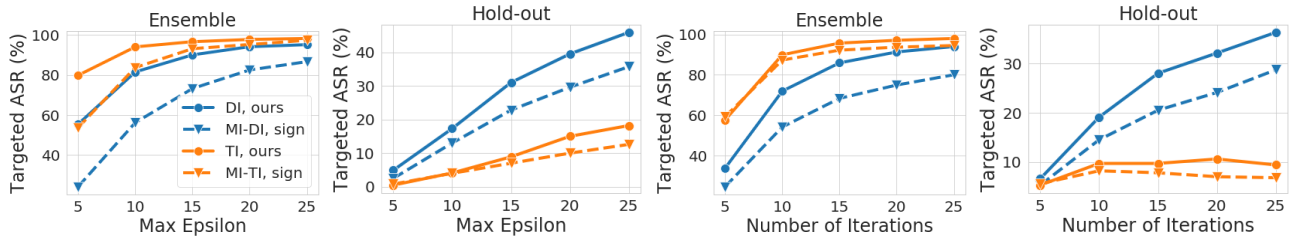

Figure 3. **Left two images:** Ablation study of a different ℓ_∞ norm constraint (ϵ) for perturbation on the targeted attack success rate (%) for white-box and hold-out scenarios with DI, TI, MI-DI, and MI-TI. **Right two images:** Ablation study of different number of iterations used for training the perturbation on the targeted attack success rate (%) for white-box and hold-out scenarios with DI, TI, MI-DI, and MI-TI.

On the effect of momentum. As discussed in Sec. A.2, except for the drawback of implicitly inducing higher APD, the momentum also makes the current perturbation update overly dependent on the historical gradients. Given this drawback, the momentum still improves the transferability of I-FGSM is because it solves the problem of unreliable sign directions in the I-FGSM. Our approach alleviates this problem directly by proportionally keeping the gradient values, and thus we conjecture that adding momentum to our approach might not improve the performance. The results in Table 2 show that adding momentum to our approach indeed decreases the transferability performance even though it increases the APD, highlighting the drawback of MI-FGSM being overly dependent on the historical gradient values.

B. Transferable Non-targeted Attack

Following the overall trend in Table 2 and Table 3 of the main manuscript, Table 3 in this supplementary shows that our method outperforms I-FGSM [3] and MI-FGSM [1] for ensemble white-box and hold-out black-box scenarios, in *non-targeted* setting. For example, with the strong baseline MI-TI-DI, the FGSM-based approach achieves an average of 91.9% transfer ASR, while our counterpart without momentum achieves a significantly higher 96.4%, with less than half APD (10.4 vs 4.8).

References

[1] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1, 2, 3

[2] Y. Dong, T. Pang, H. Su, and J. Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019. 1

[3] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 1, 3

[4] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 1

Table 2. Ablation study of applying momentum to our approach. Better result is in **bold**.

Attack	I	MI	DI	MI-DI	TI	MI-TI	TI-DI	MI-TI-DI	TI-DI-Po	MI-TI-DI-Po
White-box	**85.5** / 4.0	80.2 / 8.8	**85.3** / 4.0	80.5 / 8.8	**76.2** / 4.6	63.6 / 8.9	**76.2** / 4.6	63.8 / 8.9	**67.3** / 4.4	65.2 / 8.9
Black-box	**1.9** / 4.0	0.7 / 8.8	**2.3** / 4.0	0.7 / 8.8	**40.3** / 4.6	32.4 / 8.9	**40.2** / 4.6	33.2 / 8.9	31.6 / 4.4	**34.6** / 8.9

Table 3. The ASR/APD of non-targeted FGSM-based attack and our method. We study nine models—Inc-v3, Inc-v4, Res-152, Res-101, Res-50, IncRes-v2, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$, and adversarial examples are crafted via an ensemble of eight of them. In each column, "-" denote the hold-out model.

	Attacks	-Inc-v3$_{ens3}$	-Inc-v3$_{ens4}$	-IncRes-v2$_{ens}$	Avg.
Ensemble White-box	I	99.5 / 3.5	99.5 / 3.5	99.5 / 3.5	99.5 / 3.5
	MI	99.3 / 9.9	99.4 / 9.9	99.5 / 9.9	99.4 / 9.9
	Ours	**99.8** / 4.0	**99.8** / 4.1	**99.8** / 4.0	**99.8** / 4.0
	DI	98.0 / 3.4	97.9 / 3.4	98.1 / 3.3	98.0 / 3.4
	MI-DI	98.6 / 10.1	98.5 / 10.2	98.6 / 10.1	98.6 / 10.1
	Ours	**99.4** / 4.0	**99.1** / 4.1	**99.5** / 4.1	**99.3** / 4.1
	TI	98.2 / 4.6	98.3 / 4.6	98.6 / 4.6	98.4 / 4.6
	MI-TI	97.5 / 9.8	97.8 / 9.7	98.0 / 9.8	97.8 / 9.8
	Ours	**99.2** / 4.8	**99.1** / 4.8	**99.3** / 4.8	**99.2** / 4.8
	TI-DI	95.9 / 4.3	96.1 / 4.3	96.5 / 4.3	96.2 / 4.3
	MI-TI-DI	96.4 / 10.4	96.4 / 10.4	96.9 / 10.4	96.6 / 10.4
	Ours	**98.6** / 4.8	**98.7** / 4.8	**99.1** / 4.8	**98.8** / 4.8
Hold-out Black-box	I	31.6 / 3.5	32.9 / 3.5	19.0 / 3.5	27.8 / 3.5
	MI	**47.3** / 9.9	**50.7** / 9.9	**34.7** / 9.9	**44.2** / 9.9
	Ours	40.0 / 4.0	43.9 / 4.1	30.1 / 4.0	38.0 / 4.0
	DI	63.9 / 3.4	60.9 / 3.4	50.1 / 3.3	58.3 / 3.4
	MI-DI	75.5 / 10.1	72.9 / 10.2	66.6 / 10.1	71.7 / 10.1
	Ours	**76.2** / 4.0	**74.4** / 4.1	**71.7** / 4.1	**74.1** / 4.1
	TI	83.5 / 4.6	83.5 / 4.6	74.2 / 4.6	80.4 / 4.6
	MI-TI	90.6 / 9.8	91.1 / 9.7	87.4 / 9.8	89.7 / 9.8
	Ours	**94.0** / 4.8	**92.9** / 4.8	**88.4** / 4.8	**91.8** / 4.8
	TI-DI	88.2 / 4.3	88.0 / 4.3	83.1 / 4.3	86.4 / 4.3
	MI-TI-DI	92.6 / 10.4	93.2 / 10.4	90.0 / 10.4	91.9 / 10.4
	Ours	**96.9** / 4.8	**97.3** / 4.8	**95.0** / 4.8	**96.4** / 4.8