Supplementary Material: Strengthening the Transferability of Adversarial Examples Using Advanced Looking Ahead and Self-CutMix

Donggon Jang* Sanghyeok Son* Dae-Shik Kim
Korea Advanced Institute of Science and Technology (KAIST)
{jdg900, ssh816, daeshik}@kaist.ac.kr

1. Implementation Details

We used Tensorflow to implement the attack methods and tested them on TITAN RTX and TITAN V. Experiments were mostly conducted with three random seeds. We follow the setting in MI-FGSM with the number of iterations $T = 10$, maximum perturbation $\epsilon = 16$, and step size $\alpha = 1.6$. For M(N)I-FGSM, we adopt the decay factor $\mu = 1.0$. For VM(N)I-FGSM, we choose the upper bound factor of neighborhood $\beta = 1.5$ and the number of samples for variance tuning $v = 20$. For EMI-FGSM, we choose the number of samples $N = 11$ and sampling interval bound $\eta = 7$ and adopt the linear sampling. For DIM and TIM, we set the transformation probability to 0.5 and the filter size of the Gaussian kernel to 7×7, respectively. For SIM, we choose the number of scale copies $s = 5$. For Admix, we set the number of randomly samples images from other categories as 3 and the portion of the original image as 0.2, respectively. For our methods, we adopt the number of lookahead steps $N = 17$, the number of neighbor samples $v = 5$, the minimum patch size $P = 200$, and the number of mixed copies $c = 4$.

2. Results of Combining SCM with Other Input Transformations

We conduct additional experiments to show that SCM-P(R) achieve higher attack success rates even when combined with existing input transformation methods, *i.e*. DIM, TIM, and CTM. As shown in Tab. 1, Tab. 2, and Tab. 3, SCM-P(R) further improve the transferability of adversarial examples. Therefore, SCM-P(R) could be an attractive option to generate more transferable adversaries.

3. Results of Combining LI-FGSM with Other Input Transformations

In order to validate the effect of LI-FGSM when combined with other input transformations, we conduct additional experiments. Tab. 4, Tab. 5, Tab. 6, and Tab. 7 show the results comparing various gradient based attack methods using input transformations, *i.e*. DIM, TIM, SIM, and CTM. LI-FGSM achieves the highest attack success rates in all experiments.

4. Ablation Study on Hyper-parameters

4.1. SCM-P(R)

We conduct additional experiments to study the influence of a copied (pasted) patch size in SCM-P(R). We measure the attack success rates against four normally trained models and three adversarially trained models by varying the patch size. As shown in Fig. 1, the transferability tends to improve with increasing patch size in both SCM-P and SCM-R. Specifically, for SCM-R, the attack success rates increase steeply after patch size is 200. SCM-P also shows satisfactory consistent transferability after patch size is 200. Thus, we simply set the minimum patch size to 200 in both SCM-P and SCM-R.

4.2. LI-FGSM

We experiment with varying the number of steps looking ahead according to an inner loop algorithm. We utilize MI-FGSM, NI-FGSM, and VMI-FGSM as an inner loop except for VNI-FGSM shown in the discussion section of the main paper. As shown in Fig. 2, the attack success rate grows rapidly as the number of steps increases and then converges. However, the converged attack success rates are worse than the case of using VNI-FGSM. Consequently, we leverage VNI-FGSM as an inner loop.

5. Training Budget

5.1. Computational Complexity

It is not easy to directly compare time and memory complexity since it depends on the source model. Instead, we count the number of back-propagation at each iteration that is highly correlated to the attack method and dominantly affects computation overhead. As shown in the Tab. 8, the complexity of LI-FGSM is higher than others. However, in

*These authors contributed equally.

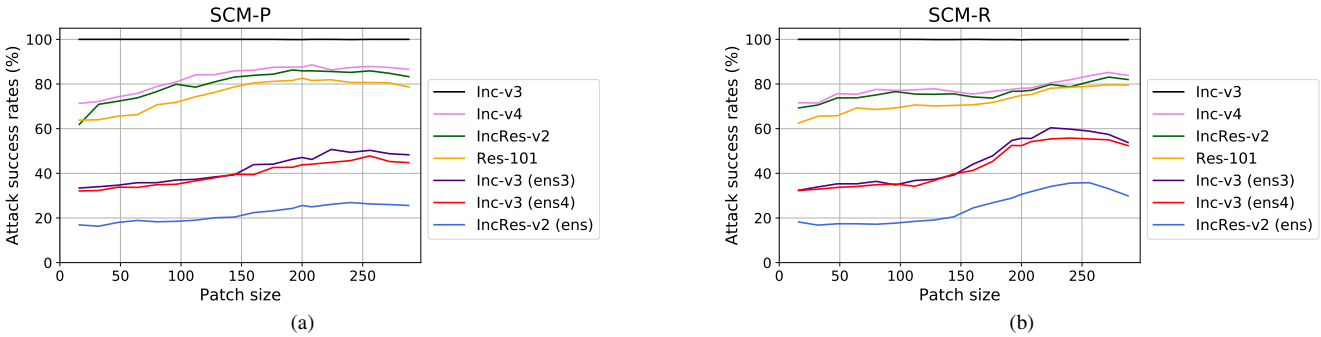

Figure 1. **(a)**: Attack success rates (%) according to varying the patch sizes of SCM-P against four normally trained models and three adversarially trained models. **(b)**: Attack success rates (%) according to varying the patch sizes of SCM-R against four normally trained models and three adversarially trained models. In this experiment, the adversarial examples are generated by Inc-v3 using MI-FGSM and SCM-P(R).

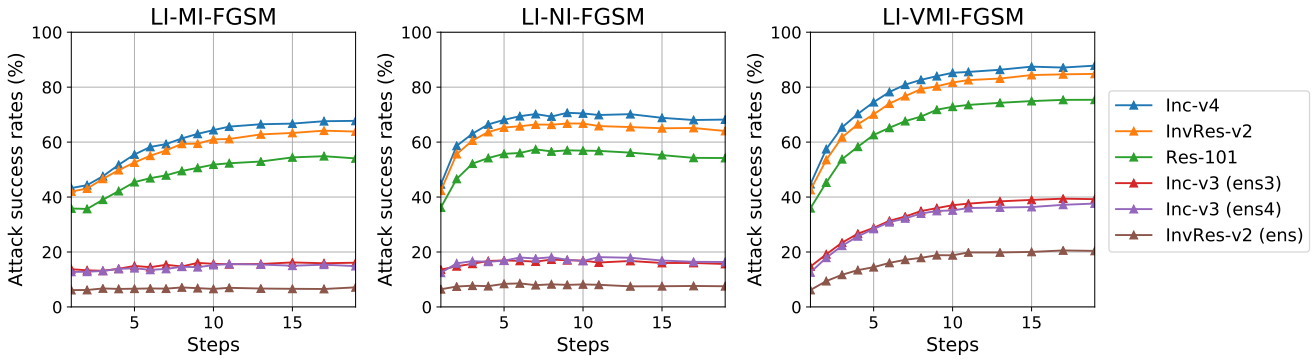

Figure 2. Attack success rates (%) according to varying the number of steps looking ahead on six models by utilizing MI-FGSM, NI-FGSM, and VMI-FGSM as an inner loop of LI-FGSM, except Inc-v3, which is the source model.

the next subsection, we exhibit that our method can outperform others with small looking ahead steps.

5.2. Attack Scenario under a Small Step Budget

We conduct additional experiments to demonstrate whether our methods can be helpful under a small step budget. We measure the performance of LI-VNI-FGSM varying the number of lookahead steps, and compare it with the existing methods as shown in Fig. 3. Our method outperforms the MI-FGSM and NI-FGSM from two steps (one additional step). Furthermore, ours achieves better performance beyond the EMI-FGSM and the VNI-FGSM from four steps (three additional steps). We believe that it is affordable overhead to improve the attack success rate. It is noteworthy that as the number of lookahead steps increases, the performance of our method also increases even after four steps. It means that if enough computing power is given, it is appropriate to use our method to have a higher attack success rates.

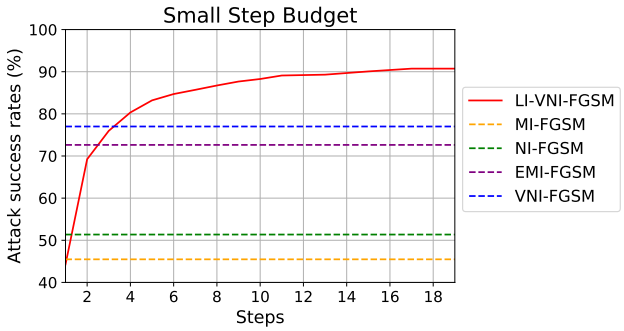

Figure 3. Attack success rates (%) according to varying the number of steps looking ahead on Inc-v4. Inc-v3 is used for the source model. We only take MI-FGSM, NI-FGSM, EMI-FGSM, and VNI-FGSM except for VMI-FGSM to compare with LI-VNI-FGSM because VNI-FGSM and VMI-FGSM show very similar performance.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	SI-DIM	99.13*	83.97	81.17	77.50	47.00	45.03	25.90
	Admix-DIM	**99.80***	90.50	**87.70**	83.60	52.20	49.90	28.60
	SCM-P-DIM	98.97*	**89.57**	81.17	84.53	52.40	48.70	28.90
	SCM-R-DIM	99.00*	88.10	86.53	**84.97**	**64.53**	**62.50**	**39.77**
Inc-v4	SI-DIM	89.27	99.13*	85.57	80.00	59.50	56.80	38.23
	Admix-DIM	**93.00**	**99.20***	**89.70**	85.20	62.40	60.30	39.70
	SCM-P-DIM	91.83	98.77*	88.67	85.00	63.13	60.40	41.20
	SCM-R-DIM	90.70	98.70*	88.47	**85.90**	**72.03**	**70.50**	**53.57**
IncRes-v2	SI-DIM	88.80	85.80	97.83*	82.43	66.80	60.93	53.93
	Admix-DIM	90.20	**88.40**	**98.00***	**85.80**	70.50	63.70	55.30
	SCM-P-DIM	**90.40**	88.03	97.63*	85.43	69.70	64.43	54.13
	SCM-R-DIM	89.27	87.17	97.17*	85.63	**75.20**	**72.17**	**63.43**
Res-101	SI-DIM	87.60	83.33	84.77	98.93*	62.30	56.23	40.40
	Admix-DIM	**91.90**	89.00	**89.60**	**99.80***	69.70	62.30	46.60
	SCM-P-DIM	91.73	**89.30**	89.07	99.00*	70.07	64.30	46.27
	SCM-R-DIM	90.33	87.53	88.70	99.03*	**76.67**	**73.80**	**57.47**

Table 1. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using SI-DIM, Admix-DIM, SCM-P-DIM, and SCM-R-DIM. Inc-v3, Inc-v4, IncRes-v2, and Res-101 using MI-FGSM are adopted as the source model respectively. * indicates that the target model is the same as the source model. In case of Admix-DIM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	SI-TIM	**100.00***	71.20	68.47	62.67	49.60	47.63	31.27
	Admix-TIM	**100.00***	83.90	80.40	74.40	59.10	57.90	39.20
	SCM-P-TIM	99.90*	**87.23**	**84.37**	80.63	67.67	64.70	47.97
	SCM-R-TIM	99.80*	85.50	83.23	**81.83**	**75.17**	**72.77**	**56.83**
Inc-v4	SI-TIM	78.07	99.60*	73.03	66.00	58.60	55.13	45.33
	Admix-TIM	87.40	**99.70***	82.30	77.00	68.10	65.30	53.10
	SCM-P-TIM	**89.40**	99.67*	**85.57**	80.37	71.10	68.40	58.77
	SCM-R-TIM	88.30	99.67*	85.03	**80.70**	**77.30**	**75.93**	**65.37**
IncRes-v2	SI-TIM	84.80	81.57	**98.87***	76.63	69.90	64.73	61.47
	Admix-TIM	90.20	**88.20**	98.60*	83.90	78.40	73.60	70.00
	SCM-P-TIM	**90.70**	88.13	98.43*	**84.90**	78.77	76.20	72.30
	SCM-R-TIM	88.63	87.13	98.53*	84.63	**82.33**	**79.30**	**75.80**
Res-101	SI-TIM	73.97	70.80	70.20	**99.80***	59.40	55.07	43.30
	Admix-TIM	83.20	78.90	80.70	99.70*	67.00	62.50	52.80
	SCM-P-TIM	**88.77**	**86.07**	**87.63**	99.77*	79.50	76.00	65.83
	SCM-R-TIM	87.43	82.20	84.83	99.73*	**81.33**	**79.10**	**70.03**

Table 2. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using SI-TIM, Admix-TIM, SCM-P-TIM, and SCM-R-TIM. Inc-v3, Inc-v4, IncRes-v2, and Res-101 using MI-FGSM are adopted as the source model respectively. * indicates that the target model is the same as the source model. In case of Admix-TIM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	CTM	99.20*	85.17	81.77	76.70	66.07	63.17	45.70
	Admix-CTM	**99.90***	89.00	**87.00**	83.10	72.20	71.10	52.40
	SCM-P-CTM	99.10*	**89.13**	86.87	**84.00**	71.87	69.77	53.43
	SCM-R-CTM	99.43*	87.67	85.83	83.80	**78.40**	**76.90**	**62.57**
Inc-v4	CTM	87.07	98.77*	83.60	78.17	71.40	68.47	57.13
	Admix-CTM	**90.40**	**99.00***	**87.30**	82.00	75.30	71.90	61.60
	SCM-P-CTM	90.20	98.63*	87.07	82.83	74.30	72.17	62.30
	SCM-R-CTM	90.03	98.47*	86.83	**83.57**	**80.00**	**78.60**	**68.27**
IncRes-v2	CTM	88.23	86.07	97.37*	82.70	77.20	75.17	72.17
	Admix-CTM	90.10	**89.60**	**97.70***	85.90	82.00	78.00	76.30
	SCM-P-CTM	**90.47**	88.57	97.43*	85.80	80.23	77.47	73.63
	SCM-R-CTM	88.87	87.57	96.83*	**85.93**	**82.50**	**80.23**	**77.07**
Res-101	CTM	85.93	82.07	83.83	98.97*	76.23	71.80	62.17
	Admix-CTM	**91.00**	87.70	89.20	**99.99***	81.10	77.40	70.10
	SCM-P-CTM	90.73	**88.07**	**89.27**	99.17*	81.90	79.03	70.70
	SCM-R-CTM	88.17	83.80	86.93	98.73*	**83.87**	**81.93**	**74.13**

Table 3. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using CTM (the combination of DIM, TIM, and SIM), Admix-CTM, SCM-P-CTM, and SCM-R-CTM. Inc-v3, Inc-v4, IncRes-v2, and Res-101 using MI-FGSM are adopted as the source model respectively. * indicates that the target model is the same as the source model. In case of Admix-CTM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	MI-DI-FGSM	98.72*	64.54	60.46	54.04	19.22	17.96	9.56
	NI-DI-FGSM	99.33*	60.20	58.07	48.77	14.93	14.67	7.23
	VMI-DI-FGSM	99.03*	76.43	73.27	66.53	39.87	37.57	22.23
	VNI-DI-FGSM	99.27*	79.43	76.73	68.60	39.37	37.67	23.33
	EMI-DI-FGSM	99.10*	83.50	78.00	70.60	27.80	26.00	13.40
	LI-DI-FGSM	**100.00***	**97.63**	**96.67**	**93.33**	**65.23**	**61.07**	**38.40**

Table 4. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using DIM. Inc-v3 is adopted as the source model. * indicates that the target model is the same as the source model. In case of EMI-FGSM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	MI-TI-FGSM	**100.00***	48.02	41.28	40.20	24.10	21.34	13.48
	NI-TI-FGSM	99.97*	44.57	38.97	35.03	31.67	29.50	22.73
	VMI-TI-FGSM	**100.00***	71.00	68.43	60.10	32.40	30.80	17.57
	VNI-TI-FGSM	**100.00***	76.80	74.80	64.93	34.70	32.80	18.80
	EMI-TI-FGSM	**100.00***	79.40	76.30	67.20	44.30	40.80	26.20
	LI-TI-FGSM	**100.00***	**89.33**	**87.87**	**79.30**	**67.40**	**63.93**	**49.43**

Table 5. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using TIM. Inc-v3 is adopted as the source model. * indicates that the target model is the same as the source model. In case of EMI-FGSM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	MI-SI-FGSM	**100.00***	69.84	67.66	62.86	32.20	31.46	17.42
	NI-SI-FGSM	**100.00***	77.50	74.97	67.20	32.30	30.07	16.23
	VMI-SI-FGSM	**100.00***	86.87	83.70	78.10	55.03	52.83	35.00
	VNI-SI-FGSM	**100.00***	89.67	88.07	81.83	58.57	55.70	36.83
	EMI-SI-FGSM	**100.00***	91.90	90.00	85.40	45.20	41.80	23.80
	LI-SI-FGSM	**100.00***	**95.50**	**93.83**	**90.17**	**63.27**	**58.73**	**38.77**

Table 6. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using SIM. Inc-v3 is adopted as the source model. * indicates that the target model is the same as the source model. In case of EMI-FGSM, we take the reported value.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3 (ens3)	Inc-v3 (ens4)	IncRes-v2 (ens)
Inc-v3	MI-CT-FGSM	99.30*	84.57	81.53	75.73	65.50	63.47	45.80
	NI-CT-FGSM	99.37*	84.33	80.63	75.53	60.80	56.07	40.40
	VMI-CT-FGSM	99.10*	88.63	86.20	82.57	77.60	76.13	64.27
	VNI-CT-FGSM	99.50*	91.80	89.57	85.87	79.90	77.90	66.93
	EMI-CT-FGSM	99.60*	94.10	92.60	89.40	78.90	75.30	60.40
	LI-CT-FGSM	**100.00***	**97.97**	**97.17**	**94.57**	**91.17**	**89.00**	**79.70**

Table 7. Attack success rates (%) against four normally trained models and three adversarially trained models under a single model setting using CTM, which is a combination of DIM, TIM, and SIM. Inc-v3 is adopted as the source model. * indicates that the target model is the same as the source model. In case of EMI-FGSM, we take the reported value.

	MI-FGSM	NI-FGSM	VMI-FGSM	VNI-FGSM	EMI-FGSM	LI-FGSM
Counts	1	1	v+1	v+1	v	N*(v+1)

Table 8. The number of back-propagation in each iteration. The v is the sampling number and the N is the number of looking ahead steps of LI-FGSM