Supplementary Material for: Boosting Adversarial Attacks with Momentum

Yinpeng Dong[1], Fangzhou Liao[1], Tianyu Pang[1], Hang Su[1], Jun Zhu[1]*, Xiaolin Hu[1], Jianguo Li[2]
[1] Department of Computer Science and Technology, Tsinghua Lab of Brain and Intelligence
[1] Beijing National Research Center for Information Science and Technology, BNRist Lab
[1] Tsinghua University, 100084 China
[2] Intel Labs China
{dyp17, liaofz13, pty17}@mails.tsinghua.edu.cn, {suhangss, dcszj, xlhu}@mail.tsinghua.edu.cn, jianguo.li@intel.com

In this supplementary material, we provide more results in our experiments. In Sec. A, we report the success rates of non-targeted attacks based on L_2 norm bound. In Sec. B, we provide the results of targeted attacks. The experiments consistently demonstrate the effectiveness of the proposed momentum-based methods.

A. Non-targeted attacks based on L_2 norm bound

We first perform non-targeted attacks based on L_2 norm bound. Since the L_2 distance between an adversarial example and a real example is defined as

$$\|\boldsymbol{x}^* - \boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{N}(x_i^* - x_i)^2}, \tag{14}$$

where N is the dimension of input $\boldsymbol{x}$ and $\boldsymbol{x}^*$, the distance measure depends on N. For example, if the distance of each dimension of an adversarial example and a real example is $|x_i^* - x_i| = \epsilon$, the L_2 norm is $\epsilon\sqrt{N}$ between them while the L_∞ norm is ϵ. Therefore, we set the L_2 norm bound as $16\sqrt{N}$ in our experiments, where N is the dimension of the input to a network.

A.1. Attacking a single model

We include seven networks in this section, which are Inc-v3, Inc-v4, IncRes-v2, Res-152, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$. We generate adversarial examples for Inc-v3, Inc-v4, IncRes-v2 and Res-152 respectively, and measure the success rates of attacks on all models. We compare three attack methods, which are the fast gradient method (FGM, defined in Eq. (2)), iterative FGM (I-FGM) and momentum iterative FGM (MI-FGM, defined in Eq. (10)). We set the number of iterations to 10 in I-FGM and MI-FGM, and the decay factor to 1.0 in MI-FGM.

The results are shown in Table 5. We can also see that MI-FGM attacks a white-box model with a near 100%

	Attack	Ensemble	Hold-out
-Inc-v3	FGM	47.3	52.7
	I-FGM	99.1	65.3
	MI-FGM	**99.2**	**89.7**
-Inc-v4	FGM	47.2	49.3
	I-FGM	99.3	56.7
	MI-FGM	**99.4**	**88.0**
-IncRes-v2	FGSM	47.3	50.4
	I-FGSM	99.4	54.3
	MI-FGSM	**99.5**	**86.1**
-Res-152	FGM	47.6	46.6
	I-FGM	99.0	44.7
	MI-FGM	**99.5**	**81.4**
-Inc-v3$_{ens3}$	FGM	51.8	35.4
	I-FGM	**99.8**	29.5
	MI-FGM	99.6	**59.8**
-Inc-v3$_{ens4}$	FGM	51.2	37.5
	I-FGM	99.2	36.4
	MI-FGM	**99.7**	**66.5**
-IncRes-v2$_{ens}$	FGSM	54.4	32.4
	I-FGSM	99.2	19.9
	MI-FGSM	**99.8**	**56.4**

Table 4: The success rates (%) of non-targeted adversarial attacks based on L_2 norm bound against an ensemble of white-box models and a hold-out black-box target model. In each row, "-" indicates the name of the hold-out model and the adversarial examples are generated for the ensemble of the other six models.

*Corresponding author.

success rate as I-FGM, and outperforms FGM and I-FGM in black-box attacks significantly. The conclusions are similar to those of L_∞ norm bound experiments, which consistently demonstrate the effectiveness of the proposed momentum-based iterative methods.

A.2. Attacking an ensemble of models

In this experiments, we also include Inc-v3, Inc-v4, IncRes-v2, Res-152, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$ models for our study. We keep one model as the hold-out black-box model and attack an ensemble of the other six models by FGM, I-FGM and MI-FGM respectively. We set the number of iterations to 20 in I-FGM and MI-FGM, the decay factor to 1.0 in MI-FGM, and the ensemble weights to $1/6$ equally.

We show the results in Table 4. Iterative methods including I-FGM and MI-FGM can obtain a near 100% success rate for an ensemble of white-box models. And MI-FGM can attack a black-box model with a much higher success rate, showing the good transferability of the adversarial examples generated by MI-FGM. For adversarially trained models, MI-FGM can fool them with about 60% success rates, revealing the great vulnerability of the adversarially trained models against our black-box attacks.

B. Targeted attacks

B.1. L_∞ norm bound

Targeted attacks are much more difficult than non-targeted attacks in the black-box manner, since they require the black-box model to output the specific target label. For DNNs trained on a dataset with thousands of output categories such as the ImageNet dataset, finding targeted adversarial examples by only one model to fool a black-box model is impossible. Thus we perform targeted attacks by integrating the ensemble-based approach.

We show the results in Table 6, where the success rate is measured by the percentage of the adversarial examples that are classified as the target label by the model. Similar to the experimental settings in Sec. 4.3, we keep one model to test the performance of black-box attacks, with the targeted adversarial examples generated for the ensemble of the other six models. We set the size of perturbation ϵ to 48, decay factor μ to 1.0 and the number of iterations to 20 for I-FGSM and MI-FGSM. We can see that one-step FGSM can hardly attack the ensemble of models as well as the target black-box models. The success rates of the adversarial examples generated by MI-FGSM are close to 100% for white-box models and higher than 10% for normally trained black-box models. Unfortunately, it cannot effectively generate targeted adversarial examples to fool adversarially trained models, which remains an open issue for future researches.

B.2. L_2 norm bound

We draw similar conclusions for targeted attacks based on L_2 norm bound. In our experiments, we also include Inc-v3, Inc-v4, IncRes-v2, Res-152, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$ models. We keep one model as the hold-out black-box model and attack an ensemble of the other six models with equal ensemble weights by FGM, I-FGM and MI-FGM respectively. We set the maximum perturbation ϵ to $48\sqrt{N}$ where N is the dimension of inputs, the number of iterations to 20 in I-FGM and MI-FGM, and the decay factor to 1.0 in MI-FGM. We report the success rates of adversarial examples against the white-box ensemble of models and the black-box target model in Table 7. MI-FGM can easily fool white-box models, but it cannot fool the adversarially trained models effectively in the targeted black-box attacks.

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FGM	76.2*	41.0	43.1	41.3	34.6	34.9	26.2
	I-FGM	**100.0***	39.9	36.4	27.5	17.5	19.2	10.9
	MI-FGM	**100.0***	**67.6**	**66.3**	**56.1**	**44.4**	**45.5**	**33.9**
Inc-v4	FGM	47.3	63.1*	37.3	39.0	35.3	33.9	27.7
	I-FGM	52.8	**100.0***	42.0	33.5	21.9	19.9	13.8
	MI-FGM	**76.9**	**100.0***	**69.6**	**59.7**	**51.2**	**51.0**	**39.4**
IncRes-v2	FGM	48.2	38.9	60.4*	39.8	36.6	35.5	30.5
	I-FGM	56.0	47.5	**99.6***	36.9	27.5	22.9	18.7
	MI-FGM	**81.7**	**75.8**	**99.6***	**66.9**	**62.7**	**57.7**	**58.8**
-Res-152	FGM	50.8	40.7	42.0	75.1*	36.5	36.0	31.6
	I-FGM	47.6	43.9	43.9	99.4*	32.7	32.3	25.2
	MI-FGM	**71.3**	**65.5**	**64.3**	**99.6***	**56.7**	**55.4**	**51.5**

Table 5: The success rates (%) of non-targeted adversarial attacks based on L_2 norm bound against all models. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-152 respectively using FGM, I-FGM and MI-FGM. * indicates the white-box attacks.

	Attack	Ensemble	Hold-out
-Inc-v3	FGSM	0.5	0.5
	I-FGSM	**99.6**	9.0
	MI-FGSM	99.5	**17.6**
-Inc-v4	FGSM	0.3	0.4
	I-FGSM	**99.9**	7.0
	MI-FGSM	99.8	**15.6**
-IncRes-v2	FGSM	0.4	0.2
	I-FGSM	**99.9**	7.3
	MI-FGSM	99.8	**16.1**
-Res-152	FGSM	0.1	0.5
	I-FGSM	**99.6**	3.3
	MI-FGSM	99.5	**11.4**
-Inc-v3$_{ens3}$	FGSM	0.3	0.1
	I-FGSM	**99.7**	0.1
	MI-FGSM	**99.7**	**0.5**
-Inc-v3$_{ens4}$	FGSM	0.2	0.1
	I-FGSM	**99.9**	0.4
	MI-FGSM	99.8	**0.9**
-IncRes-v2$_{ens}$	FGSM	0.5	0.1
	I-FGSM	99.7	0.1
	MI-FGSM	**99.8**	**0.2**

Table 6: The success rates (%) of targeted adversarial attacks based on L_∞ norm bound against an ensemble of white-box models and a hold-out black-box target model. In each row, "-" indicates the name of the hold-out model and the adversarial examples are generated for the ensemble of the other six models.

	Attack	Ensemble	Hold-out
-Inc-v3	FGM	0.7	0.4
	I-FGM	**99.7**	17.8
	MI-FGM	99.5	**21.0**
-Inc-v4	FGM	0.7	0.5
	I-FGM	**99.9**	15.2
	MI-FGM	99.8	**21.8**
-IncRes-v2	FGM	0.7	0.7
	I-FGM	99.8	16.4
	MI-FGM	**99.9**	**21.7**
-Res-152	FGM	0.5	0.4
	I-FGM	99.5	9.2
	MI-FGM	**99.6**	**17.4**
-Inc-v3$_{ens3}$	FGM	0.6	0.2
	I-FGM	**99.9**	0.7
	MI-FGM	99.6	**1.6**
-Inc-v3$_{ens4}$	FGM	0.5	0.2
	I-FGM	99.7	1.7
	MI-FGM	**100.0**	**2.0**
-IncRes-v2$_{ens}$	FGM	0.6	0.4
	I-FGM	99.6	0.5
	MI-FGM	**99.8**	**1.9**

Table 7: The success rates (%) of targeted adversarial attacks based on L_2 norm bound against an ensemble of white-box models and a hold-out black-box target model. In each row, "-" indicates the name of the hold-out model and the adversarial examples are generated for the ensemble of the other six models.