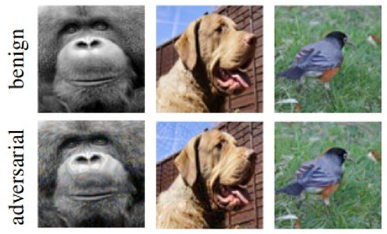
Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks

Jiadong Lin[1], Chuanbiao Song[1], Kun He[1], Liwei Wang[2], John E. Hopcroft[3]
[1]Huazhong University of Science and Technology, [2]Pecking University, [3]Cornell University

ICLR 2020

Introduction

- Deep learning models are vulnerable to adversarial examples crafted by applying human-imperceptible perturbations on benign inputs.



- Attack Methods:
  - Fast Gradient Sign Method (FGSM) [Goodfellow et al., 2015]

$$x^{adv} = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(x, y^{true}))$$

  - Iterative Fast Gradient Sign Method (I-FGSM) [Kurakin et al., 2016]

$$x^{adv}_{t+1} = \mathrm{Clip}^{\epsilon}_{x}\{x^{adv}_{t} + \alpha \cdot \mathrm{sign}(\nabla_x J(x^{adv}_{t}, y^{true}))\}$$

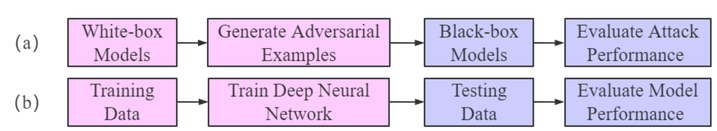
  - Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [Dong et al., 2018]

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x^{adv}_t, y^{true})}{\|\nabla_x J(x^{adv}_t, y^{true})\|_1},$$
$$x^{adv}_{t+1} = \mathrm{Clip}^{\epsilon}_{x}\{x^{adv}_{t} + \alpha \cdot \mathrm{sign}(g_{t+1})\},$$

- Defense Methods:
  - Adversarial Training Augments the training data by the adversarial examples in the training process. [Goodfellow et al., 2014; Madry et al., 2018]
  - Input Modification Mitigates the effects of adversarial perturbations by modifying the input data. [Guo et al., 2018; Liao et al., 2018]
  - Certifiable robustness studies classifiers whose prediction at any point x is verifiably constant within some set around x. [Wong et al., 2018; Cohen et al., 2019]

Motivation

- The process of generating adversarial examples (a) is similar with the process of training deep neural networks (b) .



- From the perspective of the optimization, the transferability of the adversarial examples is similar with the generalization ability of the trained models [Dong et al., 2018].
- The methods to improve the transferability of adversarial examples can be split to two aspects:
  - optimization algorithm, such as MI-FGSM [Dong et al., 2018], which applies the idea of momentum;
  - model augmentation, such as DIM [Xie et al., 2019], which performs random resizing and padding on input images to achieve model augmentation.

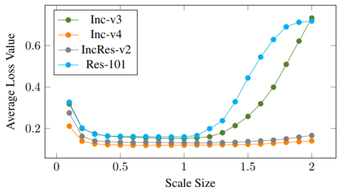
Based on above analysis, we aim to improve the transferability of adversarial examples by applying the idea of Nesterov accelerated gradient for optimization and using a set of scaled images to achieve model augmentation.

Methodology

- Nesterov iterative fast gradient sign method (NI-FGSM)
  - We integrate Nesterov Accelerated Gradient (NAG) into the iterative gradient-based attack to leverage the looking ahead property of NAG and build a robust adversarial attack.

$$x^{nes}_t = x^{adv}_t + \alpha \cdot \mu \cdot g_t,$$
$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x^{nes}_t, y^{true})}{\|\nabla_x J(x^{nes}_t, y^{true})\|_1},$$
$$x^{adv}_{t+1} = \mathrm{Clip}^{\epsilon}_{x}\{x^{adv}_{t} + \alpha \cdot \mathrm{sign}(g_{t+1})\},$$

- Scale-Invariant attack method (SIM)
  - Base on the scale-invariant property of deep neural networks. We propose a Scale-Invariant attack Method (SIM), which optimizes the adversarial perturbations over the scale copies of the input image.



$$\arg\max_{x^{adv}} \frac{1}{m}\sum_{i=0}^{m} J(S_i(x^{adv}), y^{true}),$$
$$\text{s.t. } \|x^{adv} - x\|_\infty \le \epsilon,$$

- Our two methods, NI-FGSM and SIM, can be integrated with existing gradient-based attack methods, such as DIM, TIM and TI-DIM.

Experiments

- Attacking a single model

Table 1: **Attack success rates (%) of adversarial attacks against seven models under single-model setting.** The adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101 respectively. * indicates the white-box attacks.

(a) **Comparison of TIM and the SI-NI-TIM extension.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	TIM	**100.0***	47.8	42.8	39.5	24.0	21.4	12.9
	SI-NI-TIM (**Ours**)	**100.0***	**77.2**	**75.8**	**66.5**	**51.8**	**45.9**	**33.5**
Inc-v4	TIM	58.5	99.6*	47.5	43.2	25.7	23.3	17.3
	SI-NI-TIM (**Ours**)	**83.5**	**100.0***	**76.6**	**68.9**	**57.8**	**54.3**	**42.9**
IncRes-v2	TIM	62.0	56.2	97.5*	51.3	32.8	27.9	21.9
	SI-NI-TIM (**Ours**)	**86.4**	**83.2**	**99.5***	**77.2**	**66.1**	**60.2**	**57.1**
Res-101	TIM	59.0	53.6	51.8	99.3*	36.8	32.2	23.5
	SI-NI-TIM (**Ours**)	**78.3**	**74.1**	**73.0**	**99.8***	**58.9**	**53.9**	**43.1**

(b) **Comparison of DIM and the SI-NI-DIM extension.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DIM	98.7*	67.7	62.9	54.0	20.5	18.4	9.7
	SI-NI-DIM (**Ours**)	**99.6***	**84.7**	**81.7**	**75.4**	**36.9**	**34.6**	**20.2**
Inc-v4	DIM	70.7	98.0*	63.2	55.9	21.9	22.3	11.9
	SI-NI-DIM (**Ours**)	**89.7**	**99.3***	**84.5**	**78.5**	**47.6**	**45.0**	**28.9**
IncRes-v2	DIM	69.1	63.9	93.6*	57.4	29.4	24.0	17.3
	SI-NI-DIM (**Ours**)	**89.7**	**86.4**	**99.1***	**81.2**	**55.0**	**48.2**	**38.1**
Res-101	DIM	75.9	70.0	71.0	98.3*	36.0	32.4	19.3
	SI-NI-DIM (**Ours**)	**88.7**	**84.2**	**84.4**	**99.3***	**52.4**	**48.0**	**33.2**

(c) **Comparison of TI-DIM and the SI-NI-TI-DIM extension.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	TI-DIM	98.5*	66.1	63.0	56.1	38.6	34.9	22.5
	SI-NI-TI-DIM (**Ours**)	**99.6***	**85.5**	**80.9**	**75.7**	**61.5**	**56.9**	**40.7**
Inc-v4	TI-DIM	72.5	97.8*	63.4	54.5	38.1	35.2	25.3
	SI-NI-TI-DIM (**Ours**)	**88.1**	**99.3***	**83.7**	**77.0**	**65.0**	**63.1**	**49.4**
IncRes-v2	TI-DIM	73.2	67.5	92.4*	61.3	46.4	40.2	35.8
	SI-NI-TI-DIM (**Ours**)	**89.6**	**87.0**	**99.1***	**83.9**	**74.0**	**67.9**	**63.7**
Res-101	TI-DIM	74.9	69.8	70.5	98.7*	52.6	49.1	37.8
	SI-NI-TI-DIM (**Ours**)	**86.4**	**82.6**	**84.6**	**99.0***	**72.6**	**66.8**	**56.4**

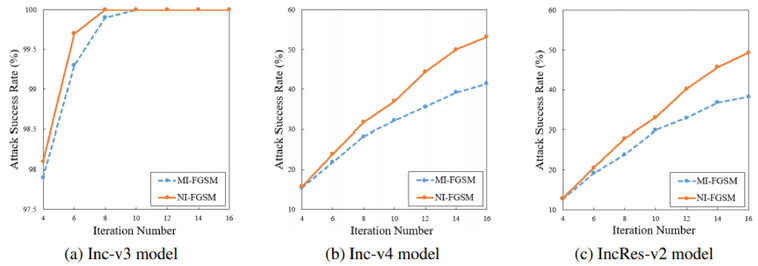
- Attacking an ensemble of models

Table 2: **Attack success rates (%) of adversarial attacks against seven models under multi-model setting.** * indicates the white-box models being attacked.

Attack	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
TIM	99.9	99.3	99.3	99.8	71.6	67.0	53.2
SI-NI-TIM (**Ours**)	**100.0**	**100.0**	**100.0**	**100.0**	**93.2**	**90.1**	**84.5**
DIM	99.7	99.2	98.9	98.9	66.4	60.9	41.6
SI-NI-DIM (**Ours**)	**100.0**	**100.0**	**100.0**	**99.9**	**88.2**	**85.1**	**69.7**
TI-DIM	99.6	98.8	98.8	98.9	85.2	80.2	73.3
SI-NI-TI-DIM (**Ours**)	**99.9**	**99.9**	**99.9**	**99.9**	**96.0**	**94.3**	**90.3**

Further Analysis

- NI-FGSM vs. MI-FGSM



(a) Inc-v3 model (b) Inc-v4 model (c) IncRes-v2 model

Figure 2: **Attack success rates (%) of NI-FGSM and MI-FGSM on various number of iterations.** The adversarial examples are crafted on Inc-v3 model against (a) Inc-v3 model, (b) Inc-v4 model and (c) IncRes-v2 model.

- Comparison with classic attacks

Table 4: **Attack success rates (%) of adversarial attacks against the models.** The adversarial examples are crafted on Inc-v3 using FGSM, I-FGSM, PGD, C&W, NI-FGSM, and SI-NI-FGSM. * indicates the white-box model being attacked.

Attack	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	**Average**
FGSM	67.1	26.7	25.0	24.4	10.5	10.0	4.5	24.0
I-FGSM	99.9	20.7	18.5	15.3	3.6	5.8	2.9	23.8
PGD	99.5	17.3	15.1	13.1	6.1	5.6	3.1	20.9
C&W	**100.0**	18.4	16.2	14.3	3.8	4.7	2.7	22.9
NI-FGSM (**Ours**)	**100.0**	52.6	51.4	41.0	12.9	12.8	6.4	39.6
SI-NI-FGSM (**Ours**)	**100.0**	**76.0**	**73.3**	**67.6**	**31.6**	**30.0**	**17.4**	**56.6**

Conclusion

- From an *optimization* perspective, we propose two new attack methods, namely Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM), to improve the transferability of adversarial examples.
- Combining our NI-FGSM and SIM with existing gradient-based attack methods can further boost the attack success rates of adversarial examples.
- Still a lot to be explored!

This work is supported by the Fundamental Research Funds for the Central Universities (2019kfyXKJC021) and Microsoft Research Asia