華中科技大学
HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Enhancing the Transferability of Adversarial Attacks through Variance Tuning

Xiaosen Wang, Kun He
Huazhong University of Science and Technology, Wuhan, China
Contact: xiaosen@hust.edu.cn
Homepage: https://xiaosen-wang.github.io/

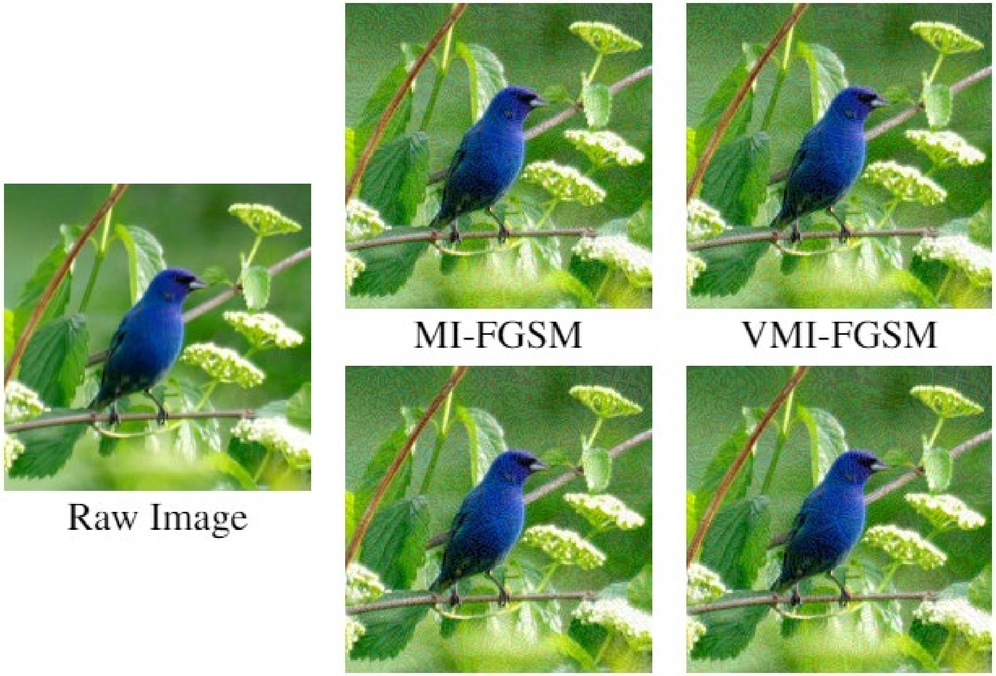
10/16/2021

Adversarial Example

Adversarial examples are **indistinguishable** from legitimate ones by adding small perturbations, but lead to **incorrect model prediction**.

Transferability: adversarial examples generated for one model can still fool other models, that enables **black-box attacks** in the real-world applications without any knowledge of target model.

Background: existing attacks (*e.g.* PGD, CW, etc.) have exhibited great effectiveness, but with **low transferability**.



Raw Image

MI-FGSM

VMI-FGSM

NI-FGSM

VNI-FGSM

Related works

Gradient-based adversarial attacks are widely used investigated:

- FGSM [Goodfellow et al., 2015]:

$$x^{adv} = x + \epsilon \cdot sign\big(\nabla_x J(x, y; \theta)\big)$$

- I-FGSM [Kurakin et al., 2016]:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(\nabla_{x_t^{adv}} J(x_t^{adv}, y; \theta)$$

- MI-FGSM [Dong et al., 2018]:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J\big(x_t^{adv}, y; \theta\big)}{||\nabla_{x_t^{adv}} J\big(x_t^{adv}, y; \theta\big)||_1}, x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$$

- NI-FGSM [Lin et al., 2020]: $\bar{x}_t^{adv} = x_t^{adv} + \alpha \cdot \mu \cdot g_t$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{\bar{x}_t^{adv}} J\big(\bar{x}_t^{adv}, y; \theta\big)}{||\nabla_{\bar{x}_t^{adv}} J\big(\bar{x}_t^{adv}, y; \theta\big)||_1}, x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$$

Multi-model attack and input transformation based attack are also shown to be effective to improve the transferability.

Motivation

Adversarial Example Generation		Standard Model Training
Input image x	⟺	Parameters
Transferability	⟺	Generalization

NI-FGSM finds that Nestorve Accelerated Gradient (NAG) that **accelerates the convergence** of optimization process, also **enhances the transferability**.

We treat the iterative gradient-based adversarial attack as **SGD optimization process,** in which at each iteration, the attacker always chooses the target model for update.

SGD introduces variance due to randomness.

Variance Tuning

Gradient Variance. Given a classifier f with parameters θ and loss function $J(x, y; \theta)$, an arbitrary image x and an upper bound ϵ' for the neighborhood, the gradient variance can be defined as:

$$V_{\epsilon'}^{g}(\mathrm{x}) = \mathbb{E}_{|x'-x|_p<\epsilon'} [\nabla_{x'} J(x', y; \theta)] - \nabla_x J(x, y; \theta)$$

In practice, we approximate the gradient variance by **sampling N examples in the neighborhood of $\boldsymbol{x}$**:

$$V(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x^i} J\left(x^i, y; \theta\right) - \nabla_x J(x, y; \theta)$$

where $x^i = x + U[-(\beta \cdot \epsilon)^d, (\beta \cdot \epsilon)^d]$.

At t-th iteration, we **tune the gradient of $\mathbf{x}_t^{\mathbf{adv}}$ with the gradient variance at (t-1)-th iteration** to stabilize the update direction.

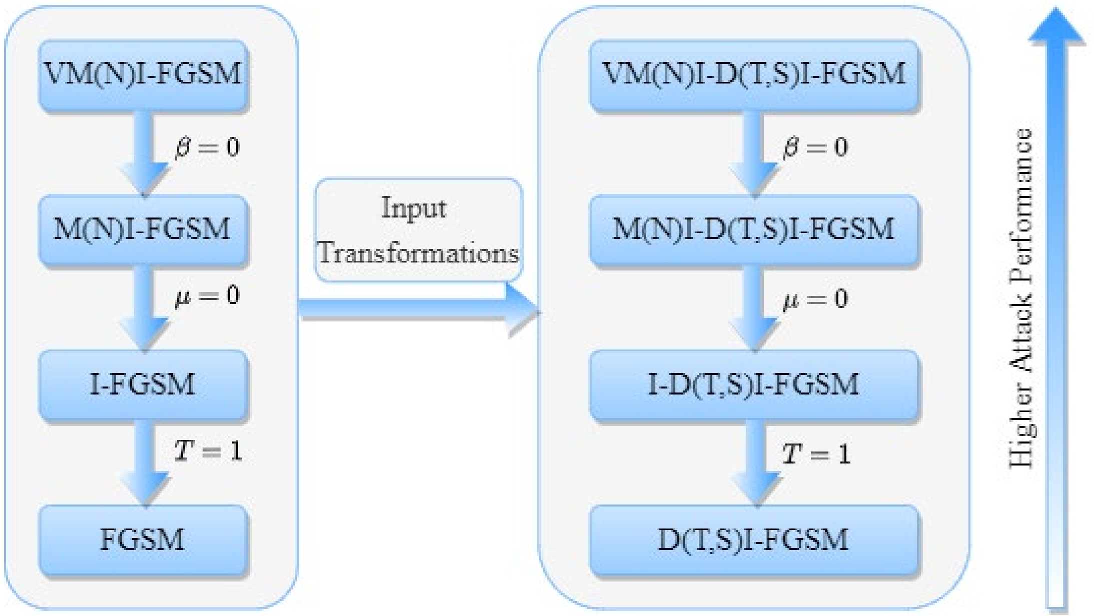
Variance Tuning

The variance tuning is generally applicable to all iterative gradient based attacks.

VMI-FGSM:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J\left(x_t^{adv}, y; \theta\right) + \boldsymbol{V(x_{t-1}^{adv})}}{||\nabla_{x_t^{adv}} J\left(x_t^{adv}, y; \theta\right) + \boldsymbol{V(x_{t-1}^{adv})}||_1},$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$$



Experimental Settings

- Dataset: 1,000 clean images from ILSVRC 2012 validation set
- Models: Inc-v3, Inc-v4, IncRes-v2, Res-v2-152
- Defense models:
  - Ensemble AT: Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, IncRes-v2$_{ens}$
  - NIPS 2017 top3 defense: HGD, R&P, NIPS-r3
  - Input transformation: JPEG, Bit-Red, FD, ComDefend
  - Certified defense: RS
  - Denoiser: NRP
- Baselines: MI-FGSM, NI-FGSM, DIM, TIM, SIM
- Attack setting: $\epsilon = 16$

Experimental Results

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	MI-FGSM	**100.0***	43.6	42.4	35.7	13.1	12.8	6.2
	VMI-FGSM	**100.0***	**71.7**	**68.1**	**60.2**	**32.8**	**31.2**	**17.5**
	NI-FGSM	**100.0***	51.7	50.3	41.3	13.5	13.2	6.0
	VNI-FGSM	**100.0***	**76.5**	**74.9**	**66.0**	**35.0**	**32.8**	**18.8**
Inc-v4	MI-FGSM	56.3	99.7*	46.6	41.0	16.3	14.8	7.5
	VMI-FGSM	**77.9**	**99.8***	**71.2**	**62.2**	**38.2**	**38.7**	**23.2**
	NI-FGSM	63.1	**100.0***	51.8	45.8	15.4	13.6	6.7
	VNI-FGSM	**83.4**	99.9*	**76.1**	**66.9**	**40.0**	**37.7**	**24.5**
IncRes-v2	MI-FGSM	60.7	51.1	**97.9***	46.8	21.2	16.0	11.9
	VMI-FGSM	**77.9**	**72.1**	**97.9***	**67.7**	**46.4**	**40.8**	**34.4**
	NI-FGSM	62.8	54.7	**99.1***	46.0	20.0	15.1	9.6
	VNI-FGSM	**80.8**	**76.9**	98.5*	**69.8**	**47.9**	**40.3**	**34.2**
Res-101	MI-FGSM	58.1	51.6	50.5	**99.3***	23.9	21.5	12.7
	VMI-FGSM	**75.1**	**68.9**	**70.5**	99.2*	**45.2**	**41.4**	**30.1**
	NI-FGSM	65.6	58.3	57.0	99.4*	24.5	21.4	11.7
	VNI-FGSM	**79.8**	**74.6**	**73.2**	**99.7***	**46.1**	**42.5**	**32.1**

Table 1: The success rates (%) on seven models in the single model setting by various gradient-based iterative attacks. The adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101 respectively. * indicates the white-box model.

Experimental Results

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	MI-CT-FGSM	98.7*	85.4	80.6	76.0	64.1	62.1	45.2
	VMI-CT-FGSM	**99.3***	**88.6**	**86.7**	**82.9**	**78.6**	**76.2**	**64.7**
	NI-CT-FGSM	98.9*	84.1	80.0	74.5	60.0	56.2	41.0
	VNI-CT-FGSM	**99.5***	**91.2**	**89.0**	**85.3**	**78.6**	**76.7**	**65.3**
Inc-v4	MI-CT-FGSM	87.2	98.6*	83.3	78.3	72.2	67.2	57.3
	VMI-CT-FGSM	**90.0**	**98.8***	**86.6**	**81.9**	**78.3**	**76.6**	**68.3**
	NI-CT-FGSM	87.8	**99.4***	82.5	75.9	65.8	62.6	51.3
	VNI-CT-FGSM	**92.1**	99.2*	**89.2**	**85.1**	**80.1**	**78.3**	**70.4**
IncRes-v2	MI-CT-FGSM	87.9	85.7	**97.1***	83.0	77.6	74.6	72.0
	VMI-CT-FGSM	**88.9**	**87.0**	97.0*	**85.0**	**83.4**	**80.5**	**79.4**
	NI-CT-FGSM	90.2	87.0	**99.4***	83.2	75.0	68.9	65.1
	VNI-CT-FGSM	**92.9**	**90.6**	99.0*	**88.2**	**85.2**	**82.5**	**81.8**
Res-101	MI-CT-FGSM	86.5	81.8	83.2	**98.9***	77.0	72.3	61.9
	VMI-CT-FGSM	**86.9**	**84.2**	**86.4**	98.6*	**81.0**	**78.6**	**71.6**
	NI-CT-FGSM	86.1	82.2	83.3	98.5*	70.0	68.5	54.6
	VNI-CT-FGSM	**90.7**	**85.5**	**87.2**	**99.1***	**82.6**	**79.7**	**73.3**

Table 2: The success rates (%) on seven models in the single model setting by various gradient-based iterative attacks enhanced by CTM. * indicates the white-box model.

Experimental Results

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MI-FGSM	**99.9***	98.2*	95.3*	**99.9***	39.4	35.3	24.2
VMI-FGSM	99.7*	**98.5***	**96.0***	**99.9***	**67.6**	**62.9**	**50.7**
NI-FGSM	99.8*	**99.8***	**98.9***	99.8*	41.0	33.5	23.1
VNI-FGSM	**99.9***	99.6*	98.6*	**99.9***	**71.3**	**66.0**	**52.9**
MI-CT-FGSM	99.6*	99.1*	97.4*	99.7*	91.3	89.6	86.8
VMI-CT-FGSM	**99.7***	**99.2***	**98.4***	**99.9***	**93.6**	**92.4**	**91.0**
NI-CT-FGSM	**100.0***	**100.0***	**100.0***	**100.0***	92.8	89.6	83.6
VNI-CT-FGSM	**100.0***	99.9*	99.6*	**100.0***	**95.5**	**94.5**	**92.3**

Table 3: The success rates (%) on seven models in the multi-model setting by various gradient-based iterative attacks. The adversarial examples are generated on the ensemble models, *i.e*. Inc-v3, Inc-v4, IncRes-v2 and Res-101.

Model	Attack	HGD	R&P	NIPS-r3	Bit-Red	JPEG	FD	ComDefend	RS	NRP	Average
Inc-v3	MI-CT-FGSM	56.6	44.9	52.5	36.2	77.3	60.0	80.1	40.3	29.3	53.0
	VMI-CT-FGSM	**73.1**	**65.1**	**70.3**	**49.5**	**85.4**	**72.4**	**86.0**	**51.9**	**45.2**	**66.5**
	NI-CT-FGSM	50.4	39.4	47.4	34.3	76.0	58.6	77.7	36.9	24.8	49.5
	VNI-CT-FGSM	**73.4**	**64.5**	**70.6**	**51.2**	**86.8**	**73.5**	**87.3**	**52.1**	**43.9**	**67.0**
Ens	MI-CT-FGSM	91.0	87.7	89.0	75.9	94.2	88.8	95.1	68.1	76.1	85.1
	VMI-CT-FGSM	**92.9**	**91.0**	**92.3**	**80.9**	**95.4**	**91.0**	**96.2**	**77.0**	**83.2**	**88.9**
	NI-CT-FGSM	91.3	85.6	89.0	72.3	95.9	89.5	95.4	63.2	69.5	83.5
	VNI-CT-FGSM	**94.7**	**92.4**	**93.4**	**82.3**	**97.1**	**92.6**	**97.4**	**77.4**	**84.0**	**90.1**

Table 4: The success rates (%) on nine models with advanced defense mechanism by various gradient-based iterative attacks enhanced by CTM. The adversarial examples are generated on Inc-v3 model and the ensemble of models respectively.

Summary

- Propose the definition of **gradient variance.**
- Introduce a broad class of iterative gradient based attacks with **variance tuning.**
- Achieve **SOTA** attack transferability on ImageNet against various models with **defenses** in different scenario.

Thanks!

Xiaosen Wang, Kun He
Huazhong University of Science and Technology, Wuhan, China
Contact: xiaosen@hust.edu.cn
Homepage: https://xiaosen-wang.github.io/