Enhancing the Transferability of Adversarial Attacks by Variance Tuning

Xiaosen Wang and Kun He
School of Computer Science and Technology, Huazhong University of Science and Technology

CVPR VIRTUAL JUNE 19-25

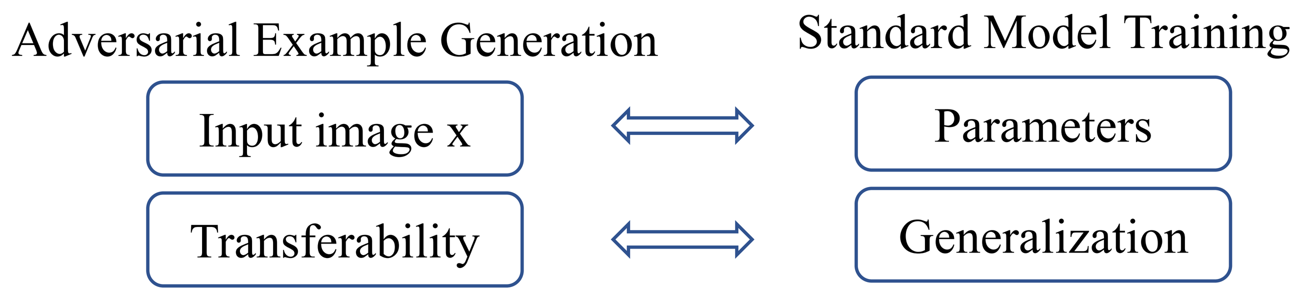
Introduction

Adversarial Examples are **imperceptible** from legitimate ones by adding tiny 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**.

Methodology



We treat the iterative gradient-based adversarial attack as a stochastic gradient decent (SGD) optimization process, in which at each iteration, the attacker always chooses the target model for update. **SGD introduces variance due to randomness**.

Definition 1 ***Gradient Variance.*** *Given a classifier f with parameters θ and loss function $J(x,y;\theta)$, an arbitrary image $x \in \mathcal{X}$ and an upper bound ϵ' for the neighborhood, the gradient variance can be defined as:*

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

In practice, however, due to the continuity of the input space, we cannot calculate $\mathbb{E}_{\|x'-x\|_p<\epsilon'}[\nabla_{x'}J(x',y;\theta)]$ directly. Therefore, we **approximate** its value by sampling N examples in the neighborhood of x to calculate $V(x)$:

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

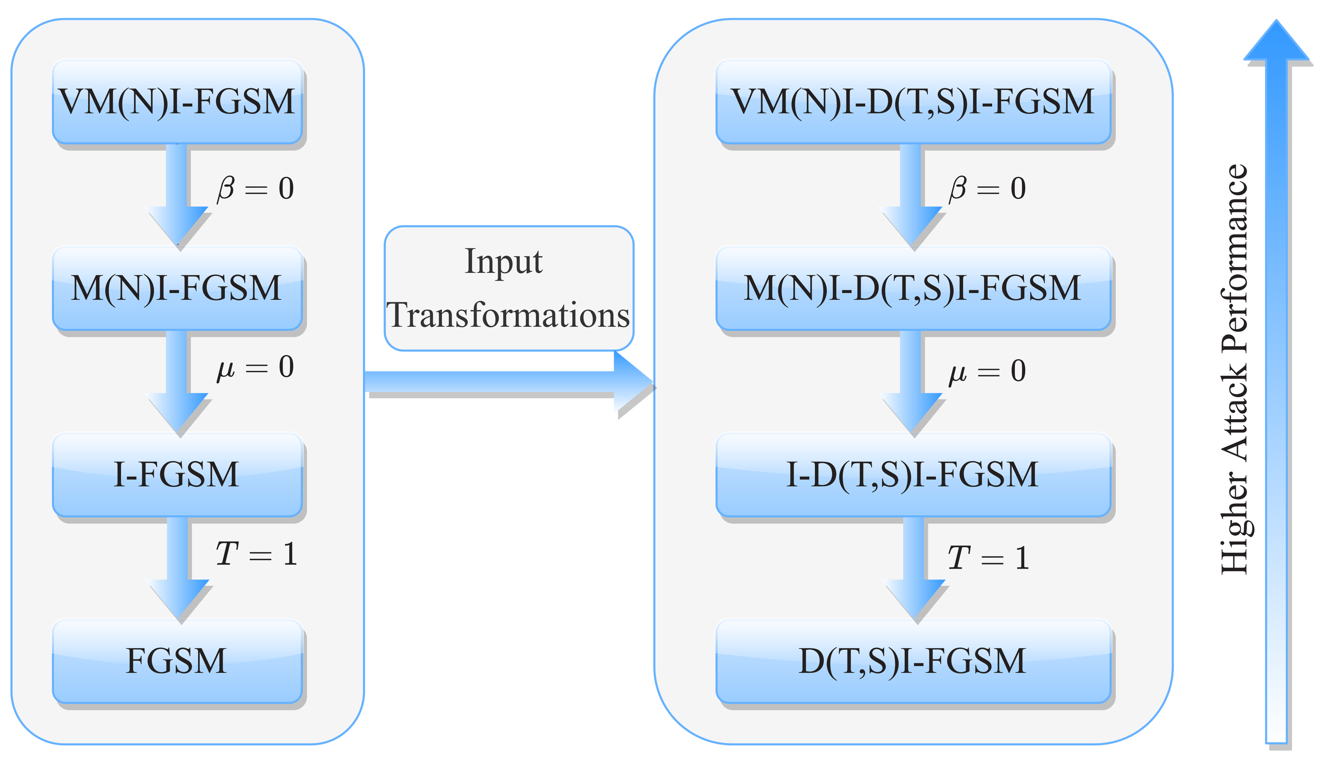


Figure 1: Relationship between various attacks.

Algorithm

Algorithm 1 VMI-FGSM

Input: A classifier f with parameters θ, loss function J. A raw example x with ground-truth label y. The magnitude of perturbation ϵ; number of iteration T and decay factor μ. The factor β for the upper bound of neighborhood and number of example N for variance tuning.

Output: An adversarial example x^{adv}

1: $\alpha = \epsilon/T$
2: $g_0 = 0$; $v_0 = 0$; $x_0^{adv} = x$
3: **for** $t = 0 \to T-1$ **do**
4: Calculate the gradient $\hat{g}_{t+1} = \nabla_{x_t^{adv}} J(x_t^{adv}, y; \theta)$
5: Update g_{t+1} by variance tuning based momentum

$$g_{t+1} = \mu \cdot g_t + \frac{\hat{g}_{t+1} + v_t}{\|\hat{g}_{t+1} + v_t\|_1} \tag{2}$$

6: Update $v_{t+1} = V(x_t^{adv})$ by Eq. (1)
7: Update x_{t+1}^{adv} by applying the sign of gradient

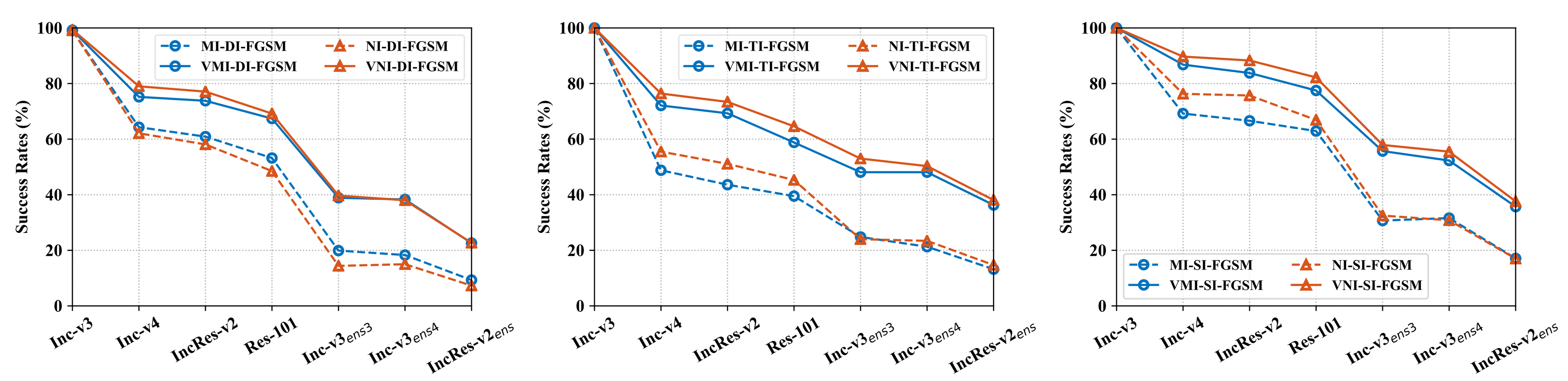
$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1}) \tag{3}$$

8: **end for**
9: **return** $x^{adv} = x_T^{adv}$

Experiments

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: Evaluations on gradient-based attacks.



(a) DIM (b) TIM (c) SIM

Figure 2: Evaluations with input transformations, *i.e.* DIM, TIM, SIM.

Experiments

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**

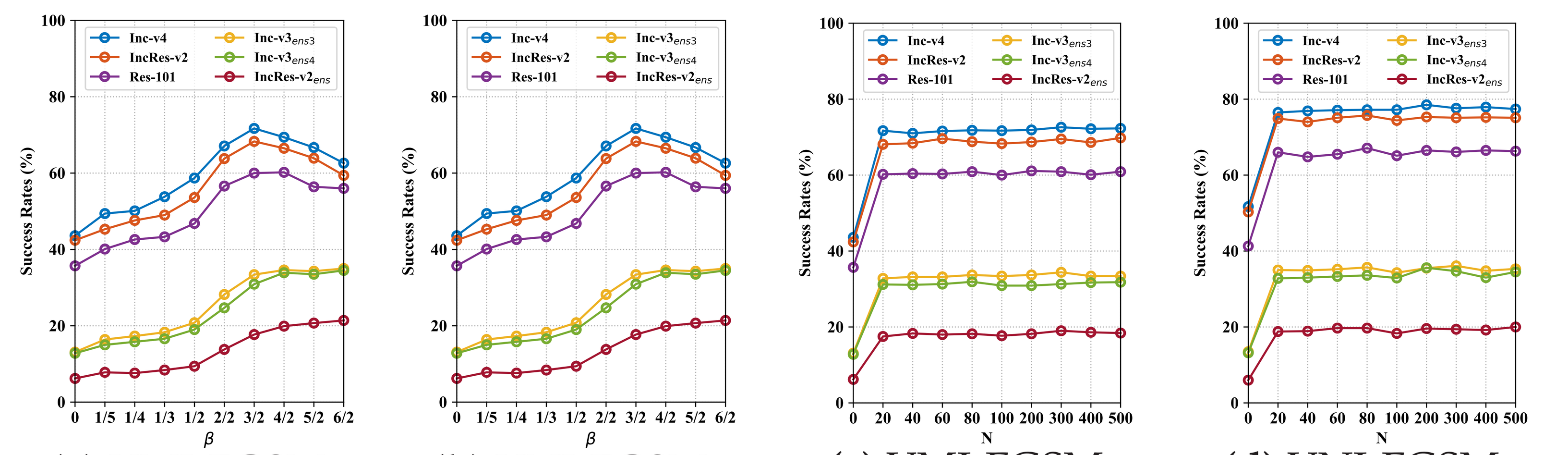
Table 2: Evaluations with the combination of DIM, TIM and SIM.

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: Evaluations in multi-model setting.

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	43.1	54.6
	VMI-CT-FGSM	**73.1**	**65.1**	**70.3**	**49.5**	**85.4**	**72.4**	**86.0**	**51.9**	**60.2**	**68.2**
	NI-CT-FGSM	50.4	39.4	47.4	34.3	76.0	58.6	77.7	36.9	39.4	51.1
	VNI-CT-FGSM	**73.4**	**64.5**	**70.6**	**51.2**	**86.8**	**73.5**	**87.3**	**52.1**	**58.9**	**68.7**
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: Evaltions on Defense models.



(a) VMI-FGSM (b) VNI-FGSM (c) VMI-FGSM (d) VNI-FGSM

Figure 3: Ablation studies.

Conclusion

- We propose a **variance tuning** method to enhance the transferability of the iterative gradient-based attacks.
- Our method is **geneally applicable** to any iterative gradient-based attacks and input transformations (*i.e.* DIM, TIM, SIM *etc.*).
- Experiments show our method could significantly **enhance the transferability** of various attacks.
- The results indicate the **insufficiency** of existing defenses and can serve as a benchamrk to evalute the robustness of future developed defense.