Admix: Enhancing the Transferability of Adversarial Attacks

Xiaosen Wang[1], Xuanran He[2], Jingdong Wang[3], Kun He[1]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology

[2]Wee Kim Wee School of Communication and Information, Nanyang Technological University [3]Microsoft Research Asia

xiaosen@hust.edu.cn, xhe015@e.ntu.edu.sg, jingdw@microsoft.com, brooklet60@hust.edu.cn

2021 ICCV OCTOBER 11-17 VIRTUAL

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

Existing input transformations are effective to improve the transferability, but they are all applied on **single input image**. Could we further improve the transferability by **incorporating the information from other categories**?

Mixup improves the model generalization by interpolating two randomly sampled samples (x, y) and (x', y') with $\lambda \in [0, 1]$ as follows:

$$\tilde{x} = \lambda \cdot x + (1 - \lambda) \cdot x', \quad \tilde{y} = \lambda \cdot y + (1 - \lambda) \cdot y'. \tag{1}$$

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MI-FGSM	100.0	43.6	42.4	35.7	13.1	12.8	6.2
Mixup	71.8	44.2	41.1	39.0	13.5	13.4	7.2

Table 1: Evaluations on MI-FGSM and *Mixup* transformation.

However, directly applying *mixup* for gradient calculation improves the transferability of crafted adversaries slightly but **degrades the attack performance significantly under white-box setting**.

To utilize the information of images from other category without harming the white-box attack performance, we propose *admix* operation that **admixes two images in a master and slave manner**.

$$\tilde{x} = \gamma \cdot x + \eta' \cdot x' = \gamma \cdot (x + \eta \cdot x'). \tag{2}$$

We further propose an *Admix* attack method to improve the attack transferability, which **calculates the average gradient on a set of admixed images** $\{\tilde{x}\}$ of the input x by changing the value of γ or picking the add-in image x' from different categories in Eq. (2).

$$\bar{g}_{t+1} = \frac{1}{m_1 \cdot m_2} \sum_{x' \in X'} \sum_{i=0}^{m_1 - 1} \nabla_{x_t^{adv}} J(\gamma_i \cdot (x_t^{adv} + \eta \cdot x'), y; \theta), \tag{3}$$

Both *admix* and *mixup* generate a mixed image from an image pair, x and x'. Here we summarize their differences as follows:

- **Different goal**: *Mixup* aims to improve the model generalization while *admix* aims to generate more transferable adversaries.
- **Different Strategy**: *Mixup* treats x and x' equally and also mixes the label while *admix* treats x as the primary component and combines a small portion of x', and maintains the label of x.
- **Different interpolated image**: *Mixup* linearly interpolates x and x' while *admix* does not have such constraint, leading to more diversed transformed images.

Algorithm

Algorithm 1 The *Admix* Attack Algorithm

Input: A classifier f with loss function J and a benign example x with ground-truth label y

Input: The maximum perturbation ϵ, number of iterations T and decay factor μ

Input: The number of admixed copies m_1 and sampled images m_2, and the strength of sampled image η

Output: An adversarial example $x^{adv} \in \mathcal{B}_\epsilon(x)$

1: $\alpha = \epsilon/T$; $g_0 = 0$; $\bar{g}_0 = 0$; $x_0^{adv} = x$

2: **for** $t = 0 \to T - 1$ **do**:

3: Randomly sample a set X' of m_2 images from another category

4: Calculate the average gradient $\bar{g}_{t+1}$ by Eq. (3)

5: Update the enhanced momentum g_t:

$$g_{t+1} = \mu \cdot g_t + \frac{\bar{g}_{t+1}}{\|\bar{g}_{t+1}\|_1}$$

6: Update x_{t+1}^{adv} by applying the gradient sign:

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

7: **end for**

8: **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	DIM	99.0*	64.3	60.9	53.2	19.9	18.3	9.3
	TIM	**100.0***	48.8	43.6	39.5	24.8	21.3	13.2
	SIM	**100.0***	69.4	67.3	62.7	32.5	30.7	17.3
	Admix	**100.0***	**82.6**	**80.9**	**75.2**	**39.0**	**39.2**	**19.2**
Inc-v4	DIM	72.9	97.4*	65.1	56.5	20.2	21.1	11.6
	TIM	58.6	**99.6***	46.5	42.3	26.2	23.4	17.2
	SIM	80.6	**99.6***	74.2	68.8	47.8	44.8	29.1
	Admix	**87.8**	99.4*	**83.2**	**78.0**	**55.9**	**50.4**	**33.7**
IncRes-v2	DIM	70.1	63.4	93.5*	58.7	30.9	23.9	17.7
	TIM	62.2	55.4	97.4*	50.5	32.8	27.6	23.3
	SIM	84.7	81.1	99.0*	76.4	56.3	48.3	42.8
	Admix	**89.9**	**87.5**	**99.1***	**81.9**	**64.2**	**56.7**	**50.0**
Res-101	DIM	75.8	69.5	70.0	98.0*	35.7	31.6	19.9
	TIM	59.3	52.1	51.8	99.3*	35.4	31.3	23.1
	SIM	75.2	68.9	69.0	**99.7***	43.7	38.5	26.3
	Admix	**85.4**	**80.8**	**79.6**	**99.7***	**51.0**	**45.3**	**30.9**

Table 2: Evaluations on various single input transformation based attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	SI-DIM	98.9*	85.0	81.3	76.3	48.0	45.1	24.9
	Admix-DIM	**99.8***	**90.5**	**87.7**	**83.5**	**52.2**	**49.9**	**28.6**
Inc-v4	SI-DIM	89.3	98.8*	85.6	79.9	58.4	55.2	39.3
	Admix-DIM	**93.0**	**99.2***	**89.7**	**85.2**	**62.4**	**60.3**	**39.7**
IncRes-v2	SI-DIM	87.9	85.1	97.5*	82.9	66.0	59.3	52.2
	Admix-DIM	**90.2**	**88.4**	**98.0***	**85.8**	**70.5**	**63.7**	**55.3**
Res-101	SI-DIM	87.9	83.4	84.0	98.6*	63.5	57.5	42.0
	Admix-DIM	**91.9**	**89.0**	**89.6**	**99.8***	**69.7**	**62.3**	**46.6**

Table 3: Evaluations on the attacks integrated with **DIM**.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	SI-TIM	**100.0***	71.8	68.6	62.2	48.2	47.4	31.3
	Admix-TIM	**100.0***	**83.9**	**80.4**	**74.4**	**59.1**	**57.9**	**39.2**
Inc-v4	SI-TIM	78.2	99.6*	71.9	66.1	58.6	55.4	45.1
	Admix-TIM	**87.4**	**99.7***	**82.3**	**77.0**	**68.1**	**65.3**	**53.1**
IncRes-v2	SI-TIM	84.5	82.2	**98.8***	77.4	71.6	64.7	61.0
	Admix-TIM	**90.2**	**88.2**	98.6*	**83.9**	**78.4**	**73.6**	**70.0**
Res-101	SI-TIM	74.2	69.9	70.2	**99.8***	59.5	54.5	42.8
	Admix-TIM	**83.2**	**78.9**	**80.7**	99.7*	**67.0**	**62.5**	**52.8**

Table 4: Evaluations on the attacks integrated with **TIM**.

Experiments

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	SI-TI-DIM	99.1*	83.6	80.8	76.7	65.2	63.3	46.5
	Admix-TI-DIM	**99.9***	**89.0**	**87.0**	**83.1**	**72.2**	**71.1**	**52.4**
Inc-v4	SI-TI-DIM	87.9	98.7*	83.0	77.7	72.4	68.2	57.5
	Admix-TI-DIM	**90.4**	**99.0***	**87.3**	**82.0**	**75.3**	**71.9**	**61.6**
IncRes-v2	SI-TI-DIM	88.8	86.8	**97.8***	83.9	78.7	74.2	72.3
	Admix-TI-DIM	**90.1**	**89.6**	97.7*	**85.9**	**82.0**	**78.0**	**76.3**
Res-101	SI-TI-DIM	84.7	82.2	84.8	99.0*	75.8	73.5	63.4
	Admix-TI-DIM	**91.0**	**87.7**	**89.2**	**99.9***	**81.1**	**77.4**	**70.1**

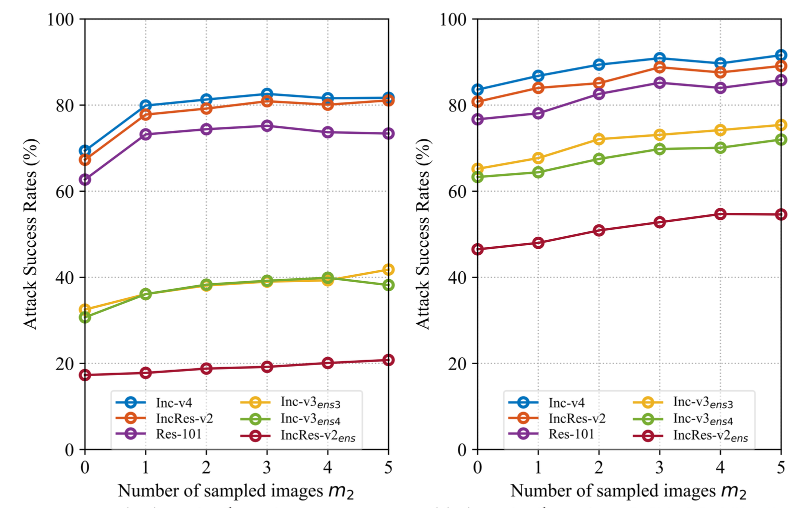
Table 5: Evaluations on the attacks integrated with **TI-DIM**.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
DIM	99.4*	97.4*	94.9*	99.8*	58.1	51.1	34.9
TIM	99.8*	97.9*	95.2*	99.8*	62.2	56.8	48.0
SIM	99.9*	99.3*	98.3*	**100.0***	78.8	73.9	59.5
Admix	**100.0***	**99.6***	**99.0***	**100.0***	**85.5**	**80.9**	**67.8**
SI-DIM	**99.7***	98.9*	97.7*	99.9*	85.2	83.3	71.3
Admix-DIM	**99.7***	**99.5***	**98.9***	**100.0***	**89.3**	**87.8**	**79.0**
SI-TIM	**99.7***	99.0*	97.6*	**100.0***	87.9	85.2	80.4
Admix-TIM	**99.7***	**99.1***	**98.1***	**100.0***	**91.8**	**89.7**	**85.8**
SI-TI-DIM	99.6*	**98.9***	97.8*	99.7*	91.1	90.3	86.8
Admix-TI-DIM	**99.7***	**98.9***	**98.3***	**100.0***	**93.9**	**92.3**	**90.0**

Table 6: Evaluations in ensemble model setting.

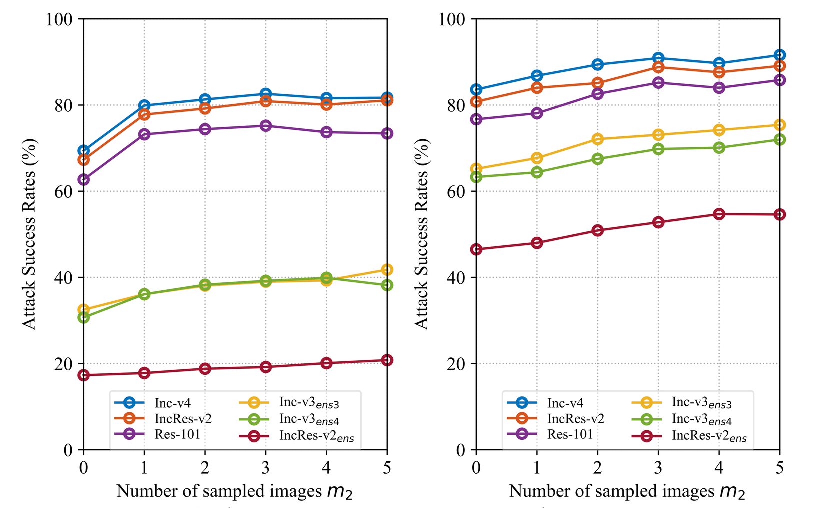
Attack	HGD	R&P	NIPS-r3	Bit-Red	FD	JPEG	RS	ARS	NRP	Average
SI-TI-DIM	91.4	88.0	90.0	75.7	88.0	93.2	69.2	46.4	77.1	79.9
Admix-TI-DIM	**93.7**	**90.3**	**92.4**	**80.1**	**91.9**	**95.4**	**74.9**	**51.4**	**80.7**	**83.3**

Table 7: Evaluations on Defense models.



(a) *Admix* **(b)** *Admix*-TI-DIM

Figure 1: Evaluations for various number of sampled image, m_2.



(a) *Admix* **(b)** *Admix*-TI-DIM

Figure 2: Evaluations for various strength of sampled image, η.

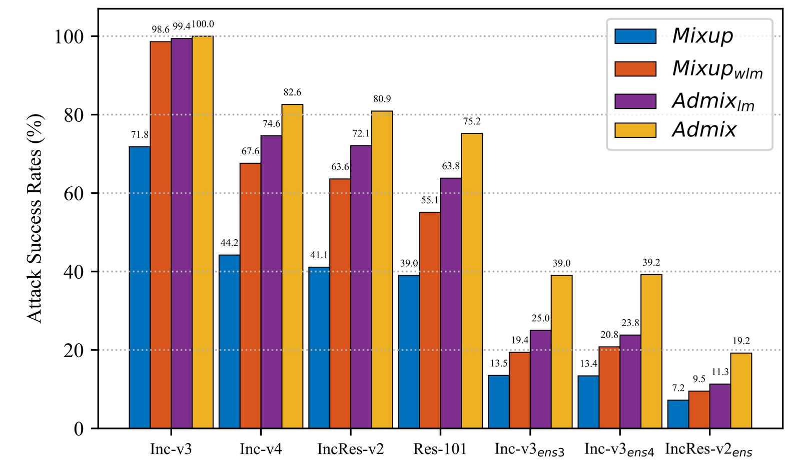


Figure 3: Evaluations on *Mixup*, *Mixup*$_{wlm}$, *Admix*$_{lm}$ and *Admix*

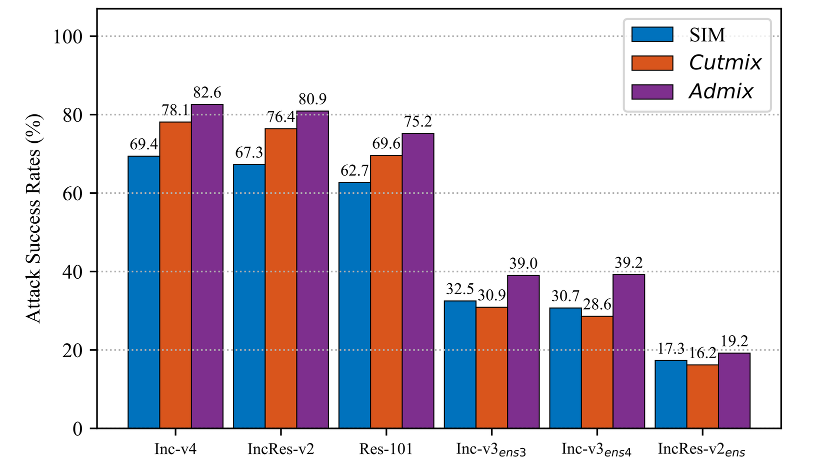


Figure 4: Evaluations on SIM, *Cutmix* and *Admix*.

Conclusion

- We propose a novel input transformation based attack, ***Admix***, utilizing the information of images from other category to enhance the transferability.
- Our method is **geneally applicable** to other input transformations (*i.e.* DIM, TIM, SIM *etc.*) and gradient-based attacks.
- Experiments show our method could significantly **enhance the transferability** of various attacks.