Adaptive Image Transformations for Transfer-based Adversarial Attack

Zheng Yuan[1,2], Jie Zhang[1,2,3], Shiguang Shan[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Institute of Intelligent Computing Technology, Suzhou, CAS
zheng.yuan@vipl.ict.ac.cn; {zhangjie, sgshan}ict.ac.cn

Project Code

ECCV TEL AVIV 2022

1. Task: Transfer-based Black-box Attack

◆ Adversarial Attack

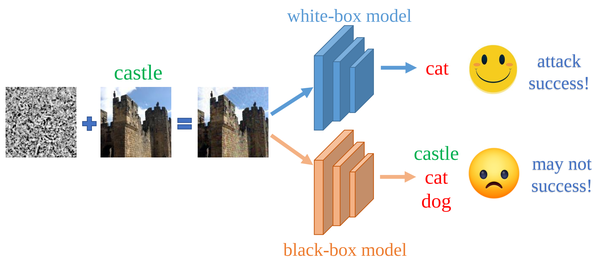
➢ Notation

clean image x, corresponding label y,
well-trained classifier model f,
adversarial perturbation budget ϵ

➢ The objective of adversarial attack

$$f(x^{adv}) \neq y, \qquad s.t. \|x - x^{adv}\|_\infty \leq \epsilon$$

◆Transfer-based black-box Attack



◆ Input-transformation-based method

$$x_0^{adv} = x, \qquad g_0 = 0$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J(f\left(T(x_t^{adv})\right), y)}{\left\|\nabla_{x_t^{adv}} J(f\left(T(x_t^{adv})\right), y)\right\|_1}$$

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

T	**Existing Work** fixed image transformation **Our AITL** adaptive transformation combination

2. Motivation

◆ Some works have studied the effect of image transformation operation on adversarial example transferability

◆ A fixed transformation is applied to all images without considering the characteristics of the different images

◆ The mutual influence of different image transform combinations is not considered

◆ Comparison of different methods

Method	Transformation	Method	Transformation
DIM	Resize	CIM	Crop
TIM	Translate	Admix	Mixup
SIM	Scale	AITL (ours)	Adaptive

3. Method

◆ Adaptive Image Transformation Learner

➢ incorporate different image transformation operations into a unified framework

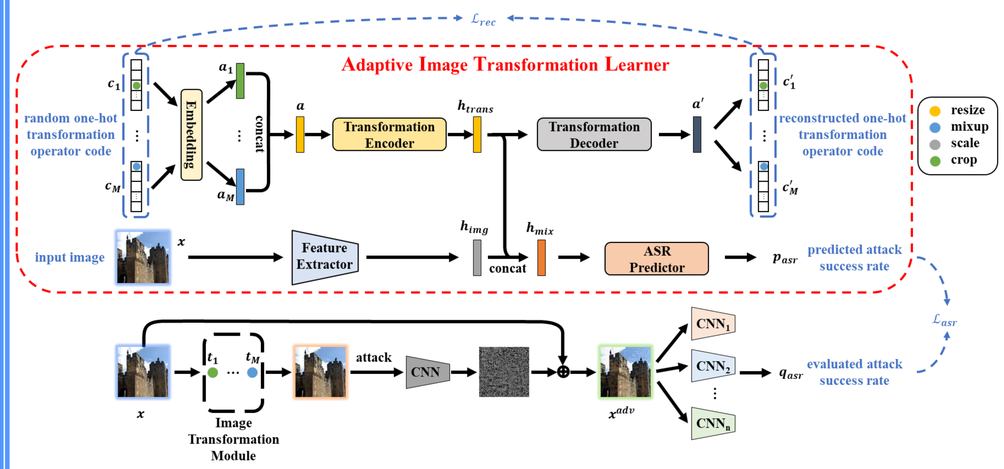
➢ adaptively select the suitable input transformations towards different input images

◆ Model Structure

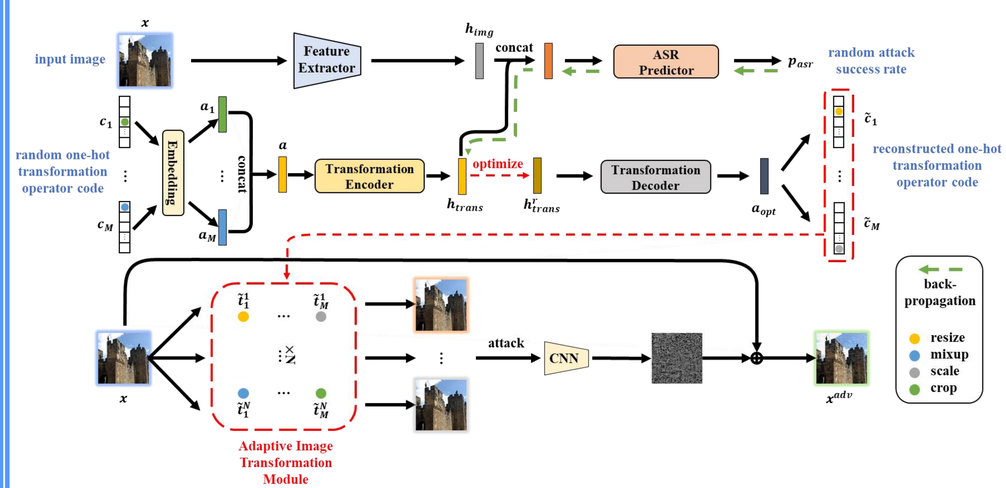
➢ Encoder & Decoder: convert the discretized image transformation operations into continuous feature embedding

➢ ASR Predictor: predict the attack success rate evaluated on black-box models when incorporating the given image transformations into MIFGSM

◆ Train



◆ Test (Attack Process)



4. Experiment

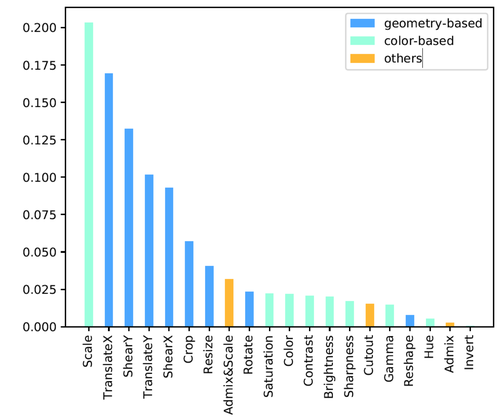
◆ The success rates under **single model** attack setting on ImageNet

	normally trained models							defense models					
	Incv3*	Incv4	IncResv2	Resv2-101	Resv2-152	PNASNet	NASNet	$Incv3_{ens3}$	$Incv3_{ens4}$	$IncResv2_{ens}$	HGD	R&P	NIPS-r3
MIFGSM	**100**	52.2	50.6	37.4	35.6	42.2	42.2	15.6	15.2	6.4	5.8	5.6	9.3
DIM	99.7	78.3	76.3	59.6	59.9	64.6	66.2	31.0	29.2	13.4	15.8	14.8	24.6
SIM	**100**	84.5	81.3	68.0	65.3	70.8	73.6	37.5	35.0	18.8	16.8	18.3	26.8
CIM	**100**	85.1	81.6	58.1	57.4	65.7	66.7	33.3	30.0	15.9	20.4	16.4	25.7
Admix	99.8	69.5	66.5	55.3	55.4	60.0	62.7	27.5	27.0	14.3	11.6	12.6	19.8
ADSCM	**100**	87.9	86.1	75.8	76.0	80.9	82.2	49.3	46.9	27.0	33.1	28.5	40.5
Random	**100**	94.0	92.0	79.7	80.0	84.6	85.5	49.8	46.7	24.5	29.2	26.4	42.2
AutoMA	98.2	91.2	91.0	82.5	-	-	-	49.2	49.0	29.1	-	-	-
AITL (ours)	99.8	**95.8**	**94.1**	**88.8**	**90.1**	**94.1**	**94.0**	**69.9**	**65.8**	**43.4**	**50.4**	**46.9**	**59.9**

◆ The success rates under **multiple models** attack setting on ImageNet

	normally trained models							defense models					
	Incv3*	Incv4*	IncResv2*	Resv2-101*	Resv2-152	PNASNet	NASNet	$Incv3_{ens3}$	$Incv3_{ens4}$	$IncResv2_{ens}$	HGD	R&P	NIPS-r3
MIFGSM	**100**	99.6	99.7	**98.5**	86.8	79.4	81.2	52.4	47.5	30.1	39.2	31.7	43.6
DIM	99.5	99.4	98.9	96.9	92.0	91.3	92.1	77.4	73.1	54.4	68.4	61.2	73.5
SIM	99.9	99.1	98.3	93.2	91.7	90.9	91.9	78.8	74.4	59.8	66.9	59.0	70.7
CIM	99.8	99.3	97.8	90.6	88.5	88.2	90.9	75.1	69.7	54.3	68.5	59.1	70.7
Admix	99.9	99.5	98.2	95.4	89.3	88.1	90.0	67.7	61.9	44.8	51.0	44.8	57.9
ADSCM	99.8	99.3	99.2	96.9	96.0	88.1	99.0	85.8	82.9	69.2	78.7	74.1	81.1
Random	**100**	99.4	98.9	96.9	94.3	94.4	95.0	83.7	80.2	64.8	73.7	67.3	77.9
AITL (ours)	99.9	**99.7**	**99.9**	97.3	**96.6**	**97.7**	**97.8**	**89.3**	**89.0**	**79.0**	**85.5**	**82.3**	**86.3**

◆ The frequency of various image transformations used in AITL



◆ Visualization of generated adversarial examples