Adaptive Gradient-based Adversarial Attacks on Deep Neural Networks

Chi-Man Pun

Outline

1. Background
2. Adversarial Attacks on Deep Neural Networks
3. Adaptive Gradient-based Perturbations Generation
4. Experimental Results
5. Conclusion

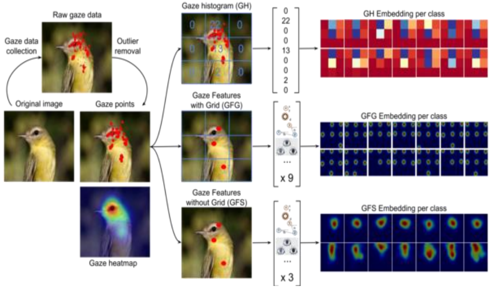
1

Background

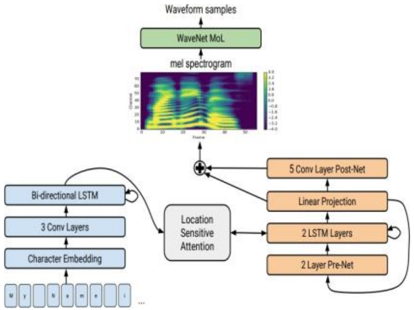
- Introduction
- Applications
- Challenges

Introduction

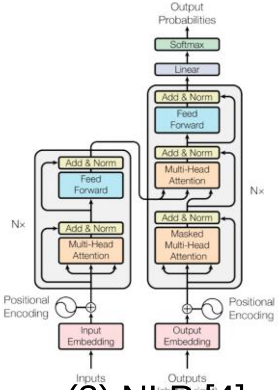
- Deep Neural Models [1] have led to a dramatic improvement on image, audio and natural language processing (NLP) tasks in recent years.



(1) Image Classification[2]



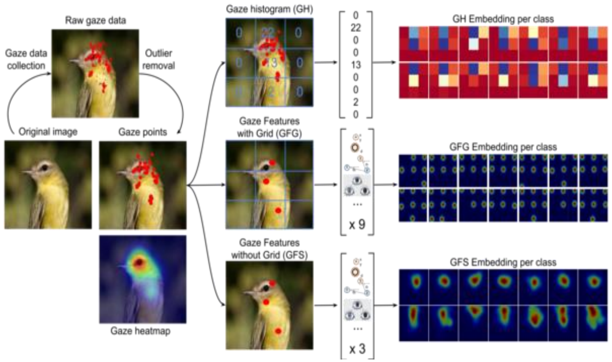
(2) Text2Audio[3]



(3) NLP [4]

Applications

- Computer Vision
  - Deep Neural Models perform well in computer version, Such as image classification, object detection, et al.

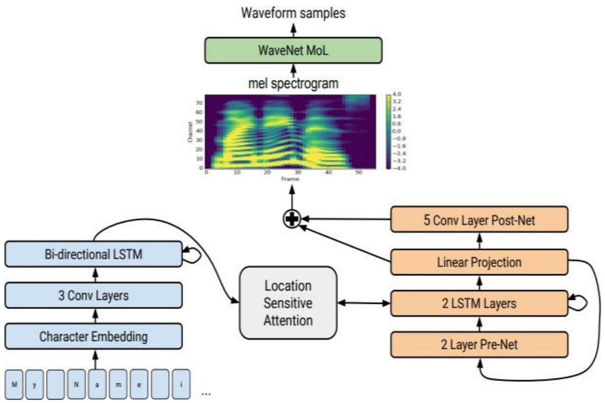


Advantages: Higher classification accuracy
Faster processing speed
et al.

Disadvantages: Time consuming on training
Vast training dataset
Vulnerable to perturbations

Applications

- Audio2Text/Text2Audio
  - Deep Neural Models can translate audio into text or text to audio.

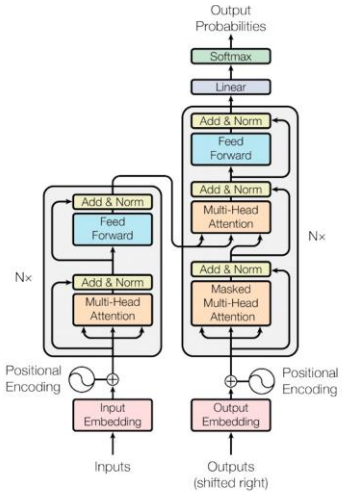


Advantages: Faster translation efficiency

Disadvantages: Vast audio and parallel text dataset
Time consuming on training
Vulnerable to perturbations

Applications

- NLP
  - One Corpus could be translated by DNNs into other corpus.

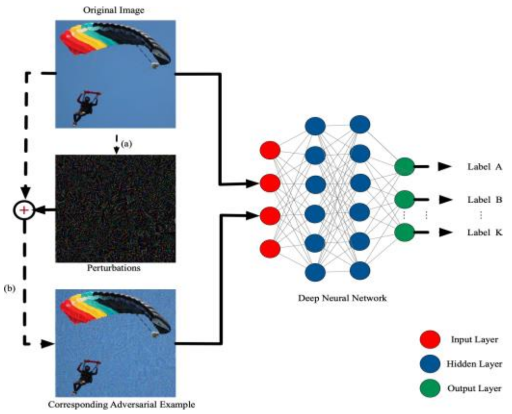


Advantages: Improve the translation efficiency
Higher translation accuracy

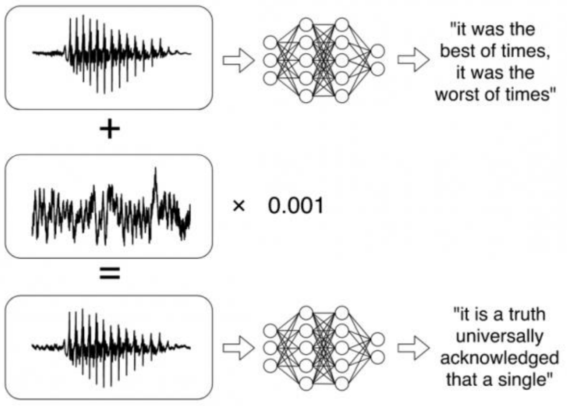
Disadvantages: Vast parallel corpus
Vulnerable to perturbations

Challenges

Shortage of DNNs: Vulnerable to Crafted Adversarial Perturbations



4: Adversarial Attacks in Image domain



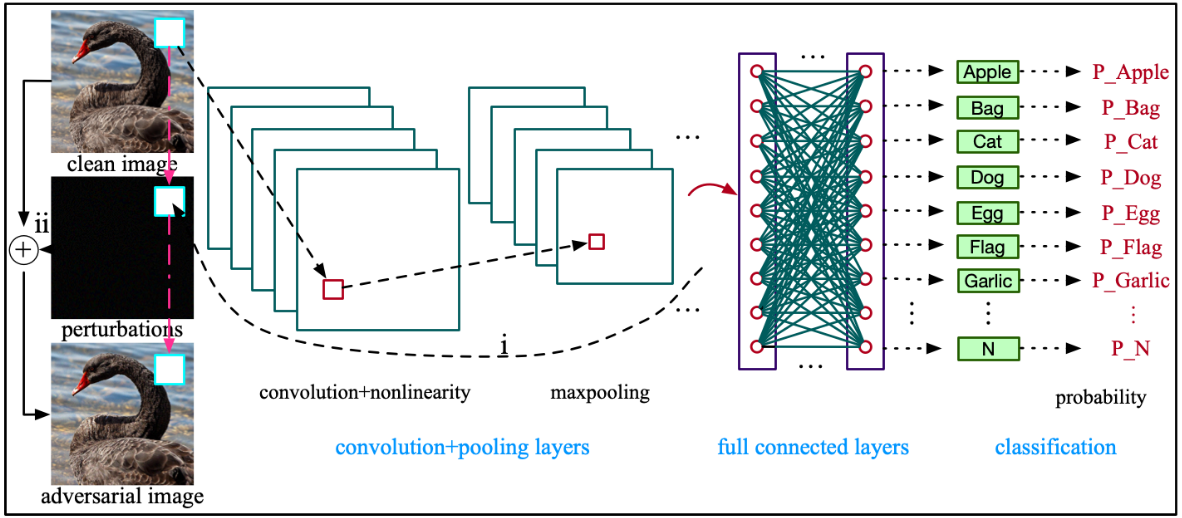
5: Adversarial Attacks in the field of Audio [5]

2

Adversarial Attacks on Deep Neural Networks

- Adversarial Attacks Principle
- Adversarial Attack Methods
- Attacks strategies
- Challenges

Adversarial Attack Principle



$$\text{Min } v \qquad \text{s.t. } f(x+v) \mathrel{!}= f(x).$$

Attack Methods：

Fast Gradient Sign Method (FGSM) [7]:

$$x* = x + \epsilon * sign(\nabla x J (\theta, x, f (x))$$

Iterative-Fast Gradient Sign Method (I-FGSM) [8]:

$$x* = x_{i-1} + \epsilon * sign(\nabla x_{i-1} J (\theta, x_{i-1}, f (x_{i-1}))$$

Carlini and Wagner Method (C&W Attack) [9]:

$$min \parallel v \parallel p + \alpha * L(x + v)$$
$$L(x + v) = max(max(Z (x + v)_i, : i != t) - Z (x + v)_t, -k)$$

Jacobian- based Saliency Map (JSMA) [10]:

$$min \parallel v \parallel p \quad \textbf{s.t. } f (x + v) = y* != y$$

Adversarial Attack Methods

Attack Methods：

Universal Perturbation [11]:

$$\| v \|_p \leq \epsilon$$

$$P(f(x+v) \,!\!= f(x)) \geq 1 - \tau$$

Projected Gradient Descent Method (PGDM) [12]:

$$min(max(J(\theta, x, f(x_{i-1}))))$$

Momentum –FGSM (MI-FGSM) [13]:

$$m_i = \alpha * m_{i-1} + \frac{(\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1})))}{\left\| (\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))) \right\|_1}$$

$$x* = x_{i-1} + \epsilon * sign(m_i)$$

ET AL.

Attack Strategies

Black-box Attack

The attacker doesn't have complete access to the policy network

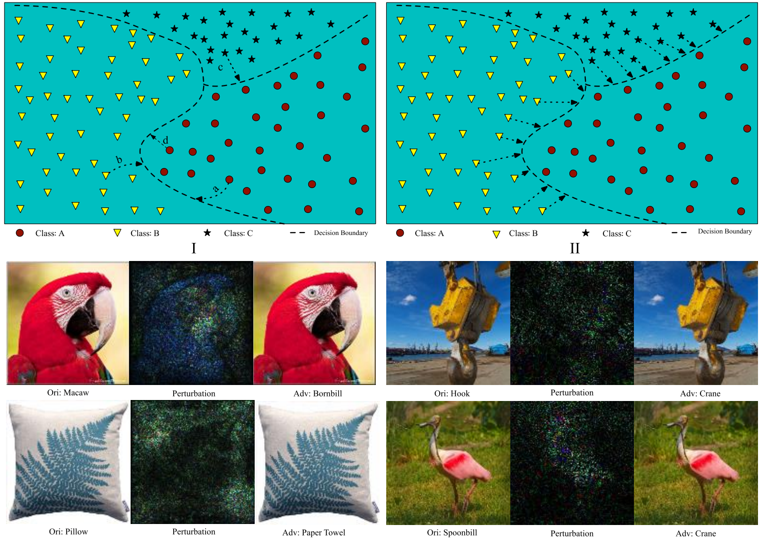
White-box Attack

The attacker has access to the policy network

Attack Strategies

Non-targeted Attack

The prediction label different from the ground truth.



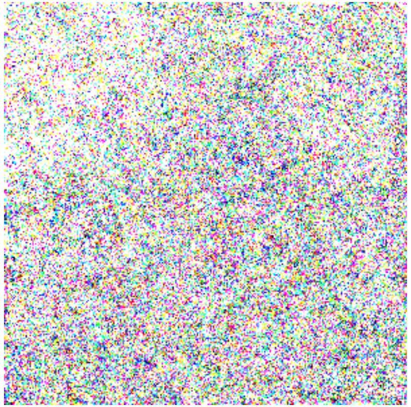
Targeted Attack

Fooling DNNs with fixed labels .

Challenges

Crafted Adversarial Perturbations result large pixel modification on clean images

+



=

MI-FGSM
L_∞=10
Iteration=10
PSNR=26.77
AMP = 0.2335
Inception-v3

Schooner(91.69%)

Perturbations

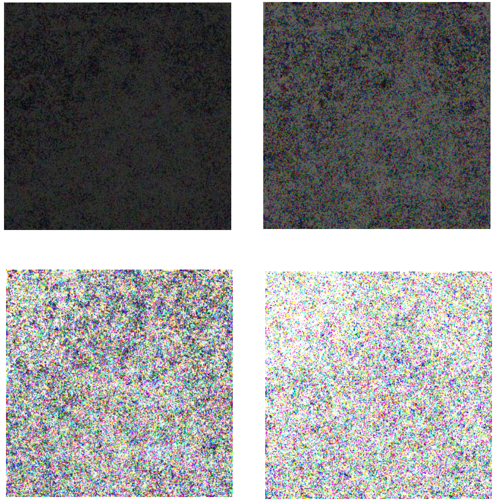
Private(99.99%)

Challenges

How to qualify the strength of crafted adversarial perturbation?

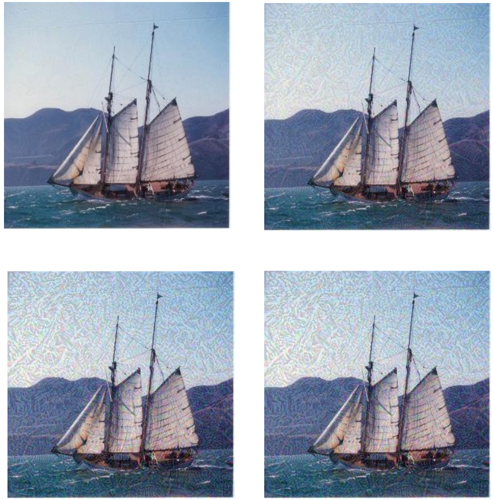
Schooner(91.69%)

+



Perturbations

=



Private

MI-FGSM
L_∞=1,2,5,10
Iteration=10
Inception-v3

3

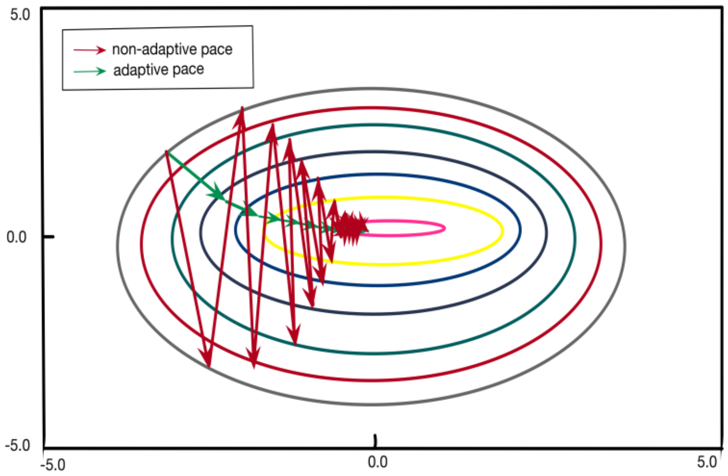
Adaptive Gradient-based Perturbations Generation

Proposed Method

Adaptive Gradient Search for Deep Neural Models

Updating gradient in a direction with a stable size may cause trapping into a local minima point.

Updating gradient in a direction with an adaptive size can reduce the rate strapping into local minima point.



Algorithm1: (White-box attack)

$$g_i = \alpha * g_{i-1} + (1 - \alpha) * (sign(\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))))^2$$

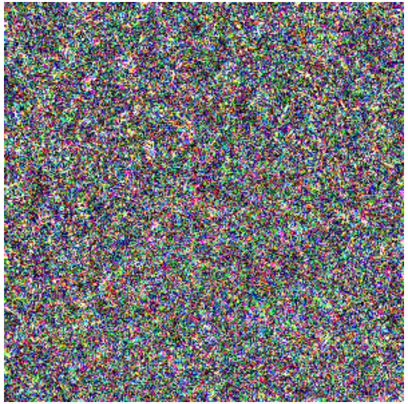
$$v_i = \frac{\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))}{\sqrt{g_i + \delta}}$$

$$x_i^* = x_{i-1} + \epsilon * v_i$$

AI-FGSM (1) :

Schooner(91.69%)

+



Perturbations

=

Private(99.23%)

AI-FGSM
L_∞=10
Iteration=10
PSNR=28.83
AMP=0.0953
Inception-v3

Algorithm2: (White-box attack)

$$g_i = \alpha * g_{i-1} + (1 - \alpha) * (sign(\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))))^2$$

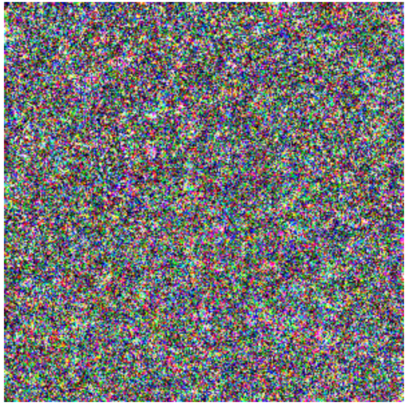
$$m_i = \alpha * m_{i-1} + (1 - \alpha) * sign(\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))$$

$$v_i = \frac{\nabla x_{i-1} J(\theta, x_{i-1}, f(x_{i-1}))}{\sqrt{g_i - m_i^2} + \delta}$$

$$x_i^* = x_{i-1} + \epsilon * v_i$$

AI-FGSM (2) ：

+



=

AI-FGSM
L_∞=10
Iteration=10
PSNR=28.84
AMP=0.0952
Inception-v3

Schooner(91.69%)

Perturbations

Private(96.84%)

Algorithm3: (White-box attack)

$$g_i = \alpha * g_{i-1} + (1-\alpha) * (sign(\nabla_{x_{i-1}} J(\theta, x_{i-1}, f(x_{i-1}))))^2$$

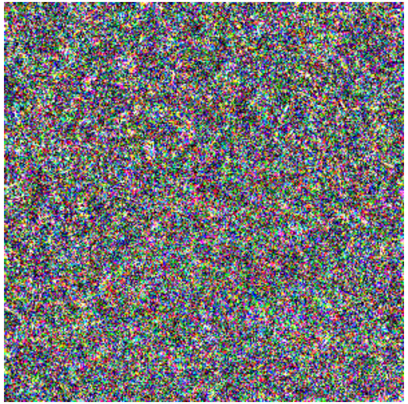
$$m_{vi} = \alpha * m_{vi-1} + (1-\alpha) * v_{i-1}$$

$$v_i = \frac{\nabla_{x_{i-1}} J(\theta, x_{i-1}, f(x_{i-1})) * \sqrt{m_{vi}^2}}{\sqrt{g_i + \delta}}$$

$$x_i^* = x_{i-1} + \epsilon * v_i$$

AI-FGSM (3)：

+



=

AI-FGSM
L_∞=10
Iteration=10
PSNR=28.84
AMP=0.0951
Inception-v3

Schooner(91.69%)

Perturbations

Private(99.24%)

Solution for Minimal Adversarial Perturbation：

Adaptive Term: $\frac{1}{\sqrt{g_i}}$, $\frac{1}{\sqrt{g_i - m_i^2}}$, $\frac{\sqrt{m_{vi}^{2}}}{\sqrt{g_i}}$

Solution for Qualify the Strength of Adversarial Perturbation :

Absolute Mean Perturbation value (AMP) = $\frac{1}{N_c * N_r} * \sum \left\| v_{c,r} \right\|_1$

4

Experimental Results

- Settings
- Results

Settings

Datasets: MNIST [14], CIFAR100 [15], IMAGENET ILSVRC2012(Val) [16]

Classifiers: MNIST, CIFAR100 (Table1), IMAGENET (Pretrained)

Evaluation Metrics: Attack Success Rate(ASR), AMP, Cosine Similarity and SSIM

Settings

Architectrue	MNIST	CIFAR100
Convolution + RELU Max pooling Convolution + RELU Max pooling Convolution + RELU Full Connected +RELU Full Connected +RELU Softmax	3x3x32 2x2 3x3x64 2x2 3x3x64 100 100 10	3x3x64 2x2 3x3x12 2x2 3x3x12 512 512 100
High-parameter	MNIST	CIFAR100
Optimization Method	SGD	SGD
Loss Function	CEL	CEL
Learning rate	0.01	0.01
Momentum	0.9	0.9
Dropout	0.5	0.5
Batch Size	128	128
Epochs	50	50

Table 1: The architecture of the DNN classifier for MNIST and CIFAR100.(CEL indicates Cross Entropy Loss, SGD stands for Stochastic Gradient Descent)

Settings

Architecture for Validation on Preprocessed ILSVRC2012(Val)

Inception-v3(Inc-v3)[17], Inception v4(Inc-v4)[18], Inception-Resnetv2(IncRes-v2)[18], Resnet-152 (Res152)[19] and other three trained by ensemble adversarial: Inc-v3*ens* 3[20], Inc- v3*ens* 4, IncRes-v2*ens* . To simplify the experiments, we choose three images in each of 1000 categories from ILSVRC2012 validation dataset.

Results

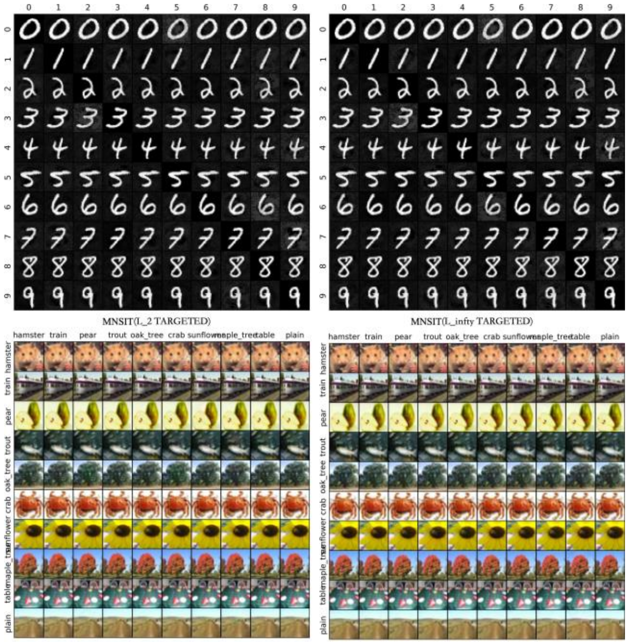
No-Targeted Results on MNIST and CIFAR100

MNIST	FGSM	I-FGSM	MI-FGSM	AI-FGSM(PI)	AI-FGSM
ASR(L∞=10)	13.44%	100.00%	100.00%	100.00%	**100.00%**
AMP(L∞=10)	0.025	0.052	0.078	0.052	**0.052**
Cosine(L∞=10)	0.825	0.805	0.788	0.805	**0.805**
SSIM(L∞=10)	0.852	0.746	0.670	0.748	**0.748**
ASR(L2 =150)	8.30%	100.00%	100.00%	100%	**100.00%**
AMP(L2 =150)	0.011	0.030	0.045	0.043	**0.043**
Cosine(L2 =150)	0.687	0.696	0.692	0.802	**0.802**
SSIM(L2 =150)	0.951	0.843	0.752	0.846	**0.846**
CIFAR100	FGSM	I-FGSM	MI-FGSM	AI-FGSM(PI)	**AI-FGSM**
ASR(L∞=10)	**94.10%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
AMP(L∞=10)	**0.028**	**0.029**	**0.048**	**0.029**	**0.029**
Cosine(L∞=10)	**0.746**	**0.750**	**0.758**	**0.750**	**0.750**
SSIM(L∞=10)	**0.975**	**0.973**	**0.951**	**0.974**	**0.974**
ASR(L2 =150)	**76.13%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
AMP(L2 =150)	**0.008**	**0.011**	**0.018**	**0.015**	**0.015**
Cosine(L2 =150)	**0.748**	**0.750**	**0.753**	**0.750**	**0.750**
SSIM(L2 =150)	**0.988**	**0.986**	**0.982**	**0.987**	**0.987**

Table.2. ASR, AMP, Cosine similarity and SSIM on MNIST and CIFAR100 with FGSM/I-FGSM/MI-FGSM and our methods on white-box and no-targeted attack strategies.

Results

Targeted Attack Results on MNIST and CIFAR100



MNSIT(L_2 TARGETED)

MNSIT(L_infty TARGETED)

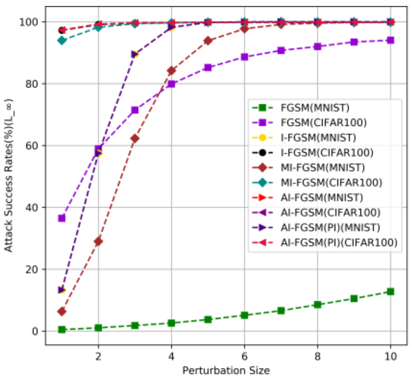
CIFAR100(L_2 TARGETED)

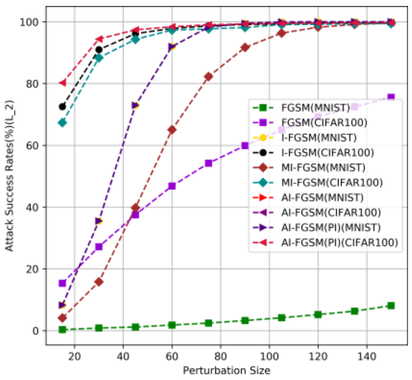
CIFAR100(L_infty TARGETED)

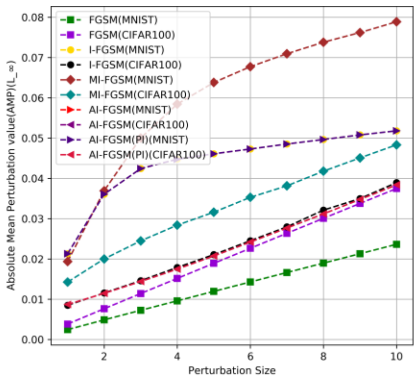
MNIST	FGSM		I-FGSM		MI-FGSM		AI-FGSM(PI)		AI-FGSM	
	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞
0	87.89%	87.89%	90.63%	98.33%	54.38%	83.54%	100.00%	98.76%	100.00%	98.76%
1	87.96%	87.78%	100.00%	100.00%	99.09%	100.00%	100.00%	100.00%	100.00%	100.00%
2	90.02%	90.21%	99.05%	100.00%	74.91%	88.21%	100.00%	100.00%	100.00%	100.00%
3	88.73%	88.51%	97.28%	99.58%	68.97%	85.95%	100.00%	99.71%	100.00%	99.71%
4	87.99%	87.99%	100.00%	100.00%	89.57%	98.04%	100.00%	100.00%	100.00%	100.00%
5	89.78%	90.02%	100.00%	100.00%	87.56%	96.59%	100.00%	100.00%	100.00%	100.00%
6	88.66%	88.87%	97.69%	100.00%	80.89%	93.91%	100.00%	100.00%	100.00%	100.00%
7	88.91%	89.11%	100.00%	100.00%	88.87%	96.36%	100.00%	100.00%	100.00%	100.00%
8	89.66%	90.30%	98.49%	100.00%	85.13%	98.71%	100.00%	100.00%	100.00%	100.00%
9	90.16%	89.96%	99.58%	100.00%	92.45%	98.11%	100.00%	100.00%	100.00%	100.00%
CIFAR100	FGSM		I-FGSM		MI-FGSM		AI-FGSM(PI)		AI-FGSM	
	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞
hamster	97.22%	97.22%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
train	100.00%	100.00%	92.31%	100.00%	92.31%	100.00%	93.44%	100.00%	93.44%	100.00%
pear	95.65%	95.65%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
trout	100.00%	100.00%	91.84%	100.00%	85.71%	97.96%	92.65%	100.00%	92.65%	100.00%
oak-tree	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
crab	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
sunflower	100.00%	100.00%	100.00%	100.00%	87.50%	100.00%	100.00%	100.00%	100.00%	100.00%
apple-tree	96.77%	100.00%	100.00%	100.00%	95.83%	100.00%	100.00%	100.00%	100.00%	100.00%
table	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
plain	87.89%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

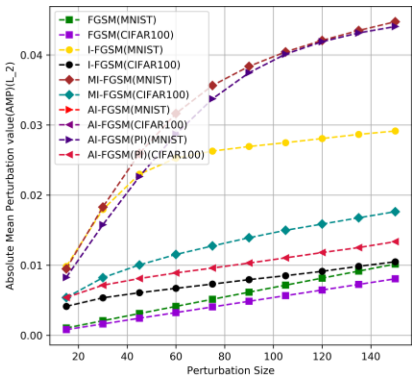
Results

Perturbation size on MNIST and CIFAR100



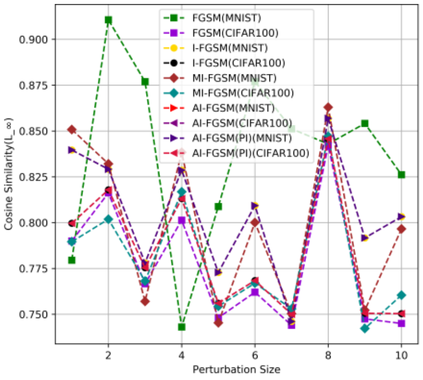


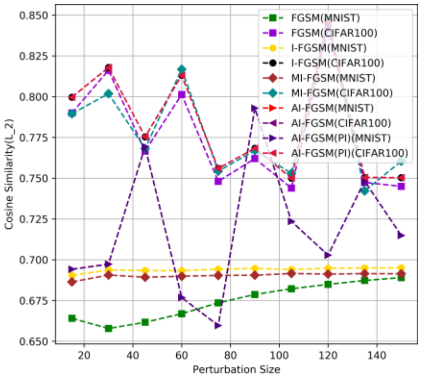


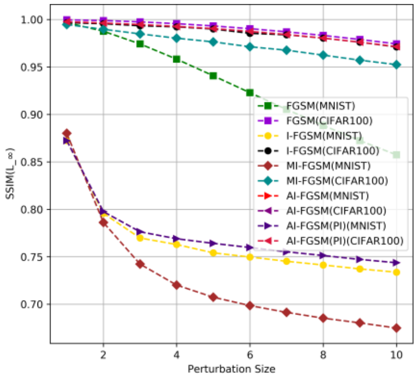


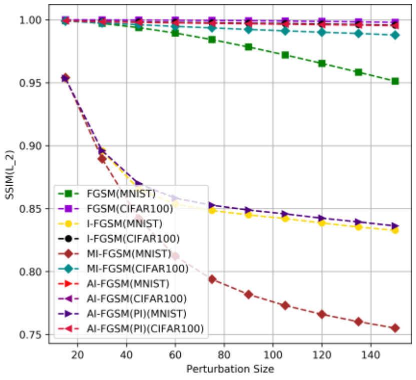
Results

Perturbation size on MNIST and CIFAR100



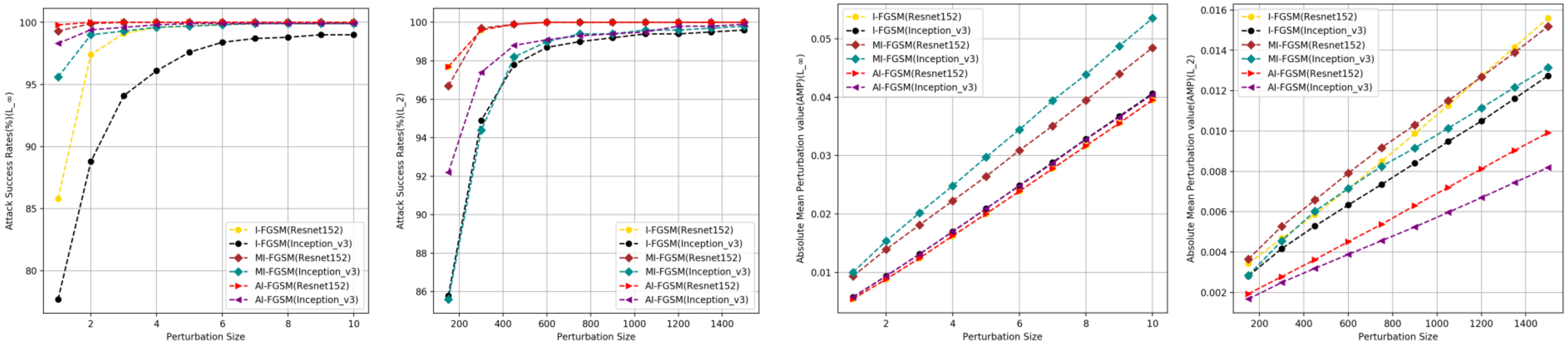






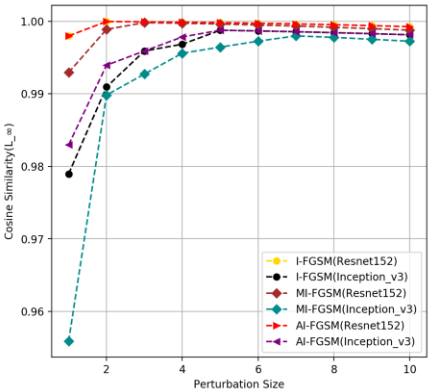
Results

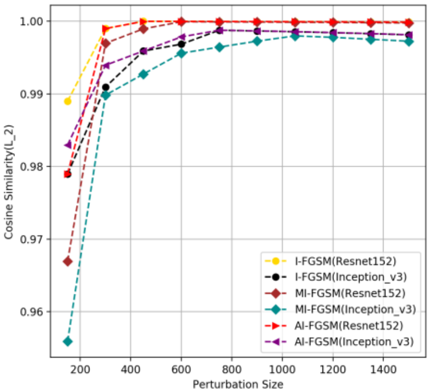
Perturbation size on IMAGENET

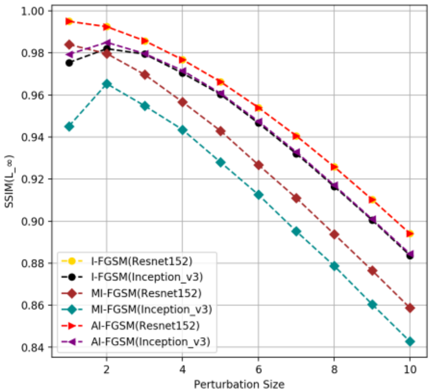


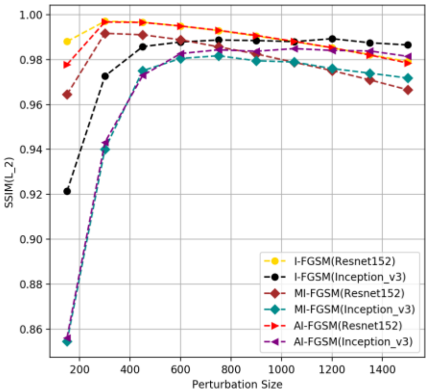
Results

Perturbation size on IMAGENET









Results

White-box and Black-box Strategies on Pre-processed ILSVRC2012(Val)

	Attacks	Inc-v3	Inc-v4	IncRes-v2	Res152	Inc-v3ens 3	Inc-v3ens 4	IncRes-v2ens
Inc-v3	FGSM I-FGSM MI-FGSM AI-FGSM(PI) AI-FGSM	79.89%* 98.41%* 99.69%* 99.80%* 99.80%*	31.03% 28.86% 25.22% 28.99% 28.99%	29.33% 27.25% 25.27% 28.81% 28.81%	27.22% 27.40% 24.72% 27.40% 27.40%	10.40% 7.16% 6.45% 7.16% 7.16%	7.56% 4.21% 4.16% 4.21% 4.21%	7.10% 3.68% 3.88% 3.68% 3.68%
Inc-v4	FGSM I-FGSM MI-FGSM AI-FGSM(PI) AI-FGSM	30.63% 29.36% 28.17% 31.55% 31.55%	72.37%* 96.72%* 95.27%* 99.11%* 99.11%*	27.85% 25.45% 26.41% 31.44% 31.44%	27.23% 27.82% 24.72% 27.85% 27.85%	9.51% 7.96% 6.45% 7.96% 7.96%	8.12% 6.17% 5.23% 6.17% 6.17%	6.23% 5.02% 4.69% 5.02% 5.02%
IncRes-v2	FGSM I-FGSM MI-FGSM AI-FGSM(PI) AI-FGSM	30.03% 28.77% 26.67% 35.16% 35.16%	29.53% 28.26% 25.65% 32.54% 32.54%	65.07%* 96.65%* 97.01%* 98.14%* 98.14%*	28.02% 28.70% 26.83% 28.83% 28.83%	10.71% 10.50% 9.07% 10.53% 10.53%	8.41% 7.80% 6.78% 7.80% 7.80%	8.03% 6.14% 6.01% 6.14% 6.14%
Res152	FGSM I-FGSM MI-FGSM AI-FGSM(PI) AI-FGSM	33.33% 33.43% 29.43% 33.43% 33.43%	35.04% 35.34% 30.03% 35.41% 35.41%	32.43% 33.03% 28.73% 33.03% 33.03%	90.30%* 100.00%* 100.00%* 100.00%* 100.00%*	11.56% 8.12% 7.46% 8.12% 8.12%	10.23% 6.09% 5.78% 6.09% 6.09%	9.72% 6.12% 6.07% 6.12% 6.12%

Table 4: Attack success rate on the ensemble models with L_∞ =10 norm constraint. * stand for white-box attacks.

Results

White-box and Black-box Strategies on Pre-processed ILSVRC2012(Val)

	Attacks	-Inc-v3	-Inc-v4	-IncRes-v2	-Res152	-Inc-v3ens 3	-Inc-v3ens 4	-IncRes-v2ens
Ensemble	FGSM	68.19%	67.52%	63.01%	59.37%	52.46%	51.44%	54.29%
	I-FGSM	96.47%	97.21%	96.58%	98.55%	98.33%	98.15%	94.37%
	MI-FGSM	95.58%	97.21%	96.20%	98.55%	95.27%	97.38%	95.62%
	AI-FGSM(PI)	96.51%	97.21%	96.63%	98.55%	96.41%	98.05%	98.22%
	AI-FGSM	96.51%	97.21%	96.63%	98.55%	96.41%	98.05%	98.22%
Hold-out	FGSM	39.27%	40.56%	40.37%	42.41%	42.02%	39.97%	35.31%
	I-FGSM	77.36%	76.21%	75.69%	78.42%	38.56%	29.58%	32.62%
	MI-FGSM	78.23%	75.58%	74.03%	77.93%	32.14%	28.54%	33.04%
	AI-FGSM(PI)	77.36%	76.23%	75.81%	78.56%	37.09%	29.40%	33.13%
	AI-FGSM	77.36%	76.23%	75.81%	78.56%	37.09%	29.40%	33.13%

Table 5: Attack success rates on the ensemble and hold-out models. In this table, ’-’ before the network indicates the hold-out network. The result shows that our proposed method can reach high success rates on black-box and white-box attacks with L_∞ =10 norm limitation.

Results

White-box and Black-box Strategies on Pre-processed ILSVRC2012(Val)

	Attacks	Inc-v3	Inc-v4	IncRes-v2	Res152	Inc-v3ens 3	Inc-v3ens 4	IncRes-v2ens
Inc-v3	FGSM	74.89%*	27.03%	26.73%	24.72%	12.17%	12.05%	11.26%
	I-FGSM	98.92%*	55.17%	57.31%	25.12%	10.07%	10.12%	9.05%
	MI-FGSM	99.89%*	56.55%	58.11%	23.92%	9.68%	8.44%	7.84%
	AI-FGSM(PI)	99.86%*	59.23%	63.01%	25.30%	10.08%	10.20%	9.05%
	AI-FGSM	99.86%*	59.23%	63.01%	25.30%	10.08%	10.20%	9.05%
Inc-v4	FGSM	26.23%	66.37%*	25.63%	24.42%	10.54%	11.03%	10.10%
	I-FGSM	59.41%	98.34%*	60.41%	26.13%	9.03%	8.16%	8.23%
	MI-FGSM	61.65%	95.88%*	61.01%	23.62%	9.07%	7.69%	7.40%
	AI-FGSM(PI)	67.44%	99.62%*	62.89%	25.03%	9.01%	7.72%	8.06%
	AI-FGSM	67.44%	99.62%*	62.89%	25.03%	9.01%	7.72%	8.06%
IncRes-v2	FGSM	27.63%	26.23%	59.86%*	25.23%	12.86%	14.70%	12.13%
	I-FGSM	63.86%	60.17%	95.15%*	26.43%	10.04%	11.17%	11.21%
	MI-FGSM	66.20%	59.88%	97.41%*	24.51%	9.12%	9.07%	8.68%
	AI-FGSM(PI)	67.61%	64.42%	99.15%*	26.64%	10.10%	11.12%	10.93%
	AI-FGSM	67.61%	64.42%	99.15%*	26.64%	10.10%	11.12%	10.93%
Res152	FGSM	29.83%	30.53%	30.83%	84.39%*	14.02%	13.46%	14.71%
	I-FGSM	30.03%	31.04%	31.53%	100.00%*	11.43%	10.89%	10.66%
	MI-FGSM	27.33%	28.43%	28.53%	100.00%*	9.17%	9.46%	10.03%
	AI-FGSM(PI)	30.13%	31.23%	31.81%	100.00%*	11.51%	11.03%	11.04%
	AI-FGSM	30.13%	31.23%	31.81%	100.00%*	11.51%	11.03%	11.04%

Table 6: We observe our methods reach the highest success rates on all black-box models and maintain higher success rates on all white-box models with L2=1500 norm limitation than other gradient-based attack methods. * stand for white-box attacks.

Results

White-box and Black-box Strategies on Pre-processed ILSVRC2012(Val)

	Attacks	-Inc-v3	-Inc-v4	-IncRes-v2	-Res152	-Inc-v3ens 3	-Inc-v3ens 4	-IncRes-v2ens
Ensemble	FGSM	62.73%	62.61%	62.41%	63.54%	62.54%	61.37%	**61.59%**
	I-FGSM	99.91%	99.95%	99.43%	100.00%	100.00%	98.21%	**99.31%**
	MI-FGSM	99.93%	99.97%	99.72%	100.00%	100.00%	98.21%	**99.31%**
	AI-FGSM(PI)	99.95%	99.98%	99.90%	100.00%	100.00%	100.00%	**100.00%**
	AI-FGSM	99.95%	99.98%	99.90%	100.00%	100.00%	100.00%	**100.00%**
Hold-out	FGSM	40.49%	40.41%	40.42%	40.52%	40.27%	40.29%	40.41%
	I-FGSM	78.46%	79.37%	78.16%	81.27%	40.16%	38.13%	39.07%
	MI-FGSM	79.21%	78.20%	76.18%	79.30%	39.97%	37.58%	38.33%
	AI-FGSM(PI)	78.46%	79.37%	78.16%	81.27%	40.16%	38.13%	39.17%
	AI-FGSM	78.46%	79.37%	78.16%	81.27%	40.16%	38.13%	39.17%

Table 7: Attack success rates on the ensemble and hold-out models. In this table, '-' before the network indicates the hold-out network. The result shows that our proposed method can reach high success rates on black-box and white-box attacks with L_∞ =10 norm limitation.

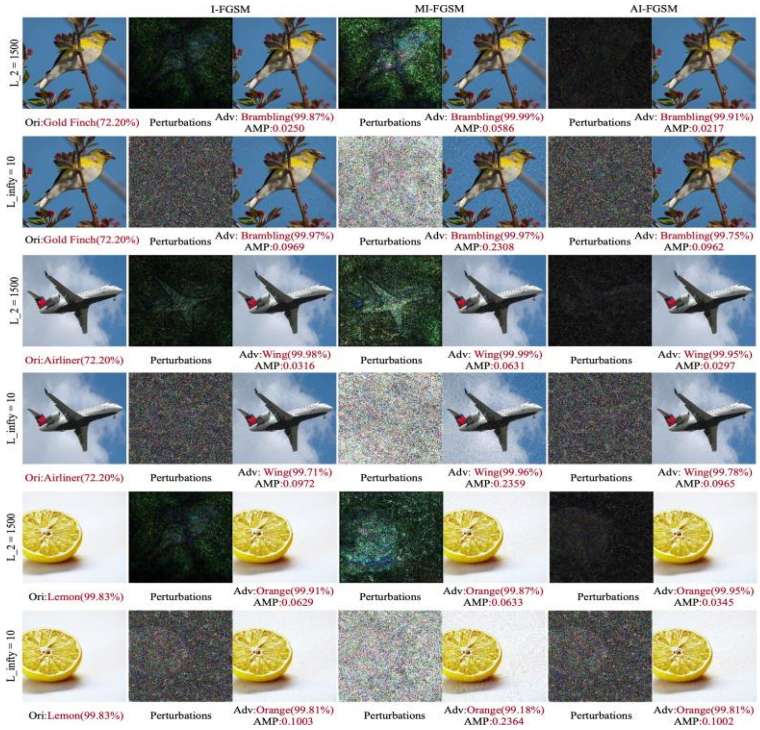
Results

White-box and Targeted Strategies on Pre-processed ILSVRC2012(Val)

	Attack	Inception-v3	Resnet152
L_∞	I-FGSM	36.41%	39.54%
	MI-FGSM	40.18%	42.20%
	AI-FGSM(PI)	38.02%	43.27%
	AI-FGSM	38.02%	43.27%
L_2	I-FGSM	42.17%	42.05%
	MI-FGSM	42.02%	42.70%
	AI-FGSM(PI)	42.10%	42.83%
	AI-FGSM	42.10%	42.83%

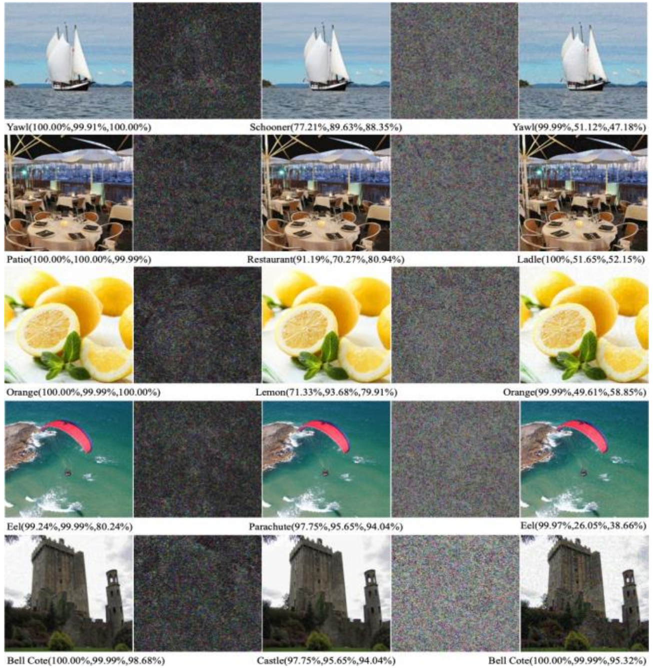
Table 8: Top-1 target accuracy rate with two norm bounds. Targeted label is crane.

Results



Adversarial examples generated by I-FGSM, MI- FGSM and our method (AI-FGSM) on Resnet152 with No- Targeted strategy and L = 10 and $L2$=1500 norm constraints. All adversarial examples are generated with 10 iterations. Perturbations are amplified by 3 times.

Results



Universal effect of our proposed method (AI- FGSM) on three different DNNs(Inception v3, Inception v4, and Inception-Resnet-v2). The left images are crafted with *L*2=1500 norm bound, and the right images are crafted with *L\infty* =10 norm bound, and the middle are clean images. All perturbations generated are amplified by 3 times.

5

Conclusion

Conclusion and Future Work

Conclusion:

1. Propose the adaptive gradient adversarial attack methods to optimize adversarial attacks, which can effectively fool the white- box models as well as the black-box models.
2. Our methods focus on adjusting gradient at a proper pace, which could escape from trapping into poor local minima for gradient searching.

Future Work:

1. We next focus our attention on how to get the path of decision boundaries to improve the success rate of the adversarial targeted attacks on general deep neural models .

Publications

- ▪ Xiao Y , Pun C M , Liu B . Adversarial example generation with adaptive gradient search for single and ensemble deep neural network[J]. Information Sciences, 2020, 528:147-167.
- ▪ Xiao Y , Pun C M , Liu B . Crafting adversarial example with adaptive root mean square gradient on deep neural networks[J]. Neurocomputing, 2020, 389:179-195.
- ▪ Y. Xiao, C.-M. Pun and J. Zhou, "Generating Adversarial Perturbation with Root Mean Square Gradient," *Proceedings of AAAI Workshops*, 2019.

THANK YOU!

Q&A