Enhance Stealthiness and Transferability of Adversarial Attacks with Class Activation Mapping Ensemble Attack

Hui Xia, Rui Zhang, Zi Kang, Shuliang Jiang, Shuo Xu

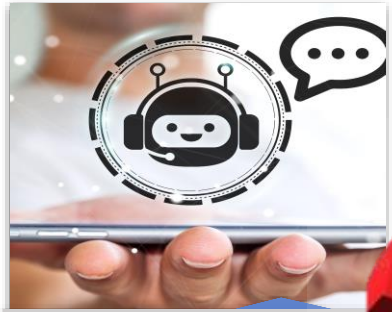
Ocean University of China

Outline

Background

Finance

Financial Accounting



Customer service

Security

Business

Risk evaluation

Fraud detection

Healthcare

Intelligent diagnosis

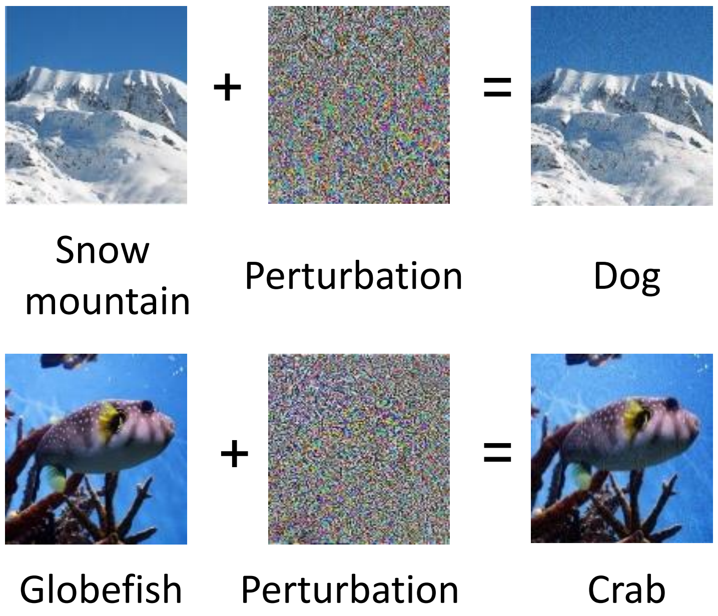
Tele medicine

Industry

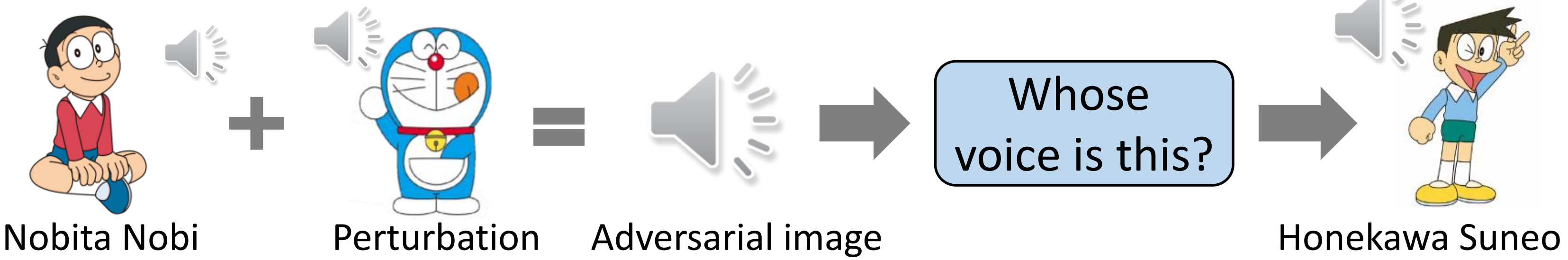
Automatic production

Quality test

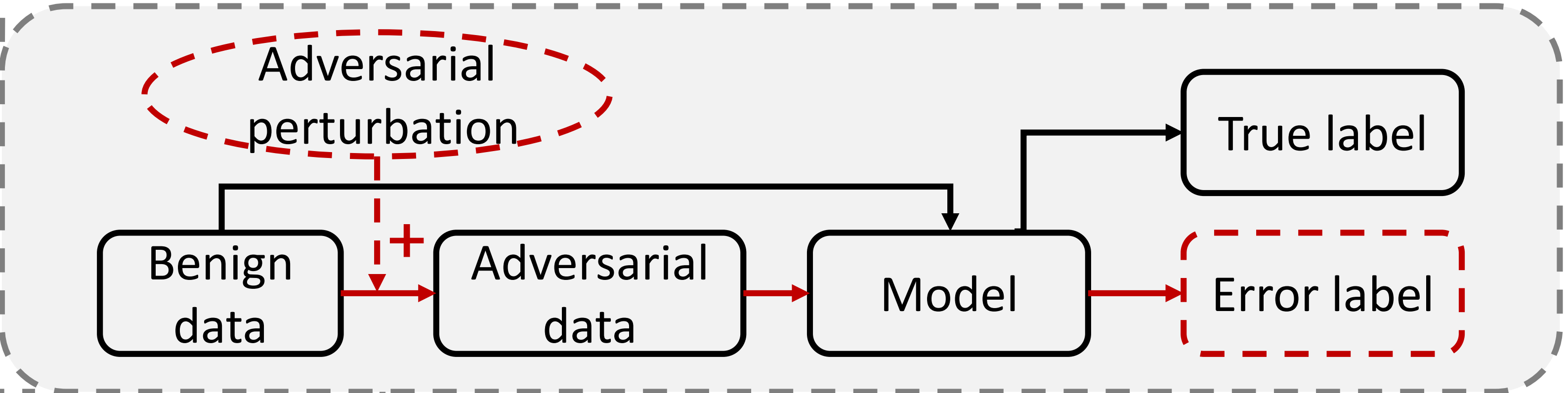
Background



Classifier



Speaker Recognition System

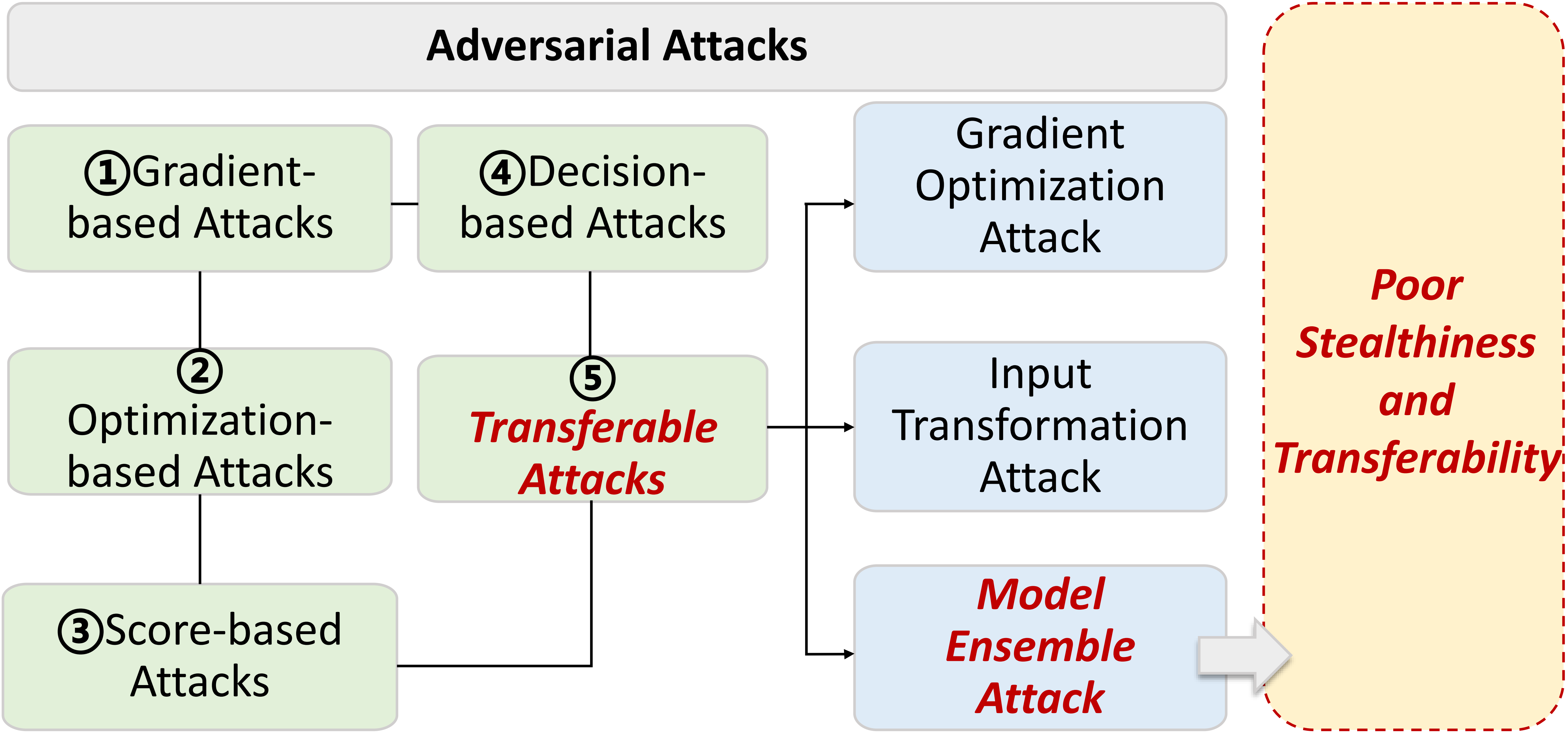


Adversarial attack

Face Recognition System

Autonomous Driving System

Related Work



◆ Methodology

Overview

- Incorporate score of class activation map as weights for adding perturbations to each pixel to enhance the attack capabilities for low attack epochs and stealthiness.
- Ensemble the score of class activation map of multiple models to ensure the transferability of adversarial examples.

Objective function for non-targeted attack

Improving transferability of adversarial examples

$$\boldsymbol{x}' = \underset{||\boldsymbol{x}-\boldsymbol{x}'||<\boldsymbol{\varepsilon}}{\mathbf{argmax}} \boldsymbol{\mathcal{L}}(\boldsymbol{x}', \boldsymbol{y}; \boldsymbol{\theta}) + ||\widetilde{\boldsymbol{x}} - \boldsymbol{x}'||_2$$

Improving the attack capability of adversarial examples

$$x'_{t+1} = x'_t + (\lambda \cdot \alpha) \cdot (w \cdot M_c) \cdot sign(g_{t+1})$$

Gradient variance optimization

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

$$\hat{g}_{t+1} = \nabla_{x'_t}\mathcal{L}(x', y; \theta) + ||\tilde{x} - x'||_2$$

$$v_{t+1} = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x^i}\mathcal{L}(x^i, y; \theta) - \nabla_x \mathcal{L}(x, y; \theta)$$

Score of class activation map

$$M_c(a,b) = \max\{m_c^1(a,b), m_c^2(a,b), \cdots, m_c^n(a,b)\}$$

$$m_c(x,y) = \sum_{k=1}^{K}\left(\alpha_c^k \mathbf{F}^{lk}(a,b)\right)$$

◆ Methodology

Score of class activation map:

$$\alpha_c^k = \sum_{x,y}\left(\frac{\mathbf{F}^{lk}(x,y)}{\sum_{x,y}\partial F^{lk}(x,y)}\frac{\partial S_c(\mathbf{F}^l)}{\partial F^{lk}(x,y)}\mathbf{F}^{lk}(x,y)\right)$$

$$S_c(\mathbf{F}^l) = \sum_{k=1}^{K}\sum_{x,y}\left(\frac{\partial S_c(\mathbf{F}^l)}{\partial F^{lk}(x,y)}\mathbf{F}^{lk}(x,y)\right) + \boldsymbol{\varepsilon}(\mathbf{F}^l)$$

$$\boldsymbol{\varepsilon}(\mathbf{F}^l) = \sum_{t=l+1}^{L}\sum_j \frac{\partial S_c(\mathbf{F}^l)}{\partial \mu_j^t} b_j^t$$

Objective function for targeted attack

$$\boldsymbol{x}' = \underset{||\boldsymbol{x}-\boldsymbol{x}'||<\boldsymbol{\varepsilon}}{\mathbf{argmax}} - \boldsymbol{\mathcal{L}}(\boldsymbol{x}', \boldsymbol{y}'; \boldsymbol{\theta}) - ||\widetilde{\boldsymbol{x}}_{\boldsymbol{tar}} - \boldsymbol{x}'||_2$$

The optimization method is the same as the non-targeted attack. Specifically,

$$\tilde{x}_{tar} = \max\{x_{tar\_cam}^1, x_{tar\_cam}^2, \cdots, x_{tar\_cam}^n\}$$

$$x'_{t+1} = x'_t + (\lambda \cdot \alpha) \cdot (w \cdot M_{tc}) \cdot sign(g_{t+1})$$

$$M_{tc}(x,y) = \max\{m_{tc}^1(x,y), m_{tc}^2(x,y), \cdots, m_{tc}^n(x,y)\}$$

$$m_{tc}(a,b) = \sum_{k=1}^{K}\left(\alpha_{tc}^k \mathbf{F}^{lk}(a,b)\right)$$

◆ Experiment

Dataset		ImageNet
Models	CAMs substitute models	WideResNet101 (WRN101), Inception v2 (Inception), and ResNet34
	Gradient substitute model	ResNet50
	Target models	AlexNet, VGG16, EfficientNet b0 (EfficientNet), WideResNet50 (WRN50), MobileNet v2 (MobileNet), ResNet18, ConvNeXt, ViT, and RegNet
Metrics	Perceptual metrics	Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Structure Similarity Index Measure (SSIM), Low_fre, CIEDE2000, L_2 norm, and L_∞ norm
	Attack capability metrics	Attack Success Rate (ASR) and Average Attack Success Rate (AASR)
Baseline Methods		PGD, TPGD, DIFGSM, TIFGSM, MFGSM, NIFGSM, SINIFGSM, VMIFGSM, VNIFGSM, and SVRE

◆ Experiment—visualization: non-targeted

Benign Image	TPGD	PGD
		
Label: 131	Label: 134	Label: 133
L_2: 0	L_2: **149**	L_2: 210
SSIM: 1.00	SSIM: **0.9185**	SSIM: 0.8261
PSNR: INF	PSNR: **30.04**	PSNR: 28.56
ASR: ----	ASR: 0.6333	ASR: 1.00

VMIFGSM	MIFGSM	OUR
		
Label:133	Label:133	Label: 133
L_2: 162	L_2: 161	L_2: **154**
SSIM: 0.9072	SSIM: 0.9048	SSIM: **0.9132**
PSNR: 29.71	PSNR: 29.69	PSNR: **29.88**
ASR: 1.00	ASR: 1.00	ASR: 1.00

◆ Experiment—visualization : targeted

Benign Image		**DIFGSM**		**PGD**	
	Ori-label: 603		Ori-label: 603		Ori-label: 603
	Tar-label: ---		Tar-label: 256		Tar-label: 256
	Epoch: ---		Epoch: 11		Epoch: 11
	Adv-label: ---		Adv-label: 603		Adv-label: 603
	ASR: ----		ASR: 0.2750		ASR: 0.2333
VMIFGSM		**VNIFGSM**		**OUR**	
	Ori-label: 603		Ori-label: 603		Ori-label: 603
	Tar-label: 256		Tar-label: 256		Tar-label: 256
	Epoch: 11		Epoch: 11		Epoch: 11
	Adv-label: 537		Adv-label: 537		Adv-label: **256**
	ASR: 0.4333		ASR: 0.4417		ASR: **0.6417**

Experiment—stealthiness: non-targeted

Metric / Method	PSNR	MSE	L_2	L_∞	LOW_FRE	SSIM	AASR
TPGD	31.54	0.0008	120.02	0.2751	47.27	0.89	37%
PGD	29.60	0.0012	175.88	0.2813	69.90	0.81	49%
TIFGSM	30.75	0.0009	139.00	0.2765	72.98	0.89	53%
DIFGSM	29.60	0.0012	175.58	0.2828	70.91	0.81	60%
NIFGSM	29.32	0.0012	185.30	0.2780	78.79	0.79	62%
MIFGSM	29.39	0.0012	182.62	0.2848	78.27	0.80	63%
SINIFGSM	29.20	0.0013	190.08	0.2816	82.14	0.79	69%
VMIFGSM	29.41	0.0012	182.64	0.2828	82.16	0.81	71%
OUR	**30.10**	**0.0011**	**158.56**	**0.2795**	**68.25**	**0.85**	**71%**

Experiment—stealthiness: targeted

Metric / Method	PSNR	MSE	L_2	L_∞	LOW_FRE	SSIM	AASR
PGD	29.53046571	0.001182558	178.0081423	0.27973857	70.71934106	0.80665016	0.39
TIFGSM	28.45646668	0.001501671	226.0435975	0.286470595	156.7158532	0.85140028	0.57
DIFGSM	29.28310092	0.001247367	187.7635999	0.281830071	79.57947454	0.801656643	0.58
NIFGSM	25.39978734	0.002909817	438.0089821	0.306960792	189.3314227	0.609225211	0.64
MIFGSM	25.28214622	0.002990327	450.1279874	0.306405236	196.8420255	0.602871796	0.67
SINIFGSM	25.24964148	0.003014128	453.7106939	0.308104582	213.6936131	0.619492346	0.65
VMIFGSM	28.43584709	0.001490813	224.4091297	0.287450986	111.7509607	0.782621774	0.65
VNIFGSM	28.25975039	0.001553464	233.8398622	0.286045759	116.9330705	0.773712301	0.67
OUR	**30.37717114**	**0.000993524**	**149.5531535**	**0.27944445**	**64.47938073**	**0.861607709**	**0.65**

◆ Experiment—attack ability: non-targeted

Method / Epoch	PGD	TPGD	DIFGSM	MIFGSM	NIFGSM	TIFGSM	SINIFGSM	VNIFGSM	VMIFGSM	OUR
1	0.66670	0.28333	0.57500	0.55833	0.55000	0.41667	0.45833	0.55000	0.55833	**0.76667**
3	0.92500	0.58333	0.86667	0.90000	0.90000	0.75833	0.80833	0.90000	0.90000	**0.98333**
5	0.99167	0.60000	0.96667	0.98333	0.98333	0.88333	0.89167	0.98333	0.99167	1
7	1	0.65000	1	1	1	0.94167	0.96667	1	1	1
9	1	0.73333	1	1	1	0.98333	0.99167	1	1	1
10	1	0.63333	1	1	1	0.99167	1	1	1	1
12	1	0.62500	1	1	1	1	1	1	1	1
14	1	0.68333	1	1	1	1	1	1	1	1
16	1	0.58333	1	1	1	1	1	1	1	1
18	1	0.59167	1	1	1	1	1	1	1	1
20	1	0.65000	1	1	1	1	1	1	1	1
22	1	0.64167	1	1	1	1	1	1	1	1
24	1	0.69167	1	1	1	1	1	1	1	1
26	1	0.70000	1	1	1	1	1	1	1	1
28	1	0.67500	1	1	1	1	1	1	1	1

◆ Experiment—attack ability: targeted

Method / Epoch	PGD	DIFGSM	MIFGSM	NIFGSM	TIFGSM	SINIFGSM	VNIFGSM	VMIFGSM	OUR
1	0.0000	0.0083	0.0083	0.0083	0.0000	0.0083	0.0083	0.0083	**0.0167**
2	0.0167	0.0000	0.0250	0.0250	0.0167	0.0250	0.0250	0.0250	0.0250
3	0.0250	0.0167	0.0250	0.0250	0.0000	0.0167	0.0250	0.0250	**0.0583**
4	0.0417	0.0333	0.1250	0.0917	0.0333	0.0500	0.0500	0.0667	**0.1667**
5	0.0750	0.0750	0.1333	0.0917	0.0250	0.0417	0.0500	0.0583	**0.1917**
6	0.1000	0.0750	0.2667	0.2083	0.0500	0.1667	0.1333	0.1417	**0.3000**
7	0.1333	0.1083	0.2833	0.2083	0.0500	0.1333	0.1417	0.1417	**0.3083**
8	0.1250	0.1917	0.4417	0.3583	0.1250	0.2917	0.3083	0.3000	**0.4167**
9	0.1417	0.2250	0.4333	0.3500	0.1167	0.2667	0.3083	0.3000	**0.4750**
10	0.1917	0.2333	0.5333	0.4917	0.2083	0.3833	0.3583	0.3667	**0.6083**
11	0.2333	0.2750	0.5250	0.4917	0.1833	0.3667	0.3667	0.3500	**0.6417**
13	0.2500	0.3250	0.7250	0.6333	0.2500	0.4167	0.4417	0.4333	**0.7333**
17	0.3250	0.4833	0.8750	0.8667	0.3833	0.6667	0.7083	0.7083	**0.8833**
19	0.3667	0.5167	0.9083	0.9167	0.4250	0.7083	0.7917	0.8167	**0.9250**
22	0.3333	0.6167	0.9583	0.9750	0.6167	0.8250	0.8833	0.8917	**0.9667**

Experiment—transferability: non-targeted

Method / Model	PGD	SVRE	TPGD	DIFGSM	MIFGSM	NIFGSM	TIFGSM	SINIFGSM	VNIFGSM	VMIFGSM	OUR
ResNet18	0.44	0.15	0.26	0.50	0.47	0.52	0.38	0.50	0.55	0.57	**0.73**
ResNet34	0.37	0.10	0.20	0.51	0.48	0.52	0.43	0.48	0.63	0.63	**0.77**
ResNet50	1.00	0.16	0.63	1.00	1.00	1.00	0.99	1.00	1.00	1.00	**1.00**
AlexNet	0.39	0.08	0.28	0.42	0.33	0.29	0.28	0.36	0.31	0.33	**0.41**
MobileNet	0.50	0.18	0.36	0.53	0.51	0.47	0.42	0.57	0.58	0.63	**0.72**
WRN 50	0.48	0.14	0.33	0.54	0.50	0.51	0.52	0.57	0.68	0.70	**0.83**
WRN101	0.30	0.08	0.25	0.38	0.42	0.41	0.33	0.41	0.53	0.56	**0.73**
VGG16	0.43	0.16	0.33	0.45	0.42	0.43	0.41	0.49	0.59	0.61	**0.74**
Inception	0.23	0.10	0.19	0.28	0.26	0.28	0.25	0.26	0.37	0.40	**0.48**
EfficientNet	0.35	0.11	0.30	0.38	0.39	0.40	0.33	0.40	0.49	0.51	**0.68**
ConvNeXt	0.28	0.16	0.25	0.32	0.29	0.28	0.27	0.28	0.333	0.36	**0.48**
ViT	0.15	0.05	0.11	0.14	0.144	0.17	0.17	0.15	0.23	0.23	**0.33**
RegNet	0.42	0.18	0.33	0.51	0.53	0.51	0.41	0.48	0.62	0.60	**0.75**

◆ Experiment—transferability: targeted

Method / Model	PGD	DIFGSM	MIFGSM	NIFGSM	TIFGSM	SINIFGSM	VNIFGSM	VMIFGSM	OUR
ResNet18	0.3667	0.3750	0.3500	0.2750	0.3833	0.4417	0.4417	0.4583	**0.7000**
ResNet34	0.2750	0.3417	0.3583	0.2750	0.3250	0.4417	0.4417	0.4583	**0.6333**
ResNet50	0.5750	0.9167	0.9000	0.7583	0.8167	0.9667	0.9667	0.9417	**0.9750**
AlexNet	0.3583	0.3083	0.3000	0.2833	0.3167	0.2917	0.2917	0.2750	**0.6500**
MobileNet	0.4167	0.3750	0.4083	0.3833	0.3917	0.5167	0.5167	0.4917	**0.7250**
WRN50	0.2667	0.3833	0.3917	0.3750	0.4000	0.5333	0.5333	0.4750	**0.6667**
WRN101	0.2583	0.2917	0.3417	0.2583	0.3083	0.3083	0.3083	0.3417	**0.5333**
VGG16	0.4083	0.4333	0.4417	0.3917	0.4000	0.5750	0.5750	0.5167	**0.9917**
Inception	0.2500	0.2167	0.2417	0.2333	0.2250	0.3417	0.3417	0.3250	**0.4500**
Efficientnet	0.3417	0.3667	0.3417	0.2917	0.3083	0.3917	0.3917	0.3750	**0.5917**
ConvNeXt	0.0000	0.0700	0.0900	0.1000	0.0600	0.1000	0.1000	0.1300	**0.3700**
ViT	0.4800	0.2600	0.5600	0.5300	0.0900	0.4500	0.4100	0.4600	**0.6800**
RegNet	0.1400	0.1800	0.5100	0.4600	0.1100	0.2400	0.2800	0.3000	**0.6500**

Conclusion

Problem

Poor stealthiness and transferability

Solution

- We first use the class activation mapping method to discover the relationship between the decision of the Deep Neural Network and the image region.
- Then we calculate the class activation score for each pixel and use it as the weight for perturbation to enhance the stealthiness of adversarial examples and improve attack performance under low attack rounds.
- In the optimization process, we also ensemble class activation maps of multiple models to ensure the transferability of the adversarial attack algorithm.

Experiment

Results show that our method generates adversarial examples with high perceptibility, transferability, and attack performance under low-round attacks.

Thank You!

Q&A