Robust Superpixel-Guided Attentional Adversarial Attack

20185153 이상영

목차

0. Background

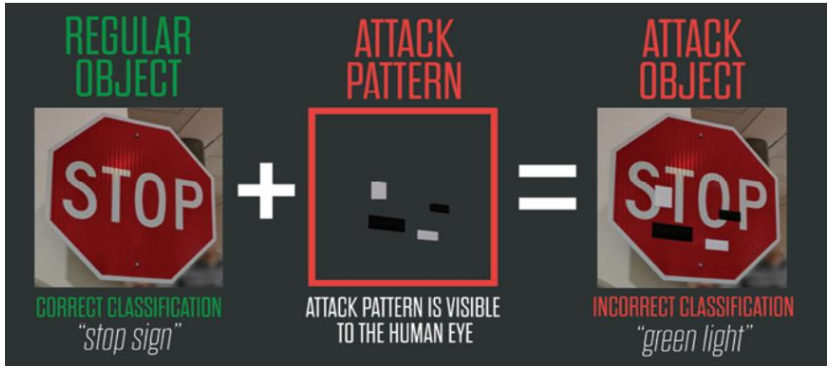
- **Adversarial Attack**

: DNN을 이용한 모델에 Adversarial Perturbation을 적용하여 오분류를 발생시키는 것.

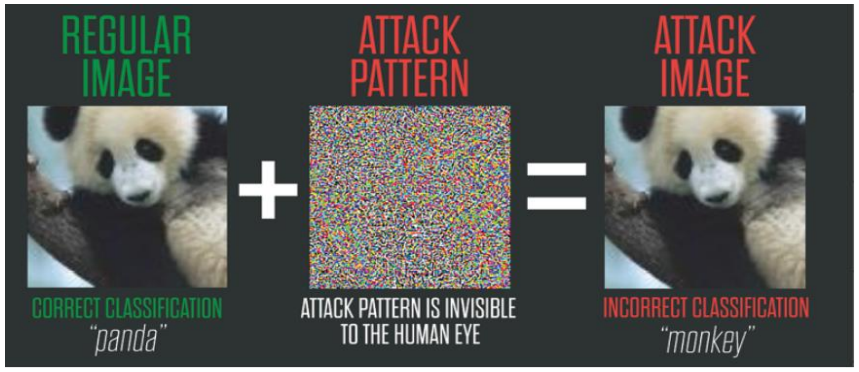
1. **회피 공격(Evasion attacks): 적대적 예제**를 사용해서 AI가 잘못된 의사결정을 하도록 하는 공격
2. 중독 공격(Poisoning attacks): 공격자가 AI 모델의 학습 과정에 관여하여 AI 시스템 자체를 손상시키는 공격
3. 탐색적 공격(Exploratory Attacks): 주어진 입력에 대해 출력되는 분류 결과와 신뢰도(Confidence)를 분석하여 역공학을 이용해 머신러닝 모델이나 학습 데이터를 탈취하는 공격

- Adversarial Examples(적대적 예제)

- 인지하기 쉬운 공격: 특정 패턴, 패치



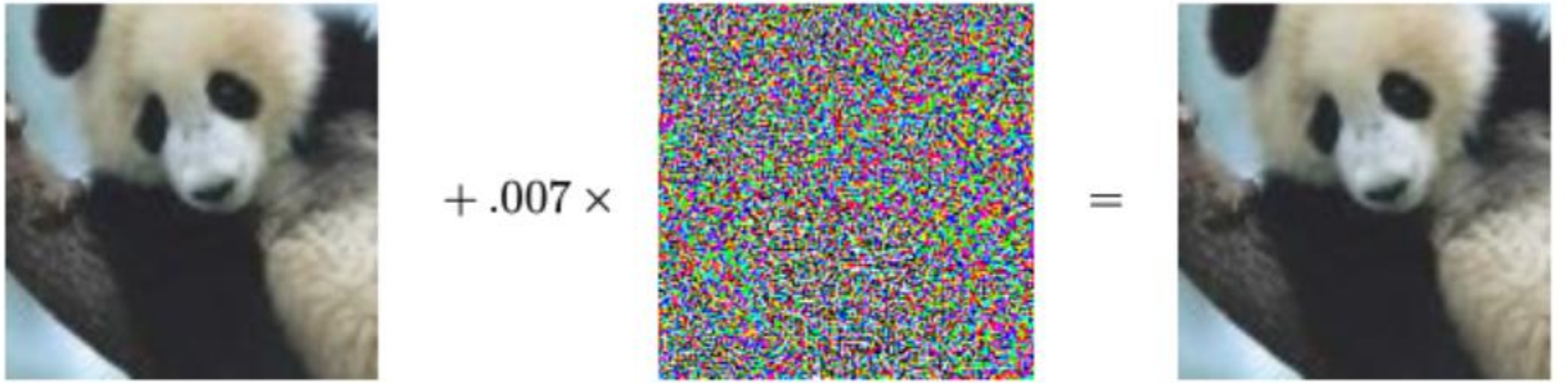
- **인지하기 어려운 공격: 육안으로는 구분할 수 없는 미세한 noise**



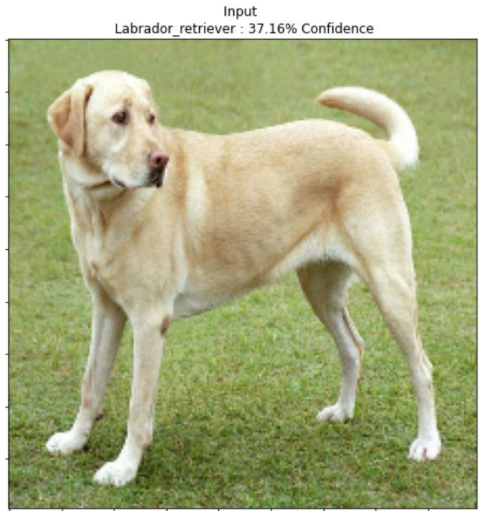
0. Background

- Adversarial Examples(적대적 예제) 생성방법 ➔ **FGSM(Fast Gradient Signed Method)**
  : 신경망의 gradient를 이용해 적대적 샘플을 생성하는 기법
  입력 이미지에 대한 loss function의 gradient를 계산하여 그 **loss를 최대화**하는 이미지를 생성

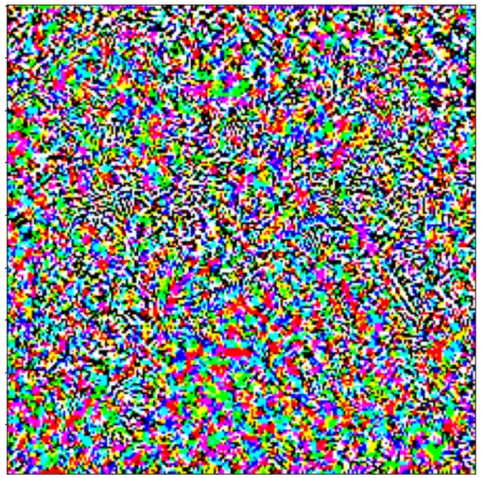
$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$



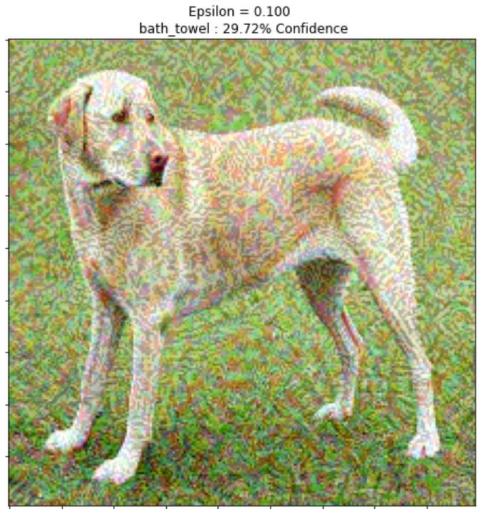
0. Background



Input
Retriever 37.16%



Epsilon = 0.01
Hound 23.81%



Epsilon = 0.1
Hound 29.72%

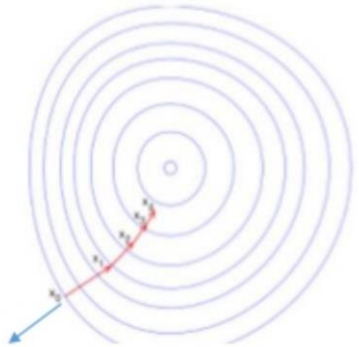
- 왜곡 승수 엡실론(epsilon)을 바꿔가며 다양한 값들을 시도
  - 엡실론의 값이 커질수록 네트워크를 혼란시키는 것이 쉬워짐.
  - 이미지의 왜곡이 점점 더 뚜렷해진다는 단점을 동반

0. Background

- **FGSM(Fast Gradient Signed Method)**

: 학습하는 방향의 반대 방향으로 노이즈 크기 한도까지 노이즈를 만든다면, 가장 잘 오작동 시킬 수 있지 않을까?

$$x^{adv} = x + \varepsilon \operatorname{sign}\left(\nabla_x J(x, y_{true})\right)$$



Loss function and gradient descent

- **I-FGSM(Iterative FGSM)**
- **Basic Iterative Method (BIM)**

: 더 나은 공격 성능을 얻기 위해(stronger) 작은 단계 크기를 **반복적**으로 수행

Algorithm 1 Basic Iterative Method

1: Input: input data x, data label y, max infinity-norm value ϵ, max iteration value n
2: $\mathbf{x}_{adv} = \mathbf{x}$
3: **for** i:=1 **do** n
4: $\quad \mathbf{x}_{adv} = \mathbf{x} + (\epsilon/n)\text{sign}(\nabla_{\mathbf{x}} J(\hat{\theta}_c, \mathbf{x}, y))$

1. Abstract

Deep Neural Networks는 원본 이미지에 작은 perturbation을 추가하여 분류기를 속일 수 있는 adversarial samples에 취약함.

adversarial samples 만드는 방법의 대부분은 "pixel-wise 및 global“ 방식으로 perturbation을 추가
→ 한계를 분석하여 **image processing** based defense and **steganalysis** based detection methods 에 robust 하지 않은 이유를 제시

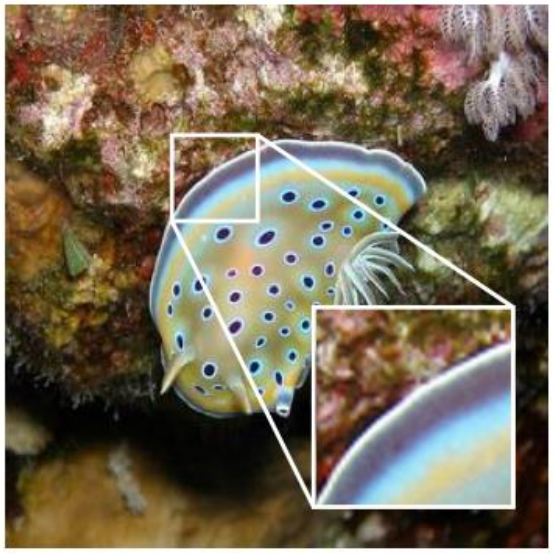
▶한계를 해결하기 위해 superpixel-guided attentional adversarial attack method을 제안
- perturbation이 foreground regions에만 추가되고 각 수퍼 픽셀 내의 픽셀이 동일한 섭동을 갖도록 제한
- 매우 제한된 perturbation 공간에서도 제안된 방법이 원래의 공격 능력을 여전히 preserve(보존) 할 수 있음을 보여줌.
- 적대적 샘플과 소스 이미지 간의 통계적 일관성이 향상 → 적대적 탐지 및 방어 모두에 대해 훨씬 더 견고성을 보여줌.

2. Introduction

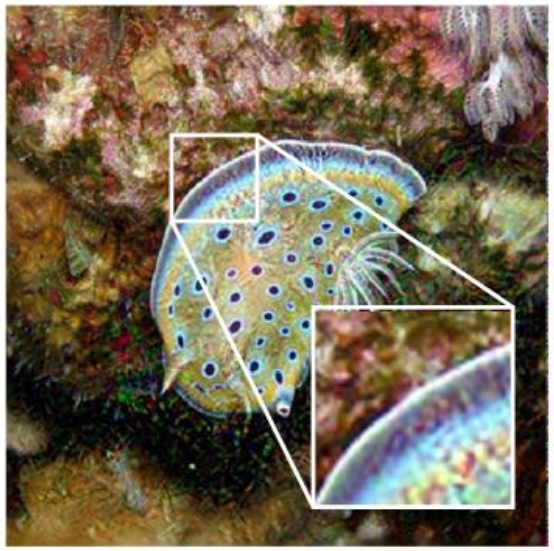
<pixel-wise>

: adversarial perturbations이 각 픽셀에 독립적으로 추가되어 very noisy in most cases.
(neighborhood information 고려x)

Vs. natural images: statistic perspective(통계적 관점) local smoothness property 가짐.
(인접 픽셀이 유사한 픽셀 값을 가짐)



Original Image



I-FGM

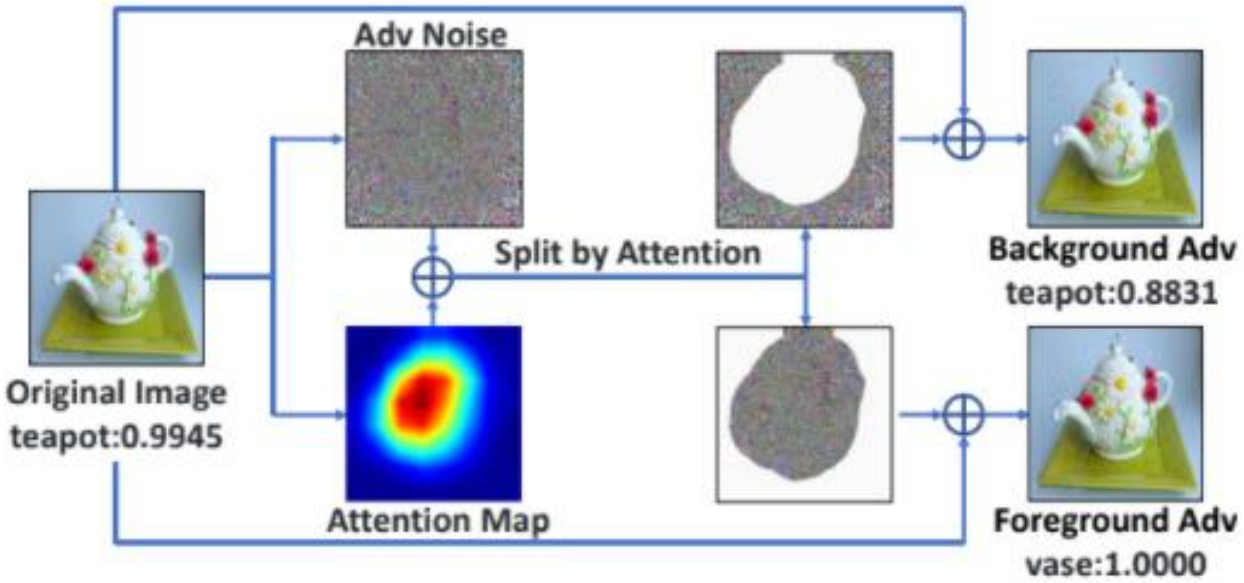
- **Image processing based defense methods**
local way (e.g., smoothing, resizing)으로 적대적 샘플을 처리하고 original noisy adversarial pattern 파괴

- steganalysis based detection methods
steganalysis 기능이 local smoothness 속성을 따르지 않는 small adversarial perturbations을 탐지할 수 있음.

※ Steganalysis란 자료에 암호화된 메시지를 숨기는 스테가노그래피 (Steganography)를 검출하는 기법

2. Introduction

<global>
: 한 이미지의 모든 픽셀을 동일하게 처리하고 모든 픽셀에 섭동을 추가 함.



적대적 잡음을 foreground object와 background 부분
→ 동일한 크기의 두 부분으로 분할

1. foreground에 적대적 잡음을 추가하는 것이 background에 추가하는 것 보다 유용

← target classifier는 activation map에서 보여주는 foreground 부분에만 초점을 맞추기 때문.

2. foreground objects 통계적으로 background 영역보다 더 많은 textures를 포함함 → 배경 영역의 섭동도 훨씬 더 쉽게 감지 할 수 있음.

← Textures가 많으면 perturbation을 더 잘 숨길 수 있기 때문.

2. Introduction

<Pixel-wise>

1. Input image → (traditional super-pixel generation method) → over-segmented superpixel map을 얻음: 각 슈퍼 픽셀의 픽셀이 비슷한 색상을 갖고 local smoothness property을 따름.

2. 슈퍼 픽셀 방식으로 적대적 섭동을 생성

- 각 슈퍼 픽셀 내의 섭동이 동일해야 함.
- local smoothness을 더욱 보장하기 위해 유사한 슈퍼 픽셀을 adaptively merging(적응적으로 병합)하여 원래의 슈퍼 픽셀 방법 [32]을 개선

<Global>

1. class activation map의 auxiliary 보조 정보를 사용하여 "attentional" 방식으로 대체
2. 섭동이 전체 이미지가 아닌 foreground object (i.e., the class activated regions) 에만 추가되도록 제한

⇒ 적대적 섭동 공간은 이전의 " pixel-wise 및 global "방법보다 훨씬 작지만 광범위한 실험을 통해 우리의 방법이 전용 설계로 공격 능력을 유지할 수 있음을 보여줌.

2. Introduction

Point!!

1. 기존의 "pixel-wise and global" 적대 공격 방법의 한계를 명확하게 분석하고, processing and steganalysis 분석에 robust 하지 않은 이유의 근본적인 이유를 밝힘.

2. 분석을 바탕으로 슈퍼 픽셀 유도 주의적 적대 공격 방법을 제안. local smoothness을 보장 할 뿐만 아니라 이미지를 보다 효과적인 방식으로 수정함.

3. 광범위한 실험을 통해 제안 된 방법이 원래의 공격 능력을 보존하고 동시에 우수한 robustness(견고성)을 달성 할 수 있음을 보여줌.

3. Related Work

1. Adversarial Attack
- Gradient based → FGSM
- 더 나은 공격 성능을 얻기 위해 작은 단계 크기를 반복적으로 수행: I-FGSM

➔ super-pixel level and foreground only adversarial sample generation method

2. Adversarial Detection
- steganalysis-based

: 이미지 통계의 변화와 작은 perturbations에 민감하기 때문에, steganalysis 기능을 사용하여 적대적 샘플을 탐지

➔ 적대적 샘플의 견고성을 평가하기 위한 하나의 중요한 지표로도 사용됨.

3. Adversarial Defense

: 적대적 공격으로부터 대상 모델을 보호하는 방법

- 대상 모델에 입력하기 전에 입력 이미지의 적대적 노이즈를 제거하기 위해 preprocessing step하나의 전처리 단계를 추가

➔ 최종 적대적 견고성을 평가

4. Superpixels

: 유사한 속성을 공유하는 픽셀을 그룹화하여 이미지를 과도하게 분할하는 것, 이미지의 중복성을 포착하고 이미지 특징을 계산하기위한 대표적인 기본 요소

➔ 이를 기반으로 적대적 섭동을 추가하여 smooth and robust 한 적대적 샘플을 생성

3. Related Work

4. Superpixel

- **Superpixel**

: 인접해 있는 픽셀들 중에 비슷한 특성(e.g.색상, 밝기)을 가지고 있는 것끼리 묶어서 커다란 픽셀을 만드는 것.

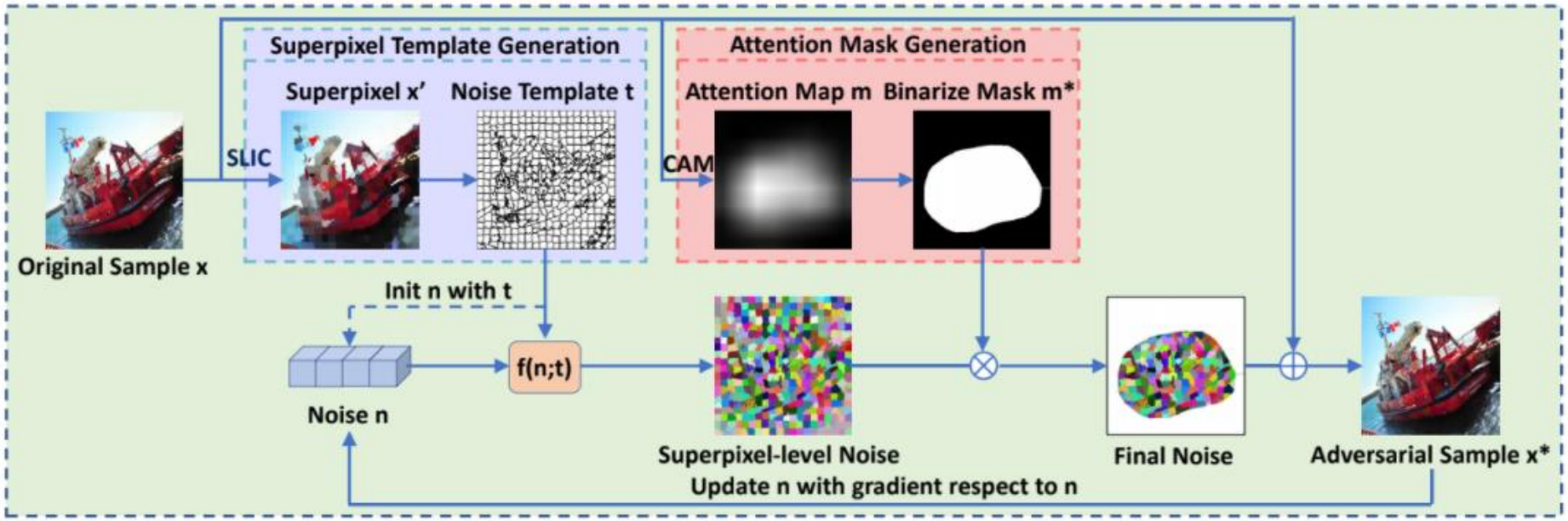
- 의미 공유

: superpixel로서 pixel은 각각 그룹의 공통점을 공유하여 의미를 가지게 됨. Pixel끼리의 상호작용이 이루어짐.

➔ local smoothness 보장

4. Method

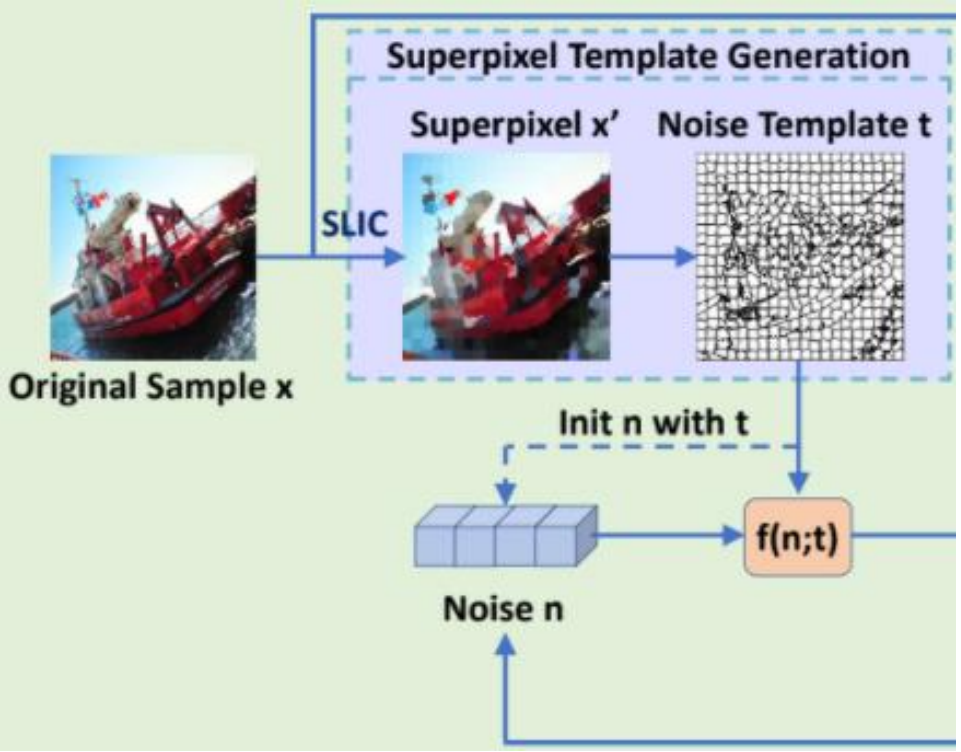
- Pixel-wise → superpixels로 perturbation 동일하게 추가: 슈퍼 픽셀 수준에서도 locally smooth함.
- Global → perturbation이 foreground object에만 추가되도록 하는 foreground attention map 사용.



1. Template Generation
2. Attention Mask
3. Adversarial Perturbation Generation

4. Method

1. Template Generation



- grid steps로 픽셀을 샘플링하여 템플릿을 생성하는 대신, superpixel algorithms(SLIC)을 사용하여 modification template으로 segmentation map을 얻음.

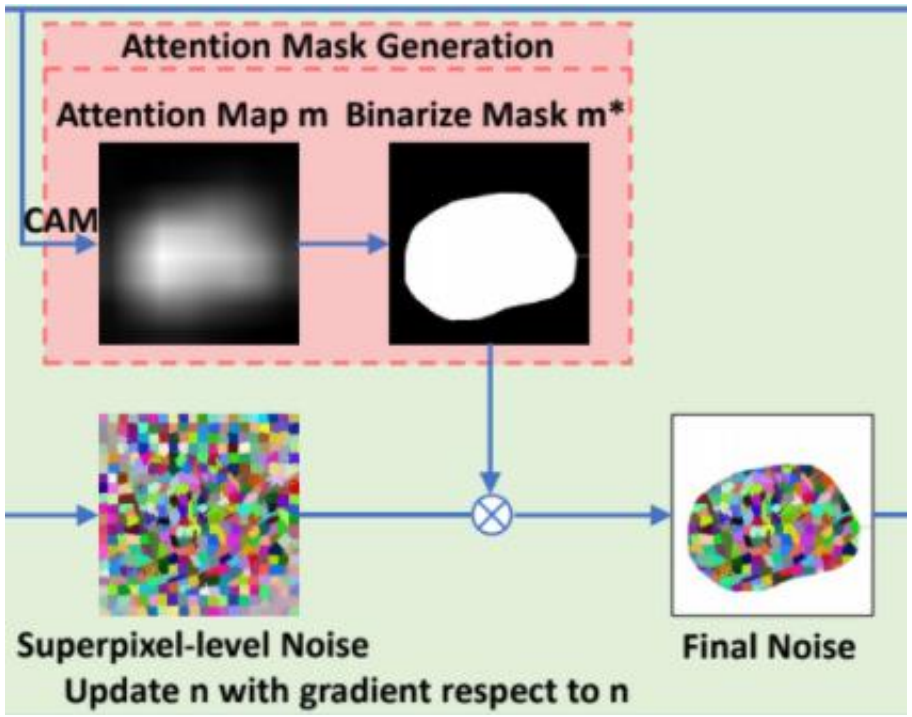
→ local smoothness을 보장하고 소스 이미지와 적대적 샘플 간의 통계적 차이를 줄이는 데 도움이 됨.

- super pixel-guided noise template **t**를 생성하고 길이가 superpixel 픽셀 수와 같은 adversarial noise **n**을 초기화 함.

※ Simple Linear Iterative Clustering (SLIC) 알고리즘
K-means 군집화 방법을 적용하여 빠르고 효율적인 슈퍼픽셀을 생성해 내는 방법 중의 하나이다. 균등한 슈퍼픽셀들을 만들어 내기 때문에 격자 구조를 유지하면서 픽셀처럼 슈퍼픽셀을 다루기가 쉽고, 적은 연산량을 갖기 때문에 영상 분할의 전처리 과정에 적합

4. Method

2. Attention Mask



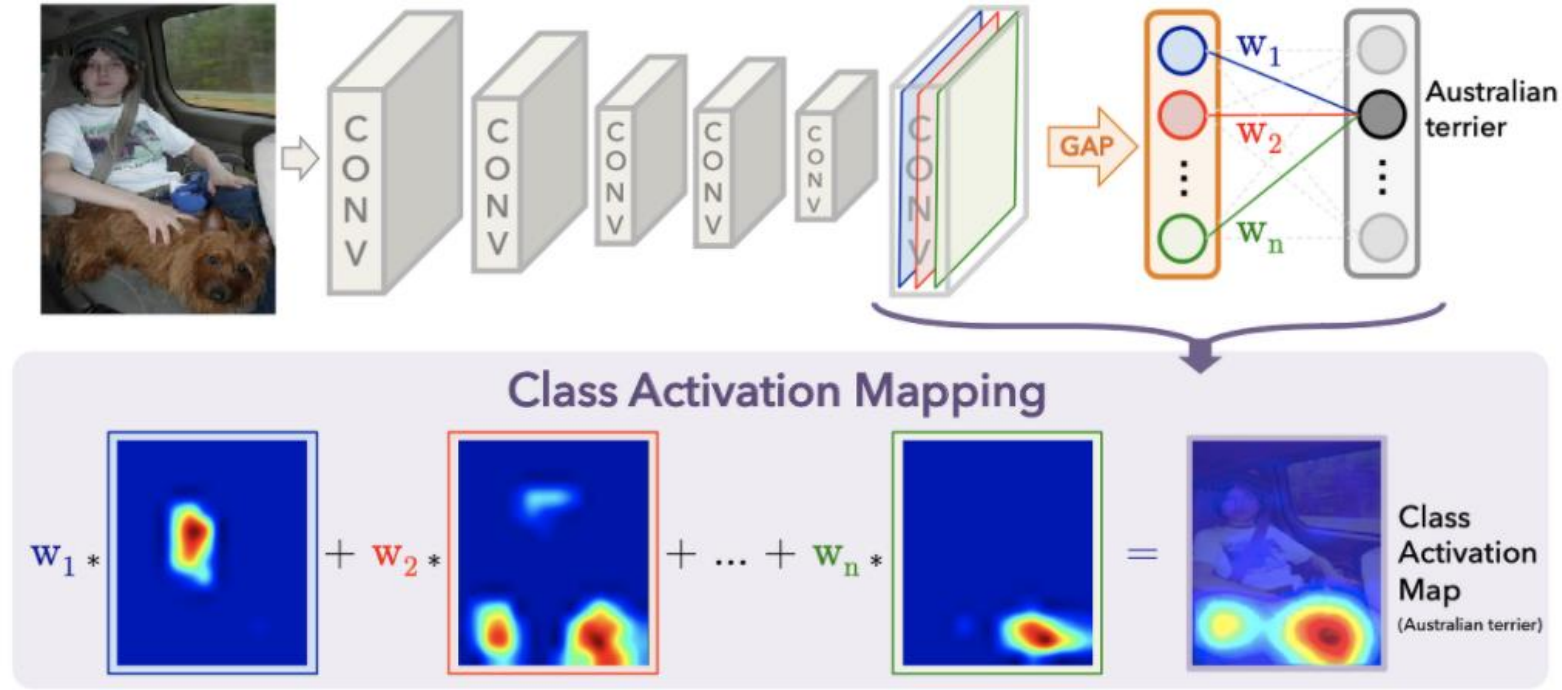
- 입력 이미지의 attention map를 생성하기 위해 Class Activation Mapping (CAM) 방법을 사용
- CAM은 convolutional feature maps에 출력 레이어를 다시 투영하여 map을 계산

- attention map → 모든 픽셀 값은 0, 1을 갖는 이진화로 변환

m: attention map
m*: 이진화 된 attention map
→ 이진화 된 attention map m *을 마스크로 사용하여 적대적 잡음을 잘라내어 attentional attack 공격 수행

4. Method

2. Attention Mask → Class Activation Mapping (CAM)

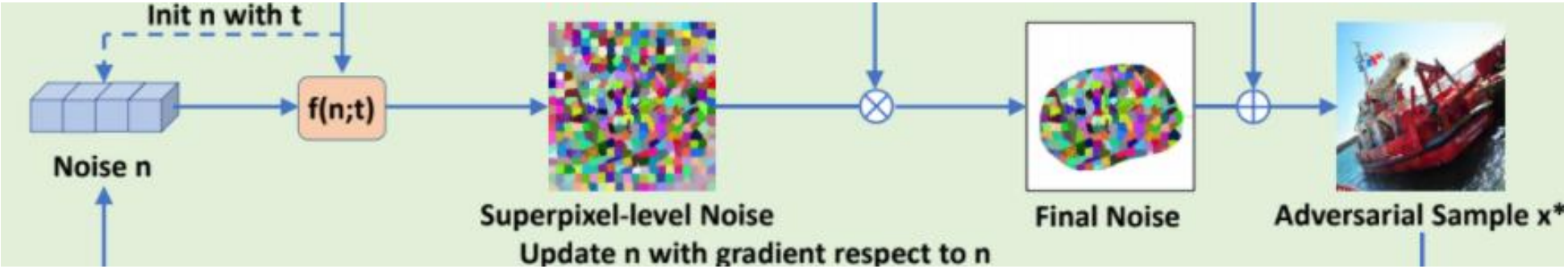


- CNN의 구조: Input - Conv Layers - FC Layers
- CNN 의 마지막 레이어를 FC-Layer 로 Flatten 시킬 때, 그 Convolution 이 가지고 있던 각 픽셀들의 위치 정보를 잃게 됨. → CNN 이 무엇을 보고 그 클래스를 그 클래스라고 판별했는지 알 수가 없음.

- 마지막 컨볼루션을 FC-layer 로 바꾸는 대신에, GAP (Global Average Pooling) 을 적용 → 별다른 추가의 지도학습 없이 CNN 이 특정 위치들을 구별하도록 만들 수 있음. Heat Map 을 통해 CNN 이 어떻게 그 이미지를 특정 클래스로 예측했는지를 이해O

4. Method

3. Adversarial Perturbation Generation

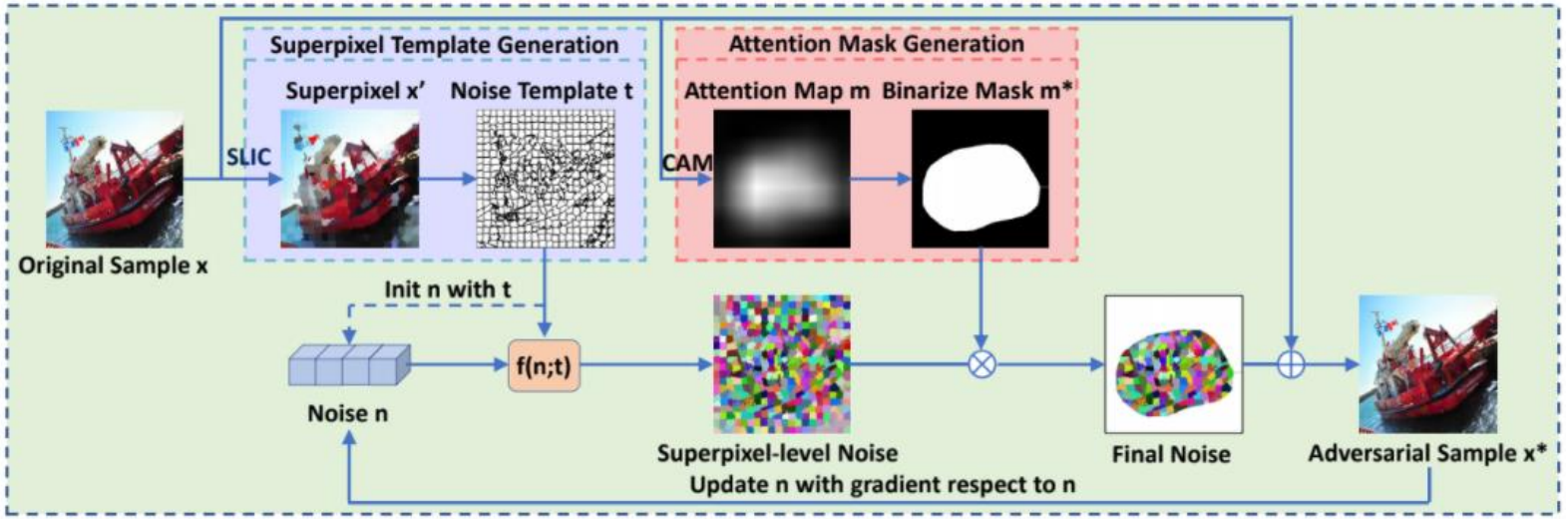


I-FGM(baseline) → superpixel및 attention map를 사용하여 SAI-FGM
I-FGM의 경우; 입력 이미지로 손실 함수의 기울기를 계산하여 적대적 노이즈가 생성
수퍼 픽셀 내의 모든 픽셀의 평균 또는 최대 그래디언트를 그래디언트로 사용하면 효율적이거나 효과적이지 않음.
→ ablation part에서 증명됨.

⇒ optimization(최적화) based methods 를 통해 적대적 잡음을 생성하는 대체 방법을 제안!

4. Method

3. Adversarial Perturbation Generation



$$\mathbf{x}_0^{\mathbf{adv}} = \mathbf{x} + f(\mathbf{n_0}; \mathbf{t})$$
$$\mathbf{x}_\mathbf{i}^{\mathbf{adv}} = \mathbf{x} + Scale_\epsilon\{Crop_\mathbf{m}(f(\mathbf{n_i}; \mathbf{t}))\} \quad (3)$$

$$\mathbf{n_{i+1}} = \mathbf{n_i} + \alpha \cdot \frac{\nabla_\mathbf{n} L(\mathbf{x}_\mathbf{i}^{\mathbf{adv}}, y, \theta)}{\|\nabla_\mathbf{n} L(\mathbf{x}_\mathbf{i}^{\mathbf{adv}}, y, \theta)\|_2} \quad (4)$$

1. 길이가 슈퍼 픽셀 수와 같은 노이즈 벡터 n을 초기화하고 n에 대한 기울기를 직접 계산
2. 각 반복 동안, 하나의 매핑 함수 f는 n을 t(슈퍼 픽셀 레벨 노이즈 템플릿)에 채우는 데 사용되어 채워진 노이즈 f (n; t)를 얻음.
3. f (n; t)는 원래 샘플 x에 추가하기 전에 자르고 임계 값 ϵ로 조정됨.

Crop 및 Scale: attention map m을 기반으로 하는 crop operation과 perturbation scale factor ϵ(섭동 배율 계수)를 기반으로 하는 배율 작업을 각각 나타냄.

5. Experiments

<두 가지 핵심 제약>

super pixel level & only added to attentional regions → 이러한 제한된 공간을 고려할 때,

"원래의 공격 능력을 유지할 수 있는지 여부"
"제안된 방법이 adversarial robustness을 높일 수 있는지 여부"

➔ 제안 된 방법을 **attack ability & attack robustness** 의 두 가지 측면에서 평가

5. Experiments

5-1. Attack Ability Comparison(공격 능력 비교)

: **white-box공격** 성공률과 **black-box 공격** 성공률을 포함한 ImageNet dataset의 기본 공격 성능을 비교

Attack	Inc-v3*	Inc-v4	IncRes-v2	Inc-v3$_{adv}$
FGM	**84.00**	54.75	**56.75**	56.05
SA-FGM(**Ours**)	73.80	**56.20**	51.55	**56.75**
I-FGM	99.80	40.65	38.00	30.70
SAI-FGM(**Ours**)	**100.00**	**68.35**	**64.95**	**65.05**
MI-FGM	99.90	66.45	**67.95**	62.40
M-SAI-FGM(**Ours**)	**99.95**	**68.50**	66.10	**67.45**

Table 1. The attack success rate (%) of adversarial attack on the ImageNet [4] dataset. * indicates the white-box attacks.

- single step attack method(단일 단계 공격 방법) FGM이 SA-FGM보다 성능이 더 우수함.

∵ 제한된 섭동 공간의 영향이 불가피함.

- multiple step attack method(다단계 공격 방식) I-FGM 또는 MI-FGM을 사용하면 SAI-FGM과 M-SAI-FGM이 유사한 화이트 박스 공격 성공률 (거의 100 %)을 달성

→ 논문에서 제시한 방법은 여러 단계를 사용하면 공격 능력을 보존 할 수 있음.

- Black-box 공격의 경우 대부분 논문에서 제시한 방법이 기준 방법보다 더 나은 성능을 보임.

5. Experiments

5-1. Attack Ability Comparison(공격 능력 비교)

: **white-box공격** 성공률과 **black-box 공격** 성공률을 포함한 ImageNet dataset의 기본 공격 성능을 비교

white-box

black-box

- White-box attack: 모델 아키텍처 및 매개 변수를 포함하여 공격 할 target model에 대한 완전한 지식이 있는 경우 (back-propagated gradients로 적대적 샘플을 직접 생성 가능)
- Gray-box attack: 모델에 대한 전체 액세스 권한이 있지만 이미지를 모델에 제공하기 전에 unknown input transformations이 있는 경우
- Black-box attack: target model에 대한 지식이 없음. Back-progagate할 수 없기에, 다른 입력에 대한 출력의 변화를 관찰해서 loss function의 기울기를 추정해야 함.

5. Experiments

5-2. Attack Robustness Comparison 공격 견고성 비교

1> Steganalysis 기반 Detection 방법에 대한 Robustness(견고성)

- Steganalysis 기반 적대적 탐지 방법; 적대적 샘플을 탐지하기 위한 주요 지표로 steganalysis 기능을 사용.
- δ(perturbation scales), β(superpixel Size)

Generation method	$\delta = 2$	$\delta = 4$	$\delta = 6$	$\delta = 8$
FGM	94.32	95.59	96.28	97.09
I-FGM	94.11	94.74	95.45	96.01
SAI-FGM($\beta = 0$)	75.50	82.10	85.00	87.20
SAI-FGM($\beta = 2$)	72.40	78.20	82.00	84.20
SAI-FGM($\beta = 4$)	**63.30**	**72.10**	**76.40**	**80.90**

Table 2. Detect rate (%) of steganalysis-based detection. Lower success rate indicate the adversarial samples are more robust.

- SAI-FGM은 다른 δ에 대해 large margin으로 기준 방법을 능가함.

⇒ 이것은 제안 된 슈퍼 픽셀 유도 적대 샘플이 픽셀 단위 샘플보다 우월함을 증명

- 인접한 유사한 슈퍼 픽셀을 결합함으로써 SAI-FGM은 모든 δ에 대한 감지 성공률을 더욱 줄일 수 있음.

→ 인접한 유사한 슈퍼 픽셀이 병합 될 때 적대적 샘플과 소스 이미지 간의 통계적 차이가 크게 감소하기 때문.

5. Experiments

2> 이미지 처리 기반 Defense 방법에 대한 Robustness(견고성)

: 이미지 처리 기술을 활용 → target model에 이미지를 제공하기 전에 adversary을 제거

섭동의 잡음 분포가 실제 이미지의 분포와 일치하지 않는다는 점 기반으로,
Resizing, JPEG 압축, DNN-Oriented JPEG Compression(DNN 지향 JPEG 압축), pooling, total variance minimization 총 분산 최소화 (TVM) 및 Bit-depth Reduction 를 포함하여 견고성 평가를 위해 다양한 이미지 처리 기술을 사용함.

→ Image processing이 아닌 적대적 샘플의 공격으로 인해 mis-classification이 발생하도록 해야 함.
→ ImageNet 데이터 셋에서 이미지 처리 후 올바르게 분류 할 수 있는 2000 개의 robust images를 선택

5. Experiments

① Resizing

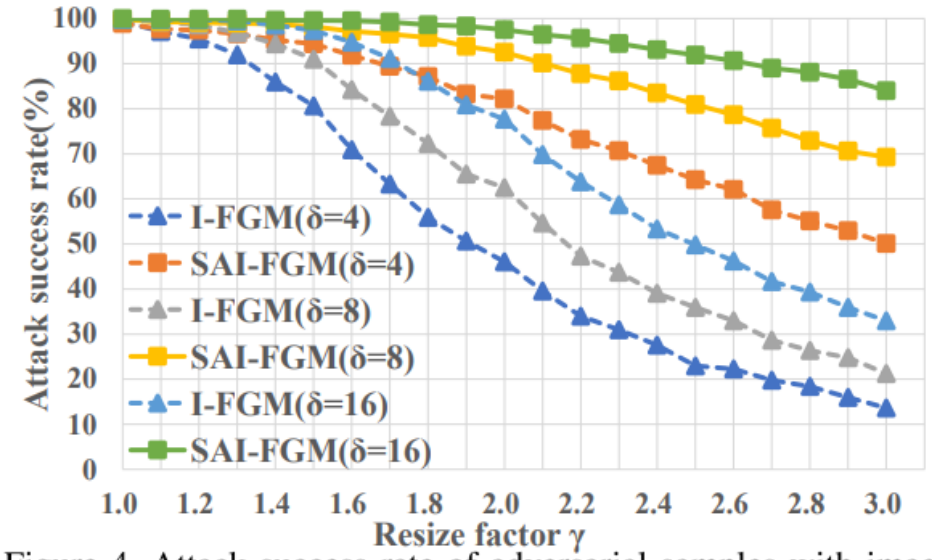


Figure 4. Attack success rate of adversarial samples with image resizing. Resize factor γ varies from 1.0 to 3.0.

- local interpolation(보간)을 통해 적대적 샘플의 효율성을 줄일 수 있음.

→ 크기가 HxW 인 adversarial sample적대적 표본이 주어지면 먼저 크기를 $\frac{H}{\gamma} \times \frac{W}{\gamma}$로 축소 한 다음 원래 크기 HxW로 다시 확대

- 그림 4는 the scale factor γ를 1에서 3으로 변경할 때 공격 성공률 곡선
- 다른 perturbation scales 섭동 척도 δ = 4,8,16 및 다른 기준 방법의 경우, 슈퍼 픽셀 유도주의 버전 "SA *"은 항상 훨씬 더 강력하고 공격 성공률이 더 높음.

5. Experiments

② JPEG 압축 및 DNN-Oriented(DNN 지향) JPEG 압축

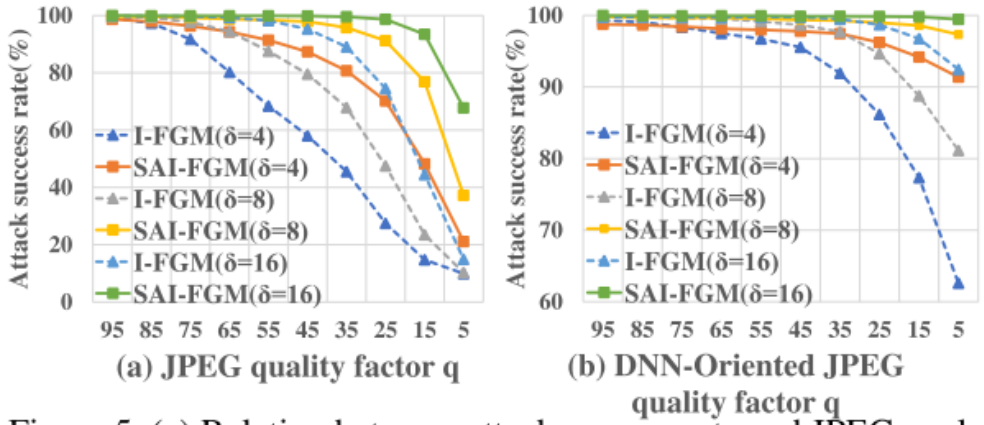


Figure 5. (a) Relation between attack success rate and JPEG quality factor q. (b) Relation between attack success rate and DNN-Oriented JPEG quality factor q.

- local smoothness property을 기반으로 JPEG 압축
유사한 시각적 품질을 보장하면서 이미지의 high-frequency 구성 요소를 크게 압축함.
→ adversarial perturbations도 high frequency이기 때문에 대부분의 경우 JPEG 압축으로 인해 공격 능력이 약화됨.

- 이미지가 high quality factor **q** = 95로 약간만 압축 된 경우, I-FGM과 SAI-FGM은 여전히 상당히 높은 공격 성공률
- q가 95에서 5로 감소함에 따라 I-FGM의 공격 성공률은 빠르게 감소하고 SAI-FGM은 여전히 높은 성공률 유지
- 특히 저품질 영역 (q < 25)에서는, 저품질 압축이 이미지의 대부분의 디테일을 제거하므로 I-FGM과 SAI-FGM의 성능이 크게 저하되지만 우리의 방법은 여전히 I-FGM보다 성능이 좋습니다.

5. Experiments

③ Pooling

Operation	Threshold	Attack method	Kernel Size		
			3	5	7
AVG	$4\sqrt{N}$	I-FGM	73.67	20.55	7.07
		SAI-FGM	**93.27**	**65.14**	**36.91**
	$8\sqrt{N}$	I-FGM	85.94	31.21	11.47
		SAI-FGM	**97.44**	**79.96**	**58.03**
	$16\sqrt{N}$	I-FGM	95.13	46.02	20.36
		SAI-FGM	**99.57**	**91.48**	**77.35**
MAX	$4\sqrt{N}$	I-FGM	41.45	16.67	13.47
		SAI-FGM	**70.37**	**39.20**	**34.67**
	$8\sqrt{N}$	I-FGM	54.79	27.29	20.39
		SAI-FGM	**85.10**	**61.64**	**55.73**
	$16\sqrt{N}$	I-FGM	68.36	37.06	30.38
		SAI-FGM	**93.85**	**83.29**	**79.82**

Table 3. The gray-box attack success rate (%) of adversarial samples after average/max pooling operation on the ImageNet [4] dataset with different ϵ and different kernel size l.

- pooling: 각 그리드 커널 내에서 최대 또는 평균 픽셀을 샘플링하여 수행
- smoothing/resizing과 유사하게 adversary의 original distribution을 변경

- 논문에서 제안한 방법은 항상 다양한 kernel sizes and perturbation levels 에서 기준 방법보다 훨씬 나음.
- 예를 들어, δ = 16이고 커널 크기가 5 일 때 SAI -FGM은 여전히 79.82 %의 공격 성공률을 달성 할 수 있는 반면 기준 I-FGM은 30.38 %의 성공률 만 가지고 있으며 약 50 %로 훨씬 뒤처짐.

5. Experiments

5-3. Visual Results

Figure 7. Some visual comparison about the adversarial samples generated by the baseline method I-FGM (top) and our SAI-FGM (bottom) with $\delta = 4$ (left) and $\delta = 16$ (right) respectively.

적대적 샘플은 전반적으로 비슷해 보이지만,
논문에서 제시한 방법은 the adversarial perturbations이 more attentional and superpixel-wise smooth.

6. Conclusion

Figure 1. Some visual results about the adversarial samples generated by our SAI-FGM with $\delta = 4$ (left) and $\delta = 8$ (right) respectively. For each column, we show the original image(left), the attention mask(middle) and adversarial samples(right) respectively.