

꼼꼼한 딥러닝 논문 리뷰와 코드 실습

Deep Learning Paper Review and Code Practice

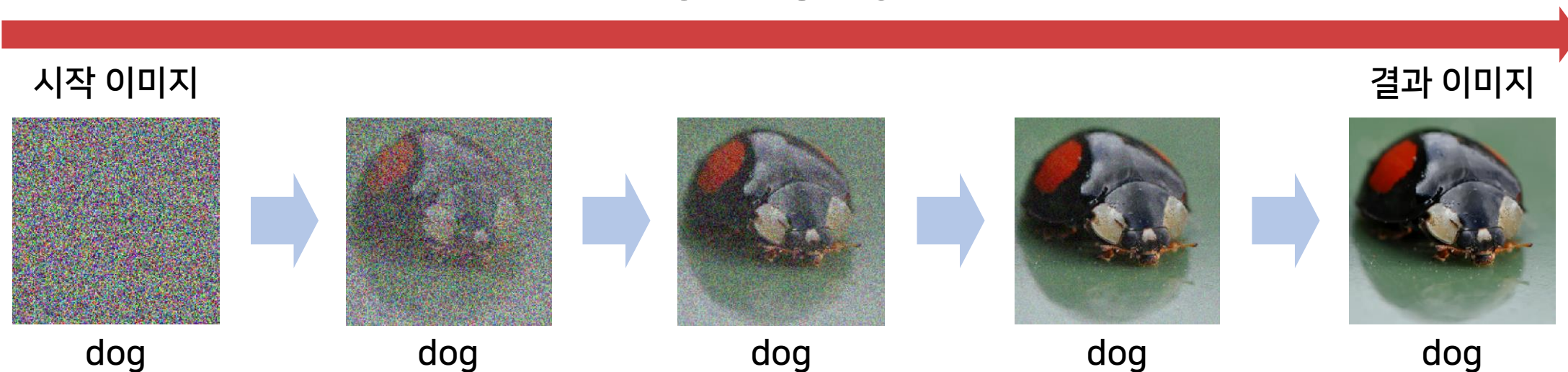
나동빈(dongbinna@postech.ac.kr)

Pohang University of Science and Technology

Boundary Attack (ICLR 2018)

- 본 논문은 뉴럴 네트워크를 공격하는 기법인 Boundary Attack (BA)을 제안합니다.
- Boundary Attack의 특징은 무엇인가요?
 1. 공격 대상 네트워크의 하나의 레이블에 대한 예측 결과만 알 수 있다면 공격할 수 있습니다.
 2. 공격자가 별도의 대체 네트워크(substitute network)를 학습하지 않아도 공격할 수 있습니다.
 3. 이전까지 제안되었던 다양한 방어 기법을 다시금 뚫을 수 있습니다.

공격 수행 과정



연구 배경: 적대적 예제 (Adversarial Examples)

- Adversarial examples
 - 인간의 눈에 띄지 않게 변형된 데이터로, 뉴럴 네트워크의 부정확한 결과를 유도합니다.
 - 기존의 많은 공격 방법은 손실(loss) 함수를 이미지(입력)로 미분하여 이미지를 변경하는 방식을 따릅니다.



x

(Tabby Cat)

$+ \epsilon *$



Perturbation (δ)

$=$

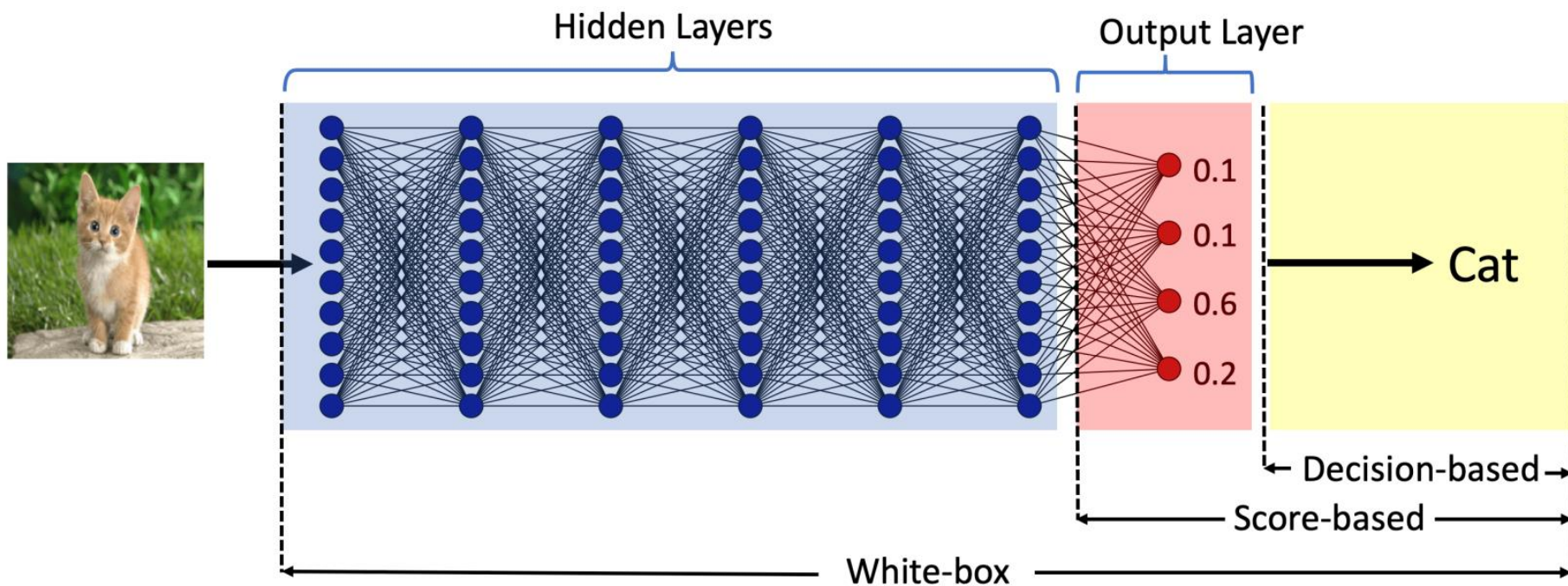


x^*

(Guacamole)

Threat Model에 따른 접근 가능한 컴포넌트(Components) 비교

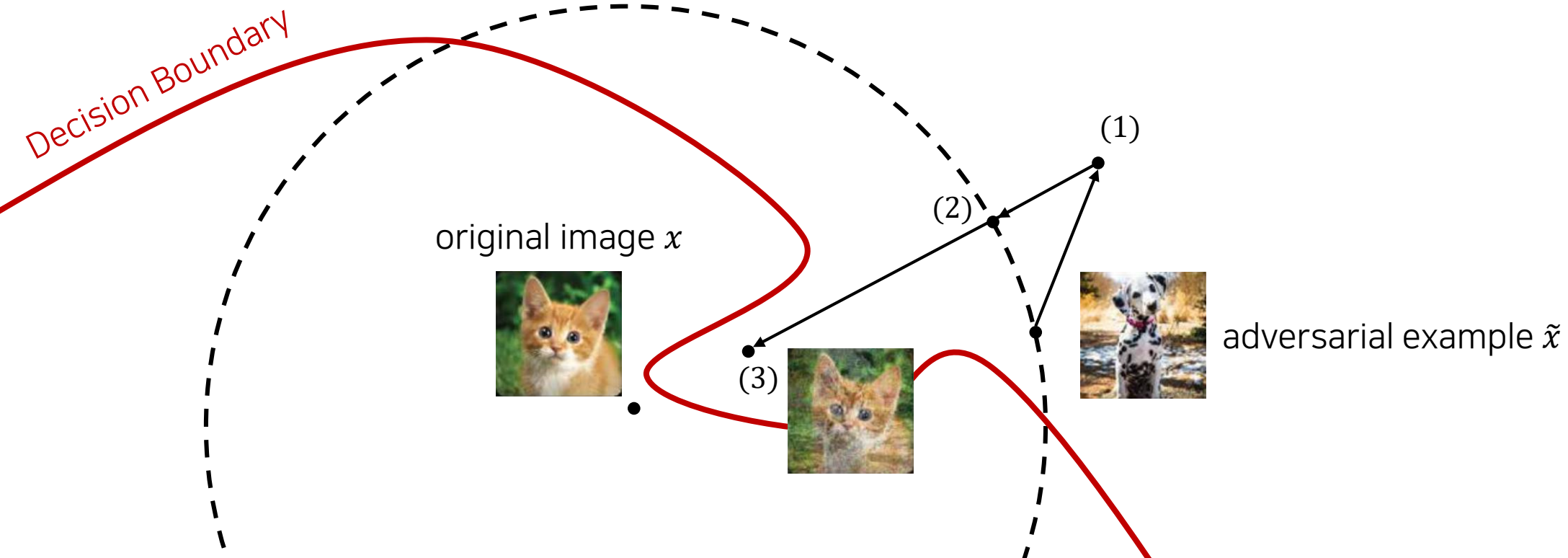
- White-box threat model: 공격자가 전체 모델에 접근이 가능하다고 가정합니다.
- Score-based threat model: 공격자가 모델의 출력 레이어에 접근이 가능하다고 가정합니다.
- Decision-based threat model: 공격자가 하나의 예측된 레이블에만 접근이 가능하다고 가정합니다.



* HopSkipJumpAttack: A Query-Efficient Decision-Based Attack (2020 S&P)

알고리즘 소개: Boundary Attack

- **(Initialization)** The Boundary Attack needs to be initialized with a sample that is already adversarial.
 1. Sample from a Gaussian distribution $\eta_i^k \sim N(0, 1)$ and then rescale and clip the sample.
 2. Project η^k onto a sphere around the original image x such that $d(x, \tilde{x}^{k-1} + \eta^k) = d(x, \tilde{x}^{k-1})$.
 3. Make a small movement towards the original image.



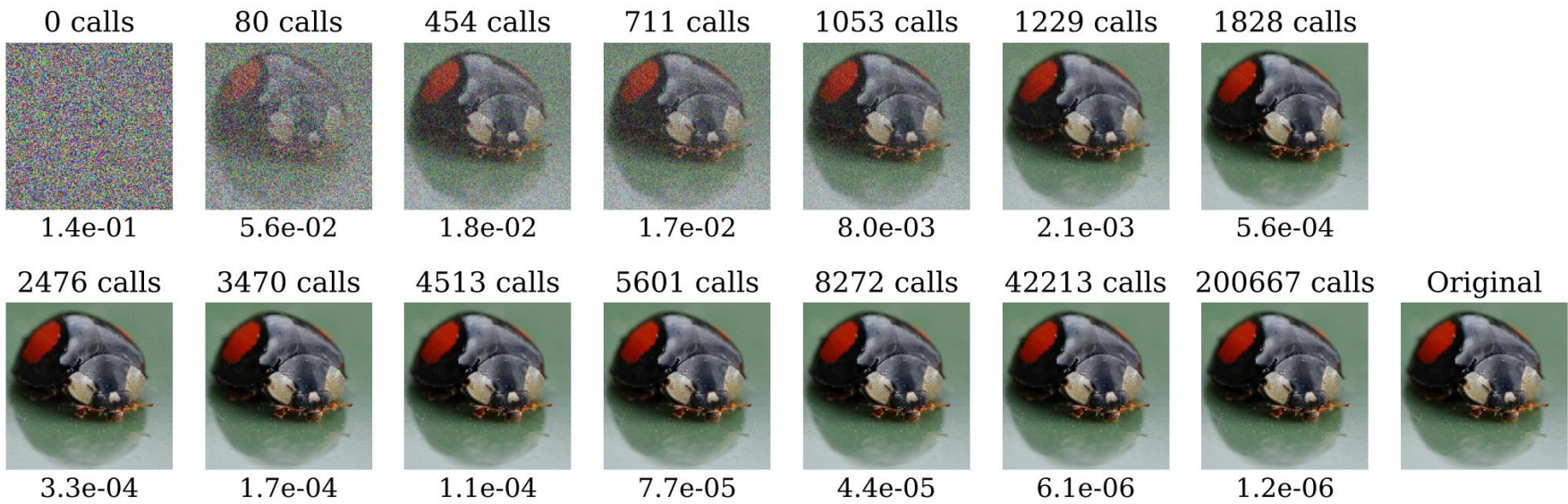
알고리즘 소개: Boundary Attack

- Boundary Attack (BA)를 간략한 형태의 소스 코드로 나타내면 다음과 같습니다.
 - 공격 도메인에 따라 다양한 proposal distribution을 사용할 수 있습니다.

```
Data: original image  $\mathbf{o}$ , adversarial criterion  $c(\cdot)$ , decision of model  $d(\cdot)$   
Result: adversarial example  $\tilde{\mathbf{o}}$  such that the distance  $d(\mathbf{o}, \tilde{\mathbf{o}}) = \|\mathbf{o} - \tilde{\mathbf{o}}\|_2^2$  is minimized  
initialization:  $k = 0$ ,  $\tilde{\mathbf{o}}^0 \sim \mathcal{U}(0, 1)$  s.t.  $\tilde{\mathbf{o}}^0$  is adversarial;  
while  $k < \text{maximum number of steps}$  do  
    draw random perturbation from proposal distribution  $\boldsymbol{\eta}_k \sim \mathcal{P}(\tilde{\mathbf{o}}^{k-1})$ ;  
    if  $\tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$  is adversarial then  
        set  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1} + \boldsymbol{\eta}_k$ ;  
    else  
        set  $\tilde{\mathbf{o}}^k = \tilde{\mathbf{o}}^{k-1}$ ;  
    end  
     $k = k + 1$   
end
```

[Algorithm] Minimal version of the Boundary Attack.

Boundary Attack: Untargeted Attack 수행 결과

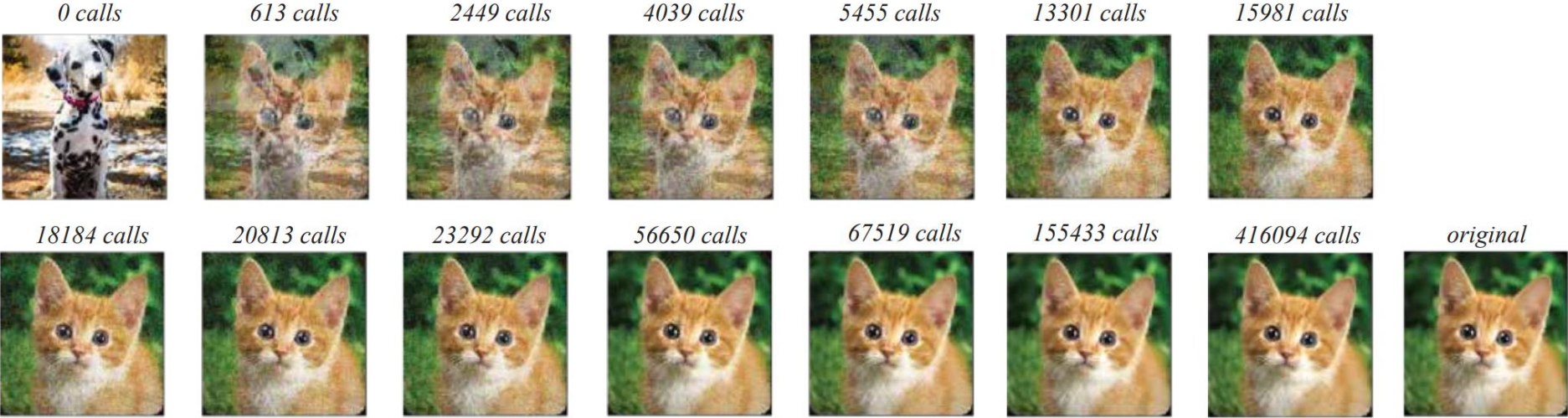


[Figure] Example of an untargeted attack.

		ImageNet				
	Attack Type	MNIST	CIFAR	VGG-19	ResNet-50	Inception-v3
FGSM	gradient-based	4.2e-02	2.5e-05	1.0e-06	1.0e-06	9.7e-07
DeepFool	gradient-based	4.3e-03	5.8e-06	1.9e-07	7.5e-08	5.2e-08
Carlini & Wagner	gradient-based	2.2e-03	7.5e-06	5.7e-07	2.2e-07	7.6e-08
Boundary (ours)	decision-based	3.6e-03	5.6e-06	2.9e-07	1.0e-07	6.5e-08

[Table] Comparison of the untargeted attacks.

Boundary Attack: Targeted Attack 수행 결과



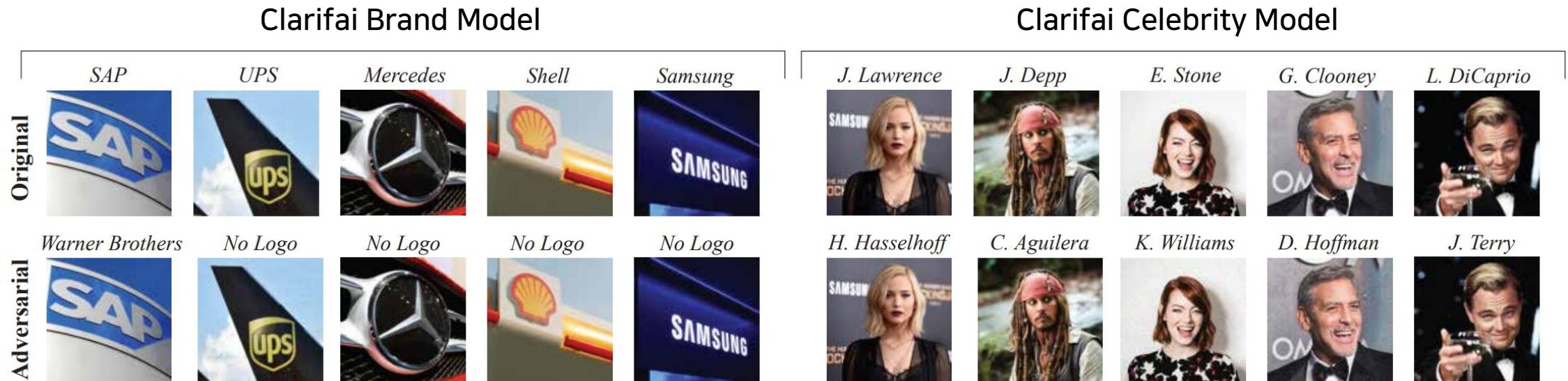
[Figure] Example of an targeted attack.

	Attack Type	MNIST	CIFAR	VGG-19
Carlini & Wagner	gradient-based	4.8e-03	3.0e-05	5.7e-06
Boundary (ours)	decision-based	6.5e-03	3.3e-05	9.9e-06

[Table] Comparison of the targeted attacks.

Real-World Applications 공격 시나리오

- Clarifai 서비스의 두 가지 black-box 모델에 대하여 공격을 수행했습니다.
 - Clarifai Brand Model: 500개 이상의 클래스를 가지는 분류 모델
 - Clarifai Celebrity Model: 10,000개 이상의 클래스를 가지는 분류 모델



[Figure] Adversarial examples generated by the Boundary Attack.

- 일반적인 ImageNet 모델에 비해 perturbation의 크기가 크게 형성되지만, 충분히 reasonable한 공격 결과입니다.