

# 꼼꼼한 딥러닝 논문 리뷰와 코드 실습

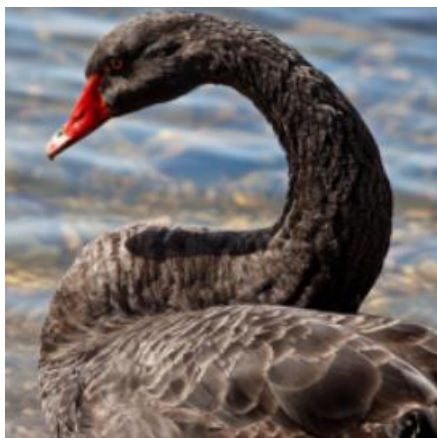
Deep Learning Paper Review and Code Practice

나동빈([dongbinna@postech.ac.kr](mailto:dongbinna@postech.ac.kr))

Pohang University of Science and Technology

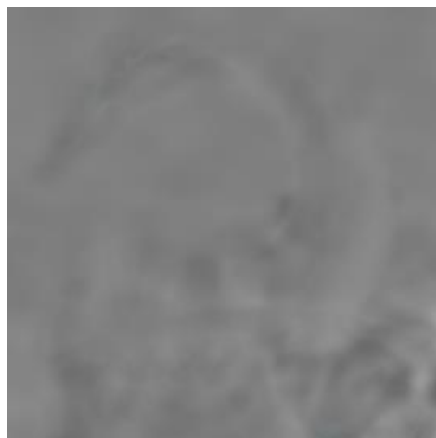
# Shadow Attack (ICLR 2020)

- 본 논문은 뉴럴 네트워크를 공격하는 새로운 공격 유형인 **Shadow attack(그림자 공격)**을 제안합니다.
- Shadow attack의 특징은 무엇인가요?
  1. Imperceptibility: 정상적인 이미지처럼 보입니다.
  2. Misclassification: 타겟 클래스로 잘못 분류하도록 유도합니다.
  3. Strongly certified: 높은 인증 반경(certificate radius)를 가집니다.



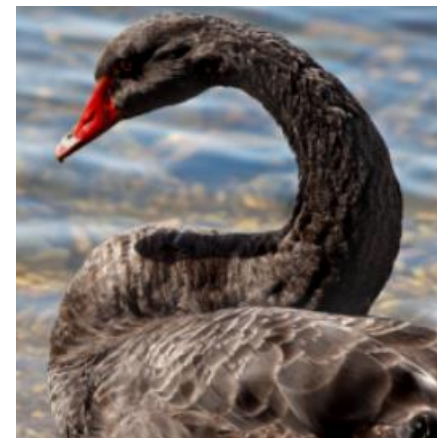
Natural image ( $x$ )

+



**Shadow** perturbation ( $\delta$ )

=



Adversarial example ( $x + \delta$ )

## 연구 배경: 적대적 예제 (Adversarial Examples)

- Adversarial examples
  - 인간의 눈에 띄지 않게 약간 변형된 데이터로, 뉴럴 네트워크의 부정확한 결과를 유도합니다.



$x$   
(Tabby Cat)

$+ \epsilon *$



*Perturbation ( $\delta$ )*

$=$



$x^*$   
(Guacamole)

## 연구 배경: 적대적 학습 (Adversarial Training)

- Adversarial training

- 뉴럴 네트워크를 강건하게 만들기 위해 adversarial example을 학습 데이터로 이용하는 방법입니다.

$$\min_{\theta} E_{(x,y) \in \mathcal{X}} \left[ \max_{\delta \in S} L(x + \delta; y; \theta) \right]$$

- How to solve the inner problem?

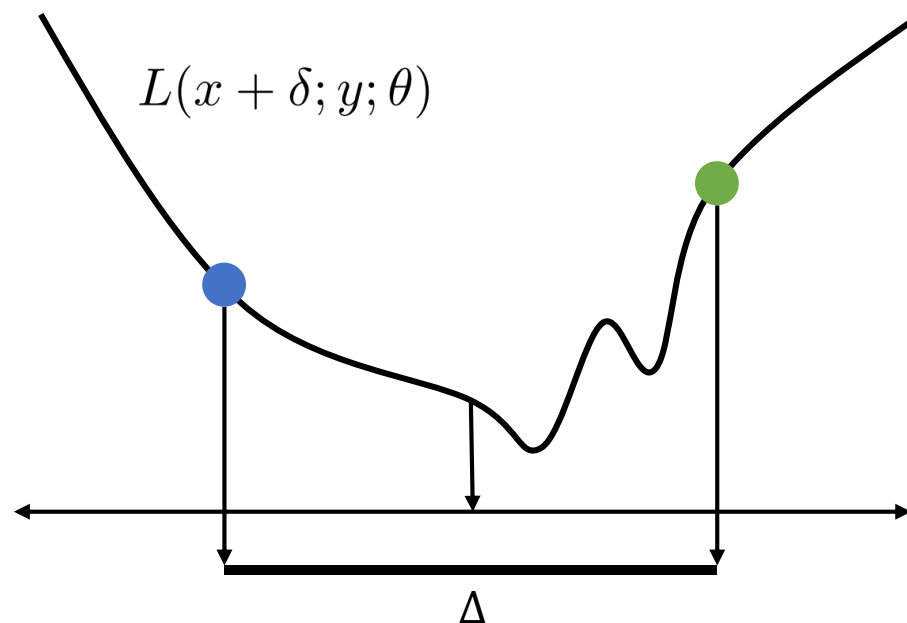
Can be calculated by *PGD*

● : Local solution

● : Global solution

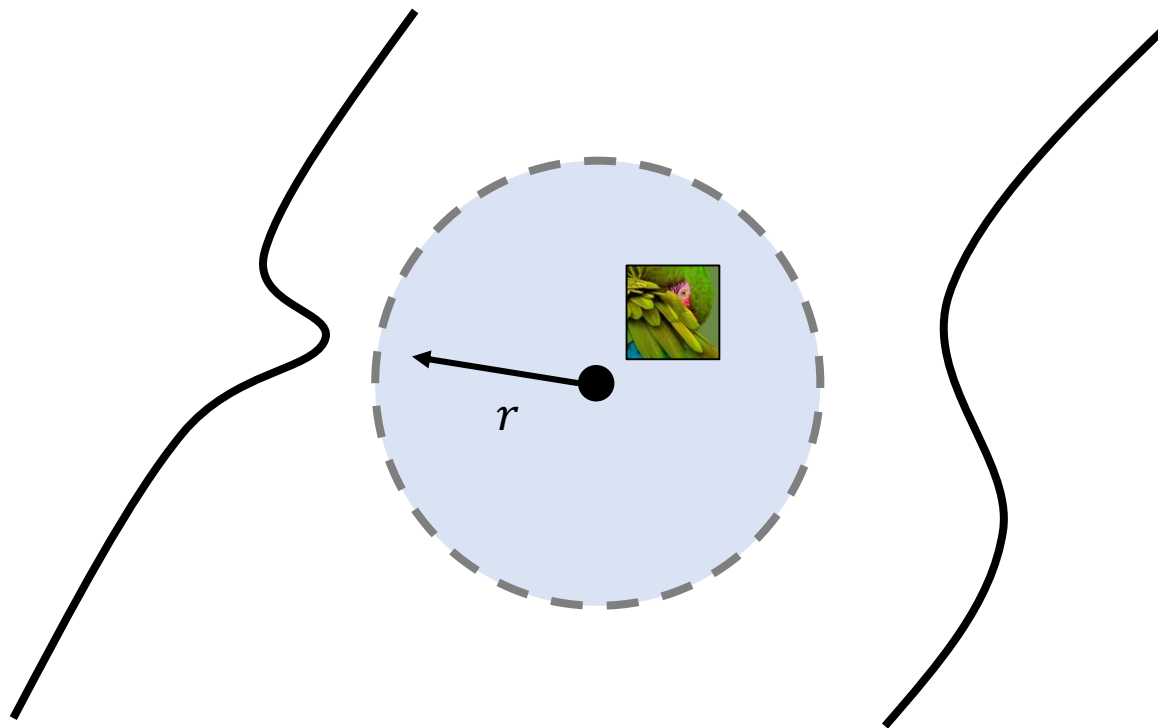


하지만, 더 강한 공격이 등장한다면?



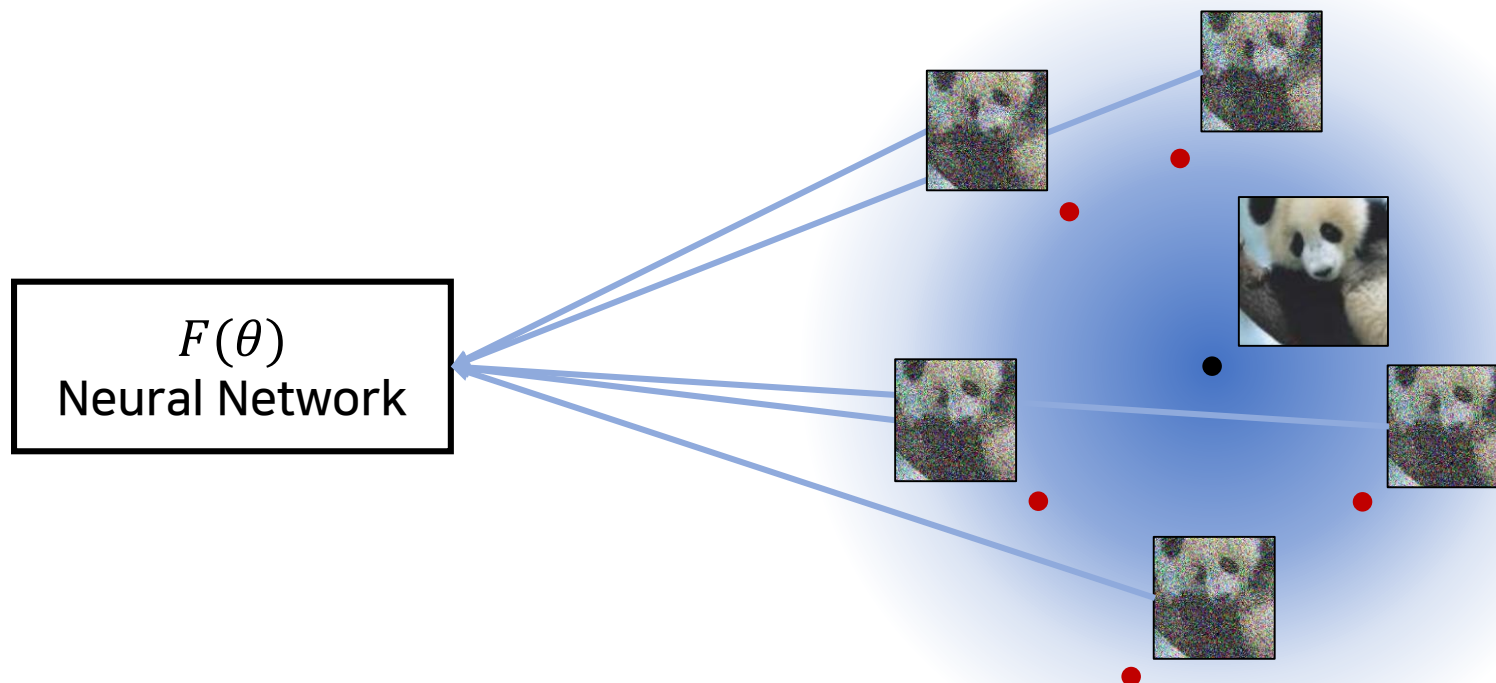
## 연구 배경: 인증된 적대적 강건성 (Certified Adversarial Robustness)

- Certified adversarial robustness
  - 입력 이미지가 주어졌을 때 특정한 크기의  $L_p$ -boundary 안에서 adversarial example이 만들어질 수 없도록 수학적으로 보장하는(guaranteeing) 방어 기법 유형입니다.



## 연구 배경: Randomized Smoothing

- **Randomized smoothing** is a provable adversarial defense in  $L_2$  norm which **scales to ImageNet**.
- In training time, it trains a neural network  $f$  with Gaussian data augmentation at variance  $\sigma^2$ .

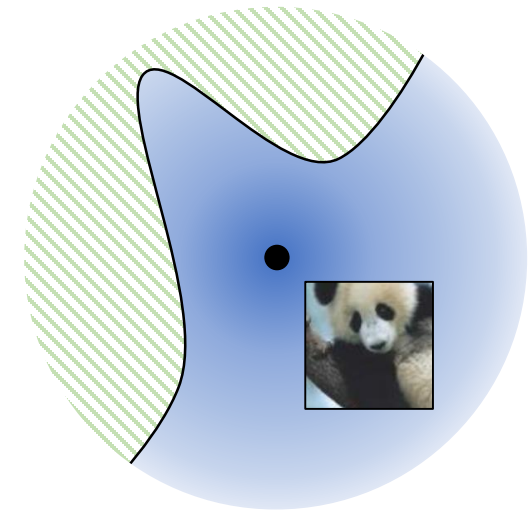


\*Certified Adversarial Robustness via Randomized Smoothing, Jeremy M Cohen, ICML 2019.

## 연구 배경: Randomized Smoothing (cont'd)

- $g(x)$  returns the class which  $f$  is most likely to return when  $x$  is corrupted by isotropic Gaussian noise with variance  $\sigma^2$ .
- In inference time, it uses  $g(x)$  by the *Monte Carlo* algorithm.
- $g$  is provably robust within an  $L_2$  radius of  $\sigma \cdot \Phi^{-1}(p)$ .
  - $\Phi^{-1}$ : A inverse CDF of the standard normal distribution.
- e.g.,  $p = 0.8, \sigma = 0.5 \rightarrow \sigma \cdot \Phi^{-1}(p) \approx 0.5 \cdot 0.842 = 0.421$

$\ell_2$ RADIUS	BEST $\sigma$	CERT. ACC (%)	STD. ACC(%)
0.5	0.25	49	67
1.0	0.50	37	57
2.0	0.50	19	57
3.0	1.00	12	44



■ : 80% ( $p$ )

■ : 20%

## 본 논문의 핵심 아이디어

- 어떠한 이미지가 결정 경계(decision boundary)로부터 멀리 떨어져 있다면, 분류기(classifier)는 이미지에 노이즈를 섞은 이미지에 대해서도 같은 레이블로 분류하게 됩니다.

