Adversarial Attacks and Defenses in Deep Learning

Yoon Sang Cho

Open DMQA Seminar

2020-07-17

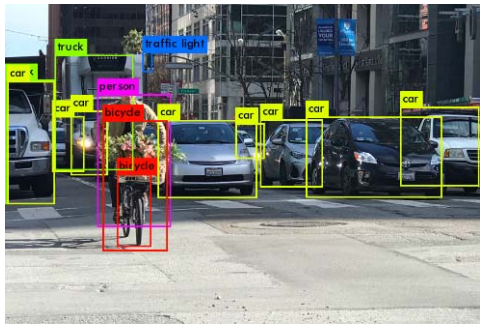
Contents

- Introduction
- Adversarial Examples
- Adversarial Attacks & Defenses
- Other Methods
- Conclusions

Introduction

- Deep learning models show remarkable performance in various machine learning tasks
- E.g. Text recognition, Signal prediction, Image classification

Object detection



Deep learning model

Human activity recognition

Smart manufacturing

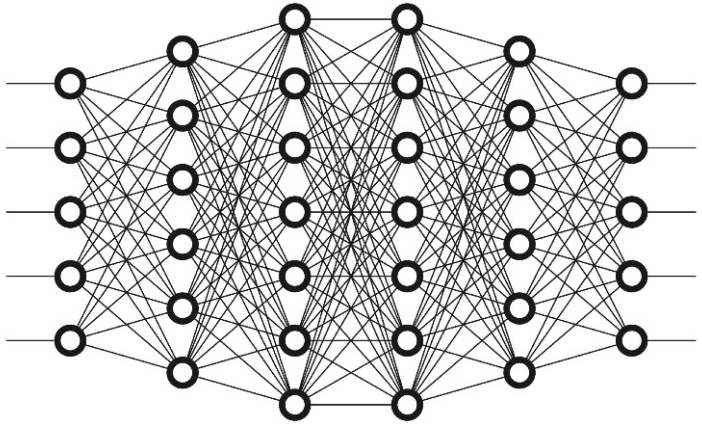
Self-driving car

Text mining

Introduction

- Deep learning models show remarkable performance in various machine learning tasks
- E.g. Text recognition, Signal prediction, Image classification

Input **Deep learning model** **Output**



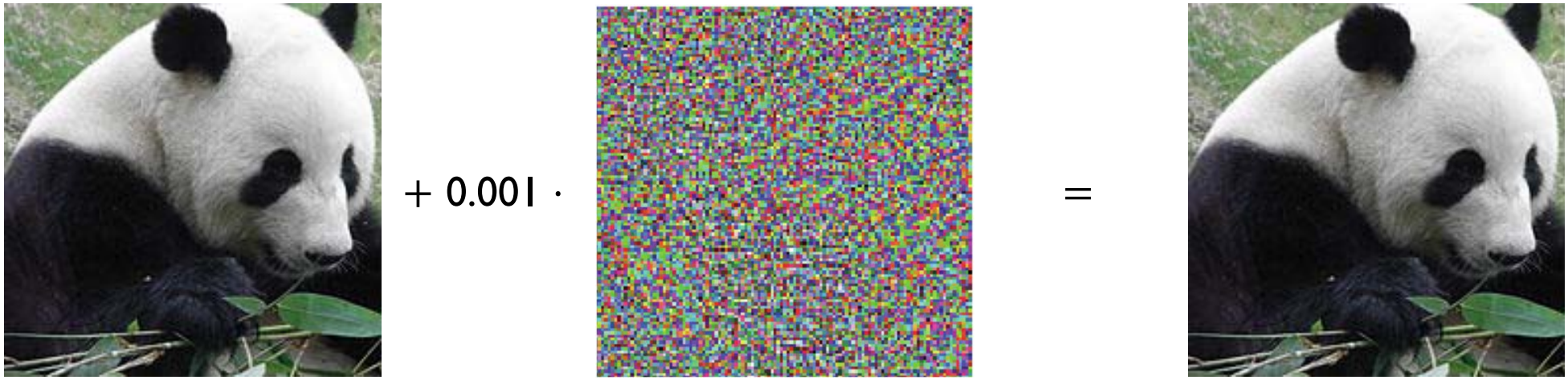
Panda

However, the AI models are vulnerable to slightly different data from input data

Introduction

- Adversarial examples are slightly different data from input data.
- Adversarial examples = Original data + Perturbation (작은 변화)

Original Data **Perturbation** **Adversarial Data**

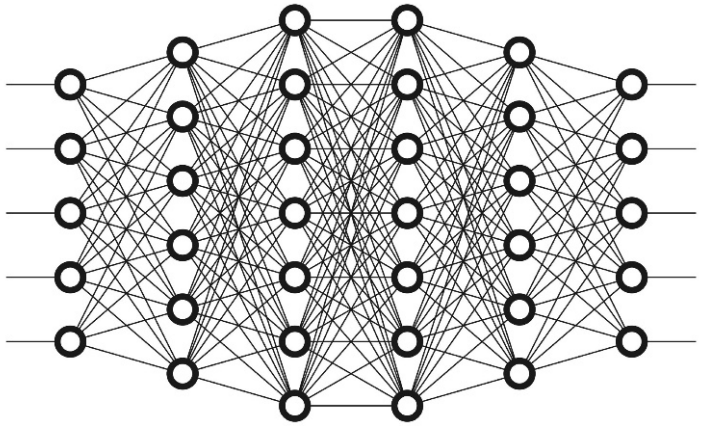


Adversarial examples fool the model (Accuracy ↓)

Adversarial Examples

- Adversarial examples are slightly different data from input data.
- Adversarial examples = Original data + Perturbation (작은 변화)

Adversarial Data | **Deep learning model** | **Output**



=

Human: It is a "Panda"

AI: It is a "Gibbon"

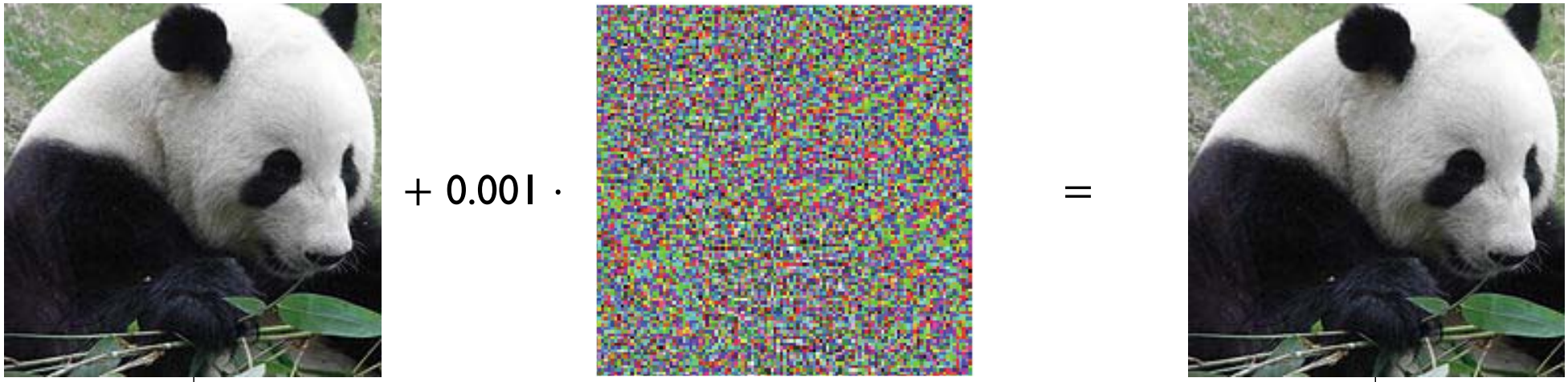
Generating adversarial examples & applying it to the AI model → Adversarial attack!

How to adversarial attack?

Adversarial Attacks

- Adversarial example = Original data + Perturbation

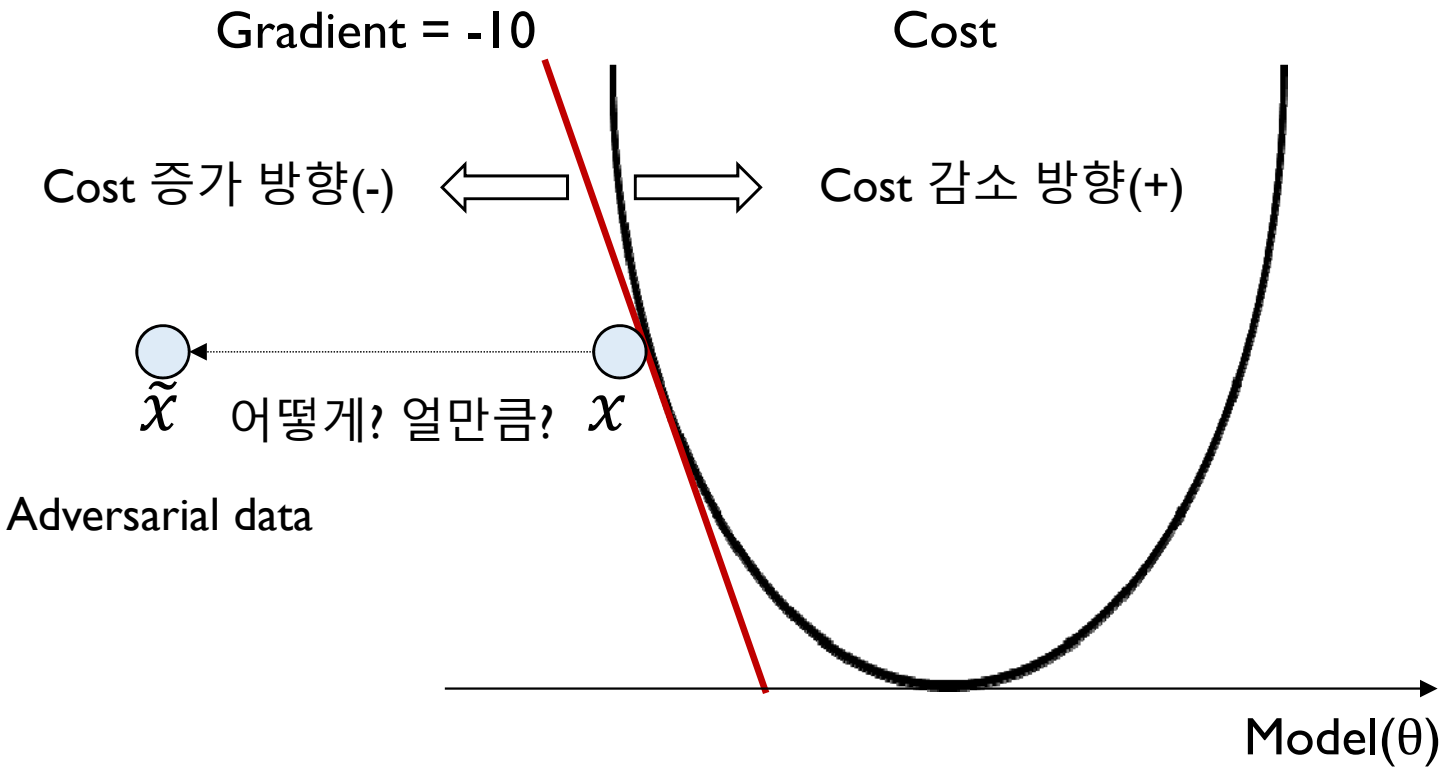
Original Data **Perturbation** **Adversarial Data**



How to generate the perturbation? → adversarial attack methods

Adversarial Attacks

- Deep neural networks → Gradient descent method (cost 감소 방향)으로 학습
  - Gradient (기울기)의 반대 방향으로 이동→ cost 감소(Model 학습)
  - Gradient (기울기)의 방향으로 이동 → cost 증가 (Adversarial attack)



Adversarial Attacks

- Adversarial attack method

$J(\theta, x, y)$➔ Cost function

Adversarial Attacks

- Adversarial attack method

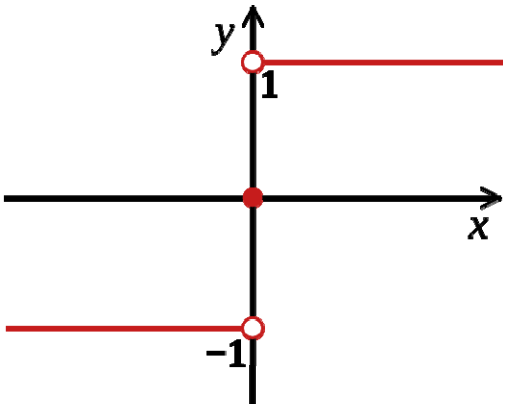
$J(\theta, x, y)$ ➔ Cost function

$\nabla_x J(\theta, x, y)$ ➔ Gradient

Adversarial Attacks

- Adversarial attack method

Sign function:



$J(\theta, x, y)$ ➔ Cost function

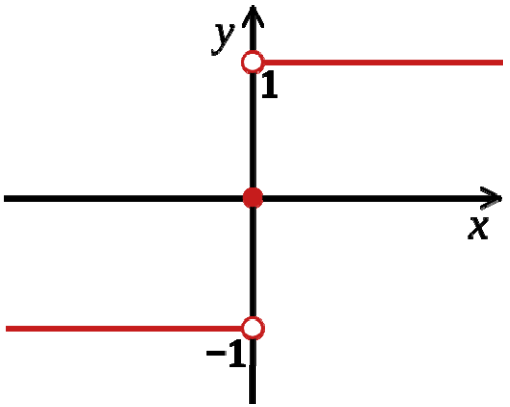
$\nabla_x J(\theta, x, y)$ ➔ Gradient

$\text{sign}(\nabla_x J(\theta, x, y))$ ➔ Gradient 방향으로

Adversarial Attacks

- Adversarial attack method

Sign function:



$J(\theta, x, y)$ ➔ Cost function

$\nabla_x J(\theta, x, y)$ ➔ Gradient

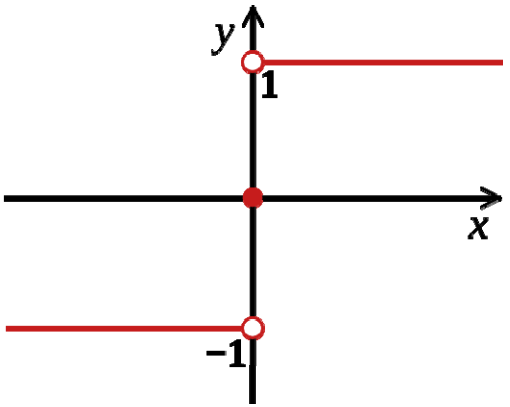
$\text{sign}(\nabla_x J(\theta, x, y))$ ➔ Gradient 방향으로

$\epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ ➔ Epsilon 만큼 이동

Adversarial Attacks

- Adversarial attack method

Sign function:



$J(\theta, x, y)$ ➔ Cost function

$\nabla_x J(\theta, x, y)$ ➔ Gradient

$\text{sign}(\nabla_x J(\theta, x, y))$ ➔ Gradient 방향으로

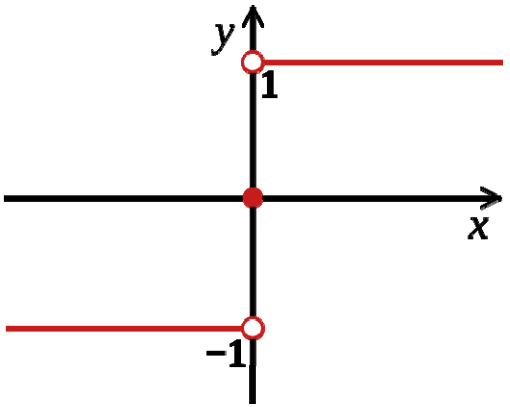
$\epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ ➔ Epsilon 만큼 이동

$x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ ➔ $\tilde{x}$ (Adversarial example)

Adversarial Attacks

- Adversarial attack method

Sign function:



$J(\theta, x, y)$ ➔ Cost function

$\nabla_x J(\theta, x, y)$ ➔ Gradient

$\text{sign}(\nabla_x J(\theta, x, y))$ ➔ Gradient 방향으로

$\epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ ➔ Epsilon 만큼 이동

$x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$ ➔ $\tilde{x}$ (Adversarial example)

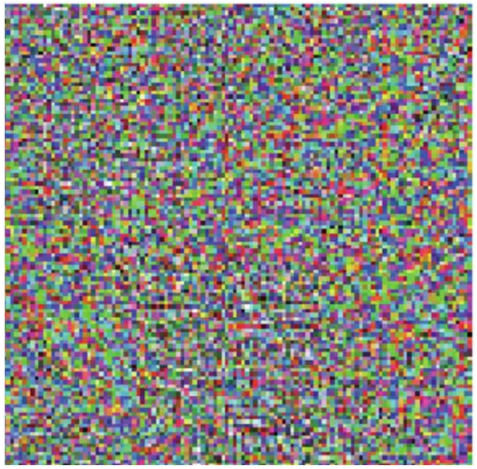
The Fast Gradient Sign Method (Goodfellow et al., 2014)

Adversarial Attacks

- Adversarial example = Original data + Perturbation

Original Data **Perturbation** **Adversarial Data**

$+ 0.001 \cdot$



$=$

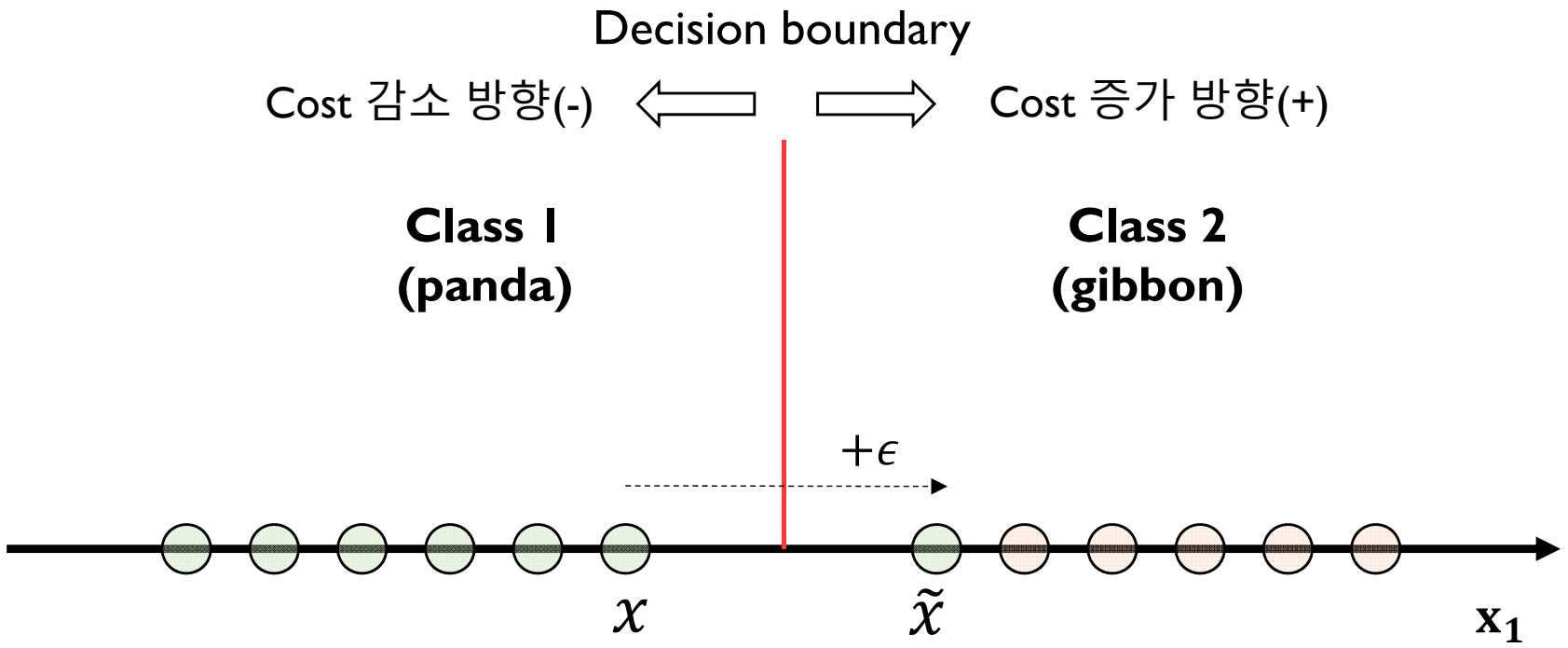
x $+\epsilon$ $\text{sign}(\nabla_x J(\theta, x, y))$ $x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$

→ Epsilon 만큼 (-) or (+) 방향으로 이동

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

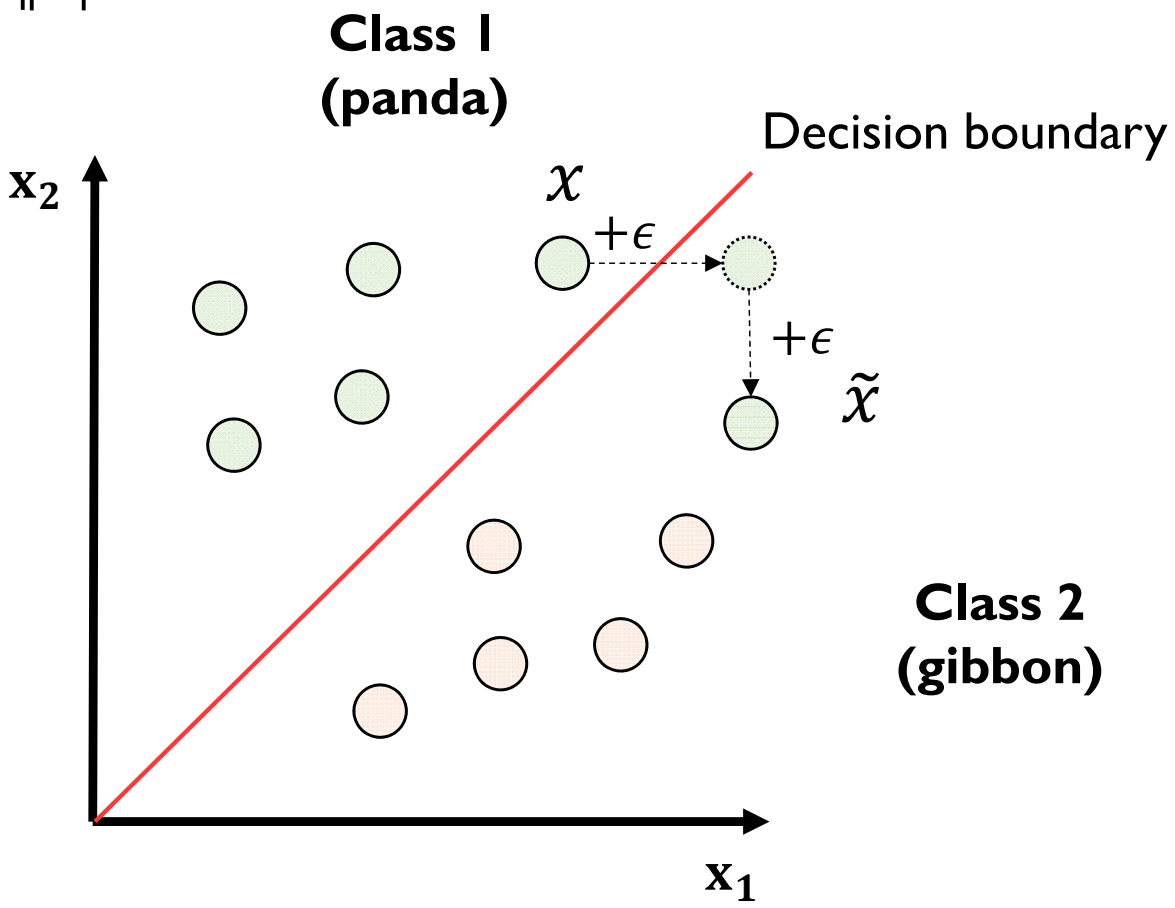
Adversarial Attacks

- The Fast Gradient Sign Method (2015)
  - 1차원 (pixel 1개) 예시



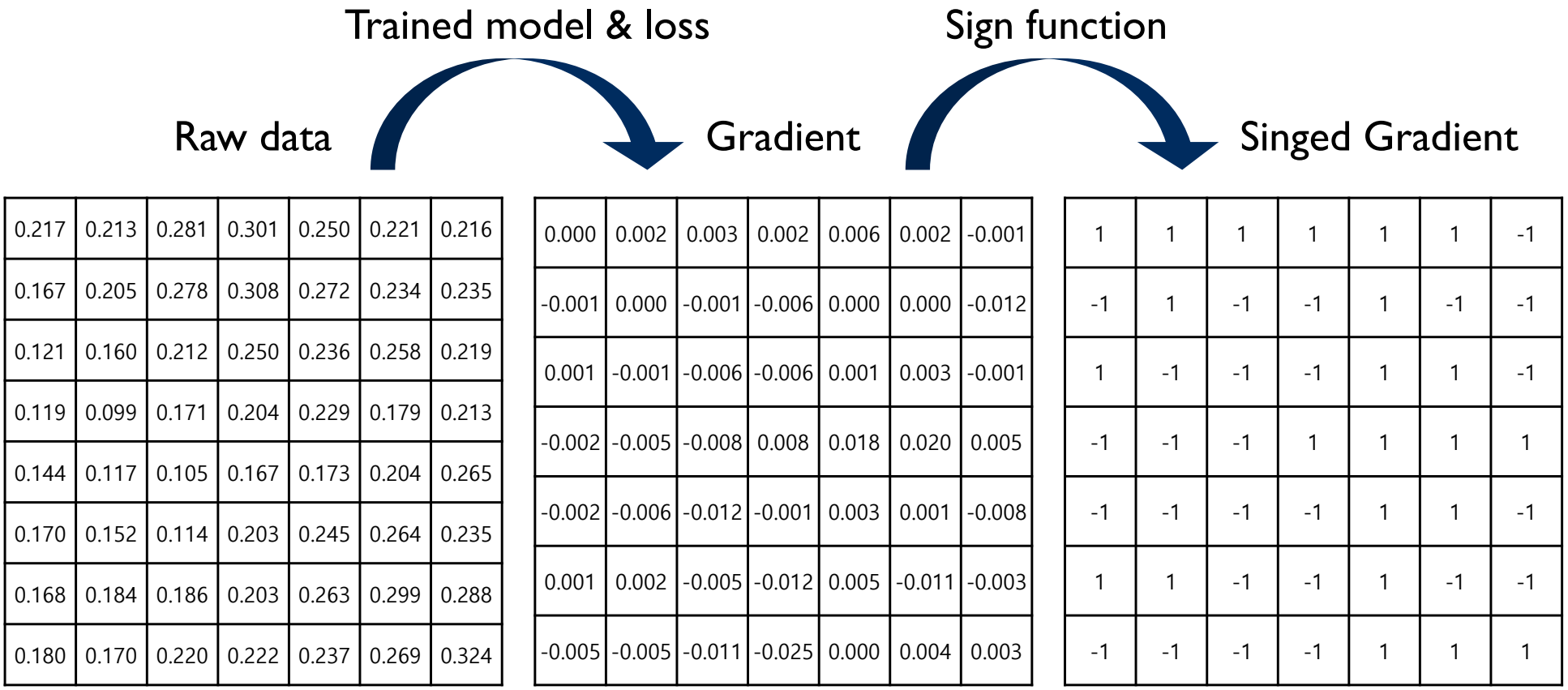
Adversarial Attacks

- The Fast Gradient Sign Method (2015)
  - 2차원 (pixel 2개) 예시



Adversarial Attacks

- The Fast Gradient Sign Method (2015)
  - 모든 차원 (모든 pixel) 예시



0.217	0.213	0.281	0.301	0.250	0.221	0.216
0.167	0.205	0.278	0.308	0.272	0.234	0.235
0.121	0.160	0.212	0.250	0.236	0.258	0.219
0.119	0.099	0.171	0.204	0.229	0.179	0.213
0.144	0.117	0.105	0.167	0.173	0.204	0.265
0.170	0.152	0.114	0.203	0.245	0.264	0.235
0.168	0.184	0.186	0.203	0.263	0.299	0.288
0.180	0.170	0.220	0.222	0.237	0.269	0.324

0.000	0.002	0.003	0.002	0.006	0.002	-0.001
-0.001	0.000	-0.001	-0.006	0.000	0.000	-0.012
0.001	-0.001	-0.006	-0.006	0.001	0.003	-0.001
-0.002	-0.005	-0.008	0.008	0.018	0.020	0.005
-0.002	-0.006	-0.012	-0.001	0.003	0.001	-0.008
0.001	0.002	-0.005	-0.012	0.005	-0.011	-0.003
-0.005	-0.005	-0.011	-0.025	0.000	0.004	0.003

1	1	1	1	1	1	-1
-1	1	-1	-1	1	-1	-1
1	-1	-1	-1	1	1	-1
-1	-1	-1	1	1	1	1
-1	-1	-1	-1	1	1	-1
1	1	-1	-1	1	-1	-1
-1	-1	-1	-1	1	1	1

Adversarial Attacks

- The Fast Gradient Sign Method (2015)
  - 모든 차원 (모든 pixel) 예시

Raw data

0.217	0.213	0.281	0.301	0.250	0.221	0.216
0.167	0.205	0.278	0.308	0.272	0.234	0.235
0.121	0.160	0.212	0.250	0.236	0.258	0.219
0.119	0.099	0.171	0.204	0.229	0.179	0.213
0.144	0.117	0.105	0.167	0.173	0.204	0.265
0.170	0.152	0.114	0.203	0.245	0.264	0.235
0.168	0.184	0.186	0.203	0.263	0.299	0.288
0.180	0.170	0.220	0.222	0.237	0.269	0.324

0.001×
(epsilon)

Singed Gradient

1	1	1	1	1	1	-1
-1	1	-1	-1	1	-1	-1
1	-1	-1	-1	1	1	-1
-1	-1	-1	1	1	1	1
-1	-1	-1	-1	1	1	-1
1	1	-1	-1	1	-1	-1
-1	-1	-1	-1	1	1	1

➔ Adversarial examples

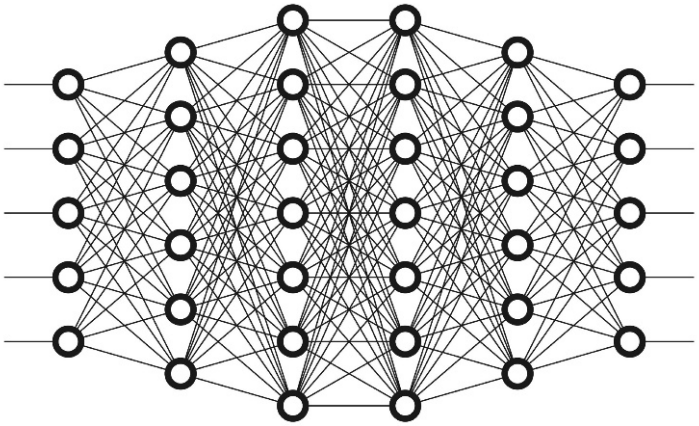
Adversarial Attacks

- Why the adversarial example works?
  - Adversarial attack (FGSM): 모든 변수(pixel) 마다 Epsilon 만큼 이동

Image ($28 \times 28 \times 3$)

2,352 pixels → 2,352 dimensional data

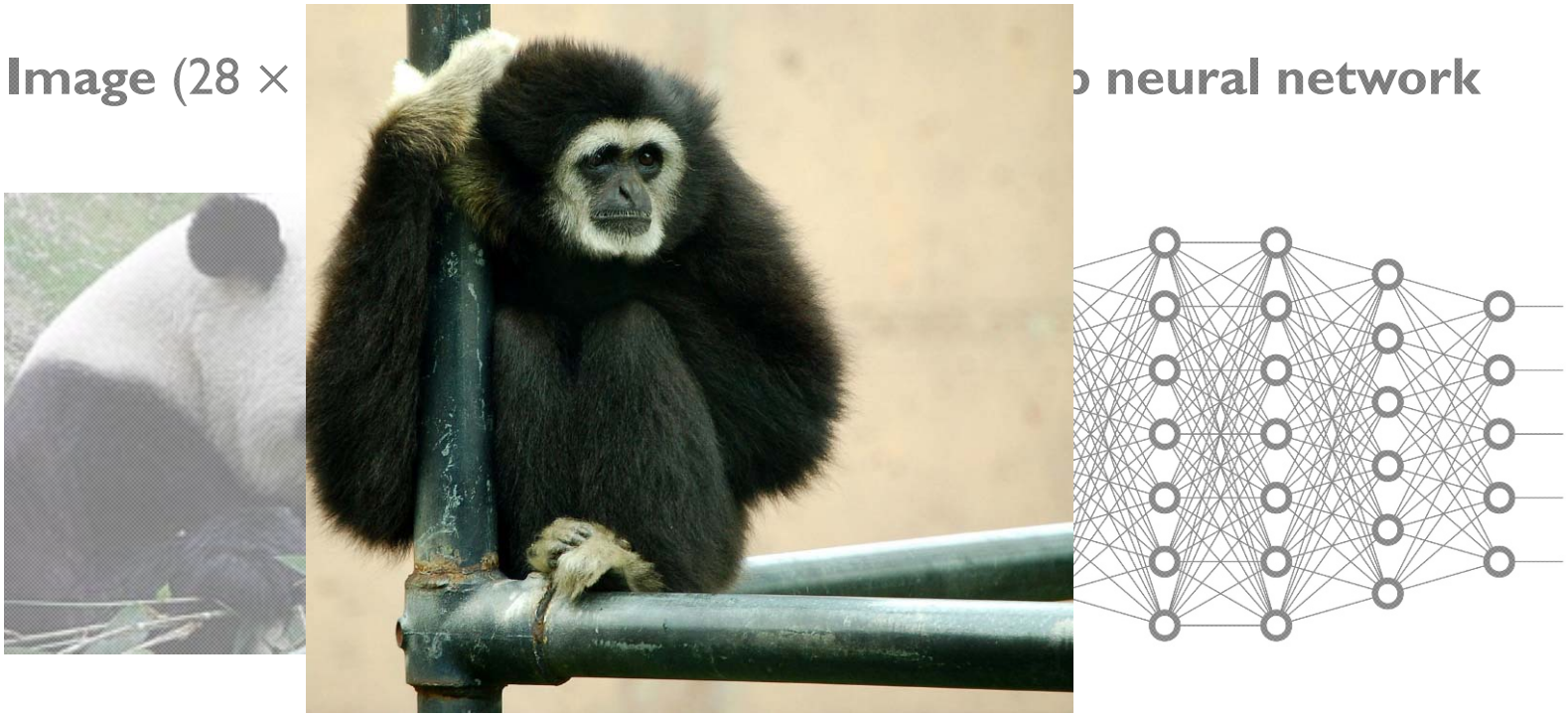
Deep neural network



Weight 곱하고, 더하고, 비선형 변환하고 , …

Adversarial Attacks

- Why the adversarial example works?
  - Adversarial attack (FGSM): 모든 변수(pixel) 마다 Epsilon 만큼 이동

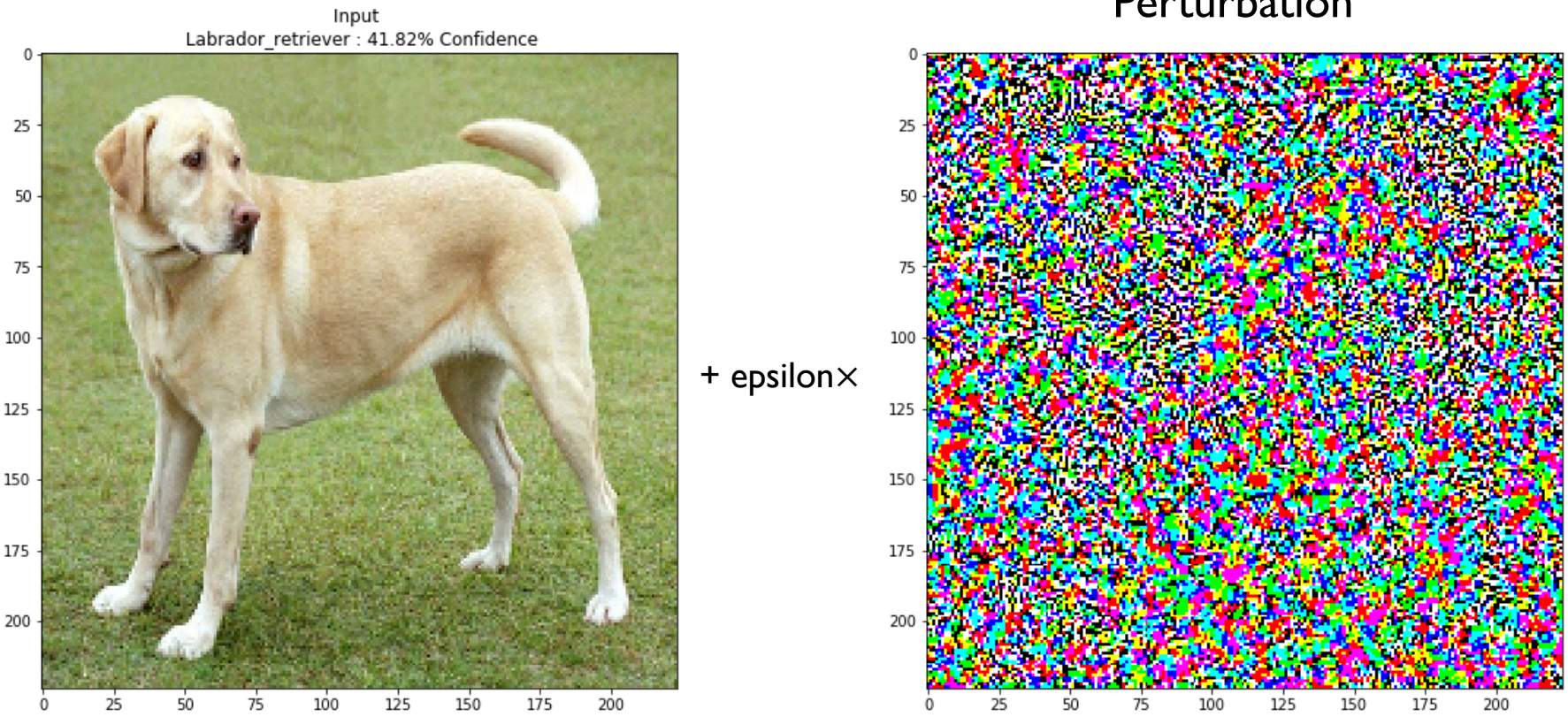


2,352 pixels → 2,352 dimensional data Weight 곱하고, 더하고, 비선형 변환하고 , …

Adversarial examples can fool the AI model simply & effectively

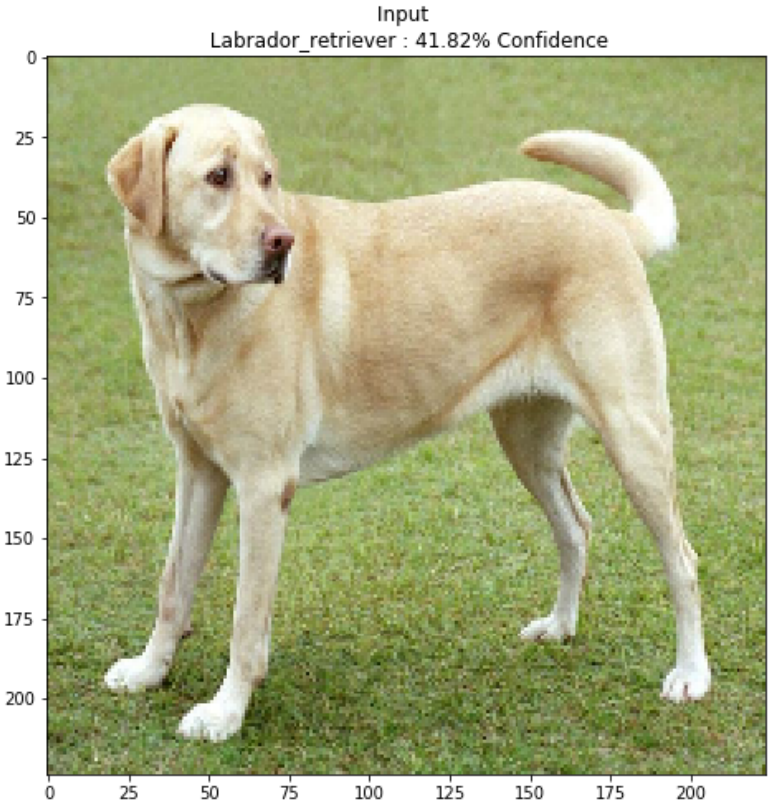
Adversarial Attacks

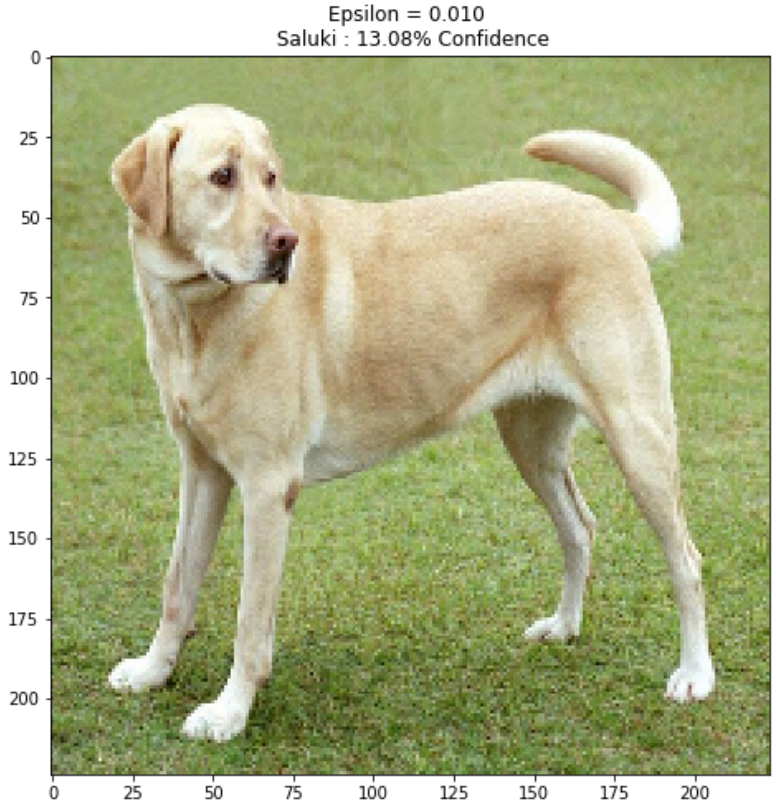
- Fast gradient sign method with TensorFlow
  - https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
  - Epsilon_list = [0.01, 0.1, 0.15, 0.5, 0.7]



Adversarial Attacks

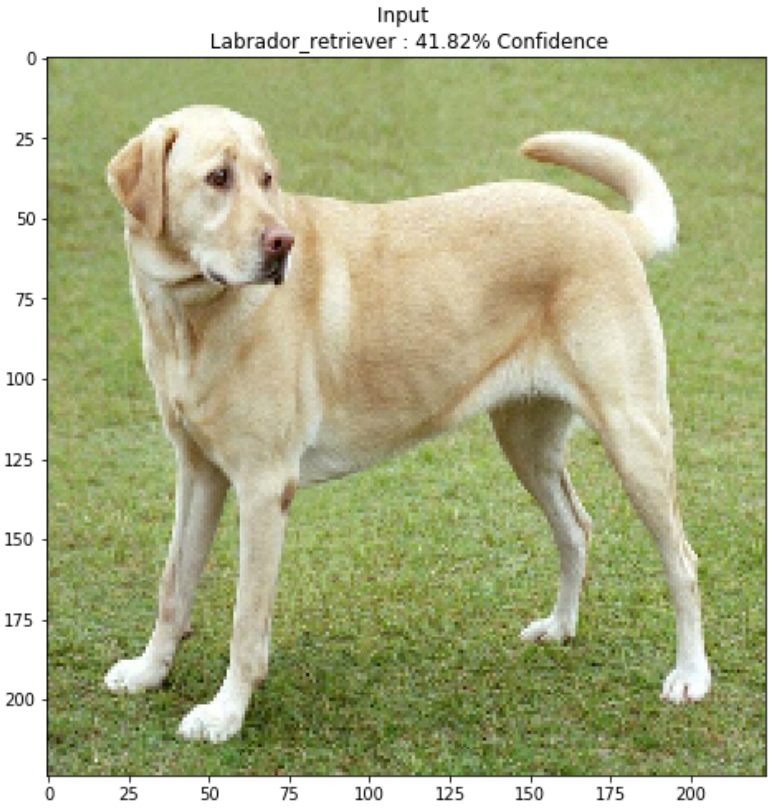
- Fast gradient sign method with TensorFlow
  - https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
  - Epsilon_list = [0.01, 0.1, 0.15, 0.5, 0.7]

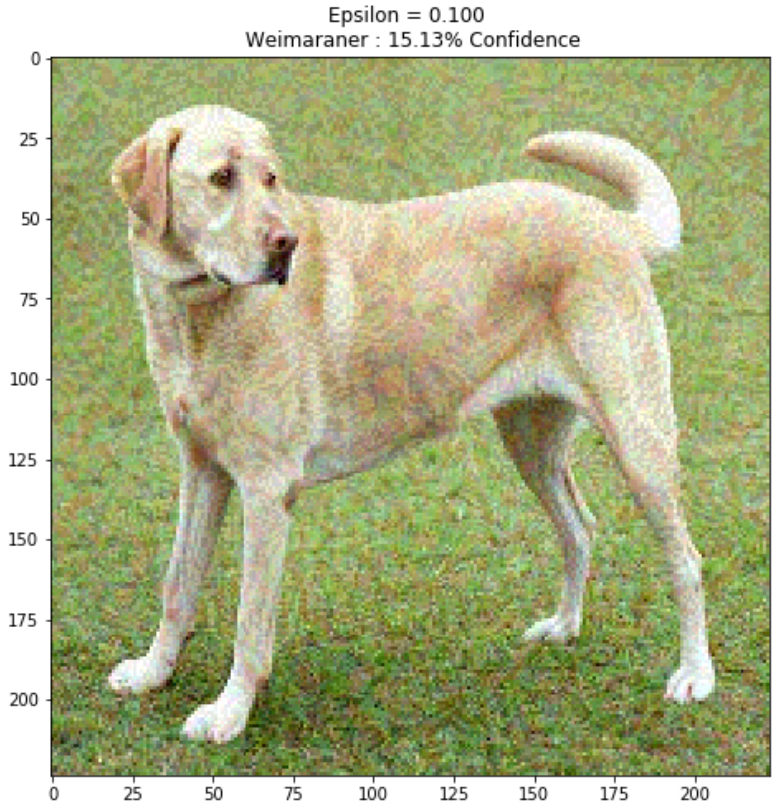




Adversarial Attacks

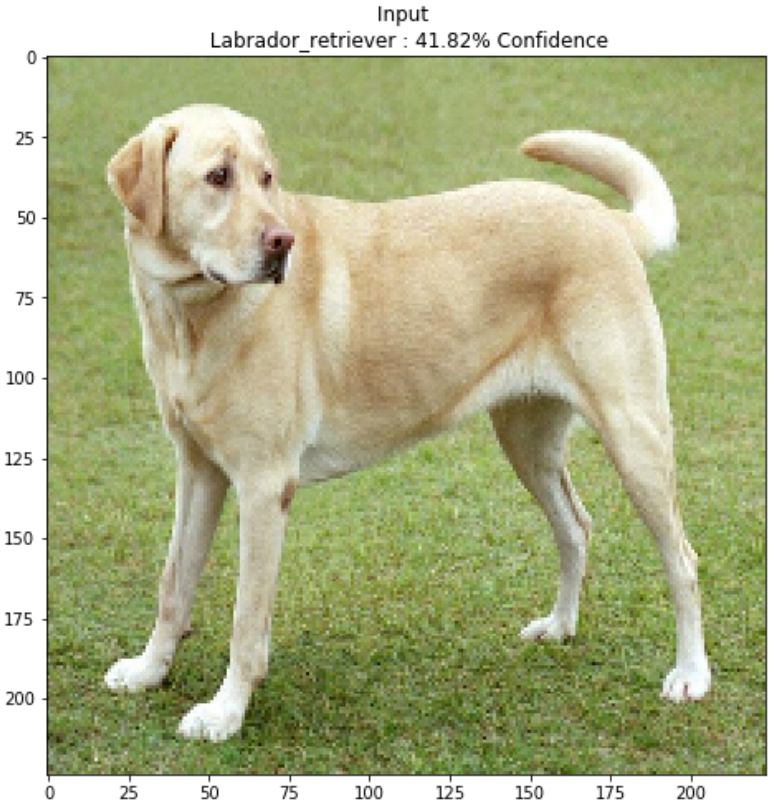
- Fast gradient sign method with TensorFlow
  - https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
  - Epsilon_list = [0.01, 0.1, 0.15, 0.5, 0.7]

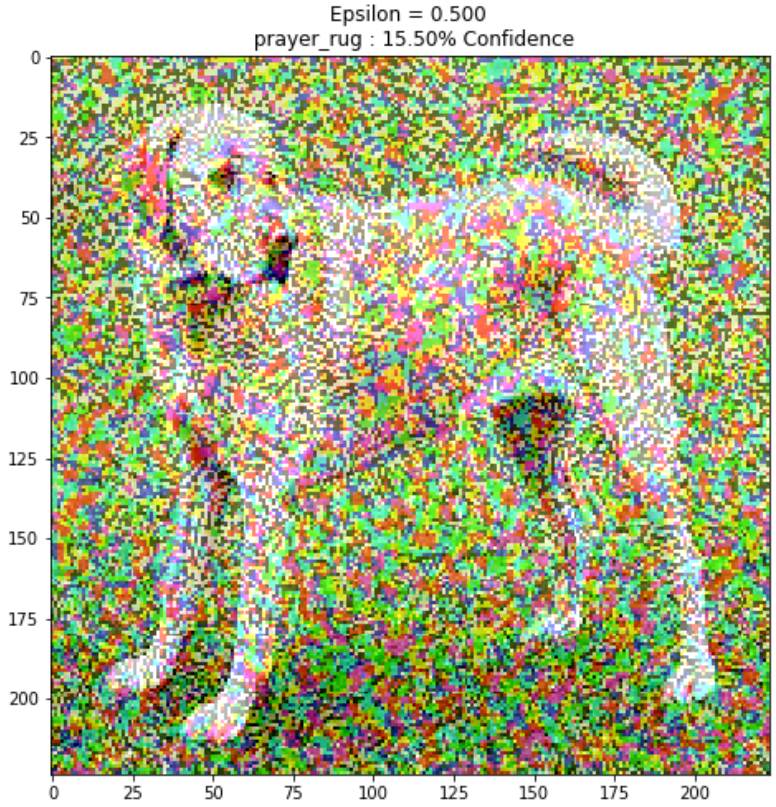




Adversarial Attacks

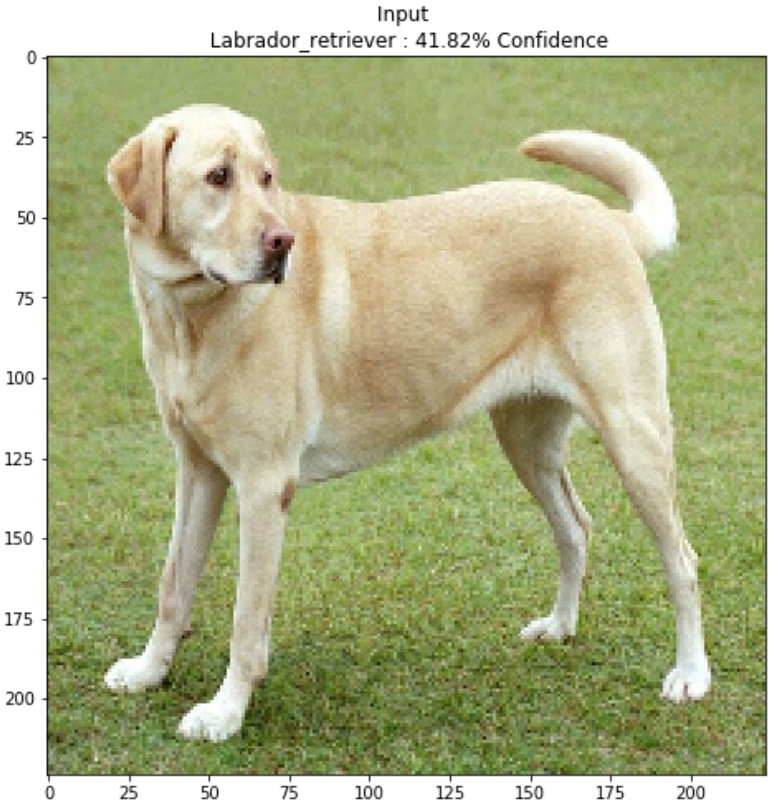
- Fast gradient sign method with TensorFlow
  - https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
  - Epsilon_list = [0.01, 0.1, 0.15, 0.5, 0.7]

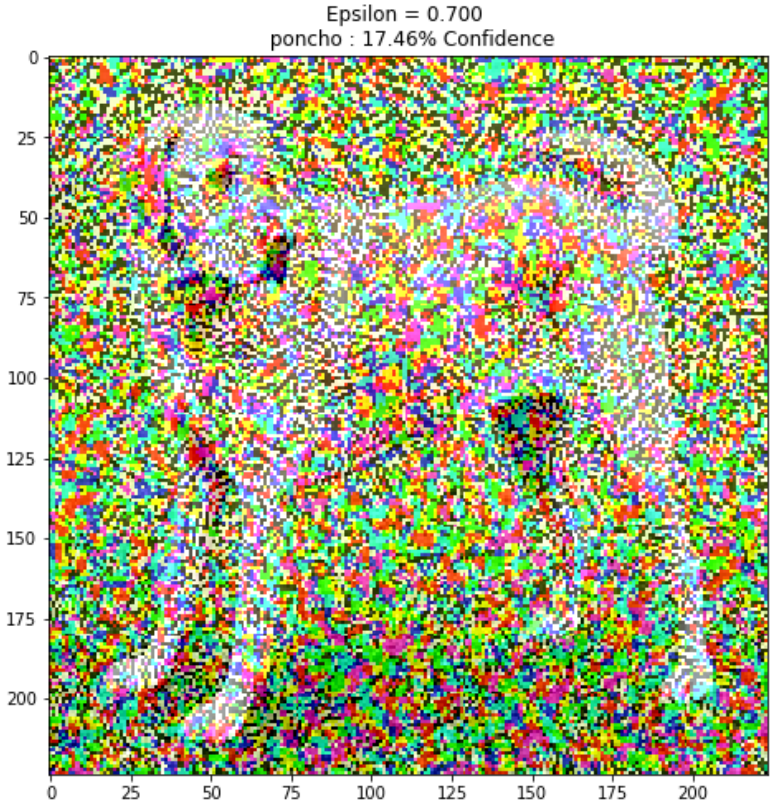




Adversarial Attacks

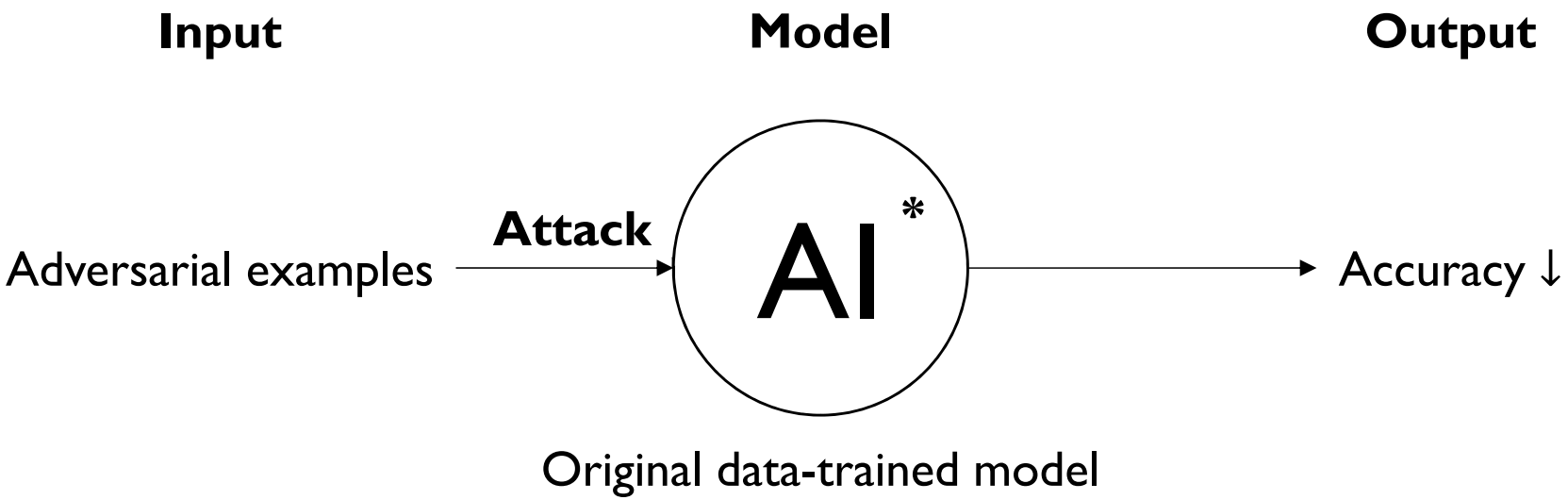
- Fast gradient sign method with TensorFlow
  - https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
  - Epsilon_list = [0.01, 0.1, 0.15, 0.5, 0.7]





Adversarial Attacks

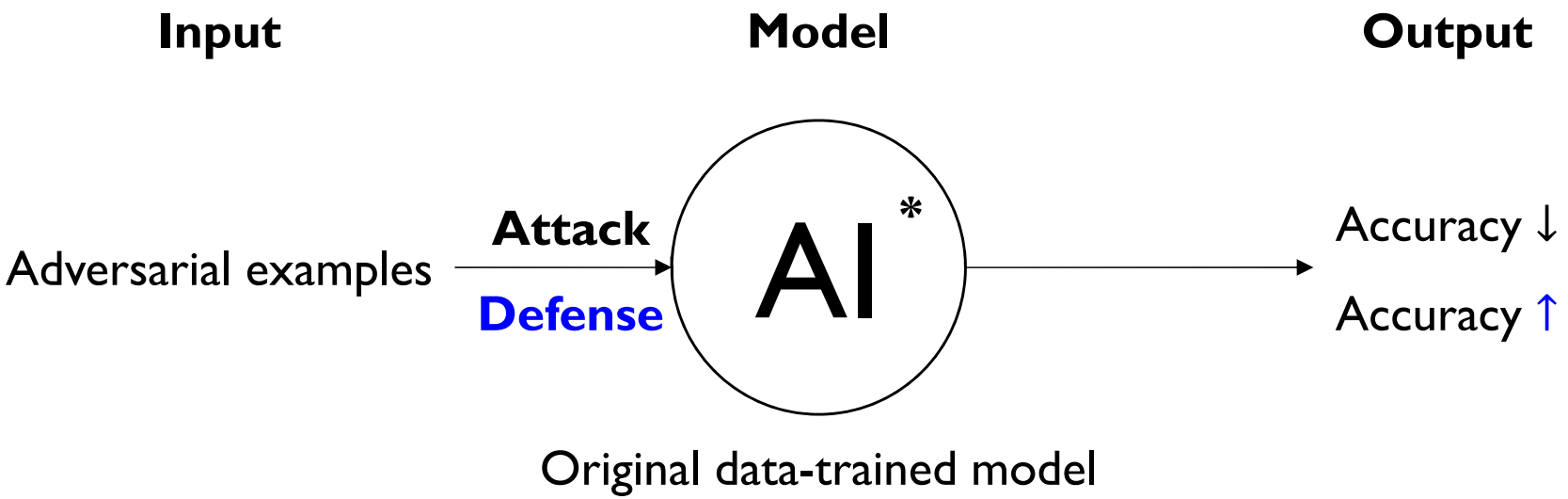
- Adversarial examples = Original data + Perturbation
- Adversarial attack fool the model using adversarial examples



Robustness ↓ ≈ generalization ability ↓

Adversarial Attacks

- Adversarial examples = Original data + Perturbation
- Adversarial attack fool the model using adversarial examples



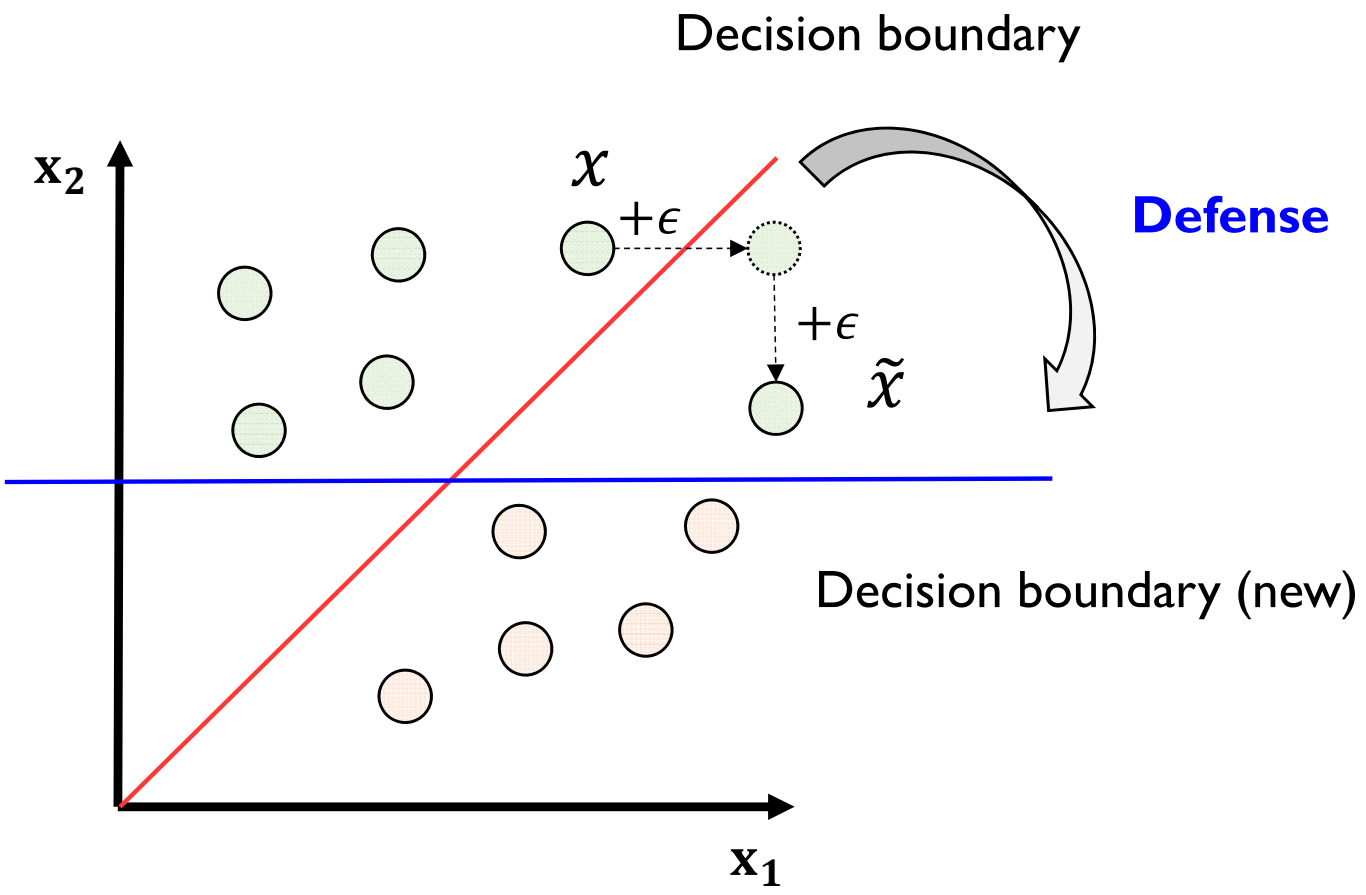
Robustness ↓ ≈ generalization ability ↓

Robustness ↑ ≈ generalization ability ↑

Defense against the adversarial attack

Defense Against Adversarial Examples

- Adversarial training



Adversarial training is the fast & simple defense method.

Defense Against Adversarial Examples

- Adversarial training with FGSM
  - Original data cost function + Perturbation cost function → adversarial training
  - 두 Model을 적절히(α) 고려하여 adversarial examples로부터 Defense
  - Hyperparameter α: Model이 Original data를 반영하는 정도

Cost function of adversarial training

$$\tilde{J}(\theta, x, y) = \alpha \cdot J(\theta, x, y) + (1 - \alpha) \cdot J(\theta, \tilde{x}, y)$$

Original data를 사용한 Model 비용 함수 Adversarial data를 사용한 Model 비용 함수

Defense Against Adversarial Examples

- Adversarial training with FGSM
  - Original data cost function + Perturbation cost function → adversarial training
  - 두 Model을 적절히(α) 고려하여 adversarial examples로부터 Defense
  - Hyperparameter α: Model이 Original data를 인식하는 정도

The Fast Gradient Sign Method

(Goodfellow et al., 2014)

Cost function of adversarial training

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha) J(\theta, \tilde{x}, y)$$

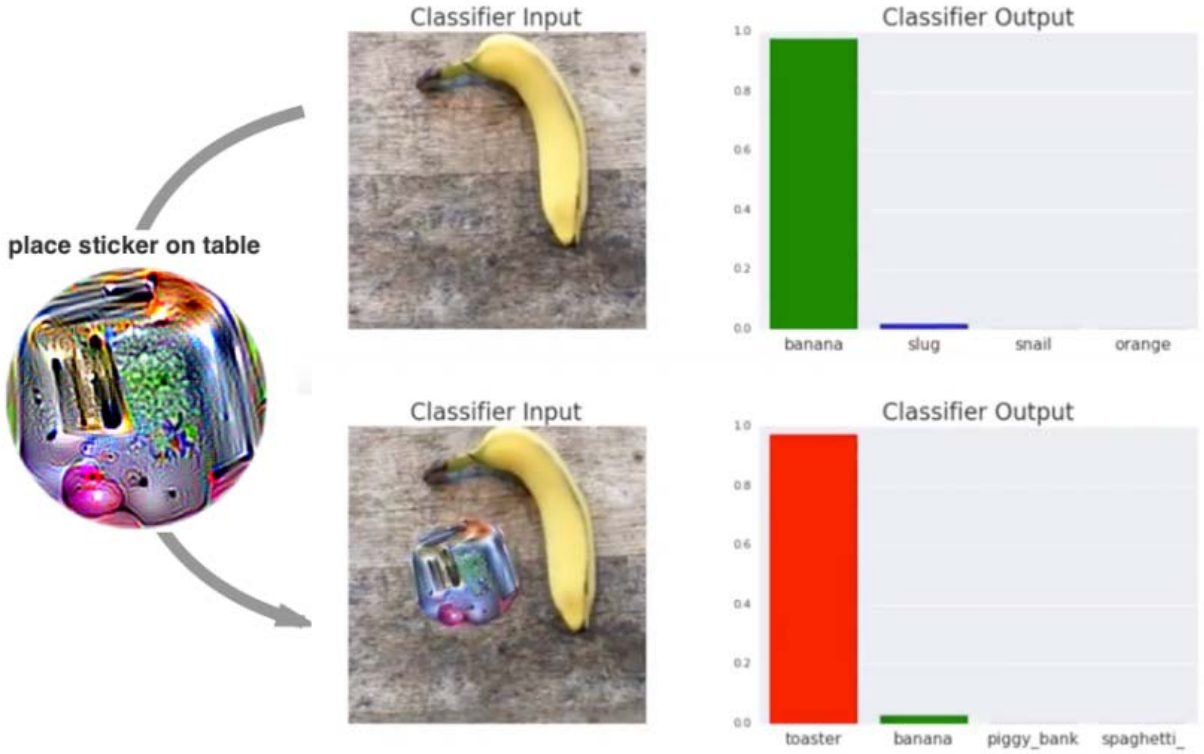
Adversarial attack & defense algorithm

Original data를 사용한 Model 비용 함수 Adversarial data를 사용한 Model 비용 함수

Other methods (attacks)

Other Adversarial Attacks

- Adversarial patch (2017)
  - FGSM은 관심대상 마다 Adversarial example을 생성
  - 그냥 sticker 붙이면 모델 성능↓

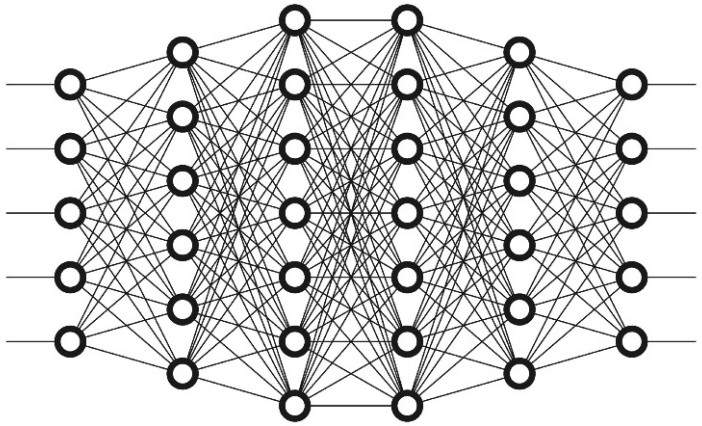


Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Other Adversarial Attacks

- Adversarial patch (2017)

Deep learning model



Cat Dog

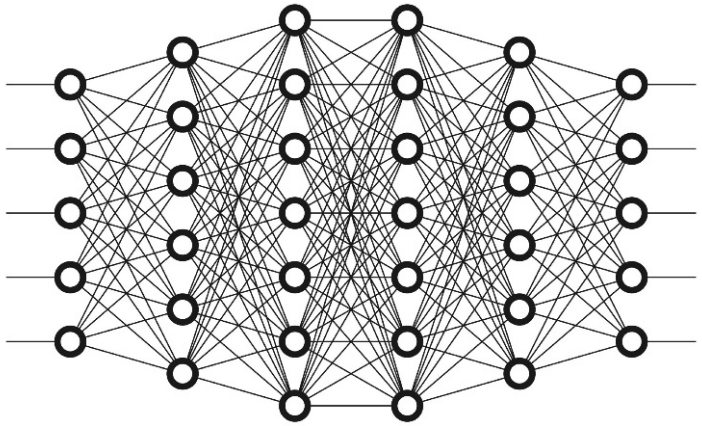
[0.90, 0.10]

https://www.anishathalye.com/2017/07/25/synthesizing-adversarial-examples/

Other Adversarial Attacks

- Adversarial patch (2017)

Deep learning model



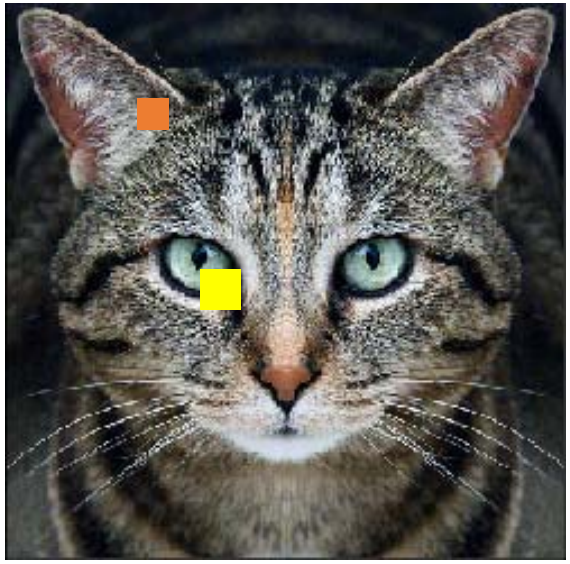
Cat Dog

[0.89, 0.11]

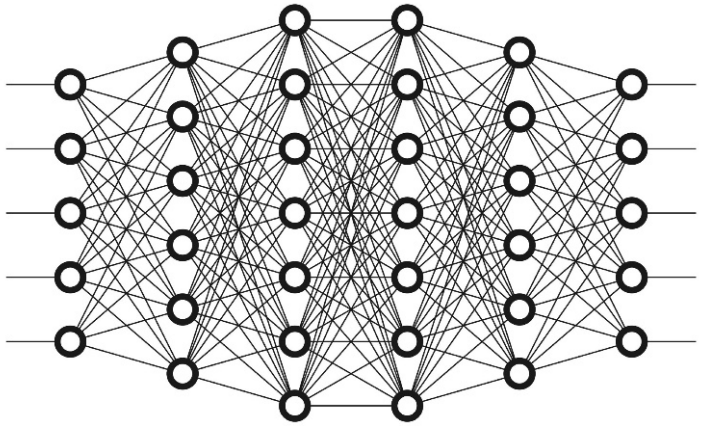
https://www.anishathalye.com/2017/07/25/synthesizing-adversarial-examples/

Other Adversarial Attacks

- Adversarial patch (2017)



Deep learning model



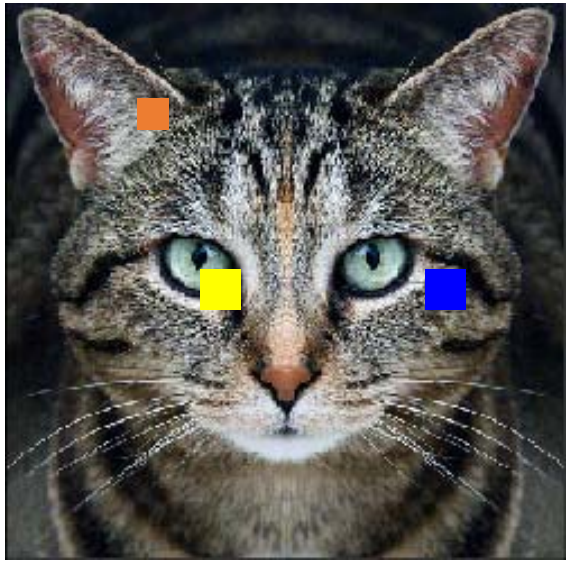
Cat Dog

[0.75, 0.25]

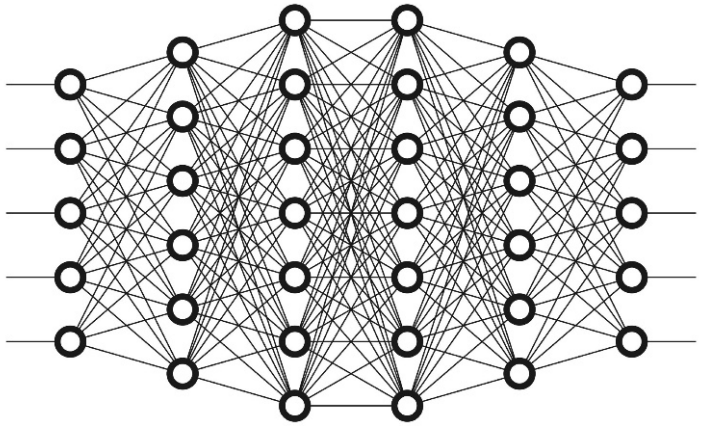
https://www.anishathalye.com/2017/07/25/synthesizing-adversarial-examples/

Other Adversarial Attacks

- Adversarial patch (2017)



Deep learning model



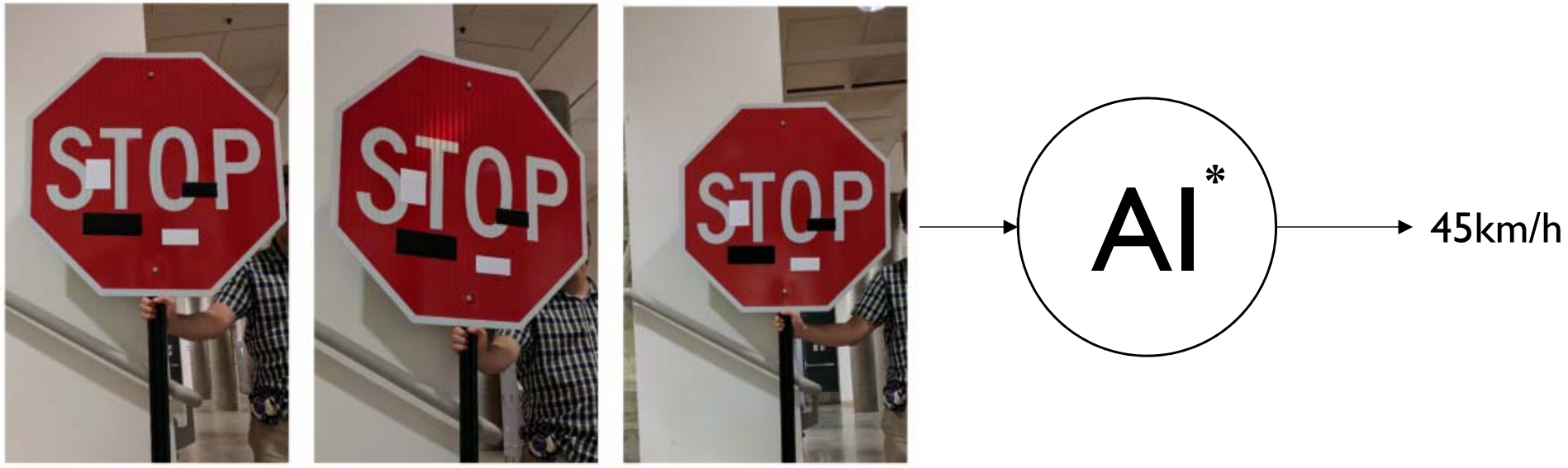
Cat Dog

[0.60, 0.40]

https://www.anishathalye.com/2017/07/25/synthesizing-adversarial-examples/

Other Adversarial Attacks

- Adversarial patch (2017)

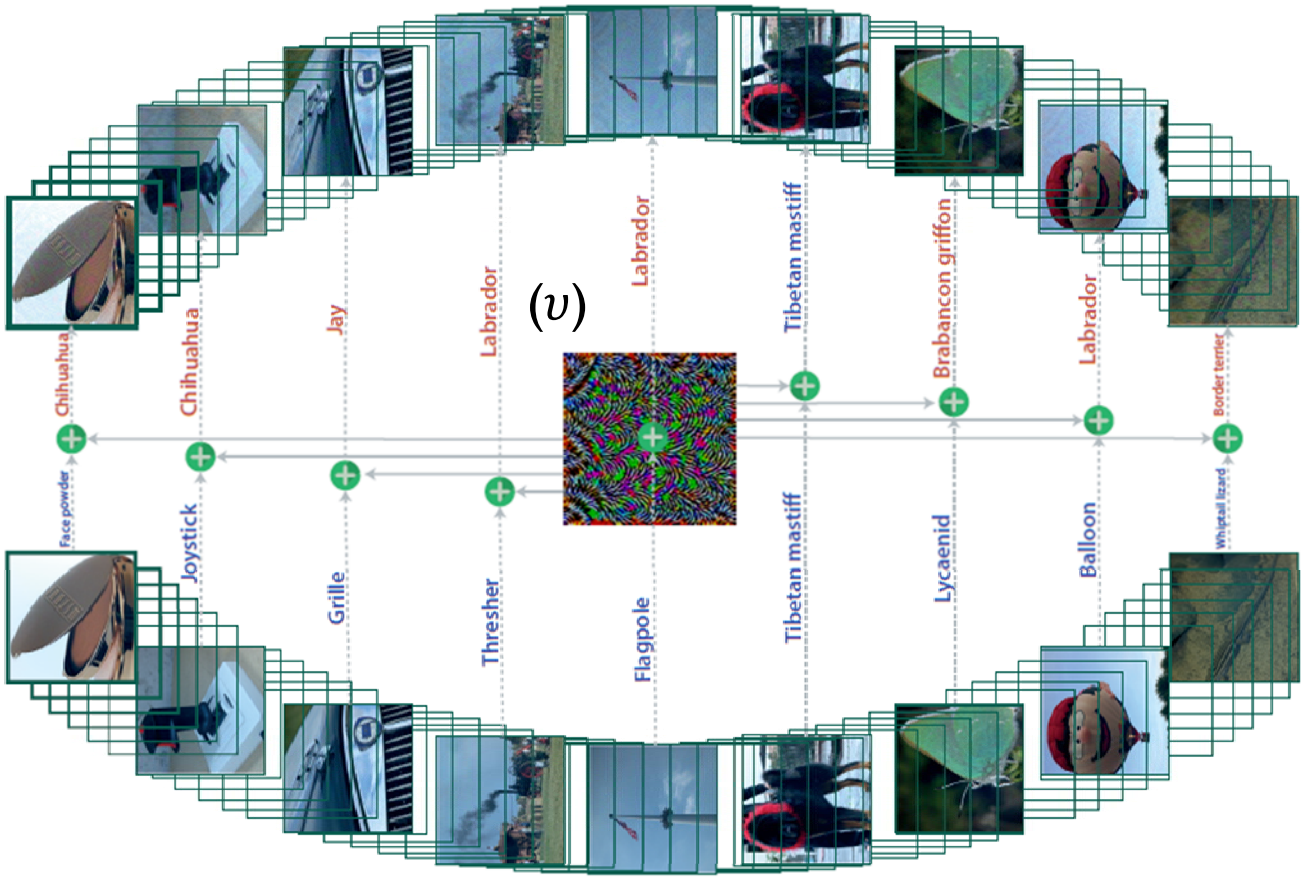


자율 주행 AI 모델

Other Adversarial Attacks

- Universal Adversarial Perturbation (2017)
  - FGSM은 관심대상 마다 Adversarial example을 생성
  - 입력데이터에 관계없이 Universal 한 perturbation을 생성해보자 → UAP

Adversarial data



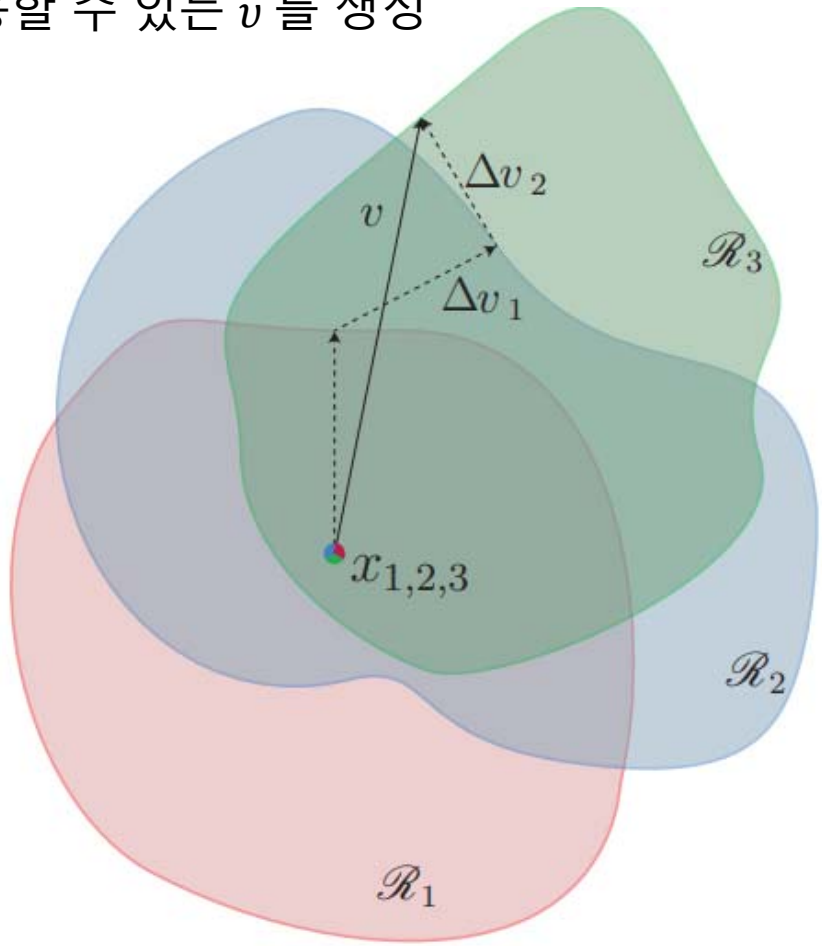
perturbation

Original data

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on CVPR* (pp. 1765-1773).

Other Adversarial Attacks

- Universal Adversarial Perturbation (2017)
  - Perturbation vector (v): 원본데이터가 Decision boundary를 벗어나는 최소한의 벡터
  - 모든 데이터에 적용할 수 있는 v 를 생성



Decision boundary of class 1
Decision boundary of class 2
Decision boundary of class 3

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on CVPR* (pp. 1765-1773).

Other methods (defenses)

Other Defense Techniques

- Adversarial training
  - 공격에서 생성된 이미지를 추가 학습데이터로 활용 (대표 사례: FGSM)
  - 공격에 좀 더 강건한 Model 구축 → Data augmentation 효과

$$LOSS = \frac{1}{(m-k)+\lambda k}\left(\sum_{i \in \text{original}} L\left(X_i, y_i\right) + \sum_{i \in \text{adversarial}} L\left(\tilde{X}_i, y_i\right)\right)$$

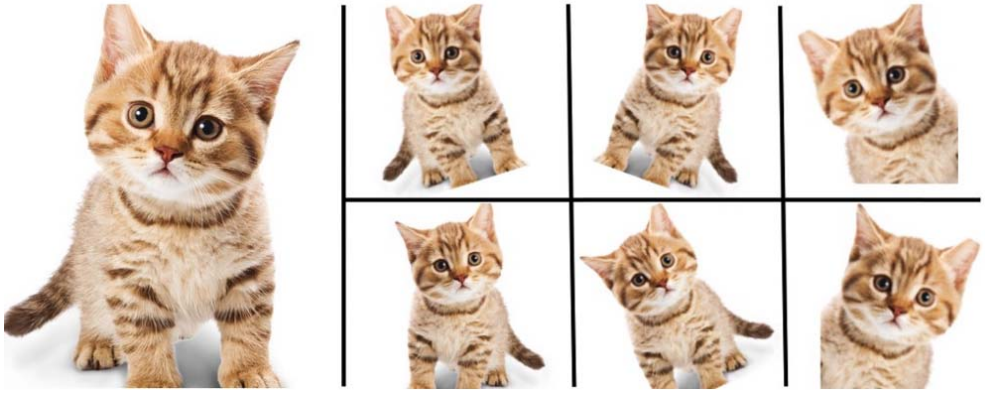
m: the total number of training data
k: the number of adversarial data
λ: a hyperparameter (adversarial data를 반영하는 정도)

Note that Data augmentation ≠ Adversarial training

Other Defense Techniques

- Adversarial training
  - 공격에서 생성된 이미지를 추가 학습데이터로 활용 (대표 사례: FGSM)
  - 공격에 좀 더 강건한 Model 구축 → Data augmentation 효과

Augmented Data



Test 데이터에 실제로 존재할 법 한 데이터

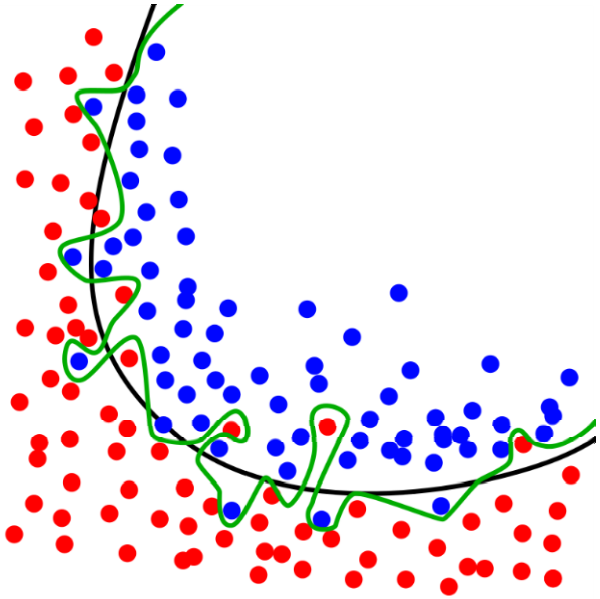
Adversarial Data

그렇진 않은데 모델에 악영향이 강한 데이터
(maliciously perturbed data)

Other Defense Techniques

- Adversarial training
  - A process of minimizing classification error rates when the data are maliciously perturbed

Overfitting

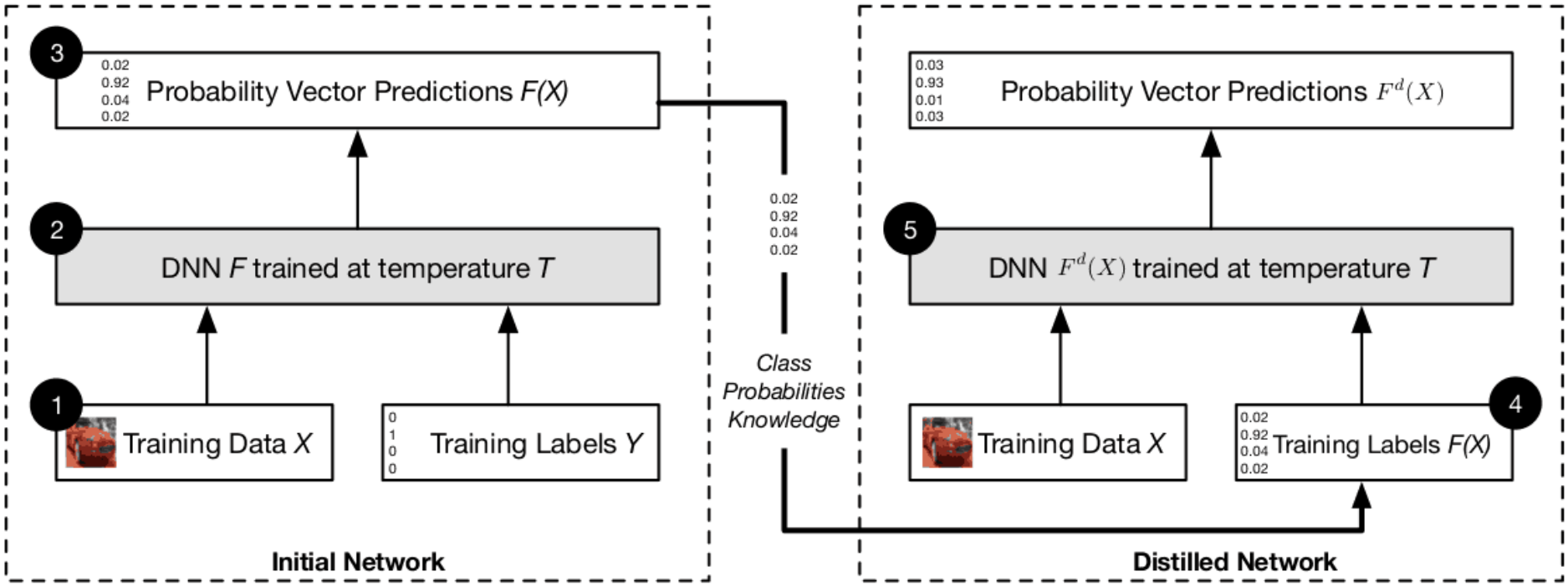


Security

https://ko.wikipedia.org/wiki/%EA%B3%BC%EC%A0%81%ED%95%A9
https://www.scnsoft.com/blog/ai-threats-cybersecurity

Other Defense Techniques

- Defensive Distillation (Knowledge distillation 사용)
  - Universal defensive method for neural network (smooth classifiers)
  - Universal defensive method reducing sensitivity of DNN to the input perturbation

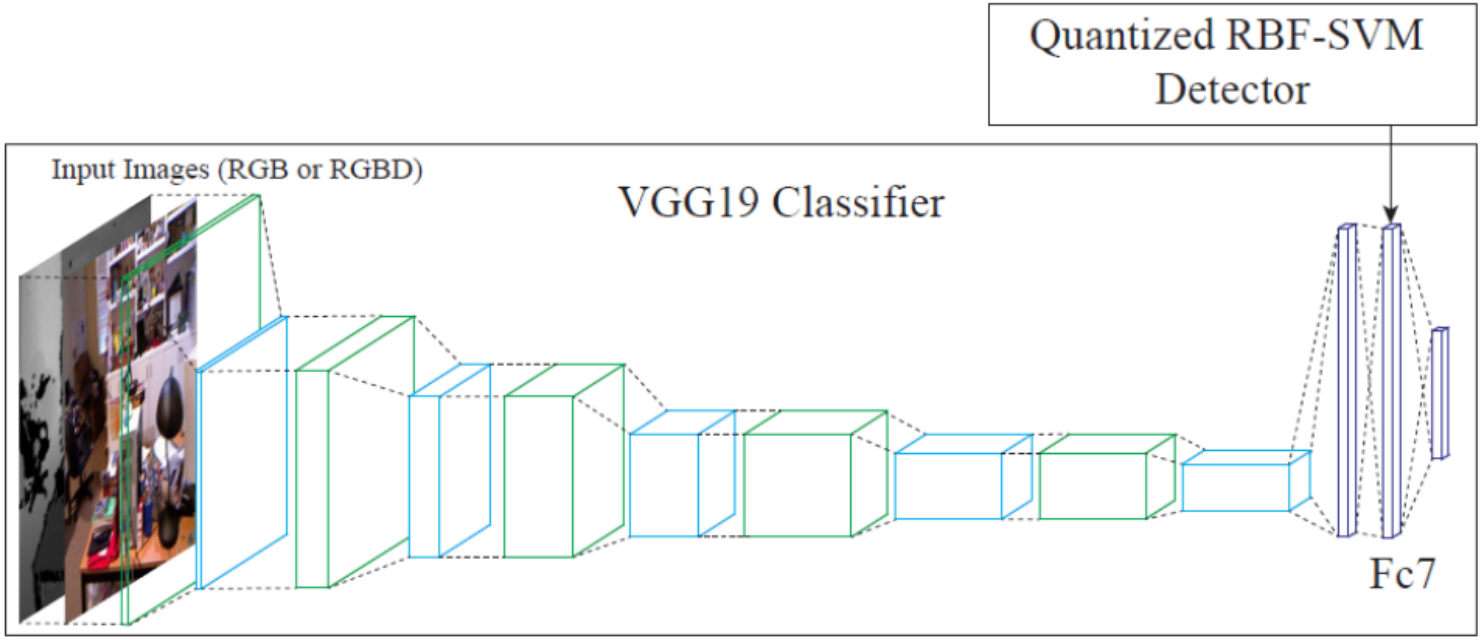


Network distillation of DNNs

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582-597). IEEE.

Other Defense Techniques

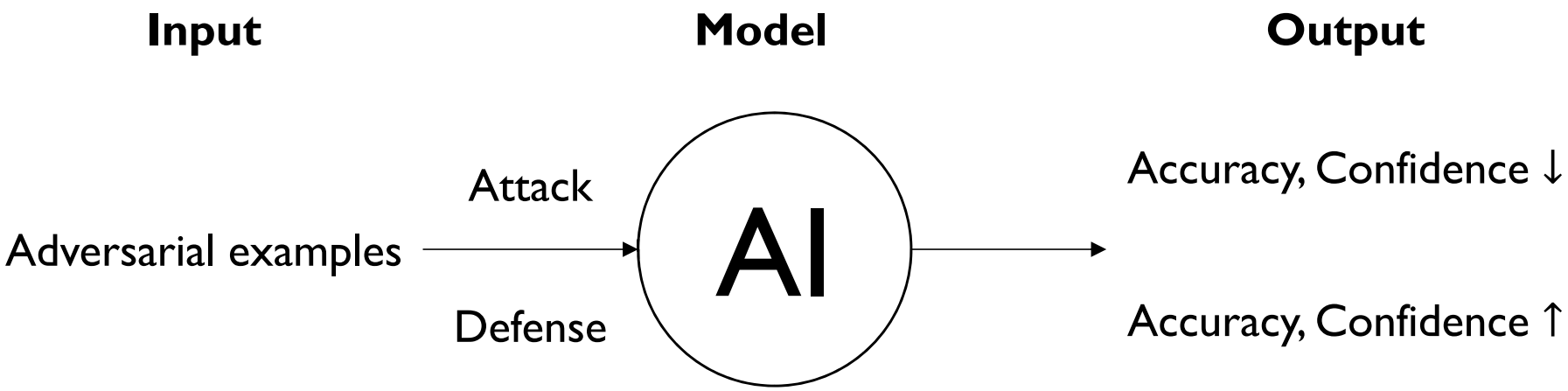
- Detector
  - 입력 데이터가 Adversarial data 인지 아닌지 판별하는 AI 사용 (추가 model 필요)
  - SafetyNet: original classifier + adversary detector



Lu, J., Issaranon, T., & Forsyth, D. (2017). Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 446-454).

Conclusions

- Adversarial attacks & defenses in deep learning
  - Attacker:“이런 데이터는 못 맞추겠지?”
  - Defender:“맞춤”



정확하고 강건한 예측모델을 구축하는 것이 목표

Conclusions

- **Adversarial examples: original data + perturbation**
- **Generating adversarial example → Adversarial attack**
  - 목적: Confidence reduction: AI의 예측 신뢰도를 낮추는 공격
  - 방법: Perturbation 생성
  - 대표 사례: The Fast Gradient Sign Method
- **Defense methods: 모델 구조, 속도에 크게 영향 없이 정확도↑**
  ① Adversarial training: original data + adversarial data
  ② Defensive distillation: original model (X, Y: knowledge distillation 을 사용)
  ③ Detector: original model + adversarial detector
- **Applications**
  - 분석 데이터 내 Noise가 존재하는 데이터셋(senor signal 등)에 강건한 예측모델을 구축

References

- Attack & defense (1): Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Attack (2) & defense (1): Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Attack (3): Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).
- Defense (2): Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582-597). IEEE.
- Defense (3): Lu, J., Issaranon, T., & Forsyth, D. (2017). Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 446-454).
- Review (1): Zhang, J., & Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*.
- Review (2): Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*.

Thank You !

Appendix

Adversarial Attacks

- 왜 인간은 육안적으로 구별이 어려운가?
  - 인간: 고해상도의 눈을 가지고 저차원 (3차원) 세계에서 보고 있는 때문..
  - AI: 모든 픽셀 변화를 감지하는 고차원 (2,352 차원) 세계에서 보고 있는 때문 ..

Human: this is panda!

가로, 세로, 색

AI: this is a gibbon!

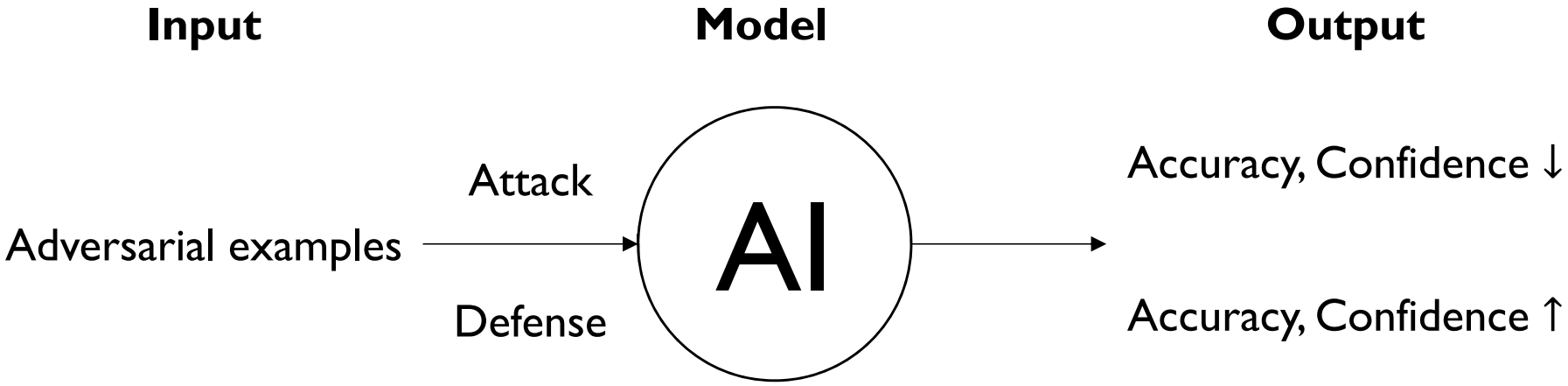
2,352 dimensional data

Appendix

- "Training on adversarial examples is somewhat different from other data augmentation; usually, one augments the data with transformations such as transformations that are expected to actually occur in the test set. This form of data augmentation instead uses inputs that are unlikely to occur naturally but that expose flaws in the ways that the model conceptualizes its decision function." (Goodfellow et al., 2015)
- Data augmentation 과 공통점
  - 원본 데이터를 변형하여 획득
- Data augmentation 과 차이점
  - Data augmentation: 실제로 발생할 법한 데이터를 원본 데이터 변형으로 획득
  - Training the adversarial examples: 실제로 발생할 법한 가능성은 낮지만 모델이 의사결정기능을 개념화 하는 방식의 결함을 노출하는 입력으로 사용

Appendix

- Adversarial attack: adversarial example generating → robustness of model ↓
  - Confidence reduction: AI의 예측 신뢰도를 낮추는 공격(90%->55%)
  - Targeted / Non-targeted misclassification: 의도한 / 오답을 유발하는 공격
  - Source-target misclassification: 입력에 따라서 오답을 유발
- Defense techniques
  ① Adversarial training: 공격에서 생성된 이미지를 추가 학습데이터로 활용
  ② Defensive distillation: knowledge distillation을 사용
  ③ Detector: 입력이 noise가 추가된 이미지인지 아닌지 판별 (additional model)



EOD