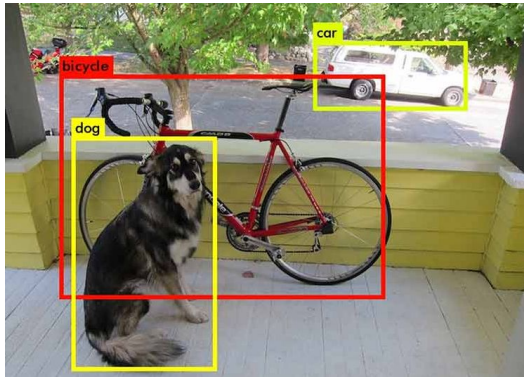
Adversarial Examples

Nguyễn Hồ Thăng Long - Nguyễn Hoàng Quân

Học máy được sử dụng thành công trong nhiều bài toán

Image Recognition



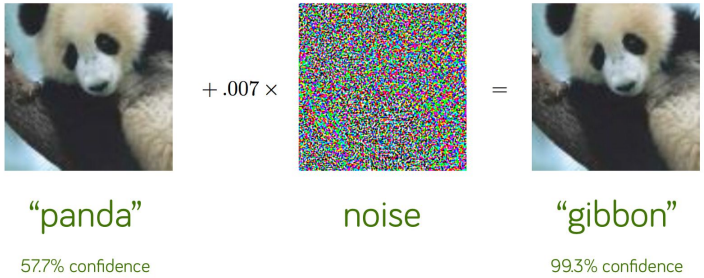
Object Detection

Visual Question Answering

Voice Recognition

và nhiều bài toán khác...

Nhiều mô hình học máy dễ bị tấn công



Dữ liệu số

Thế giới thực

Goodfellow et. al., Explaining and Harnessing Adversarial Examples, ICLR 2015
Eykholt et. al., Robust Physical-World Attacks on Deep Learning Models, CVPR 2018

Adversarial attack

Là một kỹ thuật tìm một sự thay đổi nhỏ ở đầu vào sao cho có thể thay đổi được kết quả đầu ra của mô hình máy học. Sự thay đổi này có thể rất nhỏ sao cho mắt người không thể nhận ra.

Adversarial attack - Ví dụ

ResNet → Chair 99.98%

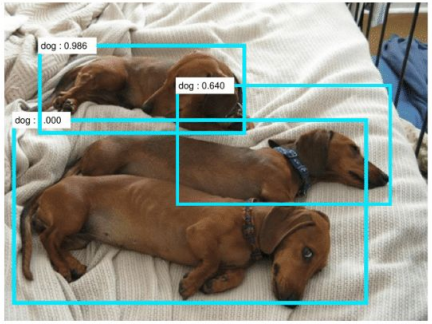
ResNet → Dining Table, 99.97%

Có điểm gì khác nhau
giữa hai tấm ảnh này?

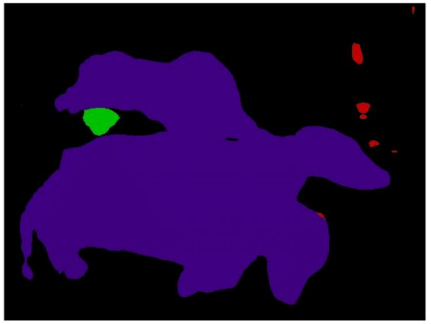
Adversarial attack trong bài toán thực tế

Original Image

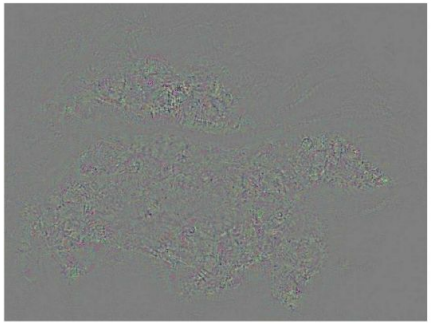
Original Image Detection



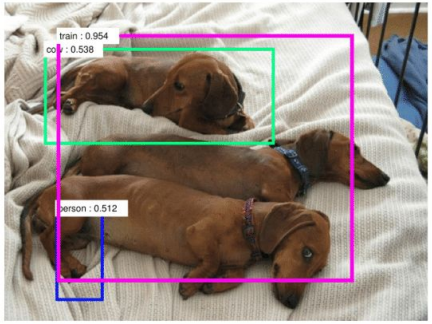
Original Image Segmentation



Adversarial Perturbation



Adversarial Image Detection



Adversarial Image Segmentation

Xinyun Chen et. al. Adversarial Examples for Semantic Segmentation and Object Detection

Phương pháp tấn công (whitebox)
Đặt vấn đề

User:

- là người sử dụng mô hình
- làm lộ thông tin mô hình

Attacker:

- biết tham số của mô hình (target)
- Không có quyền truy cập

Phương pháp tấn công (whitebox)
Mô hình hóa bài toán

Non-target adversarial example: Khiến cho mô hình dự đoán sai.

$$\max_{x'} J(\theta, x', y)$$
$$s.t. \left\| x - x' \right\|_p \leq \epsilon$$

Target adversarial example: Khiến cho mô hình dự đoán ra kết quả chỉ định khác.

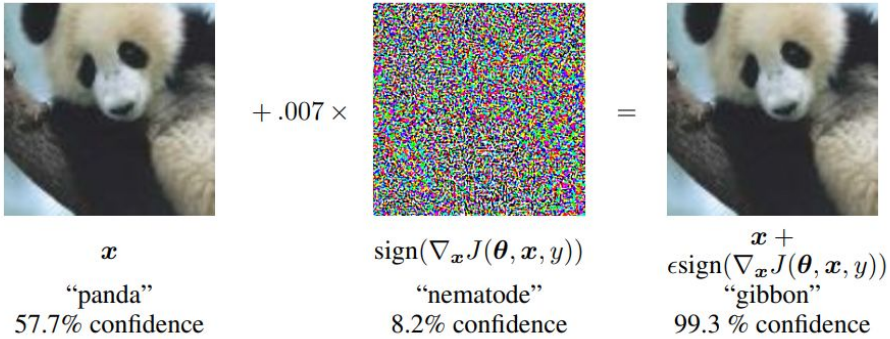
$$\min_{x'} J(\theta, x', y')$$
$$s.t. \left\| x - x' \right\|_p \leq \epsilon$$

Trong đó x là input gốc, y là nhãn, x' là adversarial example.

Phương pháp tấn công
Fast Gradient-Sign Method (FGSM)

$$\|x - x'\|_\infty \leq \epsilon$$

$$x' = x + \epsilon sign(\nabla_x J(\theta, x, y))$$



Ưu điểm:

- Là phương pháp tấn công đơn giản.
- Có thể tấn công được nhiều mô hình không có bất kỳ cơ chế bảo vệ nào.

Nhược điểm:

- Không hiệu quả đối với các mô hình có cơ chế bảo vệ.

Goodfellow et. al., Explaining and Harnessing Adversarial Examples, ICLR 2015

Phương pháp tấn công
Iterative Fast Gradient Sign Method (I-FGSM)

$$x'_0 = x$$
$$x'_{i+1} = x'_i + \epsilon sign(\nabla_{x'_i} J(\theta, x'_i, y))$$

- Hiệu quả hơn so với phương pháp trước.
- Việc xây dựng cơ chế phòng thủ trở nên khó khăn hơn.

Kurakin et. al., Adversarial Machine Learning at Scale, ICLR 2017

Phương pháp tấn công (blackbox)
Đặt vấn đề

User:

- là người sử dụng mô hình
- làm lộ thông tin mô hình

Attacker:

- Không biết tham số của mô hình (target)
- Không có quyền truy cập

Phương pháp tấn công (blackbox)
So sánh

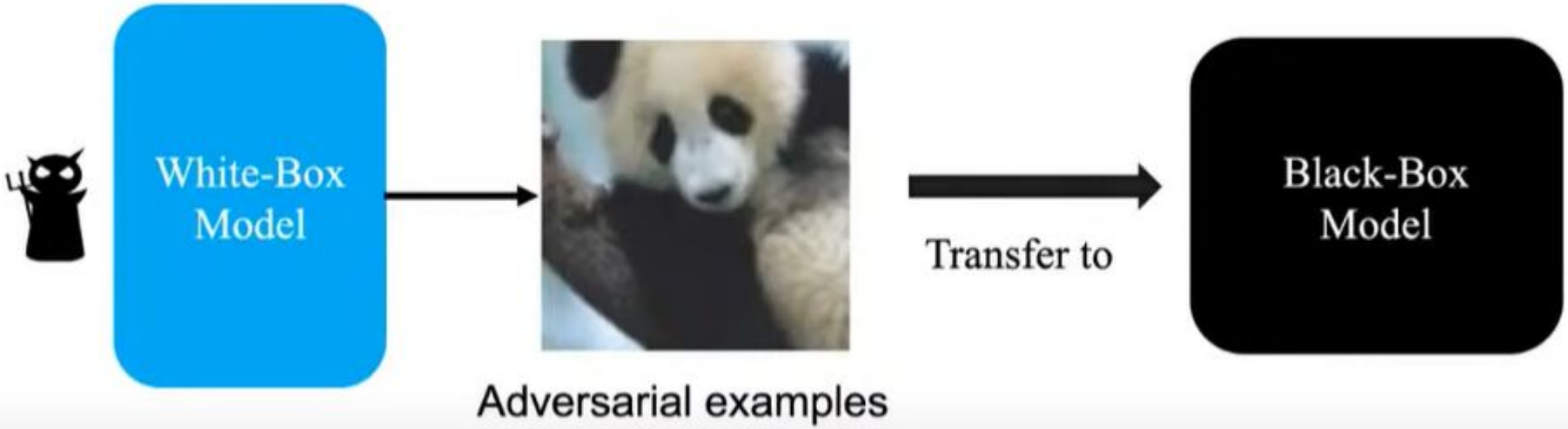
White-box attacks:

- Biết được kiến trúc, tham số của mô hình.
- Dễ tấn công.

Black-box attacks:

- Không biết bất cứ thông tin gì của mô hình.
- Việc tấn công thường thông qua các API.

Phương pháp tấn công (blackbox)
Transferable adversarial examples



Xinyun Chen et. al. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

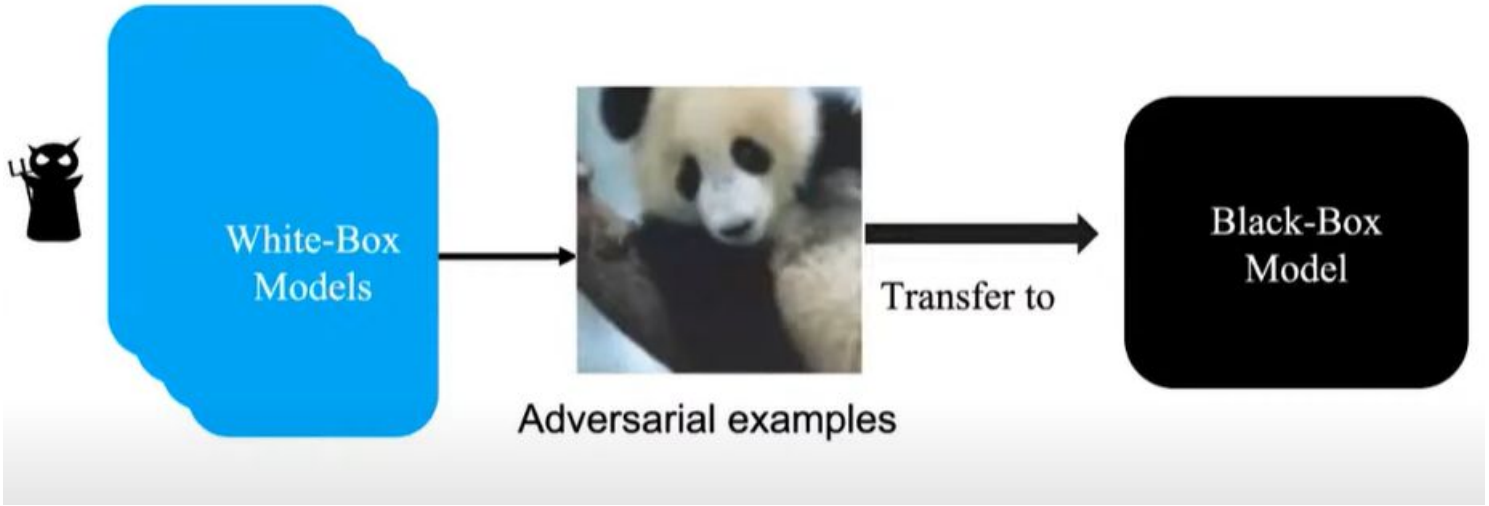
Phương pháp tấn công (blackbox)
Transferable adversarial examples

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16
ResNet-152	22.83	0%	13%	18%	19%
ResNet-101	23.81	19%	0%	21%	21%
ResNet-50	22.86	23%	20%	0%	21%
VGG-16	22.51	22%	17%	17%	0%

Trong đó RMSD là: $d(x, x^*) = \sqrt{\sum_i (x_i^* - x_i)^2 / M}$, M: image size

Xinyun Chen et. al. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

Phương pháp tấn công (blackbox)
Attacking an ensemble model



Nếu ảnh adversarial tấn công thành công trên N-1 model, liệu mẫu đó có tấn công được model thứ N hay không?

Xinyun Chen et. al. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

Phương pháp tấn công (blackbox)
Attacking an ensemble model

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

Xinyun Chen et. al. Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

Phương pháp phòng thủ

Liệu có cách nào có thể chống lại được adversarial examples hay không?

Phương pháp phòng thủ

Liệu có cách nào có thể chống lại được adversarial examples hay không?

=> Có.

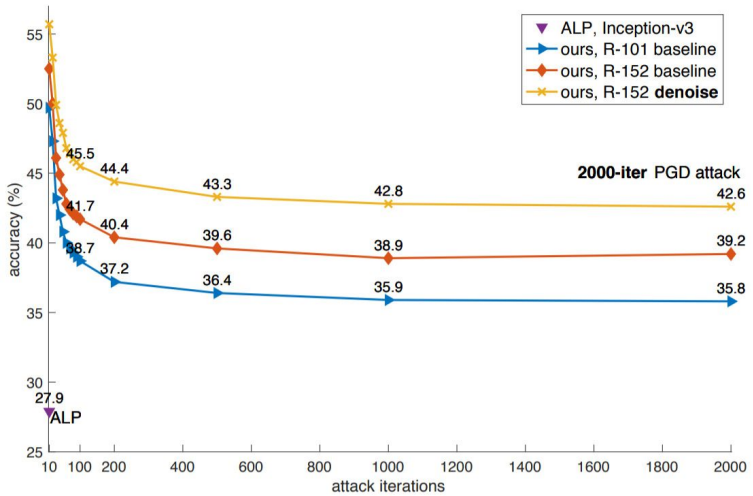
Ý tưởng của phương pháp phòng thủ:

- Tăng cường dữ liệu để mô hình có thể chống chịu được adversarial examples.
- Cố gắng loại bỏ các thay đổi làm cho mô hình dự đoán sai.

Phương pháp phòng thủ
Ý tưởng ban đầu

Liệu có thể sử dụng adversarial examples để tăng cường huấn luyện không?

=> Adversarial Training



Xie et. al., Feature Denoising for Improving Adversarial Robustness, CVPR 2019.

Phương pháp phòng thủ
Adversarial Training

Mã giả:

Repeat:

Lấy một minibatch B từ tập huấn luyện;

Tạo tập adversarial examples từ tập B và kết hợp với tập B tạo thành tập B';

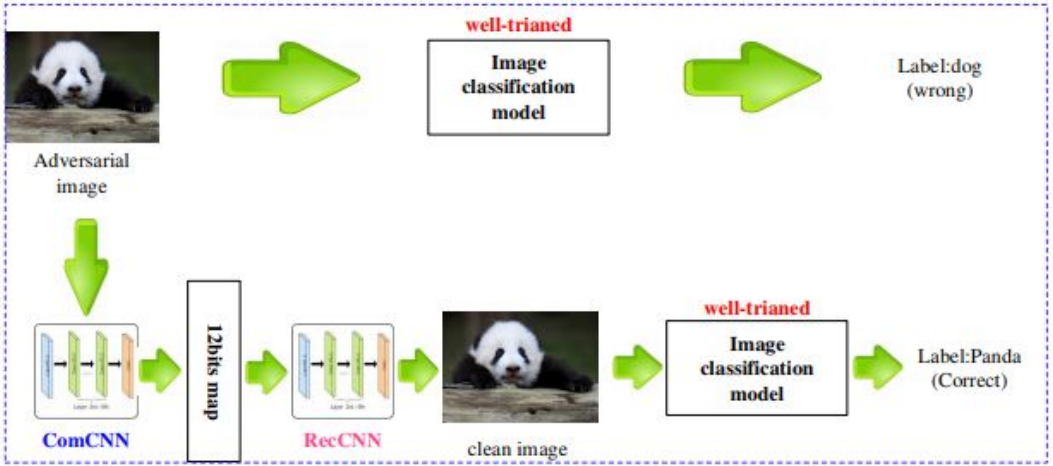
Thực hiện một bước huấn luyện từ tập B';

Until quá trình huấn luyện hội tụ.

Phương pháp phòng thủ
Images Compression

Ý tưởng: Adversarial examples được tạo từ nhiễu có chủ đích để tấn công.

=> Tái tạo lại mẫu để triệt tiêu nhiễu trước đó.



Jia et. al., ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples, CVPR 2019.

Phương pháp phòng thủ
Ensemble Models

Tăng cường dữ liệu huấn luyện từ nhiều adversarial examples lấy từ nhiều model.

Kết hợp nhiều model để dự đoán.

Tramèr et. al., Ensemble Adversarial Training: Attacks and Defenses, ICLR 2018.

Demo