Tấn công Elastic-Net vào DNN qua một số mẫu đối nghịch

Khai phá dữ liệu và học máy

Nguyễn Mạnh Linh, Đào Thị Thu Hồng

Khoa Toán-Cơ-Tin học
Đại học Khoa học Tự nhiên

2022/12

Outline

Giới thiệu

Sự dễ tổn thương của DNNs

Hình 1: Các mẫu đối nghịch bị phân loại sai bởi mô hình Inception-V3

Giới thiệu - Các loại tấn công

1. Tấn công nhắm đích - *targeted attacks*
2. Tấn công không nhắm đích - *untargeted attacks*
3. Tấn công chuyển giao - *transfer attacks*

Giới thiệu: Huấn luyện đối nghịch

Huấn luyện đối nghịch (*adverarial training*) (Madry et al. 2017): Sử dụng mẫu đối nghịch để huấn luyện một mô hình mạnh có khả năng chống chịu với các nhiễu của mẫu đối nghịch.

Giới thiệu: Tấn công phân loại ảnh

1. Tấn công phân loại ảnh dựa trên mạng neuron tích chập
2. Mẫu đối nghịch được tạo ra để làm sai lệch kết quả phân loại
3. Mẫu mới được tạo ra (gần) giống với mẫu gốc

Giới thiệu: Độ méo

$\|x\|_q = (\sum_{i=1}^{p} |x_i|^q)^{\frac{1}{q}}$ kí hiệu chuẩn L_q của vector p chiều
$x = [x_1, ..., x_p]$ với $q \geq 1$

- L_∞: Đánh giá sự thay đổi tối đa các pixel (oodfellow, Shlens, and Szegedy 2015)
- L_2: Cải thiện chất lượng hình ảnh (Carlini and Wagner 2017b)
- L_1: Sử dụng trong các bài toán phục hồi ảnh (Fu et al. 2006)

Trong bài toán mẫu đối nghịch, độ biến dạng L_1 đánh giá tổng các thay đổi trong nhiễu loạn và đóng vai trò là một thành phần (hàm) lồi đo lường số lượng pixel thay đổi (độ thưa) gây ra bởi nhiễu

Giới thiệu - Các tập dữ liệu thử nghiệm

Tấn công dựa trên L_1 với các tập dữ liệu

- MNIST
- CIFRA10
- ImageNet

Được so sánh với các tấn công dựa trên L_2 và L_∞

Outline

Nghiên cứu liên quan

This is a block

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.

This is an alert block

Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis.

Tấn công DNNs - FGM, I-FGM

- Kí hiệu $\mathbf{x_0}$ và $\mathbf{x}$ lần lượt là mẫu gốc và mẫu đối nghịch, t là lớp mục tiêu cần tấn công.
- Tấn công dựa trên L_∞

$$\mathbf{x} = \mathbf{x_0} - \epsilon \times \text{sign}(\nabla J(\mathbf{x_0}, t)) \tag{1}$$

với ϵ là độ biến dạng L_∞ giữa $\mathbf{x}$ và $\mathbf{x_0}$ và $\text{sign}(\nabla J)$ là dấu của gradient

- Tấn công dựa trên L_1 và L_2

$$\mathbf{x} = \mathbf{x_0} - \epsilon \frac{\nabla J(\mathbf{x_0}, t)}{\|\nabla J(\mathbf{x_0}, t)\|_q} \tag{2}$$

với $q = 1, 2$ và ϵ là độ méo tương quan

Tấn công DNNs - C&W

- Thay vì sử dụng hàm mất mát trên tập huấn luyện Carlini và Wagner đã thiết kế một hiệu chỉnh L_2 trong hàm mất mát dựa trên lớp logit trong DNNs để sinh ra các mẫu đối nghịch (Carlini and Wagner 2017b)
- Công thức này hóa ra là một trường hợp riêng của thuật toát EAD (sẽ được trình bày trong phần sau)

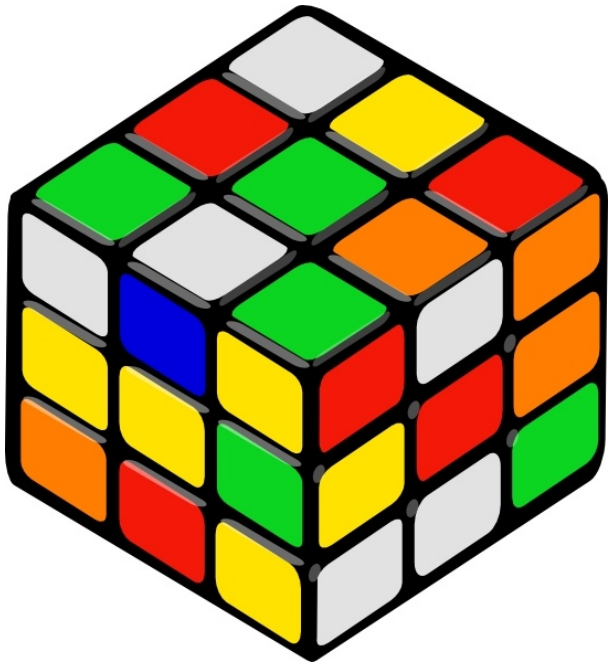
Tấn công DNNs - C&W

- Defensive distillation - Chưng cất phòng thủ (Papernot et al.2016b)
- Adversarial training - Huấn luyện đối nghịch (Zheng et al. 2016; Madry et al. 2017; Tram‘er et al. 2017; Zantedeschi, Nicolae, and Rawat 2017)
- Detection methods - Phương pháp dò tìm (Feinman et al. 2017; Grosse et al. 2017; Lu, Issaranon, and Forsyth 2017; Xu, Evans, and Qi 2017)

Outline

Text and Image in Beamer

1. This is a list
2. This has a sub-list
   - This is a sub-list
   - This has bullet points
3. The list has numbers



Hình 2: Random Image

Outline

Outline

Conclusion

This is an example block

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl.

References