

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

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 EAD: Tấn công Elastic-Net vào DNN
- 4 EAD: Tấn công Elastic-Net vào DNN
- 5 Đánh giá hiệu năng
- 6 Kết luận

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

- ➊ Tấn công nhắm đích - *targeted attacks*
- ➋ Tấn công không nhắm đích - *untargeted attacks*
- ➌ Tấn công chuyển giao - *transfer attacks*

Giới thiệu: Huấn luyện đối nghịch

Huấn luyện đối nghịch (*adversarial 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

- ➊ Tấn công phân loại ảnh dựa trên mạng neuron tích chập
- ➋ Mẫu đối nghịch được tạo ra để làm sai lệch kết quả phân loại
- ➌ 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 $x = [x_1, \dots, x_p]$ 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) lỗi đo lường số lượng pixel thay đổi (độ thưa) gây ra bởi nhiễu

Outline

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 EAD: Tấn công Elastic-Net vào DNN
- 4 EAD: Tấn công Elastic-Net vào DNN
- 5 Đánh giá hiệu năng
- 6 Kết luận

Building Up

This is a block

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

This is an alert block

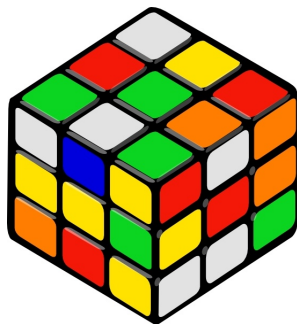
Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis.

Outline

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 EAD: Tấn công Elastic-Net vào DNN
- 4 EAD: Tấn công Elastic-Net vào DNN
- 5 Đánh giá hiệu năng
- 6 Kết luận

Text and Image in Beamer

- ❶ This is a list
- ❷ This has a sub-list
 - ▶ This is a sub-list
 - ▶ This has bullet points
- ❸ The list has numbers



Hình 2: Random Image

Outline

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 EAD: Tấn công Elastic-Net vào DNN
- 4 EAD: Tấn công Elastic-Net vào DNN**
- 5 Đánh giá hiệu năng
- 6 Kết luận

Outline

- 1 Giới thiệu
- 2 Các nghiên cứu liên quan
- 3 EAD: Tấn công Elastic-Net vào DNN
- 4 EAD: Tấn công Elastic-Net vào DNN
- 5 Đánh giá hiệu năng**
- 6 Kết luận

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