

TIỂU LUẬN

Khai phá dữ liệu và học máy
Tấn công Elastic-Net vào mạng thần kinh sâu
qua một số mẫu đối nghịch

Đại học Quốc gia Hà Nội
Đại học Khoa học Tự nhiên
Khoa Toán cơ tin

Giảng viên:

Trần Trọng Hiếu

Học viên:

Nguyễn Mạnh Linh, Đào Thị Thu Hồng

Mục lục

Tóm tắt

Các nghiên cứu gần đây đã chỉ ra tính dễ bị tổn thương của các mạng nơ ron sâu (Deep Neural Networks - DNNs) đến các mẫu đối nghịch - một cách trực quan, hình ảnh đối nghịch không thể phân biệt có thể được tạo ra một cách dễ dàng khiến các mô hình được huấn luyện tốt cũng phân loại ảnh sai. Các phương pháp hiện có để tạo ra mẫu đối nghịch thường dựa trên thông số biến dạng L_2 và L_∞ . Mặc dù trong thực tế độ biến dạng L_1 là quan trọng và quyết định đến tính thưa của nhiễu lại ít được xem xét.

Trong bài này, chúng tôi thiết kế một quy trình tấn công DNNs thông qua mẫu đối nghịch như một bài toán tối ưu hóa sử dụng hiệu chỉnh elastic-net. Tấn công DNNs bằng elastic-net (EAD) với tham số L_1 được thêm vào cùng với tấn công L_2 . Kết quả thực nghiệm trên các tập dữ liệu MNIST, CIFAR10 và ImageNet chỉ ra rằng EAD có thể mang lại một tập mẫu đối nghịch với độ nhiễu L_1 nhỏ và đạt được hiệu suất tấn công tương đương với các phương pháp hiện đại nhất qua các kịch bản tấn công khác nhau. Quan trọng hơn, EAD dẫn đến sự cải tiến trong việc tấn công DNN và gợi ý những hiểu biết mới về hiệu chỉnh L_1 để cải thiện bảo mật trong các mô hình học máy.

1 Giới thiệu

Mạng nơ ron sâu (DNNs) đã đạt hiệu quả rất tốt với các bài toán trong học máy và trí tuệ nhân tạo như phân loại ảnh, nhận diện giọng nói, dịch máy và trò chơi. Mặc dù DNNs rất hiệu quả nhưng một số nghiên cứu gần đây đã chứng minh DNNs rất dễ "tổn thương" với các mẫu đối nghịch (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015). Ví dụ, một hình ảnh với nhiều được thiết kế cẩn thận có thể làm cho một DNNs đã được huấn luyện phân loại sai. Tệ hơn nữa, các mẫu đối nghịch được tạo ra hầu như không thể phân biệt được bằng mắt người.



Hình 1.1: Minh họa trực quan về mẫu đối nghịch được sinh bởi EAD. Hình gốc (đà điểu) được lấy từ tập ImageNet. Các mẫu đối nghịch bị phân loại sai với mô hình Inception-v3.

Ví dụ trực quan trên thể hiện 3 mẫu đối nghịch của một hình con đà điểu ("ostrich") được sinh ra bằng thuật toán của chúng tôi. Các mẫu này được mô hình Inception-v3 (Szegedy et al. 2016) phân loại thành "safe", "shoe shop" và "vacuum".

Sự thiếu mạnh mẽ của DNNs thể hiện trước các mẫu đối nghịch đã làm dấy lên những lo ngại nghiêm trọng về vấn đề bảo mật các ứng dụng, bao gồm nhận dạng tín hiệu giao thông và phát hiện phần mềm độc hại. Hơn nữa, vượt ra ngoài không gian kỹ thuật số, các nhà nghiên cứu đã chỉ ra rằng những mẫu đối nghịch này vẫn có hiệu quả trong thế giới vật chất trong việc đánh lừa DNNs (Kurakin, Goodfellow, and Bengio 2016a; Evtimov et al. 2017). Do tính mạnh mẽ và ý nghĩa bảo mật, các phương tiện tạo ra các mẫu đối nghịch được gọi là các cuộc tấn công (*attacks*) vào DNNs. Cụ thể, các cuộc tấn công có chủ đích (*targeted attacks*) nhằm mục đích tạo ra các mẫu đối nghịch được phân loại nhằm thành các lớp mục tiêu cụ thể và các cuộc tấn công không có mục tiêu (*untargeted attacks*) nhằm mục đích để tạo ra các mẫu đối nghịch không được phân loại như lớp học ban đầu. Các cuộc tấn công chuyển giao (*transfer attacks*) nhằm mục đích tạo ra các mẫu đối nghịch có thể chuyển từ mô hình DNN này sang mô hình DNN khác. Ngoài việc đánh giá mức độ mạnh mẽ của DNNs, Các mẫu đối nghịch có thể được sử dụng để huấn luyện một mô hình mạnh có khả năng chống chịu với những xáo trộn của đối nghịch, được gọi

là huấn luyện đối nghịch (*adversarial training*) (Madry et al. 2017). Chúng cũng có được sử dụng để giải thích DNNs (Koh và Liang 2017; Dong et al. 2017).

Trong bài báo này, chúng tôi sử dụng các mẫu đối nghịch để tấn công phân loại ảnh dựa trên mạng nơ ron tích chập. Mẫu đối nghịch được tạo ra để làm sai lệch kết quả dự đoán và phải đảm bảo hình mới tạo ra giống với hình gốc. Trong quá khứ, sự giống nhau giữa mẫu đối nghịch được tạo ra và hình gốc được đo bằng các tham số độ méo (*distortion metrics*) khác nhau. Một metric thường được sử dụng là chuẩn L_q với $\|x\|_q = (\sum_{i=1}^p |x_i|^q)^{\frac{1}{q}}$ kí hiệu chuẩn L_q của vector p chiều $x = [x_1, \dots, x_p]$ với $q \geq 1$. Đặc biệt, khi tạo ra các mẫu đối nghịch, độ méo L_∞ được sử dụng để đánh giá sự thay đổi tối đa giá trị pixel (Goodfellow, Shlens, and Szegedy 2015), trong khi độ méo L_2 được sử dụng để cải thiện chất lượng hình ảnh (Carlini and Wagner 2017b). Tuy nhiên, trong thực tế, chuẩn L_1 được sử dụng rộng rãi trong các bài toán phục hồi ảnh (Fu et al. 2006) cũng như phục hồi tính thưa (Candès and Wakin 2008). Các mẫu đối nghịch dựa trên L_1 chưa được xem xét một cách kĩ càng. 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. Để lấp đầy khoảng trống này, chúng tôi xem xét một thuật toán tấn công dựa trên hiệu chỉnh *elastic-net*, được gọi là tấn công elastic-net vào DNNs (elastic-net attacks to DNNs - EAD). Hiệu chỉnh elastic-net là tổ hợp tuyến tính của các hàm penalty L_1 và L_2 và nó cũng là công cụ tiêu chuẩn cho bài toán lựa chọn tính chất đặc trưng cho dữ liệu nhiều chiều (Zou and Hastie 2005). Trong bài toán tấn công DNNs, EAD mở ra hướng nghiên cứu mới từ cách tấn công hiện đại dựa trên L_2 (Carlini and Wagner 2017b) và cũng đề xuất tấn công hiệu quả hơn theo định hướng L_1 so với các phương pháp tấn công hiện có.

Để khám phá hiệu quả của tấn công dựa trên L_1 , chúng tôi tiến hành các thử nghiệm trên tập dữ liệu MNIST, CIFAR10, và ImageNet trong các tình huống tấn công khác nhau. So sánh với các phương pháp tấn công hiện có dựa trên L_2 và L_∞ (Kurakin, Goodfellow, and Bengio 2016b; Carlini and Wagner 2017b), EAD có thể đạt được tỉ lệ tấn công thành công tương tự khi phá vỡ DNNs được phòng thủ hoặc không phòng thủ (Papernot et al. 2016b). Quan trọng hơn, chúng tôi chỉ ra rằng, tấn công L_1 đạt được hiệu suất vượt trội so với các cuộc tấn công L_2 và L_∞ trong các cuộc tấn công chuyển giao với các mẫu đối thủ được huấn luyện bổ sung. Với tập dữ liệu khó nhất (MNIST), các kết quả từ EAD cải thiện tấn công chuyển giao vào DNN không được phòng thủ hay được phòng thủ đạt tỉ lệ thành công gần 99%. Thêm vào đó việc huấn luyện kèm theo với mẫu đối nghịch dựa trên L_1 và L_2 có thể tăng cường khả năng phục hồi của DNNs đối với các nhiễu loạn. Những kết quả này gợi

ý rằng EAD mang lại tập mẫu đối nghịch khác biệt nhưng hiệu quả hơn. Hơn thế nữa, đánh giá các cuộc tấn công dựa trên độ méo L_1 cung cấp thêm hiểu biết mới về học máy đối kháng và sự bảo mật của DNNs, gợi ý rằng L_1 có thể bổ sung cho L_2 và L_∞ thúc đẩy các framework về học máy đối kháng hoàn thiện hơn.

2 Các nghiên cứu liên quan

Trong phần này, chúng tôi tổng hợp các nghiên cứu liên quan về tấn công và phòng thủ DNNs.

2.1 Tấn công vào DNNs

FGM và 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. Các phương pháp đạo hàm nhanh (*fast gradient methods - FGM*) sử dụng gradient ∇J của hàm mất mát trong tập huấn luyện J với \mathbf{x}_0 để tạo ra các mẫu đối nghịch (Goodfellow, Shlens, and Szegedy 2015). Với tấn công L_∞ , \mathbf{x} được tính bằng công thức:

$$\mathbf{x} = \mathbf{x}_0 - \epsilon \times \text{sign}(\nabla J(\mathbf{x}_0, t)) \quad (2.1)$$

Trong đó ϵ là độ biến dạng L_∞ giữa \mathbf{x} và \mathbf{x}_0 và $\text{sign}(\nabla J)$ là dấu của gradient. Với tấn công L_1 và L_2 , \mathbf{x} được tính bằng công thức:

$$\mathbf{x} = \mathbf{x}_0 - \epsilon \frac{\nabla J(\mathbf{x}_0, t)}{\|\nabla J(\mathbf{x}_0, t)\|_q} \quad (2.2)$$

với $q = 1, 2$ và ϵ là độ méo tương quan. Các phương pháp lặp đạo hàm nhanh (*Iterative fast gradient methods (I-FGM)*) được trình bày trong (Kurakin, Goodfellow, and Bengio 2016b) là phương pháp lặp sửa dụng FGM với một độ méo mịn hơn. Các cuộc tấn công không mục tiêu sử dụng FGM và I-FGM có thể được thực hiện theo cách tương tự nhau.

C&W attack: 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 toán EAD của chúng tôi (sẽ được trình bày trong phần sau). Tấn công C&W được coi là một trong những tấn công mạnh nhất đối với DNNs vì nó có thể phá vỡ cả những DNNs được phòng thủ và chất lọc, nó cũng có thể đạt được hiệu suất đáng kể với tấn công chuyển giao.

DeepFool: là một thuật toán tấn công không mục tiêu L_2 (Moosavi-Dezfooli, Fawzi, and Frossard 2016) dựa trên lý thuyết về phép chiếu tối siêu phẳng phân tách gần nhất trong phân lớp. Nó cũng được sử dụng để tạo ra một nhiễu loạn phổ quát nhằm đánh lừa các DNN được huấn luyện trên các hình ảnh tự nhiên (Moosavi-Dezfooli et al. 2016).

2.2 Phòng thủ trong DNNs

Defensive distillation: Chứng cất phòng thủ (Papernot et al.2016b) bảo vệ chống lại sự nhiễu loạn của đối nghịch bằng cách sử dụng kỹ thuật chứng cất trong (Hinton, Vinyals, and Dean 2015) để huấn luyện lại cùng một mạng với xác suất lớp được dự đoán bởi mạng ban đầu. Phương pháp này đưa ra tham số nhiệt độ T trong lớp softmax để tăng cường độ mạnh đối với những xáo trộn của mẫu đối nghịch.

Adversarial training: Có thể được triển khai theo 1 số cách khác nhau. Hướng tiếp cận chuẩn là thêm vào tập dữ liệu huấn luyện ban đầu các mẫu đối nghịch đã được sửa đúng nhãn, sau đó huấn luyện lại mạng. (Zheng et al. 2016; Madry et al. 2017; Tram'et et al. 2017; Zantedeschi, Nicolae, and Rawat 2017) đề xuất điều chỉnh training loss hoặc kiến trúc mạng để tăng cường sức mạnh của DNNs trước mẫu đối nghịch.

Detection methods: Sử dụng các test thống kê để phân biệt các mẫu đối nghịch với mẫu gốc (Feinman et al. 2017; Grosse et al. 2017; Lu, Issaranon, and Forsyth 2017; Xu, Evans, and Qi 2017). Tuy nhiên, 10 phương pháp phát hiện khác nhau đã không thể phát hiện được tấn công C&W (Carlini and Wagner 2017a).

3 EAD: Tấn công Elastic-Net vào mạng DNNs