ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

TIỂU LUẬN

Khai phá dữ liệu và học máy

Tấn công Elastic-Net vào mạng thần kinh sâu
qua một số mẫu đối thủ

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 đấy đã chỉ ra tính dễ bị tổn thương của các mạng nơ ron sâu (Deep Neural Networks - DNNs) đến các mẫu đối thủ - một cách trực quan, hình ảnh đối thủ 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 thủ 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 thủ 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 thủ với độ nhiều 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 mốt số nghiên cứu gần đây đã chứng minh DNNs rất dễ "tổn thương" với các mẫu đối thủ (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2015). Ví dụ, một hình ảnh với nhiễu được thiết kế cẩn thận có thể làm cho một DNNs đã được huấn luyện phân loại sai. Tệ hơn nữa, các mẫu đối thủ được tạo ra hầu như không thể phân biệt được bằng mắt người.

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

Ví dụ trực quan trên thể hiện 3 mẫu đối thủ 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 thủ 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 thủ 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 thủ 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 thủ được phân loại nhầm 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 thủ 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 thủ 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 thủ có thể được sử dụng để huấn luyện một mô hình mạnh có khả năng chống chịu với những xáo trộn của đối thủ, được gọi là huấn luyện đối thủ

(*adverarial training*) (Madry et al. 2017). Chúng cũng có được sử dụng để giải thích DNNs (Koh và Liang 2017; Dong et al. 2017).