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 thần kinh 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.

1 Giới thiệu

Có nhiều bài toán học máy liên quan đến việc cực tiểu hóa hàm mục tiêu tổng hợp (Dhurandhar et al., 2018; Lu et al., 2014; Ribeiro et al., 2016; Xie et al., 2018). Ví dụ trong việc giải thích phân loại ảnh (Dhurandhar et al.,2018; Ribeiro et al., 2016), chúng ta cần tìm một tập đủ nhỏ các biến giải thích dự đoán bằng cách giải bài toán tối ưu hóa có ràng buộc sau:

$$\begin{aligned} &\min_{x\in\mathbb{R}} l(x) + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|_2^2 \\ &\text{s.t. } |x_i| \leq c_i \text{ với } i = 1, ..., d \end{aligned} \tag{1.1}$$