

# 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

1	Giới thiệu .....	1
---	------------------	---

## 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 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  $L_\infty$ . Mặc dù trong thực tế độ biến dạng  $L_1$  là quan trọng và quyết định đến tính thừa 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 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 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".