Phương pháp học sâu giải bài toán phát hiện Malware

Sinh viên thực hiện
Đặng Mạnh Cường

Giáo viên hướng dẫn
TS.Đinh Viết Sang

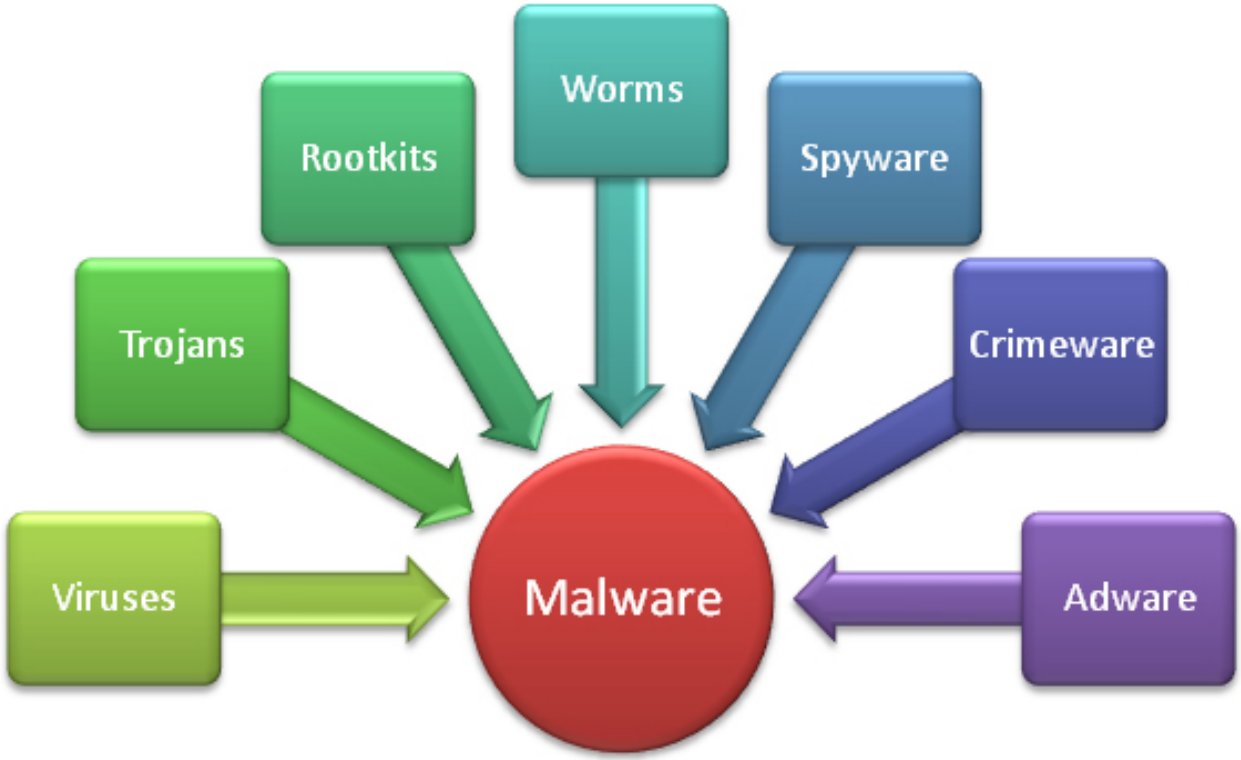
Hà Nội, 12-06-2018

Mục Lục

Malware là gì?

Giới thiệu chung

Malware - malicious software là các phần mềm độc hại, có mục đích phá hoại máy tính người dùng



Hình 1 : Các loại malware phổ biến

Bài toán phát hiện Malware

Cho tập dữ liệu đầu vào gồm các file PE, cần xác định mỗi mẫu PE có phải là malware hay không?
Về bản chất, bài toán phát hiện malware là bài toán phân loại nhị phân $\Rightarrow$ phân loại tập dữ liệu đầu vào thành 2 lớp:

- malware
- benign

Phương pháp tiếp cận

Phương pháp này đưa chia làm 2 giai đoạn chính:

- Trích chọn đặc trưng
- Huấn luyện

Mục Lục

Trích chọn đặc trưng

Phương pháp tiếp cận

0000	FF	D8	FF	E1	1D	FE	45	78	69	66	00	00	49	49	2A	00
0010	08	00	00	00	09	00	0F	01	02	00	06	00	00	00	7A	00
0020	00	00	10	01	02	00	14	00	00	00	80	00	00	00	12	01
0030	03	00	01	00	00	00	01	00	00	00	1A	01	05	00	01	00
0040	00	00	A0	00	00	00	1B	01	05	00	01	00	00	00	A8	00
0050	00	00	28	01	03	00	01	00	00	00	02	00	00	00	32	01
0060	02	00	14	00	00	00	B0	00	00	00	13	02	03	00	01	00
0070	00	00	01	00	00	00	69	87	04	00	01	00	00	00	C4	00
0080	00	00	3A	06	00	00	43	61	6E	6F	6E	00	43	61	6E	6F
0090	6E	20	50	6F	77	65	72	53	68	6F	74	20	41	36	30	00
00A0	00	00	00	00	00	00	00	00	00	00	00	00	B4	00	00	00
00B0	01	00	00	00	B4	00	00	00	01	00	00	00	32	30	30	34
00C0	3A	30	36	3A	32	35	20	31	32	3A	33	30	3A	32	35	00
00D0	1F	00	9A	82	05	00	01	00	00	00	86	03	00	00	9D	82
00E0	05	00	01	00	00	00	8E	03	00	00	00	90	07	00	04	00

Hình 2 : Biểu diễn PE dưới dạng mã máy

Trích chọn đặc trưng

Phương pháp tiếp cận

Các đặc trưng:

- N-gram
- Entropy histogram
- Biểu diễn ảnh
- Lời gọi hệ thống (trích chọn trực tiếp từ file PE)

N-gram

Trích chọn đặc trưng

- N-gram được sử dụng để biểu diễn tần suất xuất hiện của n bytes liên tiếp trong chuỗi bytes ban đầu.
- Không gian lưu trữ, độ phức tạp tính toán tỉ lệ thuận với n
  $\Rightarrow$ chọn $n = 1$
- Vector đặc trưng kích thước 256 biểu diễn tần suất xuất hiện của từng bytes.

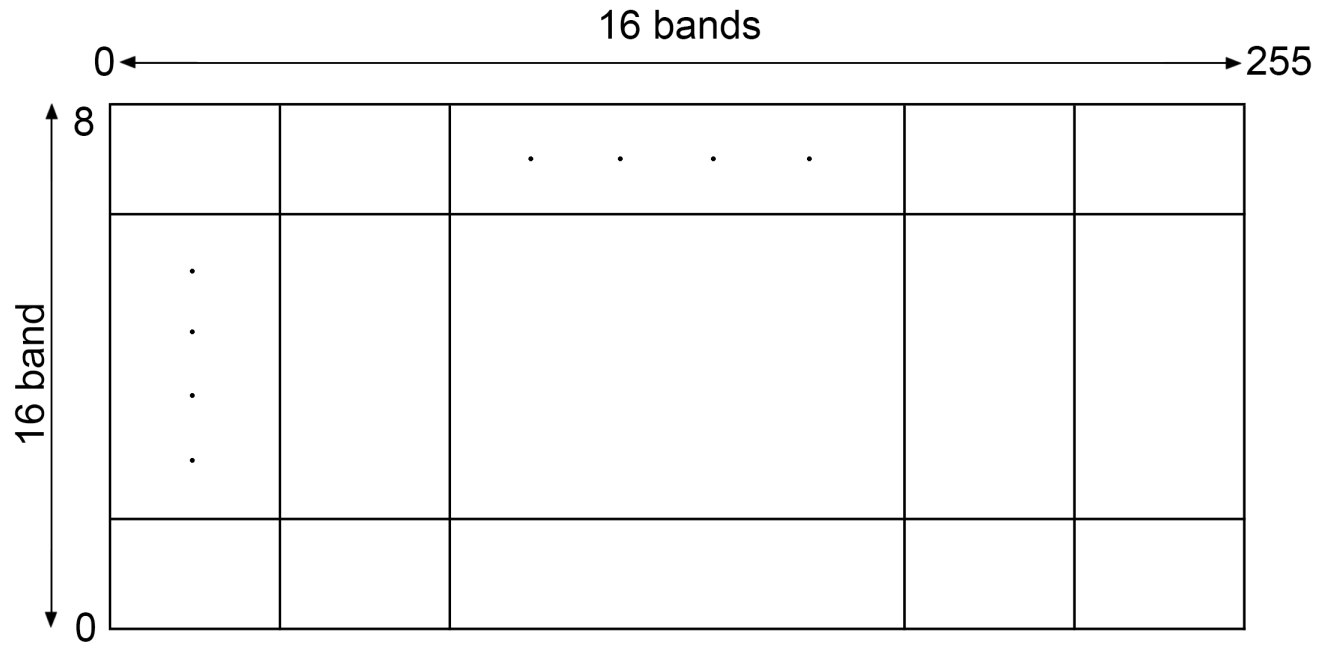
Entropy histogram

Trích chọn đặc trưng

- Entropy histogram được đề xuất bởi Joshua Saxe, đặc trưng này biểu diễn sự xuất hiện của cặp byte dữ liệu và giá trị entropy của chuỗi byte chứa nó.
- Quá trình trích xuất như sau:
  - Chia chuỗi bytes ban đầu thành n đoạn, mỗi đoạn có kích thước 1024 bytes.
  - Với mỗi đoạn, xây dựng vector v kích thước 1024, mỗi phần tử v_i là cặp giá trị (x, y), trong đó x là giá trị byte dữ liệu tương ứng với vị trí i trong đoạn, y là giá trị entropy của đoạn.
  - Sử dụng histogram để xây dựng vector đặc trưng 256 chiều dựa trên các vector tương ứng với mỗi đoạn trên.

Entropy histogram

Trích chọn đặc trưng



Hình 3 : Histogram

Giá trị của mỗi ô (x, y) trên hình biểu diễn tần suất xuất hiện của cặp giá trị (x, y) $\Rightarrow$ vector đặc trưng 256 chiều.

- Phương pháp biểu diễn file nhị phân dưới dạng ảnh được đề xuất bởi J.Nataraj.
- Mỗi byte trong file nhị phân tương ứng với giá trị một pixel trong ảnh.
- Duyệt hết toàn bộ file, thu được ảnh xám.
- Kích thước ảnh phụ thuộc vào kích thước file.

Bảng 1 : Kích thước ảnh tương ứng với kích thước file nhị phân

Kích thước file (KB)	**Chiều rộng ảnh**
< 10	32
10 - 30	64
30 - 60	128
60 - 100	256
100 - 200	384
200 - 500	512
500 - 1000	784
> 1000	1024

Để phù hợp với mô hình huấn luyện $\Rightarrow$ đưa các ảnh về cùng kích thước 64×64

Lời gọi hệ thống

Trích chọn đặc trưng

- Lời gọi hệ thống là các hàm hệ thống (api) được các file PE sử dụng để thực hiện một số hành vi nhất định.
- Sự xuất hiện của các hàm này có thể cung cấp thông tin cho việc phát hiện malware.
- Sử dụng tập 794 api được phân tích từ 500,000 mẫu malware $\Rightarrow$ xây dựng vector đặc trưng 794 chiều biểu diễn sự xuất hiện của các api trong mỗi mẫu PE.

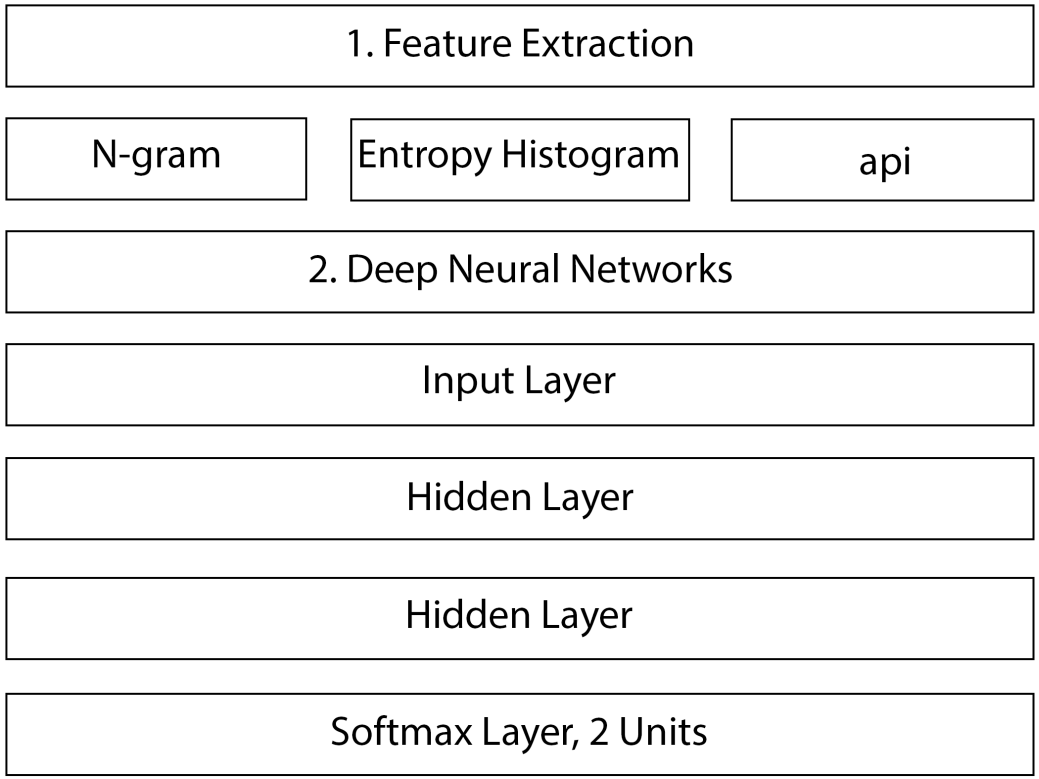
Mục Lục

Mô hình áp dụng

Phương pháp tiếp cận

Sử dụng 2 kiến trúc mạng nơ-ron nhân tạo:

- Deep Neural Nerworks (DNN): đối với 3 đặc trưng N-gram, ENT-HIS, api.
- Residual Neural Networks (RESNETS): đối với đặc trưng ảnh và sự kết hợp của cả 4 đặc trưng

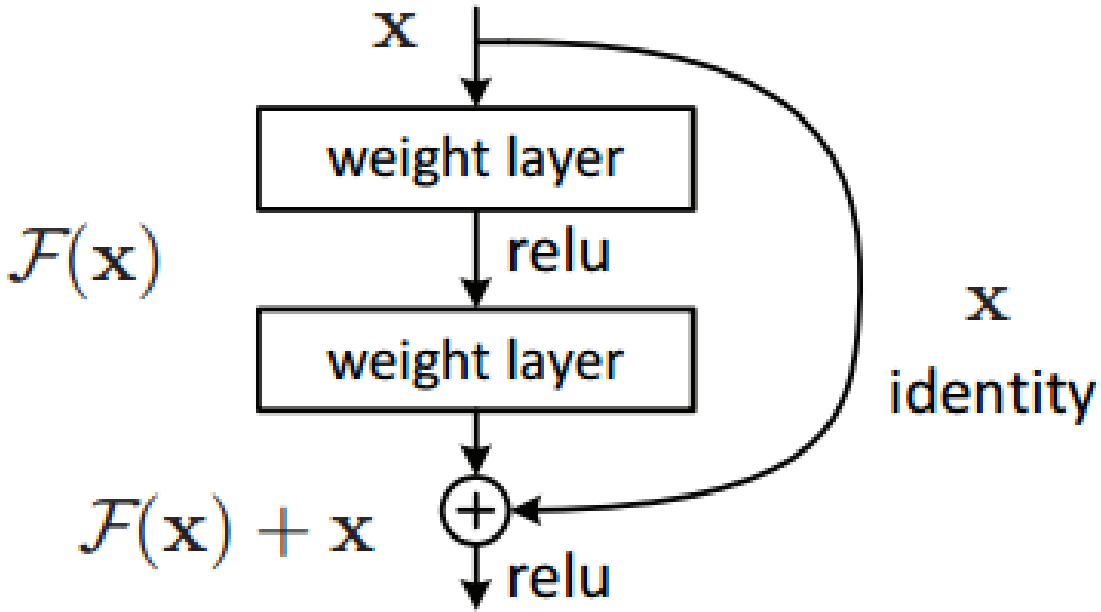


Hình 4 : Mô hình DNN

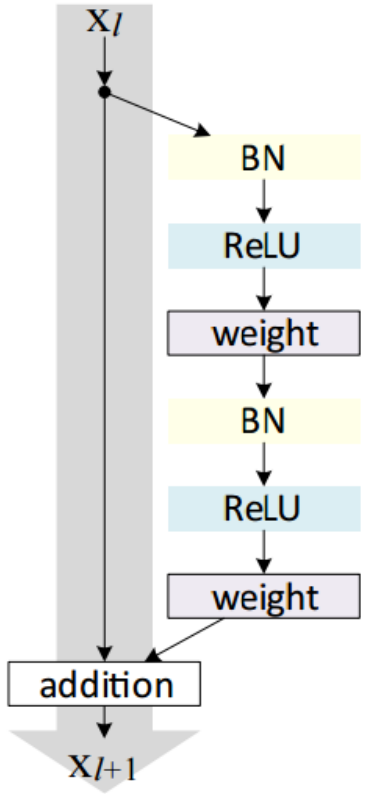
Bảng 2 : Kiến trúc mô hình với mỗi đặc trưng

Đặc trưng	Kiến trúc mô hình		
	Số lượng units lớp ẩn	Hàm kích hoạt	Batch normalization
N-gram	128	Relu	Có
ENT-HIS	128	Relu	Có
API	256	Relu	Có

- Resnets được giới thiệu lần đầu tiên vào năm 2015 bởi Kaiming He cùng nhóm nghiên cứu tại Microsoft.
- Ý tưởng chính của Resnets là sử dụng **shortcut connection**.



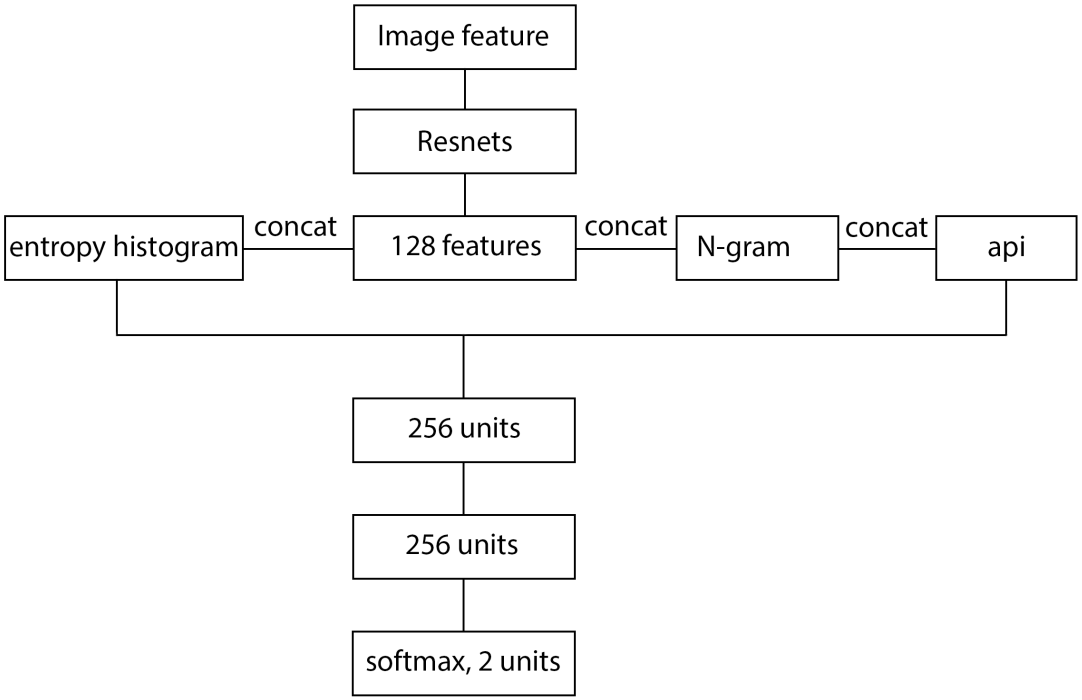
Hình 5 : Residual block



Hình 6 : Residual block

Tên lớp	Kích thước đầu vào	Kích thước đầu ra	Khối
conv1	64×64	32×32	$\begin{bmatrix} 3 \times 3 \times 32 \\ 3 \times 3 \times 32 \end{bmatrix} \times 2$
conv2	32×32	16×16	$\begin{bmatrix} 3 \times 3 \times 64 \\ 3 \times 3 \times 64 \end{bmatrix} \times 2$
conv3	16×16	8×8	$\begin{bmatrix} 3 \times 3 \times 128 \\ 3 \times 3 \times 128 \end{bmatrix} \times 2$
FC	8×8	2	average-pool, 2-fc, softmax

Hình 7 : Kiến trúc mô hình RNSIMG



Hình 8 : Mô hình RNSALL

Dữ liệu thực nghiệm

Thực nghiệm và đánh giá

Bộ dữ liệu bao gồm 28,000 mẫu PE. Trong đó, 14,000 mẫu malware được thu thập trên trang Virusshare - cộng đồng chia sẻ malware miễn phí, 14,000 mẫu benign được thu thập từ các máy tính nội bộ thuộc Tập Đoàn Công nghiệp & Viễn Thông Quân Đội Viettel. Bộ dữ liệu được chia thành 2 tập:

- Tập huấn luyện: gồm 10,500 mẫu malware và 10,500 mẫu benign
- Tập đánh giá: gồm 3,500 mẫu malware và 3,500 mẫu benign

Môi trường thực nghiệm

Thực nghiệm và đánh giá

Các thử nghiệm đánh giá được thực hiện trên hệ điều hành Ubuntu 16.04 với các thông số như sau:

- Ram 8GB.
- Ổ cứng HDD 1TB.
- Card đồ họa GeForce GTX 1080.

- So sánh hiệu năng của các mô hình DNN-Ngram, DNN-EH, DNN-API, RNSIMG, RNSALL, ENSEMBLE.
- Độ chính xác của các mô hình được đánh giá bằng k-Fold crossvalidation, k= 10.
- Các mô hình này được huấn luyện bằng Momemtum Gradient descent với các tham số được thiết lập như sau:
  - Số lần lặp: 40000
  - Batch size: 64
  - Weight decay: 0.005
  - Learning rate khởi tạo 0.001, giảm đi 1 nửa sau mỗi 10000 lần lặp

Đặc trưng	Kích thước	Mô hình	Độ chính xác trên tập validations	Độ chính xác trên tập đánh giá
N-gram	256	DNN-Ngram	94,40%	92,74%
ENT-HIS	256	DNN-EH	95,05%	94,86%
API	794	DNN-API	94,45%	93,56%
IMG	$64 \times 64 \times 1$	RNSIMG	95,48%	93,80%
ALL	1434	RESALL	95,80%	**95,14%**

Hình 9 : Hiệu năng của các mô hình

Bảng 3 : Bảng các giá trị **True/False Positive/Negative** của các mô hình

Mô hình	Dự đoán malware		Dự đoán benign	
	TP	FP	FN	TN
DNN-Ngram	3232	240	268	3260
DNN-EH	3300	160	200	3340
DNN-API	3363	314	137	3186
RNSIMG	3222	156	278	3344
RNSALL	3316	156	184	3344
ENSEMBLE	**3386**	**149**	**114**	**3351**

Kết luận và hướng phát triển

- Cài đặt, thử nghiệm, đánh giá các mô hình DNN-Ngram, DNN-EH, DNN-API, RNSIMG, RNSALL.
- Kết quả cho thấy sự kết hợp của cả 4 đặc trưng cho kết quả tốt hơn các đặc trưng riêng lẻ, đạt 95,14% trên tập đánh giá.
- Ngoài ra, việc kết hợp các mô hình với nhau cho kết quả tốt hơn, mô hình ENSEMBLE đạt 96,24% trên tập đánh giá.
- Trong tương lại cần xây dựng bộ dữ liệu lớn hơn, cải thiện độ chính xác của các mô hình để có thể đưa vào sử dụng trong thực tế.

THANK YOU!