JTE
HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Learning Spatial Features Using CNN in Network Intrusion Detection System

Thanh Van Nguyen
Ho Chi Minh City University of Technology and Education, Vietnam

Corresponding author. *Email: vanntth@hcmute.edu.vn*

ARTICLE INFO

Received: 16/03/2024
Revised: 29/03/2024
Accepted: 03/07/2024
Published: 28/08/2024

KEYWORDS

Intrusion detection system;
Learning feature;
Deep learning;
CNN;
CICIDS2017.

ABSTRACT

Today, modern communication networks and the diversity of network services have created a large growth in data transmitted through many different devices and communication protocols. This has raised serious security concerns, which in turn has increased the importance of developing advanced network intrusion detection systems (IDS). Although various techniques are applied to IDS, they face several challenges such as accuracy and efficient handling of highly variable big data. To increase the effectiveness of detecting attacks in network traffic, we need good features, but we also need to reduce the cost of feature construction techniques. Recently, Deep learning has been used as an effective way to analyze and discover knowledge in large data systems to create models with good classification capabilities. Many studies used Deep learning models to learn features automatically and effectively. In this paper, we used Convolution neural network (CNN) that exploits the visual properties of the input data to obtain features from network traffic, thereby achieving good intrusion detection performance. Our research was experimented on the CICIDS2017 dataset, achieving the highest accuracy of 91.53%.

Học Đặc Trưng Không Gian Dùng CNN trong Hệ Thống Phát Hiện Xâm Nhập Mạng

Nguyễn Thanh Vân
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh, Việt Nam

Tác giả liên hệ. *Email: vanntth@hcmute.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 16/03/2024
Ngày hoàn thiện: 29/03/2024
Ngày chấp nhận đăng: 03/07/2024
Ngày đăng: 28/08/2024

TỪ KHÓA

Hệ thống phát hiện xâm nhập mạng;
Học đặc trưng;
Deep learning;
CNN;
CICIDS2017.

TÓM TẮT

Ngày nay, các hệ thống mạng truyền thông hiện đại cùng với sự đa dạng về các loại dịch vụ mạng đã tạo ra sự tăng trưởng lớn về dữ liệu được truyền qua nhiều thiết bị và giao thức truyền thông khác nhau. Điều này đã gây ra những lo ngại nghiêm trọng về bảo mật, do đó đã làm tăng tầm quan trọng của việc phát triển các hệ thống phát hiện xâm nhập mạng (IDS) tiên tiến. Mặc dù các kỹ thuật khác nhau được áp dụng cho IDS nhưng chúng phải đối mặt với một số thách thức như độ chính xác và xử lý hiệu quả dữ liệu lớn có nhiều biến đổi. Để tăng hiệu quả phát hiện tấn công trong lưu lượng mạng, chúng ta cần các đặc trưng tốt, nhưng chúng ta cũng cần giảm chi phí kỹ thuật xây dựng đặc trưng. Gần đây, Deep learning đã được sử dụng như một cách hiệu quả để phân tích và khám phá kiến thức trong các hệ thống dữ liệu lớn nhằm tạo ra các mô hình có khả năng phân loại tốt. Có nhiều nghiên cứu đã sử dụng các mô hình Deep learning để học đặc trưng một cách tự động đem lại hiệu quả. Trong nghiên cứu này, chúng tôi đã sử dụng Convolution neural network (CNN) khai thác tính chất hình ảnh của đầu vào để thu được các đặc trưng từ lưu lượng truy cập mạng, nhờ đó việc phát hiện xâm nhập đạt hiệu quả tốt. Nghiên cứu được thực nghiệm trên tập dữ liệu CICIDS2017, đạt độ chính xác cao nhất là 91.53%.

Doi: https://doi.org/10.54644/jte.2024.1552

1. Giới thiệu

Hệ thống phát hiện xâm nhập (IDS – Intrusion Detection System) là một hệ thống các thiết bị hay ứng dụng có tính năng dò tìm và phát hiện các xâm nhập trái phép vào hệ thống mạng. IDS có thể được phân loại thành signature-based IDS và anomaly-based IDS. Signature-based IDS có thể phát hiện các cuộc tấn công dựa trên dấu hiệu của các cuộc tấn công, tuy nhiên nó không thể xác định các cuộc tấn công mới nếu không có quy tắc thủ công. Trong khi đó, anomaly-based IDS có khả năng phát hiện nhanh các tấn công chưa được biết trước đó, nó đã trở thành trọng tâm nghiên cứu chính trong lĩnh vực an ninh mạng. Nhiều giải pháp được nghiên cứu và phát triển nhằm tăng hiệu quả phát hiện tấn công và dự đoán sớm các cuộc tấn công như statistical-based, knowledge-based, và machine learning [1]. Thông thường, hiệu suất của các giải pháp phụ thuộc rất nhiều vào bộ đặc trưng của bài toán. Theo Bengio [2], một đặc trưng tốt thường là sự thể hiện các đặc điểm cơ bản và đầy đủ của các đầu vào được quan sát và nó rất hữu ích và đóng vai trò là đầu vào cho bộ dự đoán hoặc bộ phân loại có giám sát. Trong lĩnh vực an ninh mạng, các đặc trưng ứng với từng loại tấn công là không có sẵn và cần nhiều kỹ thuật để xây dựng, đòi hỏi chi phí cao. Do đó, việc cập nhật thủ công cơ sở dữ liệu đặc trưng đối với các mẫu tấn công mới được tạo ra ngày càng trở nên khó khăn. Để tăng hiệu quả phát hiện tấn công trong lưu lượng mạng, chúng ta cần các đặc trưng tốt, nhưng chúng ta cũng cần giảm chi phí kỹ thuật xây dựng đặc trưng. Điều này khiến việc sử dụng các mô hình machine learning dạng tuyến tính không còn phù hợp, các mô hình mới dựa trên deep learning đã được lựa chọn trong bối cảnh mới này.

Gần đây, Deep learning thể hiện khả năng vượt trội trong việc nắm bắt các mối quan hệ phụ thuộc phức tạp và tính phi tuyến ẩn trong dữ liệu, nó đã kết hợp các đặc trưng cấp thấp để tạo thành các danh mục hoặc đặc trưng đại diện ở mức cao hơn để khám phá các biểu diễn đặc trưng phân tán của dữ liệu. Một số mô hình học sâu phổ biến như: Autoencoder (AE), Recurrent neural network (RNN), Long Short Term Meomry (LSTM), Boltzman, Convolution neural network (CNN), Deep neural network (DNN). Trong đó, học tập đại diện với CNN đã được áp dụng rộng rãi trong nhiều lĩnh vực như ứng dụng thị giác máy tính [3]. Zeiler [4] đã chỉ ra rằng việc sử dụng giải mã và lọc các kích hoạt tối đa có thể giúp tìm ra mục tiêu gần đúng của từng bộ lọc tích chập trong mạng.

Trong phát hiện xâm nhập, các kỹ thuật học sâu có thể được áp dụng là phương pháp học đặc trưng, giúp học máy có giám sát cải thiện hiệu suất và xác định các tấn công trong hệ thống mạng. Nhiều kỹ thuật học sâu khác nhau được khảo sát trong IDS [5] như Autoencoder, LSTM, CNN. Tuy nhiên, việc áp dụng CNN vào IDS còn ít vì việc hiển thị các điểm dữ liệu mạng có vẻ khó giống với các hình ảnh thông thường. Một số nghiên cứu (được giới thiệu chi tiết ở mục 2.) đã sử dụng CNN để giải quyết bài toán IDS nhưng họ chưa mô tả một cách logic cách trực quan hóa dữ liệu mạng và hầu hết các thử nghiệm đều thực hiện trên tập dữ liệu đã cũ nên nhiều tấn công mới hiện nay chưa được cập nhật.

Trong bài báo này, chúng tôi đã sử dụng CNN để tìm hiểu các đặc trưng từ lưu lượng mạng, gồm: biểu diễn lưu lượng truy cập mạng dưới dạng hình ảnh dữ liệu và đưa chúng vào mô hình CNN để thu được các đặc trưng cụ thể. Phương pháp học biểu diễn của chuyển đổi đồ họa đã khai thác tính chất hình ảnh của đầu vào để thu được các đặc điểm nổi bật có thể xuất hiện trong ảnh. Các đặc trưng về thông tin mạng được thu từ lưu lượng mạng sau đó được dùng để việc phát hiện xâm nhập mạng một cách hiệu quả. Thực nghiệm được thực hiện trên tập dữ liệu CICIDS2017 với nhiều tấn công phổ biến hiện nay, kết quả đạt độ chính xác cao nhất là 91.53% ở một số mô hình CNN.

2. Các nghiên cứu và kiến thức liên quan

2.1. Dữ liệu mạng và tấn công: các đặc trưng

Trong môi trường mạng, lưu lượng mạng được phát sinh bởi các hoạt động truyền tải dữ liệu qua mạng Internet, do đó nó có thể được coi là bộ dữ liệu gồm một số lượng lớn các gói tin được tạo ra trong suốt quá trình truyền dữ liệu theo thời gian. Theo thời gian, các gói tin sẽ được gửi nhận giữa 2 bên với các thông tin khác nhau về giao thức, cổng, địa chỉ IP, dịch vụ… Bên cạnh các lưu lượng mạng của các hành vi bình thường còn có các hành vi bất thường hoặc tấn công của kẻ tấn công gây ra. Do đó, cần phân tích lưu lượng mạng để có thể giúp phát hiện các hoạt động tấn công mạng. Để phân tích lưu lượng mạng có hiệu quả cần có các đặc trưng mạng tốt.

Các đặc trưng của dữ liệu mạng được chia thành 4 nhóm cơ bản sau đây:

- Các đặc trưng cơ bản: đây là các đặc trưng nội tại của gói tin mạng dựa vào các trường thông tin của phần đầu gói tin mạng (packet header).
- Các đặc trưng được lấy từ 1 kết nối đơn - Single connection derived (SCD) feature: các đặc trưng được xây dựng bằng phép đo thống kê từ việc giám sát các đặc trưng cơ bản trong lưu thông mạng. Nhóm đặc trưng này rất hữu ích cho việc tìm kiếm hành vi bất thường trong một phiên duy nhất, chẳng hạn như một giao thức bất thường, kích thước dữ liệu bất thường, hoặc tần số bất thường của TCP flag.
- Các đặc trưng được lấy từ nhiều kết nối - Multiple connection derived (MCD): các đặc trưng này được xây dựng bằng cách giám sát các đặc trưng cơ bản trên nhiều dòng hoặc các kết nối, cho phép phát hiện các bất thường của các lưu thông, chẳng hạn như tấn công DoS và probe. Kiến thức chuyên gia được sử dụng để lựa chọn một cửa sổ của dữ liệu để xem xét, gồm cửa sổ thời gian (từ 5giây đến 24 giờ), cửa sổ các kết nối (ví dụ 100 kết nối). Một số kỹ thuật được dùng như: kỹ thuật về khai phá dữ liệu, luật kết hợp, phân tích trình tự chuỗi, đo tần số của các mẫu…
- Các đặc trưng về nội dung: được xây dựng từ payload của gói tin trên lưu thông mạng bằng việc dùng kiến thức chuyên gia với các kỹ thuật về khai phá dữ liệu, giải mã dữ liệu, giải nén dữ liệu… Với các đặc trưng nội dung có thể được sử dụng để phát hiện một số tấn công R2L và U2R.

Như vậy, các đặc trưng ứng với từng loại tấn công sẽ cần nhiều thao tác kỹ thuật mới có thể thu được, việc này đòi hỏi nhiều nhân lực và chi phí tốn kém. Ngoài ra, với sự thay đổi liên tục về đặc tính dữ liệu mạng, các đặc trưng cũng cần được thích ứng theo, do đó việc xây dựng các đặc trưng 1 cách thủ công, cố định không còn phù hợp.

2.2. Học đặc trưng dùng deep learning trong IDS

Deep learning có các kiến trúc có nhiều lớp (multilayer) được giám sát có thứ bậc trong các giai đoạn xử lý thông tin. Các lớp này được khai thác để việc học các đặc trưng theo cách không giám sát và để phân tích phân loại các mẫu. Học sâu có hai kiến trúc chính: Kiến trúc Generative và kiến trúc Discriminative. Sau đây chúng tôi sẽ khảo sát một số nghiên cứu IDS sử dụng một số mô hình của deep learning.

Kiến trúc Generative là dạng học sâu không giám sát mà có thể học một cách tự động từ dữ liệu thô không có nhãn để thực hiện các tác vụ khác nhau như phân loại hay dự đoán. Mục tiêu của các kiến trúc này là làm sao sinh ra được các dữ liệu giống với dữ liệu thực tế nhất. Một số mô hình Generative như: Autoencoder, LSTM, GRU.

Bảng 1. *Một số nghiên cứu dùng AE trong IDS*

Tác giả	Kỹ thuật	Dữ liệu	Kết quả
R. Can Aygun [6], 2017	Autoencoder (AE) và Denoising AE	NSL- KDD	88.28% và 88.65%
Farahnakian [7], 2018	Stacked AE (4AE).	10% KDD99	95%
S. Potluri [8], 2016	Stacked AE (2AE).	NSL-KDD	95%
Niyaz [9], 2015	Sparse AE để trích xuất đặc trưng Hồi quy softmax để phân loại 2 lớp	NSL-KDD	F1-score: 90.4%
B. Zhang [10], 2018	Stacked Sparse AE để học đặc trưng, kết hợp với cây nhị phân để phân loại	NSL-KDD	F1-score: 91.97%
Ivandro O. Lopes [11], 2022	Denoising AE để lấy đặc trưng DNN để phân loại	CICIDS2017	F1-score: 99.6%
Youngrok Song [12], 2021	Autoencoder	NSL-KDD	F1-score: 97.4
Choi [13], 2019	Autoencoder	NSL-KDD	91.70%

Autoencoder (AE) được xem như là một phương pháp rút trích đặc trưng phi tuyến mà không sử dụng nhãn lớp. Một AE là một loại mạng lưới thần kinh với vectơ đầu ra có số chiều tương tự như các vector đầu vào, các đơn vị đầu ra được kết nối trực tiếp lại cho các đơn vị đầu vào. AE có thể được dùng để học trong Neural network sâu. Quá trình huấn luyện một AE là dạng không giám sát, và mục tiêu của quá trình huấn luyện này là tìm các tham số sao cho sự khác nhau giữa các đầu vào x và tái thiết của chúng là nhỏ nhất.

Một số giải pháp Stacked AE cũng được nghiên cứu để sử dụng vào bài toán phát hiện xâm nhập mạng như trong Bảng 1. Stacked-AE là hình thức xây dựng một mạng lưới thần kinh sâu DBN từ nhiều autoencoder, trong đó kết quả đầu ra của mỗi lớp được nối với đầu vào của các lớp kế tiếp. Stacked AE được thực hiện theo cách thức greedy layerwise để học các tính năng với hồi quy softmax như một lớp phân loại để phát hiện các cuộc tấn công. Các nghiên cứu đạt kết quả khả quan trên 90%, tuy nhiên hầu hết được thực nghiệm trên tập dữ liệu khá cũ, các tấn công không được cập nhật.

Một dạng kiến trúc khác của học sâu là RNN và các phiên bản của nó là LSTM và GRU cũng được nghiên cứu đưa vào bài toán phát hiện xâm nhập, như trong Bảng 2. Kiến trúc LSTM có thể khai thác được đặc trưng về mối quan hệ giữa các thành phần trong chuỗi nhờ vào khả năng ghi nhớ được các thông tin từ một số các sự kiện trước đó và biểu diễn được mối quan hệ giữa chúng và sự kiện tại thời điểm hiện tại. Với đặc điểm này của LSTM, hệ thống có thể phát hiện được các bất thường từ sự kết hợp nhiều gói tin trên dữ liệu mạng nhằm phát hiện được các xâm nhập mạng.

Bảng 2. *Một số nghiên cứu dùng LSTM trong IDS*

Tác giả	Kỹ thuật	Dữ liệu	Kết quả
Jihyun Kim [14], 2016	Thử nghiệm kích thước lớp hidden và learning rate khác nhau	10% KDD99	Tốt nhất: 80 lớp hidden, learning rate 0.01 đạt 96.93%
Ralf C. Staudemeyer [15], 2015	Thay đổi số kích thước ở lớp hidden	10%KDD99, thử nghiệm với các số đặc trưng: 4, 8, 41.	4 đặc trưng: 93.72% 8 đặc trưng: 93.69% 41 đặc trưng: 93.82%
Loic Boitemp [16], 2017	Thử hidden size và learning rate khác nhau, tối ưu các tham số để phát hiện bất thường của nhóm	KDD99 – lựa chọn các đặc trưng cơ bản của tấn công Neptune (1 dạng của DoS)	86%-100% với các ngưỡng khác nhau
Min Cheng [17], 2016	A multi-scale LSTM . Time scale windows size=40	Raw data từ Routing Information Service (RIS): 33 đặc trưng	MS-LSTM1: 90.4% MS-LSTM2: 81.5% MS-LSTM3: 95.4%
Laghrissi [18], 2021	Dùng PCA để thu giảm chiều. Dùng 3 units LSTM	Một phần của KDD99: gộp 2 nhóm tấn công chính: DoS, R2L	LSTM: 85.65% PCA-LSTM: 99.36%

Các nghiên cứu áp dụng LSTM để phát hiện xâm nhập mạng đều có phần xử lý đặc trưng khác nhau: giữ nguyên số đặc trưng của dataset, giảm số đặc trưng. Hầu hết các nghiên cứu đạt kết quả tốt với các loại tấn công có tần số xuất hiệu lớn như: DoS, Probe, còn tấn công ít xuất hiện đạt kết quả không cao.

Kiến trúc Discriminative là dạng mô hình có điều kiện, chúng phân biệt ranh giới quyết định bằng cách suy luận kiến thức từ dữ liệu quan sát. Kiến trúc phân biệt gồm: CNN, DNN.

CNN là một trong những mô hình deep learning tiên tiến, giúp cho việc xây dựng được những hệ thống thông minh với độ chính xác cao, và được sử dụng nhiều trong các bài toán nhận dạng các object trong ảnh. Trong CNN, convolution là một cửa sổ trượt (Sliding Windows) trên một ma trận. Các lớp convolutional có các tham số (kernel) đã được học để tự điều chỉnh nhằm lấy ra những thông tin chính xác nhất mà không cần chọn các đặc trưng một cách thủ công. Gần đây, các kiến trúc Discriminative của học sâu được nghiên cứu đưa vào bài toán phát hiện xâm nhập mạng. Bảng 3 tóm tắt một số nghiên

cứu gần với hướng của chúng tôi trong việc sử dụng CNN để học đặc trưng trên tập dữ liệu phổ biến về phát hiện xâm nhập mạng như KDD, NSL-KDD, CICIDS2017.

Hầu hết các nghiên cứu được thống kê sử dụng các bộ dữ liệu để phát hiện xâm nhập mạng thiếu sự đa dạng về loài và cỡ mẫu, đồng thời một số trong đó khá cũ như KDD99, NSL-KDD, do đó chúng khác với các loại tấn công hiện tại. Ngoài ra, các bộ dữ liệu chỉ chứa thông tin tiêu đề mà thiếu thông tin về tải trọng, điều này không thể phản ánh tốt xu hướng tấn công hiện tại. Có một số nghiên cứu [19], [20], [23], [24] thực nghiệm trên tập dữ liệu mới CICIDS2017 nhưng giới hạn việc phát hiện 1 dạng tấn công Dos [19], hay phân loại hai lớp tấn công và normal, do đó kết quả khá cao; còn [23], [24] phát hiện được các loại tấn công.

Bảng 3. *Một số nghiên cứu dùng CNN trong IDS*

Tác giả	Mô hình CNN	Dữ liệu	Kết quả
Kim, J. [19], 2020	Thử nghiệm 1,2,3 lớp CNN	CICIDS2017	99.97% (tấn công Dos)
Venkata R. [20], 2021	3 Conv2D	CICIDS2017	99% (tấn công và normal)
Sinh-NN, [21], 2018	2 lớp CNN 2D	1/5 KDD99	99.87% (tấn công Dos)
Li.Z [22], 2017	Resnet50 và GoogLeNet.	NSL-KDD	Test+: 79.14%,; 77.04% Test- : 81.57%, 81.84%.
Taejoon Kim [23], 2018	GoogLeNet	NSL-KDD và CICIDS2017	88-89% và 82% (các loại tấn công
Yong Zhang [24], 2019	2 CNN song song	CICIDS2017	99.92% (các loại tấn công)

**** Chú ý***: các trích dẫn trong bài báo này, thực tế là do tên tác giả đứng đầu và đồng nghiệp (chi tiết ở mục Tài liệu tham khảo), để cho gọn chúng tôi lược bớt "và đồng nghiệp", chỉ dùng tên của tác giả đầu tiên.

3. Phương pháp đề xuất

Trong phần này, chúng tôi đề xuất việc hình ảnh hóa các gói dữ liệu mạng, sau đó dùng kiến trúc mạng CNN để tìm hiểu sự phụ thuộc không gian giữa các vùng ảnh của dữ liệu mạng nhằm nâng cao sức mạnh phân biệt của biểu diễn ảnh.

Dữ liệu mạng là một tập hợp các gói tin mạng chứa các đặc trưng khác nhau, trong mỗi đặc trưng tồn tại các bộ thông số tương ứng với các trạng thái không gian và thời gian khác nhau của dữ liệu mạng. Chúng tôi cho rằng việc hình ảnh hóa dữ liệu mạng sẽ hiện ra được các kết nối không gian giữa các đặc trưng mạng, sau đó nhờ khả năng của CNN để thu được các đặc trưng không gian trong dữ liệu mạng. Vector đặc trưng thu được sẽ đưa vào bộ phân loại để phân biệt dữ liệu mạng bình thường hay bất thường, hoặc phân loại ra các loại tấn công.

3.1. Hình ảnh hóa các dữ liệu mạng

Từng packet dữ liệu mạng đi vào được tiền xử lý, gồm các bước:

- Min-Max [0,1].
- Rời rạc các giá trị liên tục vào 10 khoảng.
- Dùng mã hóa one-hot để mã hóa dữ liệu số của 10 khoảng thành các vector nhị phân.
- Chuyển đổi vector nhị phân thành các giá trị pixel của ảnh grayscale 8bit.

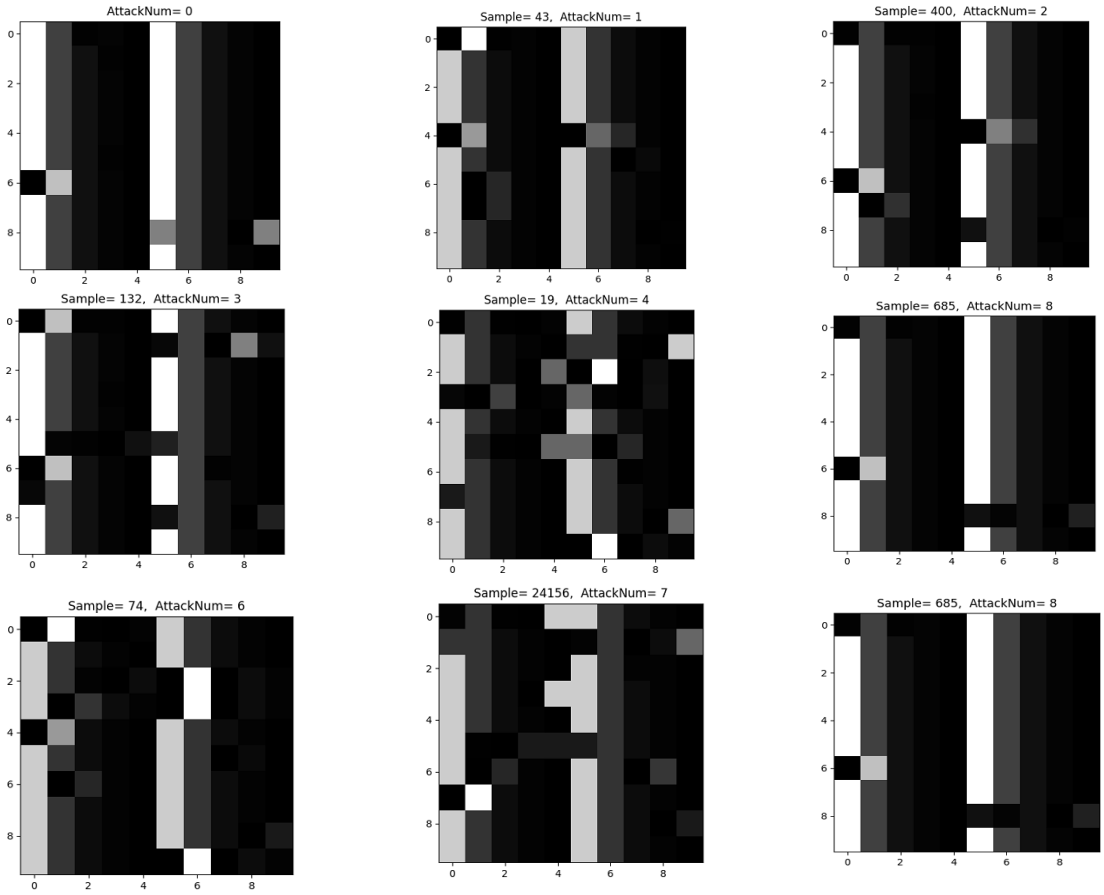
Trong thực nghiệm, chúng tôi dùng tập dữ liệu CICIDS2017 [25], gồm các dữ liệu bình thường và 14 loại tấn công khác nhau được biểu diễn bởi 80 đặc trưng. Các đặc trưng được chuyển thành các vector nhị phân, sau đó được chuyển đổi tiếp thành giá trị pixel của các hình ảnh grayscale 8x8bit. Cuối cùng, chúng tôi thu được hình ảnh là các ma trận 10x10, như Hình 1.

JTE
HCMUTE
ISSN: 1859-1272

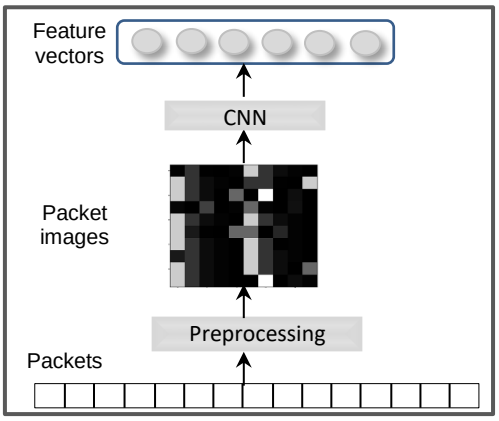
TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

3.2. Học đặc trưng không gian dùng CNN

CNN được sử dụng nhiều trong các bài toán nhận dạng các đối tượng trong ảnh, còn trong lĩnh vực an ninh mạng hay phát hiện xâm nhập còn rất ít các thử nghiệm. Chúng tôi đã xem xét các hình ảnh dữ liệu mạng sau khi tiền xử lý ở bước trên (mục 3.1) để đưa vào 1 mạng CNN nhằm mục đích cho hệ thống học các đặc trưng không gian để phục vụ cho việc phân biệt hình ảnh dữ liệu mạng bình thường hay các tấn công, như trong Hình 2.



Hình 1. *Một số hình ảnh dữ liệu tấn công trong tập dữ liệu CICIDS2017*

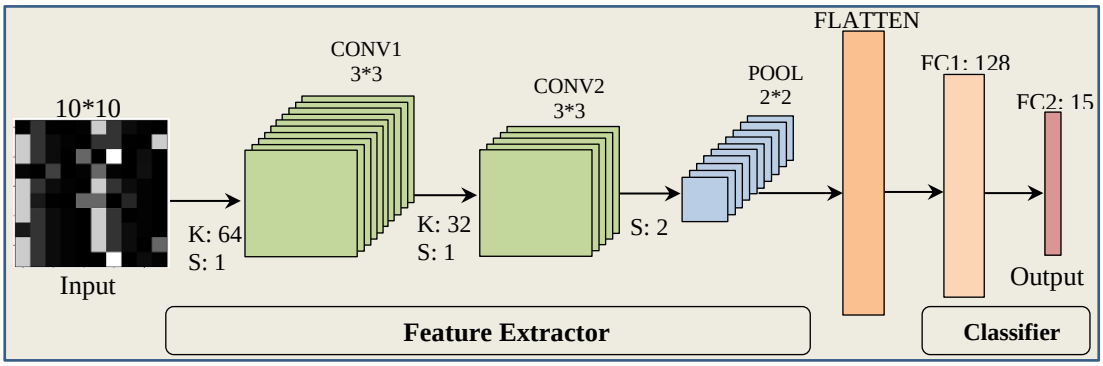


Hình 2. *Quá trình học đặc trưng không gian từ dữ liệu mạng*

CNN biến đổi thông tin đầu vào thông qua một phép tích chập với bộ lọc để trả về đầu ra là một tín hiệu mới. Tín hiệu này sẽ giữ những đặc trưng chính, giảm những đặc trưng phụ, từ đó mạng CNN học được đặc trưng ở những cấp độ có độ phức tạp tăng dần. Vì bộ lọc quét qua toàn bộ bức ảnh, nên những

đặc trưng này có thể nằm ở vị trí bất kì trong bức ảnh, cho dù ảnh bị xoáy trái/phải thì những đặc trưng này vẫn bị phát hiện. Tập hợp nhiều bộ lọc sẽ cho phép phát hiện được nhiều loại đặc trưng khác nhau, và giúp định danh được đối tượng tốt hơn. Trong hệ thống này, gói dữ liệu mạng đi vào sau khi được chuyển thành hình ảnh 2 chiều sẽ được đưa vào mô hình CNN để học các đặc trưng về không gian. Vector đặc trưng thu được sẽ đưa vào bộ phân loại là các lớp Full Connected để phân biệt dữ liệu mạng bình thường hay các dạng tấn công.

Mô hình CNN_Simple được triển khai như Hình 3.



Hình 3. *Mô hình CNN_Simple: 2 conv + 1 pooling*

ResNet50 [26] và GoogleNet [27] là các mô hình phân loại hình ảnh có thể được huấn luyện trên các tập dữ liệu lớn đạt được kết quả tốt. Chúng tôi sử dụng dạng thu gọn của hai mô hình này, cụ thể:

- Resnet_short: gồm 1 conv block + 1 identity
- GoogleNet_short: 1 stem block + 1 inception block + 1 reduction block

4. Thực nghiệm và kết quả

Tập dữ liệu

Bộ dữ liệu CICIDS2017 là bộ dữ liệu ngăn chặn và phát hiện xâm nhập mạng nguồn mở được viện nghiên cứu an ninh mạng Canada thu thập vào năm 2017. CICIDS2017 chứa các cuộc tấn công phổ biến và các hành vi lành tính, có thể đáp ứng tốt hơn xu hướng tấn công hiện nay. Các luồng dữ liệu được gắn nhãn dựa trên dấu thời gian, IP nguồn và IP đích, cổng nguồn và đích, giao thức và tấn công. Một số tấn công phổ biến như: Web based, Brute force, DoS, DDoS, Infiltration, Heart-bleed, Bot, Scan... CICIDS2017 có tỉ lệ bình thường là 83.4% và tấn công là 16.6% (bao gồm 14 loại tấn công). Dữ liệu thực nghiệm được phân chia như Bảng 4. Các loại tấn công trong bộ dữ liệu CICIDS2017 có số hiệu được biểu diễn trong Bảng 5, còn dữ liệu bình thường có giá trị 0.

Bảng 4. *Tập dữ liệu CICIDS2017*

Phân chia tập dữ liệu	Tấn công	Bình thường	Tổng
Train (80%)	446,117	1,818,479	2,264,596
Validation (20% train)	89,223	363,696	452,919
Test (20%)	111,527	454,620	566, 147

Bảng 5. *Các loại tấn công trong bộ dữ liệu CICIDS2017*

Tấn công	Số hiệu
FTP-Patator, SSH-Patator, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed, Brute Force, Sql Injection, XSS, Infiltration, Bot, DdoS, PortScan	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

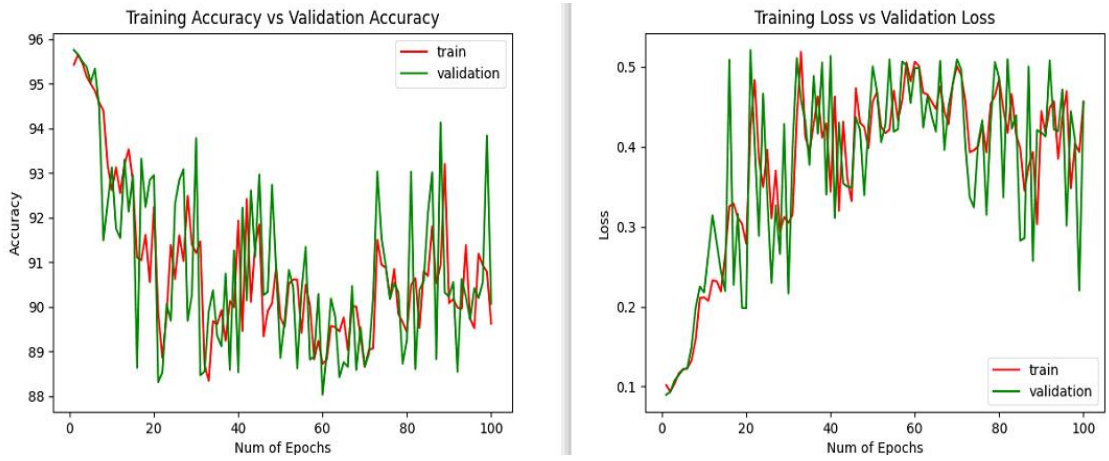
JTE
HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Kết quả và thảo luận

Các dữ liệu trong tập CICIDS2017 được xử lý dưới dạng hình ảnh như mục 3.1. Sau khi có được hình ảnh các tấn công, chúng tôi dùng các mô hình CNN để học các đặc trưng không gian nhằm phục vụ cho việc phân loại hình ảnh dữ liệu mạng bình thường hay các tấn công. Chúng tôi thử nghiệm 3 mô hình CNN khác nhau như mô tả ở 3.2, và tiến hành phân loại dữ liệu bình thường và 14 loại tấn công.

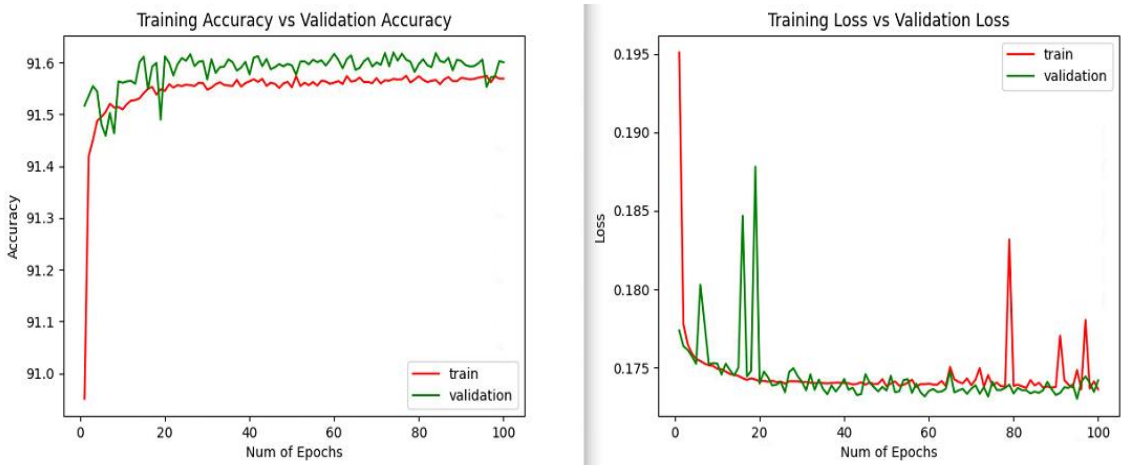
Mô hình CNN_Simple có tỉ lệ phát hiện dữ liệu bình thường cao đến 100% so với thực tế. Các loại tấn công 4, 13 phát hiện đúng cao đến 99% so với tổng tấn công phát hiện được, đây cũng là các tấn công có số mẫu lớn. Trong khi đó các loại tấn công 1, 2, 3, 5 đến 14 phân loại thấp, đây là các tấn công có số mẫu ít. Kết quả tổng thể trong quá trình training và validation của mô hình CNN_Simple được thể hiện ở Hình 4 với độ chính xác và loss khá dao động.



(a) *(b)*

Hình 4. *Kết quả quá trình training và validation của mô hình CNN_Simple: (a) Độ chính xác; (b) Loss*

Mô hình Resnet_short có tỉ lệ phát hiện dữ liệu bình thường đạt 98% so với thực tế, tỉ lệ phân loại tấn công sai giảm đáng kể so với mô hình CNN_Simple. Các loại tấn công 3, 11 phát hiện chính xác 100%, các loại tấn công 4, 6, 12, 13 đạt 90-94% so với số tấn công phát hiện được. Tỉ lệ phát hiện được so với thực tế cao đạt 100% đối với tấn công 7 và đạt 81% đối với tấn công 4, còn lại còn khá thấp. Các loại tấn công 4, 13 phát hiện chính xác cao đạt 90-95% so với tổng mẫu phát hiện được, và tỉ lệ phát hiện được so với thực tế đạt khá 83%, 72%. Còn các tấn công phát hiện kém gồm 1, 2, 8, 9, 10. Kết quả tổng thể trong quá trình training và validation của mô hình Resnet_short được thể hiện ở Hình 5, đạt độ chính xác 91.53%.

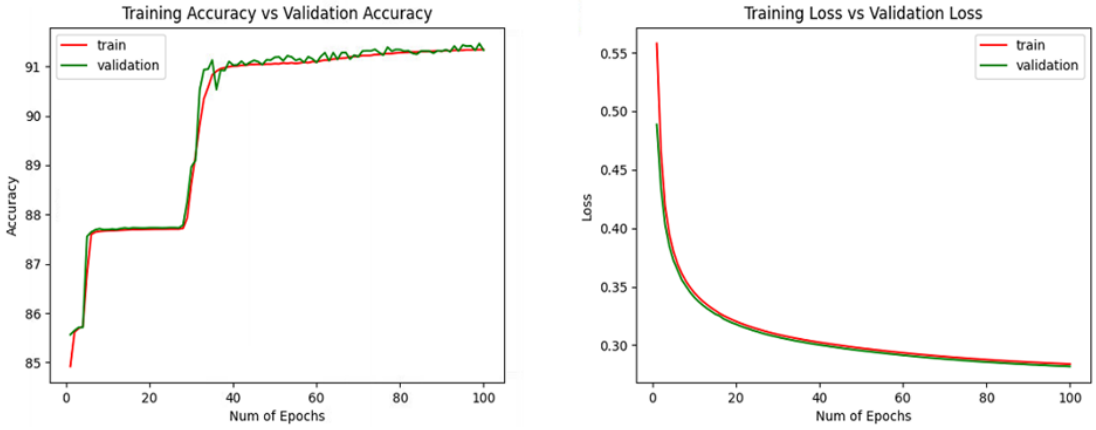


(a) *(b)*

Hình 5. *Kết quả quá trình training và validation của mô hình Resnet_short: (a) Độ chính xác; (b) Loss*

JTE
HCMUTE
ISSN: 1859-1272

TẠP CHÍ KHOA HỌC GIÁO DỤC KỸ THUẬT
Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn



(a) *(b)*

Hình 6. *Kết quả quá trình training và validation của GoogLeNet_short: (a) Độ chính xác; (b) Loss*

Mô hình GoogLeNet_Short có tỉ lệ phát hiện dữ liệu bình thường đạt 97% so với thực tế. Một số tấn công phát hiện tốt hơn Resnet-Short, như: 1, 3, 4, 5 ,6, 13, 14. Một số tấn công phát hiện kém như: 2, 7 đến 12. Kết quả tổng thể trong quá trình training và validation của mô hình GoogLeNet _short được thể hiện ở Hình 6 với độ chính xác đạt tới 91.24% và loss có xu hướng giảm tốt, ít dao động.

Chúng tôi lập Bảng 6 để so sánh phương pháp đề xuất với các nghiên cứu khác với cùng tập dữ liệu CICIDS2017. Hai nghiên cứu [19], [20] có cùng cách tiền xử lý dữ liệu đầu vào chuyển trực tiếp các đặc trưng (78 đặc trưng) thành hình ảnh có kích thước 13 x 6, 11 x 7, sau đó đưa vào mô hình CNN để học đặc trưng và phân loại đạt độ chính xác cao, tuy nhiên 2 nghiên cứu này chỉ tập trung phát hiện 1 dạng tấn công Dos [19], hoặc chỉ phân loại 2 lớp tấn công và normal. Ở nghiên cứu [24] sử dụng cách tiền xử lý dữ liệu dưới dạng từng packet 1600byte chuyển thành hình ảnh 40x40, sau đó đưa vào 2 CNN song song đạt kết quả cao khi phát hiện các loại tấn công. Xét nghiên cứu gần nhất [23], với cùng cách tiền xử lý là đặc trưng được chuyển thành các vector nhị phân, và chuyển đổi tiếp thành giá trị pixel của các hình ảnh grayscale 8x8bit để được hình ảnh 10x10, cùng mô hình CNN là GoogLeNet, kết quả chúng tôi đạt độ chính xác cao hơn khi phát hiện các loại tấn công khác nhau. Rất khó tìm được các nghiên cứu tương tự để so sánh, vì môi trường thử nghiệm, dữ liệu, cách xử lý dữ liệu đầu vào, mô hình CNN khác nhau, chúng tôi chỉ so sánh một cách tương đối với một số nghiên cứu gần với hệ thống của chúng tôi như Bảng 6.

Bảng 6. *So sánh các mô hình đề xuất với một số nghiên cứu khác*

Tác giả	Tập dữ liệu	Tiền xử lý	Mô hình	Độ chính xác
Kim, J. [19], 2020	CICIDS2017	Grayscale 13 x 6	3 CNN	99.97 % (Dos)
Venkata R. [20], 2021	CICIDS2017	Grayscale 11 x 7	3 Conv2D	99% (tấn công và normal)
Taejoon Kim [23], 2018	CICIDS2017	Grayscale 8bit 10 x10	GoogLeNet	88-89% (tấn công và normal)
Yong Zhang [24], 2019	CICIDS2017	1 packet 1600byte shape thành 40x40	2 CNN song song: 2 x (4conv)	99.92% (các loại tấn công)
Phương pháp đề xuất	CICIDS2017	Grayscale 8bit 10 x 10	CNN_Simple Resnet50 – Short GoogLeNet - Short	3 mô hình: 90.09%, 91.53%, 91.24% (các loại tấn công)

5. Kết luận

Trong bài báo này chúng tôi đã sử dụng CNN để tìm hiểu các đặc trưng không gian của các hình ảnh dữ liệu mạng đã được chuyển đổi và sau đó đưa vào bộ phân loại là các lớp Full Connected để phân biệt dữ liệu mạng bình thường hay các loại tấn công. Phương pháp đề xuất của chúng tôi tìm hiểu các biểu diễn đặc trưng tốt một cách hiệu quả từ một lượng lớn dữ liệu mạng, và được thực nghiệm trên tập dữ liệu CICIDS2017 đạt kết quả tốt. So với các bộ phân loại khác thì CNN có thể không vượt trội, nhưng CNN sử dụng dạng dữ liệu hình ảnh làm đầu vào mà không cần lựa chọn tính năng, đó là một ưu điểm lớn của phương pháp deep learning.Trong công việc trong tương lai, chúng tôi sẽ nghiên cứu một số phương pháp khác để có được nhiều đặc trưng tốt về lưu lượng truy cập mạng và thực nghiệm trên các tập dữ liệu mới hơn, nhằm cải thiện tỷ lệ phát hiện các loại tấn công mới. Ngoài ra, chúng tôi sẽ xem xét sử dụng gói dữ liệu thô để thử nghiệm.

Lời cảm ơn

Tác giả cảm ơn Khoa Công Nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật TP.HCM, Việt Nam đã hỗ trợ và giúp đỡ thực hiện các thực nghiệm.

Xung đột lợi ích

Tác giả tuyên bố không có xung đột lợi ích

TÀI LIỆU THAM KHẢO

[1] H. Liu and B. Lang, “Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey,” *Appl. Sci.*, vol. 9, p. 4396, 2019, doi: 10.3390/app9204396.

[2] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep CNN,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, Jun. 2017, doi: 10.1145/3065386.

[4] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision – ECCV 2014*, vol. 8689, Lecture Notes in Computer Science, Springer, 2014, pp. 818-833, doi: 10.1007/978-3-319-10590-1_53.

[5] G. Gilberto, “A comprehensive survey on network anomaly detection,” *Telecommun. Syst.*, vol. 70, pp. 447-489, 2019, doi: 10.1007/s11235-018-0414-7.

[6] R. C. Aygun and A. G. Yavuz, "Network Anomaly Detection with Stochastically Improved Autoencoder Based Model," in *IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2017, pp. 193-198, doi: 10.1109/CSCloud.2017.32.

[7] S. Farahnakian and J. Heikkonen, "A deep auto-encoder based approach for intrusion detection system," in *Int. Conf. on Advanced Communication Technology (ICACT)*, 2018, pp. 178-183, doi: 10.23919/ICACT.2018.8323744.

[8] S. Potluri and C. Diedrich, "Accelerated deep neural networks for enhanced intrusion detection system," in *Int. Conf. on Emerging Technologies and Factory Automation (ETFA)*, 2016, pp. 1-8, doi: 10.1109/ETFA.2016.7733704.

[9] Q. Niyaz, W. Sun, A. Javaid, and M. Alam, "A Deep Learning Approach for NIDS," in *Bio-inspired Information and Communications Technologies (BIONETICS)*, Brussels, Belgium, 2015, pp. 21-26, doi: 10.4108/eai.3-12-2015.2262516.

[10] B. Zhang, Y. Yu, and J. Li, "Network intrusion detection based on stacked sparse autoencoder and binary tree ensemble method," in *IEEE Int. Conf. on Communications (ICC)*, 2018, pp. 1-6, doi: 10.1109/ICC.2018.8422406.

[11] I. O. Lopes, “Effective network intrusion detection via representation learning: A Denoising AutoEncoder approach,” *Computer Communications*, vol. 194, pp. 55-65, Oct. 2022, doi: 10.1016/j.comcom.2022.07.004.

[12] Y. Song, S. Hyun, and Y.-G. Cheong, “Analysis of Autoencoders for Network Intrusion Detection,” *Sensors*, vol. 21, no. 4294, 2021, doi: 10.3390/s21134294.

[13] H. Choi, M. Kim, G. Lee, et al., “Unsupervised learning approach for NIDS using autoencoders,” *Journal of Supercomputing*, vol. 75, pp. 5597–5621, 2019, doi: 10.1007/s11227-019-02873-2.

[14] K. Ji, J. Kim, L.T.T. Huong, et al., "LSTM - RNN Classifier for Intrusion Detection," in *International Conference Platform Technology and Service (PlatCon)*, South Korea, 2016, pp. 1-5, doi: 10.1109/PlatCon.2016.7456801.

[15] R. C. Staudemeyer, "Applying LSTM RNN to intrusion detection," *South African Computer Journal*, vol. 56, pp. 6-15, 2015, doi: 10.18489/sacj.v56i0.225.

[16] L. Bontemps, C. Van Cao, J. McDermott, et al., "Collective Anomaly Detection based on LSTM RNN," in *International Conference on Future Data and Security Engineering*, 2016, pp. 141-152, doi: 10.1007/978-3-319-49358-9_10.

[17] M. Chen, X. Qi, J. Liu, et al., "MS-LSTM: a Multi-Scale LSTM Model for BGP anomaly detection," in *24th International Conference on Network Protocols (ICNP)*, 2016, pp. 1-6, doi: 10.1109/ICNP.2016.7784448.

[18] F. Laghrissi, et al., “Intrusion detection systems using long short-term memory (LSTM),” *Journal of Big Data*, vol. 8, no. 65, pp. 1-14, 2021, doi: 10.1186/s40537-021-00466-3.

[19] J. Kim, J. Kim, H. Kim, M. Shim, and E. Choi, “CNN-Based Network Intrusion Detection against Denial-of-Service Attacks,” *Electronics*, vol. 9, no. 916, 2020, doi: 10.3390/electronics9060916.

[20] V. R. Varanasi and S. Razia, “CNN Implementation for IDS,” in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1585-1589, doi: 10.1109/ICAC3N53548.2021.9725426.

[21] S. N. Nguyen, Q. V. Nguyen, and K. Kim, "Design and implementation of intrusion detection system using CNN for DoS detection," in *International Conference on Machine Learning and Soft Computing*, 2018, pp. 34-38, doi: 10.1145/3184066.3184094.

[22] Z. Li, Z. Qin, K. Huang, and X. Yang, "Intrusion Detection Using CNNs for Representation Learning," in *Neural Information Processing (ICONIP), Lecture Notes in Computer Science*, vol. 10638, Springer, Cham, 2017, pp. 103-111, doi: 10.1007/978-3-319-70096-0_11.

[23] T. Kim, S. C. Suh, H. Kim, et al., "An Encoding Technique for CNN-based Network Anomaly Detection," in *IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2960-2963, doi: 10.1109/BigData.2018.8622337.

[24] Y. Zhang, X. Chen, D. Guo, et al., "Parallel Cross CNN for Abnormal Network Traffic flows Detection in multi-class imbalanced," *IEEE Access*, vol. 7, pp. 119904-119916, 2019, doi: 10.1109/ACCESS.2019.2936982.

[25] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *ICISSP*, pp. 108–116, 2018, doi: 10.5220/0006639801080116.

[26] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

Nguyen Thanh Van graduated from university in Infomatics in 1998 at Hue University's College of Education, Hue University, and received a master's degree in computer science in 2005 at Da Nang University. She is currently working at the Faculty of Information Technology, HCMC University of Technology and Education. Her research interests include Information and Network security, machine learning technologies. Email: vanntth@hcmute.edu.vn. ORCID: https://orcid.org/0009-0003-9686-606X