PHƯƠNG PHÁP PHÁT HIỆN WEBSITE LỪA ĐẢO DỰA TRÊN HỌC SÂU ĐA PHƯƠNG THỨC KHÁNG MẪU TRỐN TRÁNH

Võ Quang Minh
Trường Đại học Công nghệ Thông tin - ĐHQG HCM

Giới thiệu

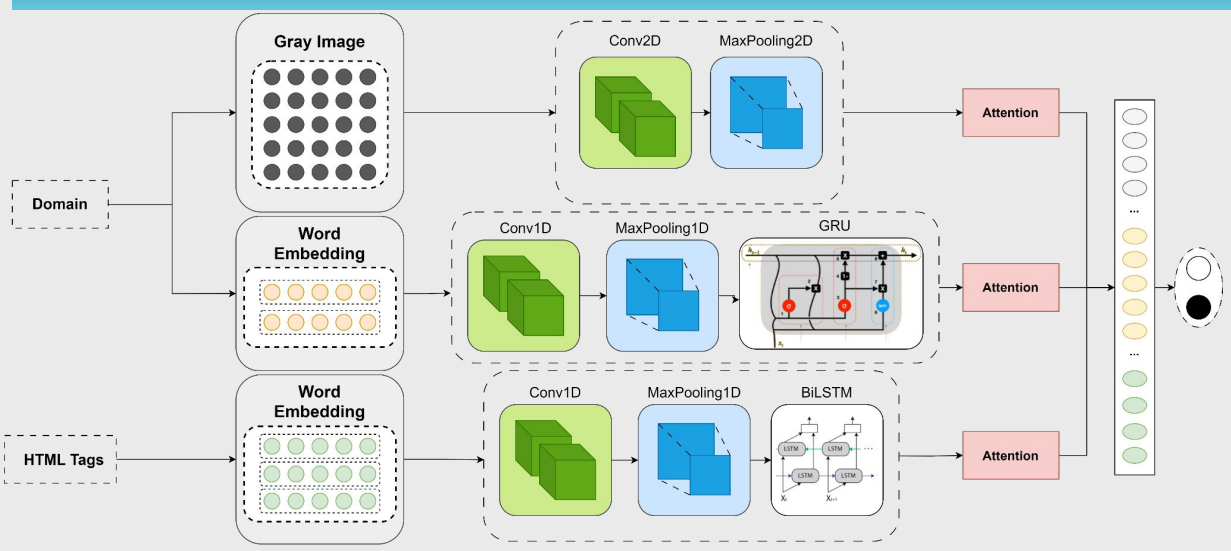
Tấn công Phishing là một vấn đề đã xuất hiện từ lâu nhưng vẫn chưa được giải quyết triệt để.

- Mục tiêu thường là các người dùng bất cẩn và thiếu kiến thức về công nghệ.
- Thường đánh cắp tài sản, thông tin cá nhân của cá nhân, tổ chức.
- Các phương pháp phát hiện vẫn còn tồn tại nhiều khuyết điểm.

Mục tiêu

- Nghiên cứu xây dựng **một mô hình học sâu đa phương thức** (Multimodal model) phát hiện trang web lừa đảo hiệu quả có tên là **Shark-eyes.**
- Mô hình **không phụ thuộc vào một nhóm thuộc tính** được trích xuất từ một khía cạnh duy nhất của trang web.
- **Tiếp cận nhiều đối tượng của trang web** để trích xuất thông tin, giúp mô hình tăng khả năng kháng các nhiễu loạn tốt hơn.

Tổng quan mô hình



Hình 1. Sơ đồ tổng quan mô hình Shark-eyes.

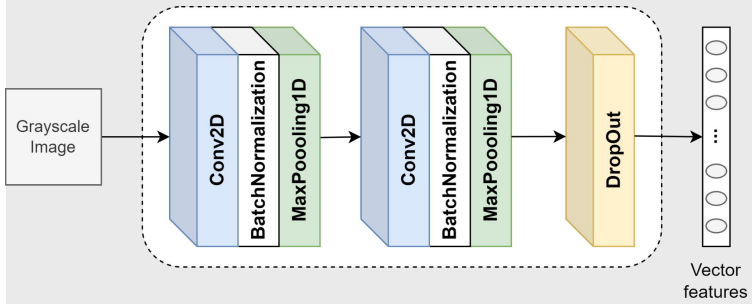
Chúng tôi đề xuất mô hình học sâu đa phương thức, tập trung vào trích xuất thuộc tính từ tên miền hiện diện trong URL và phân tích cấu trúc các thẻ có trong tệp HTML của trang web. Cấu trúc tổng quan của mô hình Shark-eyes được thể hiện ở *Hình 1*:

Mô hình sử dụng kỹ thuật học sâu để trích xuất các thuộc tính từ các đối tượng của trang web, tránh được sự thiên vị và hiệu quả trong chuyên sâu trích xuất.

Phương pháp

1. Nhánh hình ảnh tên miền

Ánh xạ tên miền thành ảnh xám



Hình 2. Kiến trúc của mạng Convolutional Neural Network

Ảnh xám đầu vào sẽ qua các lớp Conv2D và MaxPooling2D để trích xuất thuộc tính.

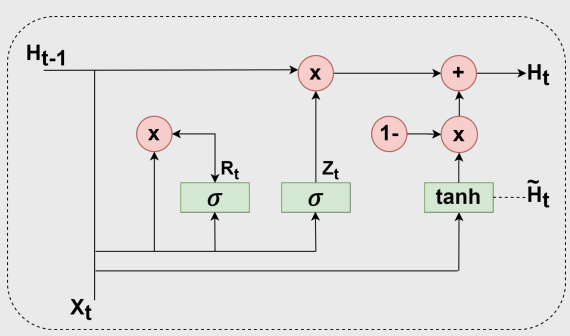
Sau đó, lớp BatchNormalization ổn định quá trình huấn luyện, lớp Dropout giảm Overfitting.

Cuối cùng, vectơ thuộc tính được đưa vào lớp Attention để làm nổi bật các thuộc tính quan trọng cho việc phân loại trang web lừa đảo.

2. Nhánh cấu trúc từ ngữ tên miền

Các tên miền đầu vào được qua lớp Embedding để giảm chiều dữ liệu và overfitting. Sau đó, lớp Conv1D và MaxPooling1D thực hiện tích chập và giảm dữ liệu. Cuối cùng, lớp GRU trích xuất các thuộc tính liên tục từ ký tự trong tên miền. Cấu trúc mạng của nhánh này bao gồm:

- Lớp Embedding: Chuyển đổi vector một chiều thành ma trận giá trị thực.
- Lớp Conv1D: Thực hiện tích chập một chiều.
- Lớp MaxPooling1D: Tóm tắt dữ liệu.
- Lớp GRU: Trích xuất các thuộc tính liên tục trong tên miền.

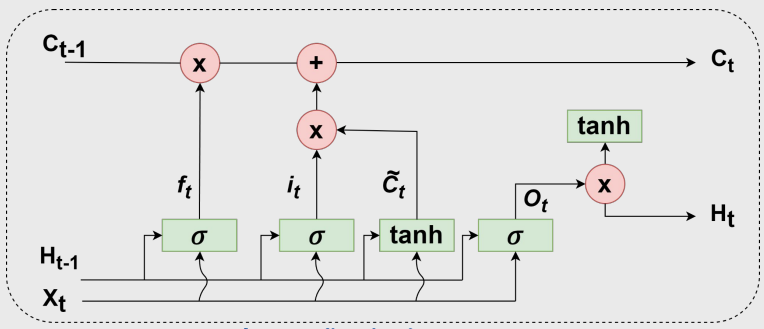


Hình 3. Kiến trúc của mạng Gated Recurrent Units

3. Nhánh cấu trúc DOM

Vector một chiều từ DOM của website được qua lớp Embedding để chuyển thành ma trận giá trị thực, cải thiện trích xuất đặc trưng. Ma trận này được đưa vào lớp Conv1D và MaxPooling1D, sau đó là lớp BiLSTM. BiLSTM, được hình thành từ LSTM theo hai hướng tiến và lùi, giúp mô hình học các phụ thuộc dài hạn hiệu quả hơn. Cấu trúc mạng của nhánh này bao gồm:

- Embedding: Chuyển vector một chiều từ DOM thành ma trận giá trị thực.
- Conv1D: Tích chập một chiều.
- MaxPooling1D: Giảm dữ liệu.
- BiLSTM: Trích xuất các thuộc tính liên tục và quan hệ giữa các thẻ.



Hình 4. Kiến trúc của mạng LSTM

Các thuộc tính đặc trưng được chấm điểm ở các lớp Attention sẽ kết nối lại với nhau thành một vectơ một chiều sau đó đổ vào các lớp ẩn. Lớp ẩn cuối cùng sẽ đưa ra quyết định trong khoảng từ 0 đến 1 với 0 cho trang web lành tính và 1 cho trang web lừa đảo.