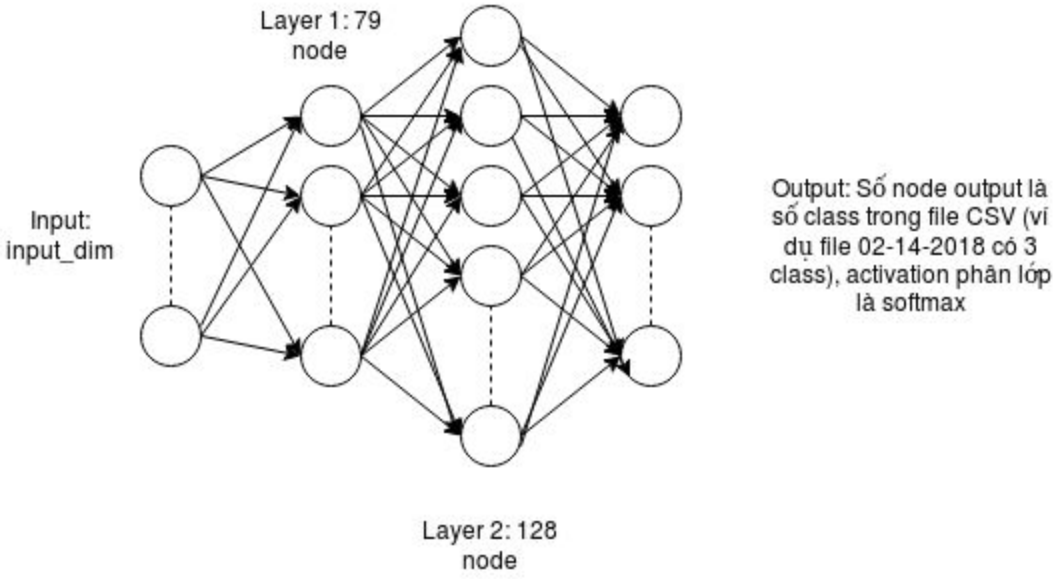
1. GIỚI THIỆU TỔNG QUAN PROJECT

Project khảo sát hiệu suất của các deep learning framework khác nhau trong bài toán phát hiện và phân loại lưu lượng truy cập mạng với mục đích thiết kế hệ thống phát hiện xâm nhập dựa trên Machine Learning.

Đầu tiên, tải data IDS-2018 xuống, sau đó clean data như sau:

- Loại bỏ hàng với giá trị Infinitiy.
- Một số file có tiêu đề lặp đi lặp lại, loại bỏ chúng.
- Chuyển định dạng timestamp: từ 15-2-2018 sang epoch UNIX từ 1/1/1970.
- Phân tách dữ liệu dựa trên attack types cho mỗi file dữ liệu.

Gần 20K hàng đã bị xóa như một phần của việc clean data. Sau đó data đã xử lý được sử dụng để huấn luyện mô hình phát hiện và phân loại lưu lượng truy cập mạng trong ***Hình 1.1***.



Hình 1.1. Mô hình phát hiện và phân loại lưu lượng truy cập mạng.

Mô hình có đầu vào là input_dim (với input_dim là độ dài của vector đầu vào), sau đó qua layer 1 với 79 node, layer 2 với 128 node và sử dụng activation ReLu. Activation softmax được sử dụng để phân lớp với hàm mất mát là cross entropy. Lưu ý rằng số node tại output phụ thuộc vào class (attack types) trong từng file csv và các node trong mô hình được kết nối đầy đủ (Fully connected).

Các deep learning framework được sử dụng so sánh trong project: keras (với backend lần lượt là tensorflow, cntk, theano) và fastai.

Kỹ thuật K-fold cross-validation (với K=10 trong project này) được sử dụng để huấn luyện mô hình.

2. CÁC KIẾN THỨC LIÊN QUAN

2.1. Tổng quan về quá trình huấn luyện dữ liệu

Xem lý thuyết tại: https://nguyenvanhieu.vn/machine-learning-la-gi/#21-xu-ly-anh

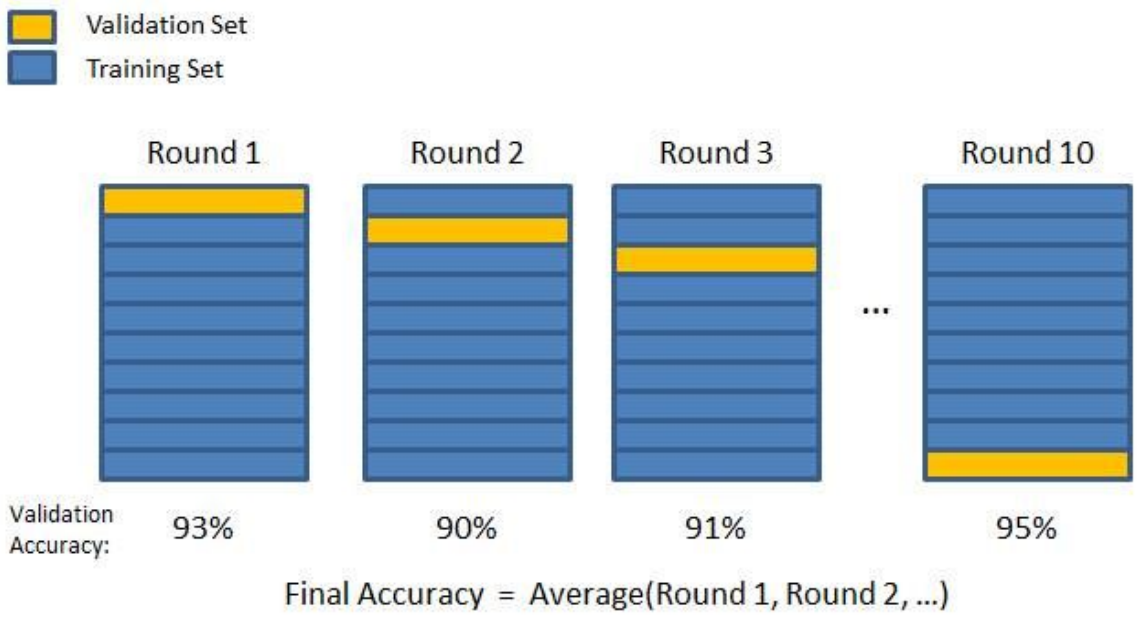
2.2. ANN, ReLU, sofmax, Cross entropy

Xem lý thuyết tại: https://viblo.asia/p/handbook-cv-with-dl-phan-1-cac-khai-niem-co-ban-trong-computer-vision-va-deep-learning-maGK7p2MZj2#_rectified-linear-unit-relu-5

2.3. K-fold cross-validation

Xem lý thuyết tại: http://cuonglv1109.blogspot.com/2018/11/cross-validation.html và

https://machinelearningcoban.com/2017/03/04/overfitting/#-cross-validation



Hình 2.1. K-fold cross-validation