CHƯƠNG 4: CÀI ĐẶT VÀ TRIỂN KHAI

4.1 GIỚI THIỆU VỀ PYTHON VÀ THƯ VIỆN TENSORFLOW

4.1.1 Python

Mã nguồn xây dựng mô hình huấn luyện của đề tài được phát triển dựa trên Python. Python là một ngôn ngữ lập trình thông dịch được sử dụng rất phổ biến trong lĩnh vực khoa học máy tính nhờ những ưu điểm sau:

❖ Đa nền tảng

Python có thể chạy trên nhiều hệ điều hành như Windows, MacOS, Linux/Unix và một số hệ điều hành khác trên máy tính. Ngoài ra, Python còn có cả những phiên bản chạy được trên .NET, máy ảo Java. Tất cả chỉ với cùng một mã nguồn cho một công việc.

❖ Đơn giản

Python có cú pháp rất đơn giản, rõ ràng. Cú pháp của Python dễ viết và dễ đọc hơn rất nhiều khi so sánh với những ngôn ngữ lập trình khác như Java, C/C++, C#, JavaScript, ... Điều này cũng giúp cho nhà phát triển tập trung vào việc phát triển giải pháp thay vì cú pháp.

❖ Mã nguồn mở

Python là một dự án mã nguồn mở nên nhà phát triển có thể thoải mái sử dụng cho các mục đích cá nhân và vì vậy nên cộng đồng phát triển Python thường xuyên đưa ra những bản cập nhật mới nhằm tăng trải nghiệm cũng như tối ưu hoá Python.

❖ Nhiều thư viện hỗ trợ

Python có một khối lượng lớn các thư viện tiêu chuẩn giúp cho công việc của nhà phát triển trở nên dễ dàng hơn rất nhiều, đặc biệt là các thư viện xử lý toán học của Python cực kỳ đa dạng và mạnh mẽ.

4.1.2 Tensorflow

Thư viện Tensorflow được sử dụng trong việc tính toán các biểu đồ và các dữ liệu dưới dạng số hoá trong sản phẩm khoá luận. Là một thư viện mã nguồn mở hỗ trợ mạnh mẽ các phép toán học để tính toán trong máy học. Để xây dựng một mô hình huấn luyện cho đề tài, nhóm sử dụng các giao diện lập trình cấp thấp (low level APIs) mà Tensorflow cung cấp:

❖ Tensor

Đây là một sự khái quát hóa các vectơ và ma trận cho các kích thước có khả năng cao hơn. Là cấu trúc dữ liệu đại diện cho tất cả các loại dữ liệu trong Tensorflow. Một tensor sẽ có 3 thuộc tính cơ bản nhất bao gồm:

- Số bậc (rank): giúp phân loại dữ liệu của tensor. (Scalar, Vector, Matrix, N- Tensor)

- Số chiều (shape): giúp xác định mức độ tương hợp giữa các tensor khi thực hiện tính toán.
- Kiểu dữ liệu (type): kiểu dữ liệu cho toàn bộ các thành phần (elements) trong tensor.

❖ Graph

Đây là một loại đồ thị với các đỉnh (node) là đại diện cho biến đầu vào hoặc một phép tính toán và các cạnh (edge) là đại diện cho dữ liệu truyền bên trong đồ thị tức dữ liệu đầu vào và đầu ra của các phép tính tại một đỉnh. Và trong tensorflow, tất cả thành phần bên trong một đồ thị đều ở dạng tensor. Cách xử lý tính toán theo hướng đồ thị này có thể giúp tensorflow tận dụng được khả năng tính toán song song bằng việc chia tách các phép toán độc lập và khả năng phân tán khi chia nhỏ công việc xử lý cho nhiều CPU, GPU khác nhau.

❖ Session

Đây là một phiên xử lý được định nghĩa trong thư viện tensorflow. Một đối tượng phiên (session) cung cấp quyền truy cập vào các thiết bị trong máy cục bộ và các thiết bị từ xa bằng cách sử dụng thời gian chạy phân tán. Nó cũng lưu trữ thông tin về đồ thị (graph) để có thể chạy cùng một tính toán hiệu quả nhiều lần. Nếu không có phiên (session), mọi tính toán trong đồ thị (graph) sẽ gần như không được triển khai.

4.2 CÀI ĐẶT

4.2.1 Giới thiệu

4.2.2 Cài đặt

4.3 DỮ LIỆU HUẤN LUYỆN MÔ HÌNH

Để huấn luyện một hệ thống dịch máy từ tiếng Anh sang tiếng Việt dựa trên mô hình nhóm sinh viên phát triển thì bộ dữ liệu phải có hai thành phần chính bao gồm:

❖ Văn bản tiếng Anh

Để huấn luyện hệ thống dịch máy từ tiếng Anh sang tiếng Việt đủ tốt thì lượng dữ liệu văn bản dùng để huấn luyện cũng phải đủ nhiều và đủ tốt. Nhóm sinh viên đã thu thập được khoảng 2.500.000 câu song ngữ để tiến hành huấn luyện. Ngoài ra dữ liệu khi huấn luyện cũng cần điều chỉnh sao cho độ dài bé hơn 100 từ để bảo đảm mô hình huấn luyện tốt nhất có thể.

❖ Văn bản tiếng Việt

Là bản dịch tương ứng với nội dung của câu tiếng Anh. Dữ liệu tiếng Việt với khoảng 2.500.000 câu và để bảo đảm mô hình huấn luyện tốt ta cũng nên hạn chế độ dài lớn hơn 100 từ vì nhóm sinh viên giới hạn độ dài câu. Nếu câu dài hơn sẽ bị cắt bỏ và mất đi các nội dung quan trọng, mô hình sẽ huấn luyện lâu hơn, sai sót.

Trong quá trình thu thập dữ liệu, nhóm sinh viên đã gặp rất nhiều vấn đề về chất lượng dữ liệu như: các tập dữ liệu song ngữ English – Vietnamese có khá nhiều với các dự án như là …. Tuy nhiên các dự án này lại không công khai dữ liệu nên nhóm sinh viên phải thu thập khắp nơi. Đối với những mẫu có mức độ sai lệch nhỏ nhóm cố gắng tinh chỉnh sao cho phù hợp nhất. Những mẫu bị sai lệch nhiều hoặc chất lượng quá thấp buộc nhóm sinh viên phải bỏ. Việc này một phần sẽ giảm bớt tình trạng gây nhiễu cho mô hình trong quá trình huấn luyện. Điều này dẫn đến thời gian huấn luyện mô hình còn khoảng ? giờ cho khoảng 2.500.000 câu song ngữ. Trong đó, dữ liệu được chia nhỏ thành 3 bộ train, dev, test với kích thước như sau:

❖ Bộ train:

❖ Bộ dev:

❖ Bộ test:

4.4 HUẤN LUYỆN MÔ HÌNH

4.5 ĐÓNG GÓI MÔ HÌNH

4.6 XÂY DỰNG MÁY CHỦ (SERVER)

Flask Framework và AWS EC2 (hoặc AWS Elastic Beanstalk) là hai nền tảng được nhóm sinh viên chọn để xây dựng hệ thống máy chủ nhằm đóng vai trò làm cầu nối giữa ứng dụng và mô hình nhận dạng âm thanh. Với các yếu tô như tốc độ triển khai nhanh gọn, sự tiện ích và tính thông dụng

nên việc chọn hai nền tảng này để xây dựng máy chủ là quyết định phù hợp với nhu cầu đặt ra của nhóm sinh viên.

Hệ thống máy chủ trong giới hạn luận văn này sẽ cung cấp ra bên ngoài duy nhất một giao diện lập trình ứng dụng (Application Programming Interface-API) để chuyển đổi văn bản tiếng Anh (dạng text) nhận được và trả về dữ liệu văn bản tiếng Việt tương ứng.

4.7 TỔNG KẾT