TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Nghiên cứu ứng dụng mạng đồ thị giải bài toán tự động trích xuất thông tin chữ viết từ hình ảnh

NGUYỄN MẠNH HIỆP
hiep.nm176750@sis.hust.edu.vn

Ngành: Công nghệ thông tin

Giảng viên hướng dẫn: TS. Nguyễn Phi Lê _____

PGS.TS. Đỗ Phan Thuận _____

Chữ kí GVHD

Trường: Công nghệ thông tin và Truyền thông

HÀ NỘI, 08/2022

LỜI CẢM ƠN

Để hoàn thành đồ án tốt nghiệp này, em xin cảm ơn đến các thầy cô, gia đình và bạn bè, những người đã giúp đỡ và động viên em trong thời gian vừa qua. Nay đã sắp kết thúc quãng đường là sinh viên đại học Bách khoa Hà Nội.

Đầu tiên, em xin gửi lời cảm ơn đến cô TS. Nguyễn Phi Lê – giảng viên trực tiếp hướng dẫn em hoàn thành đồ án này. Cô là người đam mê với công việc, luôn quan tâm đến sinh viên. Trong kỳ cuối cùng để thực hiện đồ án tốt nghiệp này, cô luôn khuyến khích sinh viên lên lab nhiều buổi nhất có thể để tập trung hơn. Nhờ sự hướng dẫn nhiệt tình của cô, đồ án tốt nghiệp (ĐATN) của em đã hoàn thành như kỳ vọng.

Tiếp theo, con xin gửi lời cảm ơn đến bố mẹ, gia đình. Thời gian học đại học có rất nhiều khó khăn, tuy nhiên với sự động viên của bố mẹ đã tiếp thêm động lực để con hoàn thành quãng đường sinh viên.

Cuối cùng, tôi xin gửi lời cảm ơn đến tất cả những người bạn đã trải qua quãng đời sinh viên với bao niềm vui, nỗi buồn. Việc gặp gỡ và quen biết các bạn là niềm vinh hạnh đối với tôi.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Trong lĩnh vực thị giác máy tính, những tác vụ như định vị chữ viết và nhận diện chữ viết đã trở nên tốt với những mô hình mạng nơ-ron học sâu. Trong ứng dụng của OCR (Optical Character Recognition - nhận diện kí tự quang học), một lớp bài toán trích xuất thông tin trong tài liệu có nhiều ứng dụng trong thực tế, chẳng hạn như hóa đơn, đơn thuốc, etc. Đối với lớp bài toán này, có những hướng tiếp cận khác nhau như phương pháp sử dụng tập luật, phương pháp sử dụng những mô hình ngôn ngữ tương tự bài toán phân loại văn bản. Tuy nhiên, những phương pháp trên không có tính tổng quát đối với nhiều định dạng của tài liệu. Vì vậy, em đề xuất một giải pháp hoàn chỉnh cho bài toán trích xuất thông tin từ ảnh văn bản, cụ thể là đơn thuốc, nhận đầu vào là hình ảnh định dạng bất kỳ, trả về thông tin về các trường chuẩn đoán bệnh, tên thuốc, số lượng, cách dùng, ngày tháng. Giải pháp của em bao gồm 3 khối chính: khối tiền xử lý gồm mô hình BERT, khối mạng nơ-ron đồ thị GraphSAGE và khối hậu xử lý gồm các lớp mạng phân loại. Em sử dụng hàm mất mát tiêu điểm để huấn luyện mô hình. Ngoài ra, để tăng độ chính xác của phần nhận diện chữ viết, em đề xuất một thuật toán tiền xử lý tài liệu, giúp tài liệu có thể được nắn chỉnh thẳng, sử dụng bài toán tìm bao lồi và tìm hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi.

Đóng góp chính của đồ án tốt nghiệp gồm 3 phần: mô hình trích xuất thông tin trong tài liệu, phương pháp tiền xử lý dữ liệu nghiêng và trang web trích xuất thông tin từ đơn thuốc tiếng Việt. Để đánh giá hiệu quả của mô hình đề xuất, em so sánh mô hình đề xuất trên 3 bộ dữ liệu liên quan đến bài toán trích xuất thông tin trong tài liệu như là bộ dữ liệu SROIE, bộ dữ liệu FUNSD và bộ dữ liệu đơn thuốc bệnh viện của Việt Nam. Kết quả cho thấy, mô hình đề xuất có hiệu suất vượt trội mô hình PICK [1] 3/3 bộ dữ liệu.

ABSTRACT

In the field of computer vision, tasks such as text localization, text recognition have become better with deep learning models. In application of OCR (Optical Character Recognition), a class of problems of extracting information from documents (such as invoices, prescriptions, etc) have many practical use. In order to deal with this class of problems, there are different approaches, for example the rule-base method, the NLP-base method. However, the above methods are not universal to many document formats. Therefore, i propose a complete solution to the problem of extracting information from text images, namely prescriptions, taking as input an image of any format, returning information about diagnose, medical name, quantity, usage, date. My solution consists of 3 main blocks: a pre-processing block consisting of a BERT model, a graph neural network block consisting of GraphSAGE Conv and a post-processing block consisting of classifier network layers. I use the focal loss function to train the model. In addition, to increase the accuracy of the text recognition model, I propose a document pre-processing algorithm, which helps the document to be straightened based on the problem of finding convex hull and a minimum area rectangle around the convex hull.

The main contribution of this graduation project consists of 3 parts: an information extraction model from documents, a data pre-processing method, a website for extracting information from Vietnamese prescriptions. To evaluate the effectiveness of this deep learning model, I compared the proposed model on 3 datasets related to the information extraction problem in documents such as the SROIE dataset, the FUNSD dataset and the Vietnamese prescription dataset. The results show that the proposed model has superior performance to the PICK model in 3/3 datasets.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
API	Giao diện lập trình ứng dụng (Application Programming Interface)
BERT	Mô hình biểu diễn mã hóa 2 chiều dựa trên biến đổi (Bidirectional Encoder Representation from Transformer)
CNN	Mạng thần kinh tích chập (Convolutional Neural Network)
GRU	Mạng hồi tiếp nút có cổng (Gated Recurrent Units)
HTML	Ngôn ngữ đánh dấu siêu văn bản (HyperText Markup Language)
LSTM	Bộ nhớ dài-ngắn hạn (Long short-term memory)
RNN	Mạng thần kinh hồi quy (Recurrent neural network)

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Hiện nay, chuyển đổi số trong cuộc Cách mạng công nghiệp 4.0 đang diễn ra trong tất cả các lĩnh vực kinh tế xã hội nhằm thay đổi phương thức quản lý, vận hành, mô hình sản xuất, kinh doanh. Việc số hóa những giấy tờ như thủ tục hành chính, hóa đơn bán hàng, đơn thuốc sẽ giúp nhà nước, doanh nghiệp có thể tận dụng được nền tảng công nghệ thông tin được xây dựng cho dữ liệu điện tử một cách thuận tiện và nhanh chóng, giảm được thời gian so với việc xem xét, rà soát những thông tin đó trên giấy. Trong phạm vi của đồ án này, em tập trung vào giải quyết việc số hóa những thông tin quan trọng trong một đơn thuốc, được định nghĩa là bài toán trích xuất thông tin từ ảnh tài liệu.

Bài toán trích xuất thông tin từ ảnh tài liệu là một lớp bài toán con trong lĩnh vực nhận diện ký tự quang học (thường được gọi là bài toán OCR - Optical Character Recognition). Mục tiêu của bài toán này là việc phân loại các hộp văn bản vào những trường thông tin tương ứng được định nghĩa trong từng loại tài liệu. Trong bộ dữ liệu đơn thuốc thu thập từ bệnh viện Việt Nam, những trường thông tin bao gồm: diagnose (chuẩn đoán), medical name (tên thuốc), quantity (số lượng), usage (cách dùng), date (ngày tháng). Để thực hiện được bài toán này, cần phải thực hiện được hai bài toán trước đó là định vị và nhận diện văn bản. Đầu ra của hai bài toán là đầu vào của bài toán trích xuất thông tin trong tài liệu.

1.2 Các giải pháp hiện tại và hạn chế

Trong lĩnh vực thị giác máy tính, có nhiều mô hình đạt được những thành tựu lớn với những tác vụ như định vị văn bản và nhận diện văn bản. Tuy nhiên, tác vụ trích xuất thông tin trong tài liệu là một thử thách lớn vì nó phụ thuộc vào cách bố trí của mẫu tài liệu. Trong bài toán này, thực tế có nhiều hướng tiếp cận khác nhau như: Phương pháp dựa trên mẫu tài liệu, sử dụng các tập luật được xây dựng dựa trên các hộp văn bản và sự phân bố của các hộp văn bản trên tài liệu, phương pháp dựa trên mô hình xử lý ngôn ngữ tự nhiên, sử dụng những đặc trưng về văn bản của hộp văn bản, phương pháp dựa trên mạng lưới thần kinh đồ thị, sử dụng những đặc trưng về văn bản, hình ảnh và đặc trưng về mặt không gian giữa các hộp văn bản.

Phương pháp dựa trên mẫu tài liệu [2] [3] [4] : phương pháp sử dụng các tập luật áp dụng lên tài liệu có cấu trúc cố định, không thay đổi nhiều về cách bố trí cách trình bày, tiếp theo áp dụng những kỹ thuật regrex, text/ keyword matching để xác định nhãn của những trường thông tin tương ứng. Ưu điểm của phương pháp dựa vào mẫu tài liệu là dễ tiếp cận nếu chỉ áp dụng với số lượng mẫu nhỏ, cố định.

Ngược lại, nhược điểm của phương pháp là cần tạo tập luật riêng biệt cho từng loại mẫu tài liệu, mất thời gian tạo thêm luật mới với một mẫu tài liệu mới xuất hiện. Ngoài ra còn một nhược điểm khi tiếp cận theo phương pháp này là tập luật sẽ phụ thuộc vào cách viết của từng người.

Đầu vào: Văn bản + Tọa độ của hộp văn bản $\rightarrow$ Tập luật $\rightarrow$ Đầu ra: nhãn của hộp văn bản

Phương pháp dựa trên mô hình xử lý ngôn ngữ tự nhiên [5]: phương pháp sử dụng các thông tin văn bản vào mô hình phân loại văn bản để phân loại, xác định nhãn của những trường thông tin tương ứng. Một mô hình phân loại văn bản hay được sử dụng trong học máy là mô hình Naïve Bayes Classifier for Multinomial Models [5], sử dụng như một phương pháp cơ sở trong phân loại văn bản. Ưu điểm của phương pháp dựa trên mô hình xử lý ngôn ngữ tự nhiên là khả năng áp dụng đối với dữ liệu mới cao hơn so với phương pháp dựa trên mẫu tài liệu đề cập trên. Nhược điểm của phương pháp này là phụ thuộc nhiều vào cách bố trí của mẫu tài liệu, hạn chế nhiều đối với dữ liệu dạng bảng biểu, không tận dụng được những thông tin liên quan đến tọa độ của văn bản trên tài liệu. Ví dụ, trong một đơn thuốc có bảng chứa danh sách thuốc cần uống, 2 trường thông tin là STT, quantity (số lượng) có nội dung tương đối giống nhau, với phương pháp dựa trên mô hình xử lý ngôn ngữ tự nhiên, chỉ sử dụng thông tin liên quan đến nội dung văn bản mà bỏ qua thông tin về tọa độ của văn bản, sẽ khó phân loại, xác định được nhãn của các trường thông tin tương ứng.

Đầu vào: Văn bản $\rightarrow$ Mô hình phân loại văn bản $\rightarrow$ Đầu ra: nhãn của hộp văn bản

Phương pháp dựa trên mạng nơ-ron đồ thị [1]: phương pháp sử dụng các thông tin văn bản và các thông tin tọa độ vào mô hình mạng nơ-ron đồ thị để phân loại, xác định nhãn của những trường thông tin tương ứng. Trong bài toán trích xuất thông tin quan trọng từ tài liệu bán cấu trúc như đơn thuốc, hóa đơn... thì những thông tin về tọa độ của văn bản (những thông tin về không gian của văn bản) là các thông tin quan trọng để có thể hiểu được ngữ cảnh của văn bản. Từ đó, những mạng nơ-ron tích chập hay mạng nơ-ron đồ thị trong lĩnh vực thị giác máy tính được sử dụng để xử lý và trích rút mối quan hệ giữa các hộp văn bản tài liệu. Sự kết hợp của mô hình xử lý ngôn ngữ tự nhiên và mô hình mạng nơ-ron tích chập, mạng nơ-ron đồ thị giúp khai thác được thêm thông tin từ các hộp văn bản, khắc phục được nhược điểm của phương pháp dựa trên mô hình xử lý ngôn ngữ tự nhiên đã đề cập ở trên. Những nghiên cứu trước đó sử dụng mạng nơ-ron đồ thị để giải quyết bài toán trích xuất thông tin từ tài liệu, đồng thời sử dụng hai thông tin về

ngữ nghĩa của văn bản và hình ảnh trong tài liệu. Do vậy, đây là một mô hình cơ sở để chúng tôi có thể so sánh và cải thiện.

Đầu vào: Văn bản + Tọa độ của hộp văn bản $\rightarrow$ Mô hình mạng nơ-ron đồ thị $\rightarrow$ Đầu ra: nhãn của hộp văn bản

1.3 Mục tiêu và định hướng giải pháp

Kỹ thuật huấn luyện sẵn (pre-training) được sử dụng phổ biến trong tác vụ xử lý ngôn ngữ tự nhiên và thị giác máy tính. Mô hình huấn luyện sẵn BERT [6] cũng sử dụng kỹ thuật này và có kết quả vượt trội trong tất cả các tác vụ liên quan đến ngôn ngữ tự nhiên. Từ đó, em tận dụng điểm mạnh này, đưa văn bản trong hộp văn bản đã được định vị và nhận diện bởi mô hình định vị và nhận diện văn bản trong ảnh ngữ cảnh, qua mô hình huấn luyện sẵn BERT để thu được những vec-tơ nhúng ngữ nghĩa chứa nhiều thông tin hơn so với những phương pháp trước đó như RNN [7], LSTM [8]. Bên cạnh đó, mô hình đề xuất sử dụng mạng nơ-ron đồ thị nhằm học thông tin của nút bao gồm đặc trưng về văn bản và tọa độ.

Để đánh giá hiệu năng của mô hình đề xuất, em đã thực hiện huấn luyện trên 3 bộ dữ liệu (bộ dữ liệu SROIE, bộ dữ liệu FUNSD, bộ dữ liệu đơn thuốc). Em cũng so sánh kết quả của mô hình đề xuất với mô hình PICK [1] và mô hình không chứa mô-đun mạng nơ-ron đồ thị, so sánh độ hội tụ của hàm mất mát tiêu điểm và hàm mất mát balanced cross entropy. Từ kết quả so sánh, mô hình đề xuất cho thấy sự hiệu quả tốt hơn khi so sánh với mô hình cơ sở 3/3 bộ dữ liệu chuẩn, mô hình đề xuất sử dụng hàm mất mát tiêu điểm có tốc độ hội tụ nhanh hơn so với hàm mất mát balanced cross entropy. Ngoài thay đổi về mặt mô hình của bài toán trích xuất thông tin trong tài liệu, em đã đề xuất phương pháp sử dụng bao lồi để xử lý những trường hợp ảnh tài liệu nghiêng trong thực tế.

1.4 Đóng góp của đồ án

Đồ án của em có ba đóng góp chính sau:

1. Đề xuất một phương pháp tiền xử lý tài liệu nghiêng. Em nhận thấy, trong những tài liệu như hóa đơn, đơn thuốc, etc, phần chính giữa của tài liệu là nội dung của văn bản cần xử lý, còn lại phần còn lại của tài liệu không thực sự cần thiết trong bài toán trích xuất thông tin. Do đó, phương pháp cắt ảnh theo bao lồi của những hộp văn bản được phát hiện trong ảnh tài liệu được đề xuất.
2. Đề xuất một mô hình mạng nơ-ron đồ thị trích xuất thông tin trong tài liệu. Mô hình đề xuất sử dụng những thông tin như văn bản và tọa độ của văn bản một cách đầy đủ và hiệu quả như một biểu diễn ngữ nghĩa tốt giúp phân loại nhãn của các trường thông tin tương ứng.

3. Xây dựng trang web trích xuất thông tin của đơn thuốc tiếng Việt

1.5 Bố cục đồ án

Phần còn lại của đồ án được tổ chức như sau. Trong chương 2, em giới thiệu một số nghiên cứu liên quan trong bài toán trích xuất thông tin trong tài liệu và các kiến thức nền tảng và cơ sở lý thuyết phục vụ cho quá trình nghiên cứu đồ án. Chương 3 mô tả chi tiết mô hình đề xuất và phương pháp tiền xử lý tài liệu nghiêng. Trong mô hình đề xuất trích xuất thông tin trong tài liệu bao gồm mô hình xử lý ngôn ngữ tự nhiên BERT, mạng nơ-ron đồ thị và hàm mất mát tiêu điểm. Phương pháp tiền xử lý tài liệu nghiêng dựa trên khái niệm về bao lồi và bài toán tìm hình chữ nhật nhỏ nhất bao quanh bao lồi. Chương 4, em đưa ra các phân tích lý thuyết về mặt toán học nhằm chứng minh sự hiệu quả của hàm mất mát tiêu điểm trong việc xử lý vấn đề mất cân bằng dữ liệu. Chương 5 trình bày kết quả thí nghiệm đánh giá hiệu năng của mô hình đề xuất. Cuối cùng, em trình bài các kết luận và hướng phát triển của đồ án ở chương 6.

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

Trong phần này, em trình bày các kiến thức nền tảng liên quan đến đồ án của em. Cụ thể, trong chương này, em trình bày về học máy cơ bản, mô hình huấn luyện sẵn BERT, mô hình mạng nơ-ron đồ thị GraphSAGE, thư viện xử lý hình ảnh OpenCV, bài toán nhận diện ký tự quang học (OCR). Trong chương tiếp theo, em sẽ trình bày về giải pháp đề xuất của em.

2.1 Ngữ cảnh của bài toán

Không chỉ dừng lại ở việc định vị và nhận diện văn bản, mà quan trọng hơn là phải bóc tách được các thông tin. Ví dụ, đối với một hóa đơn thì cần bóc tách được tên từng loại hàng hóa, giá cả, thuế, etc. Đây là một bài toán khó và trước đây thường được giải quyết bằng cách xây dựng các tập luật. Ví dụ, mô hình [2] [3] [4] với cách tiếp cận bằng tập luật giải quyết bài toán trích xuất thông tin từ tài liệu. Tuy nhiên, việc dựa vào những tập luật như vậy sẽ không có tính bao quát, không đối ứng được khi thay đổi các mẫu giấy tờ. Khi đi sâu nghiên cứu bài toán, em nhận thấy, mặc dù các chi tiết cụ thể trong các mẫu giấy tờ thay đổi, nhưng quan hệ tương đối về mặt không gian, vị trí của các thành phần trong một loại văn bản thường có tính nhất quán. Ví dụ: chẩn đoán bệnh thường nằm phía trên của danh sách các thuốc trong ảnh đơn thuốc. Chính vì thế, trong đồ án của em, em tập trung nghiên cứu ứng dụng mạng nơ-ron đồ thị để giải quyết bài toán trích xuất thông tin từ ảnh văn bản.

2.2 Các kết quả nghiên cứu tương tự

Trong lĩnh vực thị giác máy tính, bài toán OCR đã quen thuộc với giới nghiên cứu, có nhiều mô hình đạt kết quả tốt cho bài toán nhận diện văn bản trong ảnh khung cảnh [9], [10], [11], bài toán phát hiện văn bản trong ảnh khung cảnh [12], [13], [14]. Và một bài toán tiếp theo trong OCR cũng được quan tâm đó là bài toán trích xuất thông tin văn bản từ ảnh tài liệu [15], [16], [17], [18], [1]. Những nghiên cứu hiện tại nhận thấy vai trò quan trọng của việc sử dụng đặc trưng về văn bản lẫn đặc trưng về tọa độ văn bản của tài liệu để cải thiện độ chính xác của bài toán trích xuất thông tin từ tài liệu. Tuy nhiên, phần lớn các phương pháp tập trung vào đặc trưng về văn bản thông qua những bộ trích xuất đặc trưng khác nhau như mạng nơ-ron hồi quy (RNN) hay mạng nơ-ron tích chập (CNN) [19], [20]. Mô hình [21] dùng mạng mã hóa- giải mã (encoder - decoder) dự đoán mặt nạ phân đoạn (segmentation mask) và hộp giới hạn (bounding box). Điều này cho thấy [21] đang chỉ tập trung vào những đặc trưng về mặt hình ảnh, không chú ý đến đặc trưng về văn bản. Bên cạnh đó, [3] là một phương pháp sử dụng đầy đủ các đặc trưng để hỗ

trợ việc trích xuất thông tin, tuy nhiên những đặc trưng đó thường được con người tự thiết kế, hoặc phù hợp với nhiệm vụ cụ thể, vì vậy không có tính mở rộng cho các loại tài liệu khác nhau. Trong bài báo [1] đề xuất một mô hình sử dụng kết hợp mô hình Transformer để thu được vec-tơ nhúng chứa thông tin ngữ nghĩa của văn bản bên trong các hộp văn bản, mô hình mạng nơ-ron tích chập để thu được vec-tơ đặc trưng về hình ảnh của tài liệu, mô hình mạng nơ-ron đồ thị với đầu vào là vec-tơ tổng hợp bởi vec-tơ văn bản và vec-tơ hình ảnh. Tuy nhiên, hiệu quả của mô hình [1] đối với những tài liệu có nhiều hộp văn bản (bộ dữ liệu FUNSD, bộ dữ liệu đơn thuốc Việt Nam) không tốt.

2.3 Học máy cơ bản

Học máy (machine learning) là một lĩnh vực của ngành trí tuệ nhân tạo giúp giải quyết những vấn đề trong mọi lĩnh vực bằng cách cho phép hệ thống tự học từ dữ liệu. Trong học máy, chia thành hai nhánh chính:

1. Học có giám sát: mô hình hóa một tập dữ liệu có sẵn những ví dụ đã được gán nhãn
2. Học không giám sát: mô hình hóa một tập dữ liệu không có sẵn những ví dụ đã được gán nhãn

2.3.1 Độ đo Precision, Recall, F1 score

Trong bài toán phân loại nhị phân có khái niệm lớp dữ liệu cần xác định đúng là lớp Positive (dương tính) và lớp dữ liệu còn lại là lớp Negative (âm tính). Có 4 định nghĩa True Positive (dự đoán đúng là dương tính), False Negative (dự đoán sai là âm tính), False Positive (dự đoán sai là dương tính), True Negative (dự đoán đúng là âm tính).

Độ đo Precision: tính bằng tỷ lệ giữa True Positive và tổng của True Positive và False Positive – tổng số điểm được phân loại là lớp Positive.

Công thức toán học:

$$PRECISION = \frac{TP}{TP + FP} \tag{2.1}$$

Độ đo Recall: tính bằng tỷ lệ giữa True Positive và tổng của True Positive và False Negative – tổng số điểm thực sự thuộc lớp Positive. Công thức toán học:

$$RECALL = \frac{TP}{TP + FN} \tag{2.2}$$

Độ đo F1 score: còn được gọi là trung bình điều hòa của hai đại lượng Precision

và Recall. Công thức toán học:

$$F1 = \frac{2}{\frac{1}{PRECISION} + \frac{1}{RECALL}} \tag{2.3}$$

Trong bài toán phân loại nhiều lớp, coi lần lượt một lớp là Positive và các lớp còn lại là Negative để tính các Precision, Recall tương ứng.

2.3.2 Batch Normalization

Batch Normalization là một kỹ thuật huấn luyện mạng nơ-ron bằng cách chuẩn hóa các đầu vào thành một lớp mạng theo từng lô nhỏ. Phương pháp chuẩn hóa này giúp chuẩn hóa các vec-tơ đặc trưng (đầu ra của một lớp mạng sau khi qua các hàm kích hoạt) về phân phối chuẩn (giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1).

Đầu vào: Lô nhỏ $B = \{x_{1..m}\}$ và tham số γ,β

Đầu ra: tập hợp những vec-tơ đã được chuẩn hóa $\{y_i = BN_{\gamma,\beta}(x_i)\}$

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Hình 2.1: Mô tả thuật toán batch normalization

Tác động của phương pháp Batch Normalization lên việc huấn luyện mô hình mạng nơ-ron:

- Giả sử một lớp mạng $y = Wx + b$, đạo hàm của đầu ra y theo trọng số W bị ảnh hưởng bởi đầu vào x. Nếu đầu vào x có độ lớn không ổn định thì mô hình sẽ không ổn định. Batch Normalization có thể tăng độ ổn định khi huấn luyện mô hình mạng nơ-ron.
- Giảm được hiện tượng quá khớp (overfitting)

- Không cần sử dụng nhiều lớp mạng Dropout
- Tăng tốc độ học (learning rate) khi huấn luyện mô hình nơ-ron

2.3.3 Hàm softmax

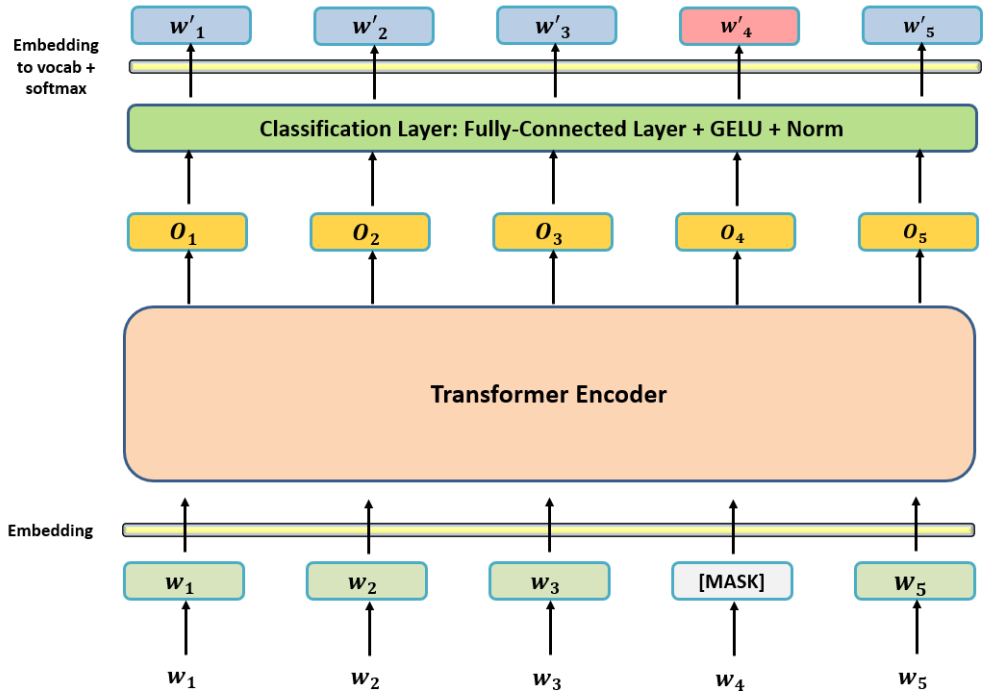
Hàm softmax là hàm số tính toán khả năng xuất hiện của một lớp trong tổng số tất cả các lớp có thể xuất hiện. Xác suất này được dùng để xác định lớp mục tiêu cho đầu vào của bài toán. Một vec-tơ n chiều có giá trị thực bất kỳ sau khi đi qua hàm softmax, thành vec-tơ n chiều có giá trị thực trong khoảng (0, 1] có tổng bằng 1. Trong học sâu, hàm softmax được sử dụng ở nhiều lớp mạng, trong đó có lớp mạng cuối cùng của bài toán phân loại nhiều lớp.

Công thức toán học:

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.4}$$

Trong đó: σ là hàm softmax; $\overrightarrow{z}$ là vec-tơ đầu vào; e là hàm lũy thừa tiêu chuẩn; K là số lớp cần phân loại

2.4 Mô hình huấn luyện sẵn BERT



Hình 2.2: Mô hình BERT

Mô hình BERT [6] là mô hình biểu diễn mã hóa 2 chiều từ kĩ thuật Transformer. Mô hình hoặc được sử dụng để trích xuất đặc trưng ngôn ngữ chất lượng cao từ dữ liệu văn bản, hoặc được tinh chỉnh (fine-tune) phù hợp với những nhiệm vụ khác nhau như phân loại, trả lời câu hỏi,etc. Hình 2.2 mô tả kiến trúc của mô hình BERT.

Cơ chế chú ý (attention) của kiến trúc Transformer là cơ chế truyền toàn bộ các từ trong câu văn bản đồng thời. Do đó, Transformer được xem là mô hình huấn luyện đầu vào theo hai chiều, giúp mô hình học bối cảnh của một từ dựa trên tất cả những từ xung quanh.

Với mô hình huấn luyện sẵn BERT, có thể lấy vec-tơ nhúng của một câu văn bản đầu vào để thực hiện tác vụ khác nhau của bài toán cụ thể.

Hiện nay, mô hình huấn luyện sẵn BERT có rất nhiều phiên bản khác nhau phụ thuộc theo 3 tham số trong kiến trúc Transformer là số lượng những khối lớp mạng con (block sub-layers) (L), kích thước của vec-tơ nhúng (hidden size) (H), Số lượng head trong multi-head layer (A). Tên gọi của 2 kiến trúc bao gồm:

- BERT BASE (L=12, H=768, A=12)
- BERT LARGE (L=24, H=1024, A=16)

2.5 Mô hình mạng nơ-ron đồ thị GraphSAGE

Mô hình mạng nơ-ron đồ thị GraphSAGE là một framework cho việc học biểu diễn quy nạp trên đồ thị cỡ lớn. GraphSAGE được sử dụng để sinh ra vec-tơ biểu diễn thấp chiều từ các nút.

Inductive learning là quá trình học từ các tập huấn luyện suy ra các quy luật chung rồi áp dụng các quy luật chung đó vào tập kiểm thử. Với cách học như này, mô hình không cần huấn luyện lại khi có dữ liệu mới, chưa từng xuất hiện trong tập huấn luyện nên phù hợp với việc giải các bài toán tổng quát hóa cao.

Các đặc trưng của đối tượng được biểu diễn thành các đặc trưng của nút và nhúng vào đồ thị đã cho thấy được hiệu quả trong nhiều tác vụ dự đoán. Tuy vậy những phương pháp transductive learning như GCN kém hiệu quả khi xuất hiện những nút mới, chưa xuất hiện trong tập huấn luyện.

Vấn đề này đã được mô hình mạng nơ-ron đồ thị GraphSAGE [22] giải quyết. Ý tưởng chính của mô hình là thuật toán tạo ra các vec-tơ nhúng cho nút mới chưa huấn luyện. Giải thuật thực hiện huấn luyện hàm tổng hợp *(mean, max, lstm)* giúp nút hiện tại có thể tổng hợp được thông tin từ các nút lân cận.

Giải thuật sinh ma trận embedding:

Algorithm 1: GraphSAGE embedding generation (i.e., forward propagation) algorithm

Input : Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$; input features $\{\mathbf{x}_v, \forall v \in \mathcal{V}\}$; depth K; weight matrices $\mathbf{W}^k, \forall k \in \{1, ..., K\}$; non-linearity σ; differentiable aggregator functions $\text{AGGREGATE}_k, \forall k \in \{1, ..., K\}$; neighborhood function $\mathcal{N} : v \rightarrow 2^{\mathcal{V}}$

Output : Vector representations $\mathbf{z}_v$ for all $v \in \mathcal{V}$

$\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$;
for $k = 1...K$ **do**
  for $v \in \mathcal{V}$ **do**
    $\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$;
    $\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$
  end
  $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / \|\mathbf{h}_v^k\|_2, \forall v \in \mathcal{V}$
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in \mathcal{V}$

Hình 2.3: Mô tả thuật toán sinh ma trận embedding của mô hình mạng nơ-ron đồ thị GraphSAGE

Có thể lý giải hình vẽ 2.3 như sau:
Ở vòng lặp k:

Với mỗi nút v thuộc V:

1. Đặc trưng của các nút lân cận của nút v tại bước k ($h_{\mathcal{N}(v)}^k$) bằng đầu ra của hàm tổng hợp với đầu vào là đặc trưng của nút đó tại bước $k-1$
2. Thực hiện kết hợp vec-tơ đặc trưng nút v ở bước $k-1$ (h_v^{k-1}) với vec-tơ đặc trưng của các nút lân cận của nút v tại bước k ($h_{\mathcal{N}(v)}^k$) và cho ma trận sau khi kết hợp vào một lớp mạng kết nối đầy đủ (fully connected) với một hàm kích hoạt phi tuyến σ
3. Kết quả của bước thứ 2 được sử dụng cho vòng tiếp theo

Đặc trưng của mỗi nút v ở bước k được chuẩn hóa.
Cuối cùng, sau K vòng lặp, thu được đặc trưng của nút v (z_v).

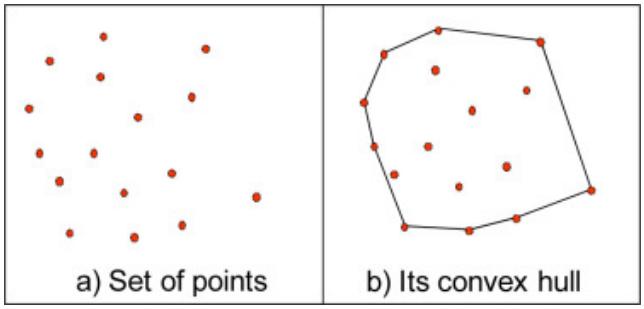
2.6 Thư viện xử lý hình ảnh OpenCV

Thư viện OpenCV (open source computer vision library) là kho lưu trữ các mã nguồn mở được dùng để xử lý hình ảnh, phát triển các ứng dụng đồ họa trong thời gian thực.

2.6.1 Convex Hull

Trong hình học tính toán, bao lồi (convex hull) của một tập điểm là tập lồi nhỏ nhất (theo diện tích, thể tích, ...) mà tất cả các điểm đều nằm trong tập đó.

Thuật toán bọc gói (Gift wrapping) là thuật toán đơn giản, dễ hiểu của bài



Hình 2.4: Minh họa bao lồi (convex hull)

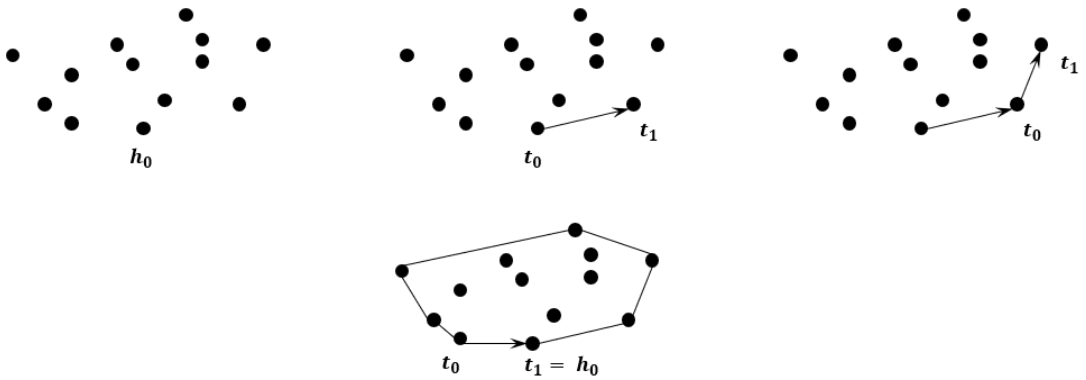
toán tìm bao lồi. Ý tưởng của thuật toán như việc đi qua các điểm với một dải băng.

Đầu vào: Cho một tập hợp n điểm $S = s_0, s_1, ..., s_n$

Đầu ra: Một tập hợp điểm nằm trên bao lồi H

Thuật toán như sau:

1. Chọn điểm h_0 là điểm thấp nhất bên phải của tập S. Thêm điểm h_0 vào tập H. Gán điểm t_0 bằng điểm h_0 (2.5a)
2. Gán điểm t_1 bằng điểm s_0. Với mỗi điểm $p \in S$, nếu p nằm ở bên phải của hướng thẳng từ t_0 đến t_1, thì gán điểm t_1 bằng điểm p. Sau bước 2, không có điểm nào nằm bên phải của hướng thẳng từ t_0 đến t_1 (2.5b)
3. Nếu điểm t_1 bằng điểm h_0 (2.5d) thì đã hoàn thành việc tìm tập H. Ngược lại, thêm điểm t_1 vào tập H, gán điểm t_0 bằng điểm t_1, và quay lại bước 2 (2.5c)



Hình 2.5: Mô tả giải thuật bọc gói

2.6.2 Perspective Transformation

Perspective Transformation (chuyển đổi góc nhìn) là một phương pháp thay đổi góc nhìn của một ảnh hoặc một video nhất định để hiểu rõ hơn về thông tin cần thiết.

Đối với những ảnh tài liệu nghiêng, cần thiết phải thực hiện thay đổi góc nhìn, giúp mô hình định vị và nhận diện văn bản hoạt động tốt hơn, Và mô hình mạng

nơ-ron đồ thị cũng sẽ học được thông tin về tọa độ một cách đơn giản hơn với những hộp giới hạn có độ nghiêng nhỏ.

Công thức toán học:

$$\begin{bmatrix} t_i\acute{x} \\ t_i\acute{y} \\ t_i \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ c_1 & c_2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.5}$$

Trong đó: (x, y) là điểm đầu vào; $(\acute{x}, \acute{y})$ là điểm thu được sau khi biến đổi; $\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$ định nghĩa sự biến đổi (xoay, mở rộng); $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ là vector dịch (translation vector); $\begin{bmatrix} c_1 & c_2 \end{bmatrix}$ là vector chiếu (projection vector)

2.7 Bài toán nhận diện ký tự quang học (OCR)

OCR là thuật ngữ về nhận diện ký tự quang học có nhiều ứng dụng thực tế như số hóa các tài liệu, xe tự hành, etc. OCR có những bài toán điển hình như mô hình định vị văn bản và mô hình nhận diện văn bản.

2.7.1 Mô hình định vị văn bản

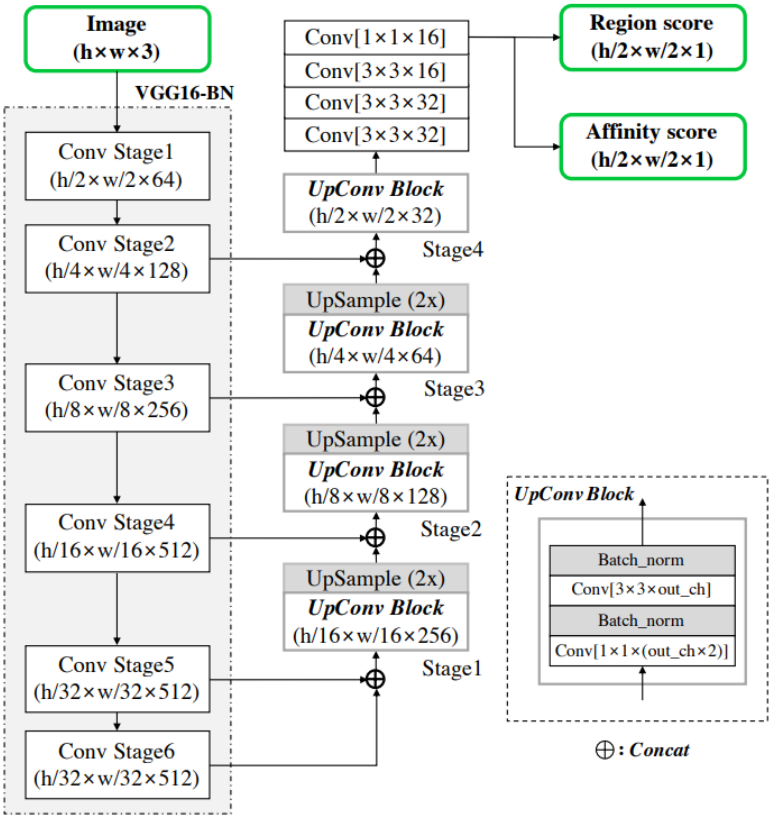
CRAFT [12] là một mô hình dùng để phát hiện vùng văn bản trong ảnh. Trong việc phát hiện văn bản do mô hình này thực hiện thì vùng văn bản được tính từ vùng ký tự và vùng kết nối giữa các ký tự thay vì phát hiện bằng phương pháp phát hiện đối tượng thông thường.

Hình 2.6 mô tả kiến trúc của mô hình CRAFT. Mô hình CRAFT sử dụng mạng VGG-16 và các lớp *skip connection* và đầu ra là *region score* và *affinity score*.

Mô hình CRAFT được huấn luyện bằng ảnh tổng hợp với cấp độ ký tự. Tạo bản đồ nhiệt (heatmap) để biểu diễn nhãn của sự thật nền tảng (ground truth label) là region score và affinity score

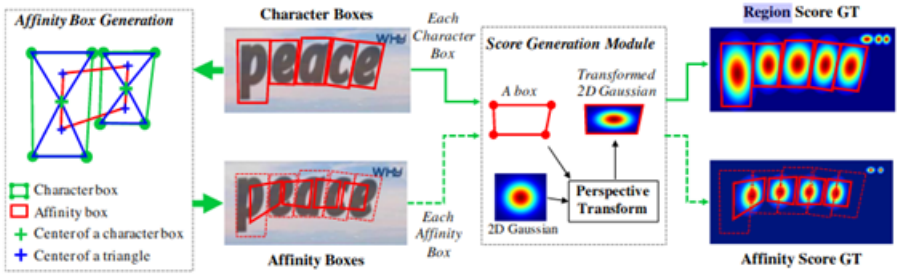
Có 3 bước để tạo bản đồ nhiệt (heatmap):

1. Chuẩn bị bản đồ hai chiều isotropic Gaussian.
2. Tính toán ma trận chuyển đổi góc nhìn (perspective transform) giữa Gaussian map và mỗi hộp ký tự.
3. Dùng ma trận tìm được để chuyển bản đồ Gaussian về hình dạng của hộp ký tự. Đối với affinity score, nối hai đường chéo của hộp ký tự tạo thành 2 tam giác phía trên và phía dưới. Và một affinity box có đỉnh là tâm của 4 cái tam



Hình 2.6: Kiến trúc của mô hình phát hiện văn bản CRAFT

giác của 2 hộp ký tự liền kề.

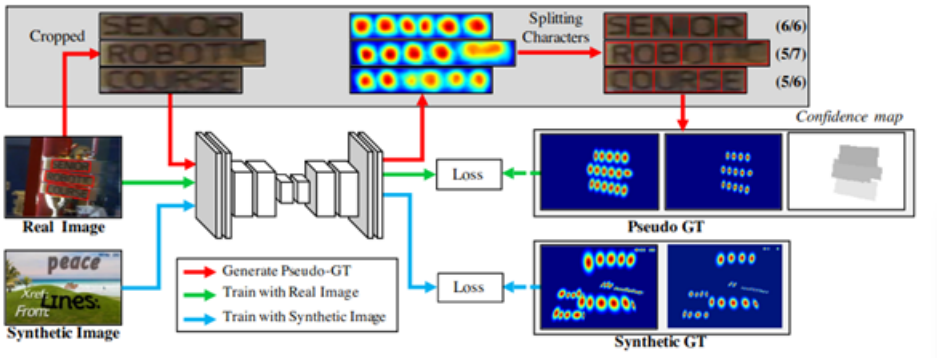


Hình 2.7: Hình ảnh mô tả quá trình sinh dữ liệu cho mô hình CRAFT

Phương pháp học giám sát yếu: không giống như bộ dữ liệu tổng hợp, ảnh thực có nhãn ở cấp độ từ. Mục tiêu là tách các ký tự từ mỗi wordbox này. Khi một ảnh thực tế được đưa vào mô hình, tạm thời sẽ dự đoán điểm số của vùng ký tự của các từ đã được cắt ra nhằm tạo ra hộp giới hạn mức độ ký tự. Để biểu diễn độ tin cậy dự đoán của mô hình tạm thời, giá trị của bản đồ tin cậy (confidence map) ở mỗi hộp từ (word-box) là tỷ lệ của số lượng ký tự bóc tách được và số lượng ký tự của sự thật nền tảng (ground truth), giá trị này được dùng để tính độ mất mát và cập nhật trọng số.

Các bước thực hiện:

1. Đầu tiên, cắt những hộp từ (word-box) ra, sau đó dùng mô hình tạm thời để dự đoán region score.
2. Tiếp theo, dùng thuật toán *watershed* để tách các ký tự ra nhằm tạo ra hộp giới hạn mức độ ký tự.
3. Cuối cùng, tọa độ của các hộp ký tự sẽ được chuyển lại về tạo độ trong ảnh thực. Pseudo-ground truth cho region score và affinity score sẽ được tạo ra giống như ở bước trên.



Hình 2.8: Hình ảnh mô tả quá trình huấn luyện của mô hình CRAFT

Với phương pháp học giám sát yếu, mô hình được huấn luyện với các pseudo-GTs không hoàn hảo, tức là các nhãn bị sai, điều này dễ dẫn đến đầu ra vùng ký tự bị mờ. Để giải quyết điều này, chất lượng của pseudo-GTs được tính toán theo độ dài của mỗi từ. Với $R(w)$ và $l(w)$ là vùng của hộp giới hạn và độ dài của từ của mẫu w. $l^c(w)$ là tổng số lượng ký tự được dự đoán, thì độ tin cậy $s_{conf}(x)$ được tính dựa trên công thức:

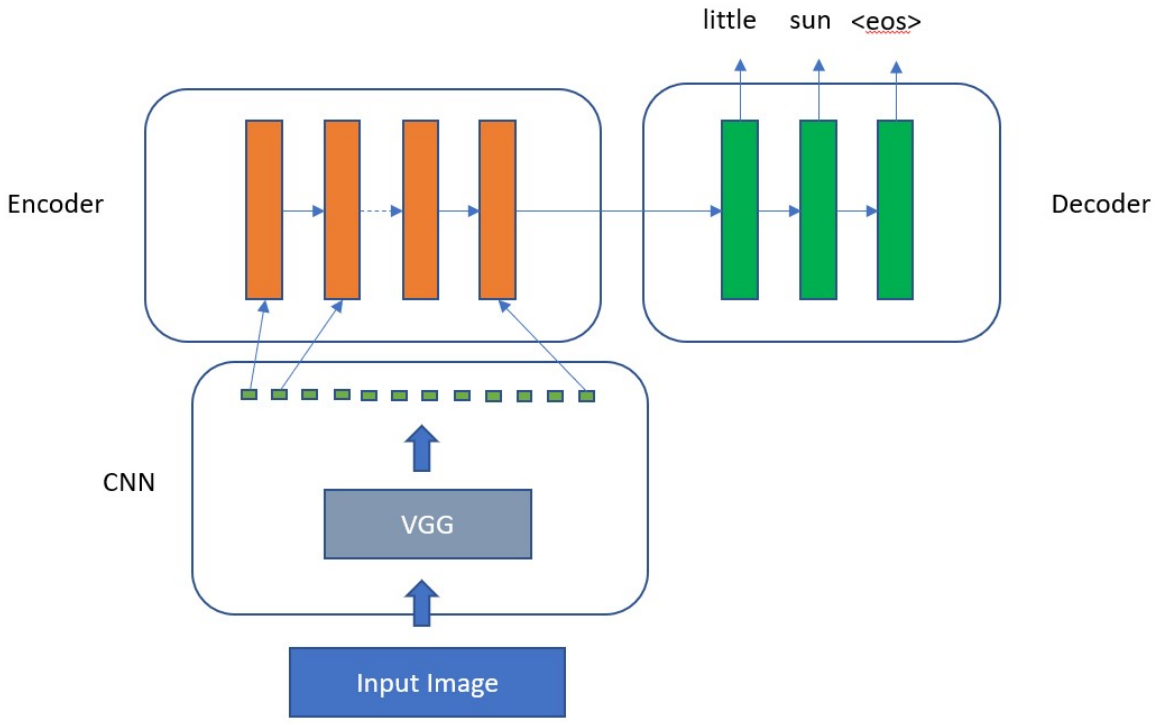
$$s_{conf}(w) = \frac{l(w) - min(l(w), |l(w) - l^c(w)|)}{l(w)} \tag{2.6}$$

Với phương pháp huấn luyện, mục tiêu là region score và affinity score được dự đoán từ mạng nơ-ron sẽ càng gần vớ sự thật nền tảng (ground truth), nên hàm mất mát MSE được sử dụng trong mô hình CRAFT:

$$L = \sum_p S_c(p).(\|S_r(p) - S_r^*(p)\|_2^2 + \|S_a(p) - S_a^*(p)\|_2^2) \tag{2.7}$$

Trong đó $S_r(p)$, $S_a(p)$ là region/affinity score groundtruth; $S_r^*(p)$, $S_a^*(p)$ là region/affinity score dự đoán từ mô hình. Bên cạnh đó, ta chỉ muốn học ít đối với các groundtruth không có độ tin cậy cao, nên đặt $S_c(p)$ như một trọng số kiểm soát việc này.

2.7.2 Mô hình nhận diện văn bản



Hình 2.9: Minh họa mô hình nhận diện văn bản

Mô hình nhận diện văn bản (hình 2.9) được đề cập đến ở phần này là VGG-Seq2seq. Mô hình bao gồm 2 phần:

1. Phần bóc tách đặc trưng từ ảnh, thành các vec-tơ đầu vào của phần tiếp theo
2. Phần sử dụng mô hình mạng sequence to sequence được lấy ý tưởng từ bài toán dịch máy.

Phần thứ nhất: sử dụng cấu trúc mạng VGG đơn giản phù hợp với kích thước ảnh nhỏ

Phần thứ hai: sử dụng mô hình sequence to sequence gồm:

- Với bộ mã hóa nhận đầu vào là đầu ra của mạng VGG sau đó được đi qua lớp BiLSTM. BiLSTM giúp các vec-tơ đầu ra có được thông tin ngữ cảnh. Đầu ra của mạng BiLSTM được cho đi qua lớp mạng tầng kết nối đầy đủ (fully connected layer) trở thành đầu ra của bộ mã hóa.
- Với bộ giải mã sử dụng cơ chế chú ý (attention). Chữ cái ở vị trí thứ t được tính bằng cách: Sử dụng trạng thái ẩn (hidden state) của bước thời gian thứ $t-1$ đóng vai trò như *query* và đầu ra của bộ mã hóa đóng vai trò vừa là *key* vừa là *value* trong cơ chế chú ý (attention). Đầu ra của cơ chế chú ý (attention) ta được một vec-tơ được nối với vec-tơ nhúng của từ thứ $t-1$ sau đó qua mạng hồi quy (RNN) thu được *rnn output* và *hidden state*. Hidden state được sử

dụng trong quá trình sinh ra từ thứ $t + 1$ còn rnn output được đi qua lớp mạng linear, softmax và từ thứ t chính là từ có xác suất lớn nhất.

2.8 Kết chương

Trong chương nay, em trình bày ngữ cảnh của bài toán trích xuất thông tin trong tài liệu dựa vào hướng tiếp cận mạng nơ-ron đồ thị. Em cũng trình bày những nghiên cứu tương tự và các kiến thức cơ bản liên quan đến học máy cơ bản như tham số đánh giá, độ đo Precision, Recall, F1-score, lớp mạng Batch Normalization, hàm softmax. Bên cạnh đó, mô hình huấn luyện sẵn BERT, mô hình mạng đồ thị GraphSage, convex hull (bao lồi), perspective transformation (chuyển đổi góc nhìn), mô hình định vị văn bản CRAFT, mô hình nhận diện văn bản VGG-Seq2seq. Những kiến thức này sẽ liên quan đến phương pháp đề xuất trong chương tiếp theo.

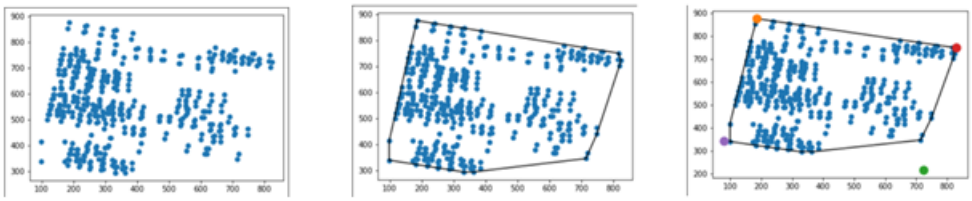
CHƯƠNG 3. GIẢI PHÁP TRÍCH XUẤT THÔNG TIN TỪ ẢNH VĂN BẢN

Luồng xử lý đối với bài toán tự động trích xuất thông tin từ ảnh văn bản là từ ảnh đơn thuốc đầu vào, em sử dụng mô hình định vị văn bản để lấy được tọa độ của những hộp văn bản, sau đó, nếu ảnh bị nghiêng sẽ đi qua thuật toán tiền xử lý ảnh nghiêng để thu được văn bản và tọa độ của văn bản của ảnh sau khi nắn chỉnh. Thuật toán tiền xử lý ảnh nghiêng dựa trên bài toán tìm bao lồi và hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi. Tiếp theo, những thông tin trên được đưa qua mô hình trích xuất thông tin trong tài liệu sử dụng mạng nơ-ron đồ thị giúp phân loại văn bản vào các trường thông tin tương ứng. Trong mô hình đề xuất, em sử dụng mô hình BERT, mô hình mạng nơ-ron đồ thị GraphSAGE và hàm mất mát tiêu điểm. Bên cạnh việc xây dựng mô hình học sâu, em cũng phát triển một trang web đơn giản giúp trích xuất thông tin từ đơn thuốc tiếng Việt.

3.1 Thuật toán tiền xử lý ảnh nghiêng

Hình 3.1: Minh họa ảnh sau khi thực hiện thuật toán đề xuất

Đối với những tài liệu như hóa đơn hay đơn thuốc, trên thực tế tồn tại nhiều ảnh tài liệu nghiêng và có những phần sau (background) phức tạp xung quanh, tạo ra thử thách lớn cho việc cải thiện độ chính xác của mô hình phát hiện, nhận diện văn bản. Không chỉ vậy, tài liệu nghiêng cũng ảnh hưởng nhiều đến việc xây dựng cạnh của đồ thị bằng vị trí tương đối giữa các hộp văn bản trong mô hình mạng nơ-ron đồ thị. Trước đó, những phương pháp sử dụng mô hình học sâu như [23], [24] đã tiếp cận theo hướng *định vị tài liệu như phát hiện điểm đặc trưng* sao cho góc của tài liệu là bốn điểm góc phía trên bên trái, phía trên bên phải, phía dưới bên phải, phía dưới bên trái. Những phương pháp trên thường được dùng để xử lý đối với tài liệu bị nghiêng trong những hệ thống trích xuất thông tin, tuy nhiên, trong trường hợp ảnh chụp tài liệu bị nghiêng và mất đi một vài góc, mô hình học sâu phát hiện 4 góc không còn hiệu quả.



Hình 3.2: Mô tả thuật toán

Trong hình bên trái của 3.1, có thể thấy ảnh đã bị mất 2 điểm TL (góc phía trên bên trái), BL (góc phía dưới bên trái), khiến các mô hình học sâu phát hiện 4 góc gặp khó khăn. Nhận thấy, đặc trưng của ảnh chứa tài liệu (hóa đơn, đơn thuốc ...) là phần văn bản nằm gọn ở bên trong tài liệu. Vì vậy, em đã đề xuất một phương pháp dựa vào những văn bản bên trong tài liệu để thực hiện phép biến đổi góc nhìn (perspective transformation). Phương pháp được tóm lược bằng 4 bước như sau:

1. Dùng mô hình học sâu phát hiện văn bản để xác định vị trí của hộp văn bản (giả sử tài liệu có n hộp văn bản, có tất cả $4 * n$ điểm thuộc vùng chứa văn bản)
2. Tìm một bao lồi chứa $4 * n$ điểm trên
3. Tìm một hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi tìm được tại bước 2.
4. Từ 4 điểm của hình chữ nhật có diện tích nhỏ nhất, thực hiện biến đổi góc nhìn để nắn chỉnh ảnh

Hình bên trái của 3.2 giúp trực quan hóa $4 * n$ điểm của n hộp văn bản được mô hình phát hiện văn bản tìm được. Hình chính giữa của 3.2 trực quan hóa bao lồi của $4 * n$ điểm. Hình bên phải của 3.2 trực quan hóa 4 điểm góc của hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi.

Sau khi xác định được 4 điểm trên tài liệu nghiêng, hình bên phải của 3.1 là kết quả của phép biến đổi góc nhìn. Sau khi sử dụng phép biến đổi góc nhìn, thu được ảnh chỉ chứa vùng văn bản, bỏ qua bối cảnh phức tạp bên ngoài.

Nhận thấy, phương pháp này rất phù hợp với những tài liệu giàu thông tin về văn bản, bởi vì nếu tài liệu thỏa mãn điều đó sẽ có số lượng điểm đủ lớn, dẫn đến việc tìm bao lồi và hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi có tính hiệu quả và độ chính xác cao.

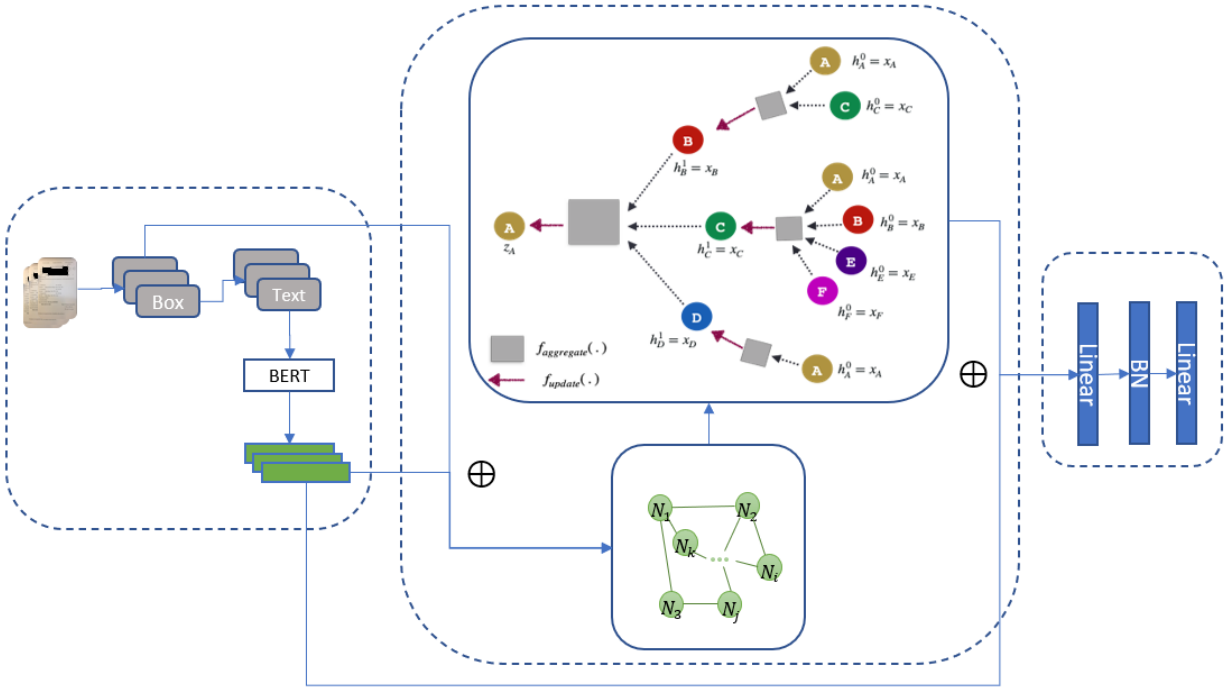
Phương pháp tìm hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi:

1. Tính toán bao lồi của các điểm
2. Đối với mỗi cạnh của bao lồi:

- Tính toán hướng cạnh với hàm arctan
- Xoay phần bao lồi băng cách sử dụng hướng này để dễ dàng tính toán diện tích hình chữ nhật giới hạn với tối thiểu/tối đa của x/y của phần bao lồi đã xoay.
- Lưu trữ hướng tương ứng với diện tích tối thiểu được tìm thấy

3. Trả lại hình chữ nhật tương ứng với diện tích tối thiểu tìm được

3.2 Mô hình trích xuất thông tin trong tài liệu



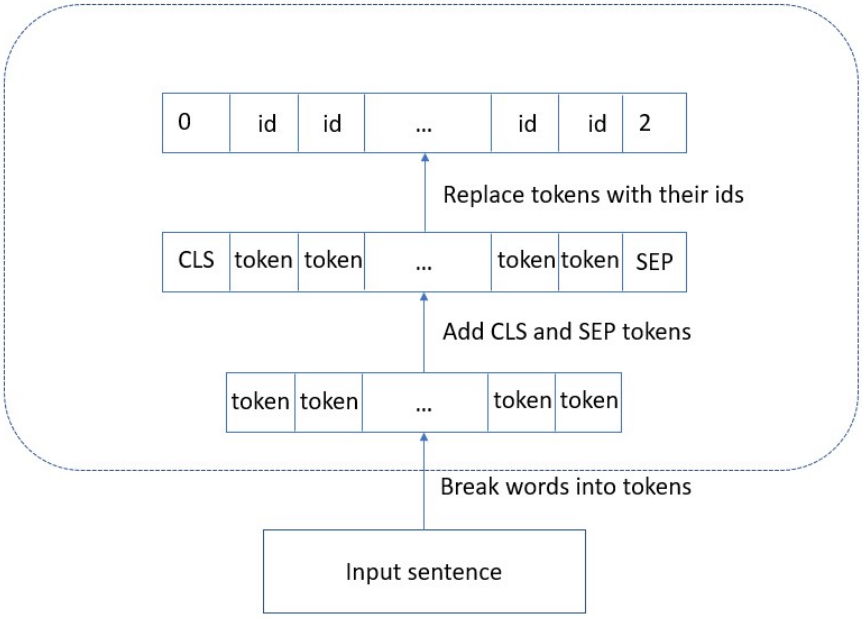
Hình 3.3: Mô hình mạng đồ thị đề xuất

Mô hình sẽ lấy dữ liệu đầu vào là giá trị của các hộp văn bản bao gồm văn bản, tọa độ của văn bản đi qua mô hình huấn luyện sẵn BERT, thu được vec-tơ nhúng biểu diễn ngữ nghĩa của những văn bản trong tài liệu. Đối với thông tin liên quan đến tọa độ của văn bản, em đã chuẩn hóa bằng cách chia cho độ rộng và chiều cao của ảnh tài liệu. Tạo vec-tơ đầu vào cho mô-đun mạng nơ-ron đồ thị giúp phân tích và học mối quan hệ của những hộp văn bản trong tài liệu. Và sau đó, qua khối cuối cùng để phân loại nhiều lớp.

3.2.1 Mô hình huấn luyện sẵn BERT

Đối với một mô hình liên quan đến lĩnh vực xử lý ngôn ngữ tự nhiên thì kiến trúc Transformer đang rất phổ biến. Cụ thể ở đây là mô hình BERT – mô hình biểu diễn từ theo 2 chiều sử dụng kiến trúc Transformer. Mô hình BERT đạt rất nhiều kết quả tốt trên các tác vụ xử lý ngôn ngữ tự nhiên khi so sánh với những phương pháp trước đó như RNN [7], LSTM [8], GRU [25].

Hình 3.4 mô tả quá trình hoạt động của BERT tokenizer. Đầu tiên, sử dụng BERT tokenizer chia câu văn thành những mã thông báo (tokens). Sau đó, các ký tự đặc biệt như ký tự CLS ở vị trí ban đầu, ký tự SEP ở vị trí cuối cùng được thêm vào.Bước tiếp theo là sẽ thay thế những mã thông báo đó bằng những id đại diện cho mã thông báo trong bảng nhúng (embedding table).



Hình 3.4: Mô tả cách hoạt động của BERT tokenizer

Một vec-tơ sau khi được xử lý bởi tokenizer có thể qua mô hình BERT thu được đầu ra là một vec-tơ của mỗi mã thông báo đầu vào, được tạo thành từ 768 số thực đối với phiên bản BASE.

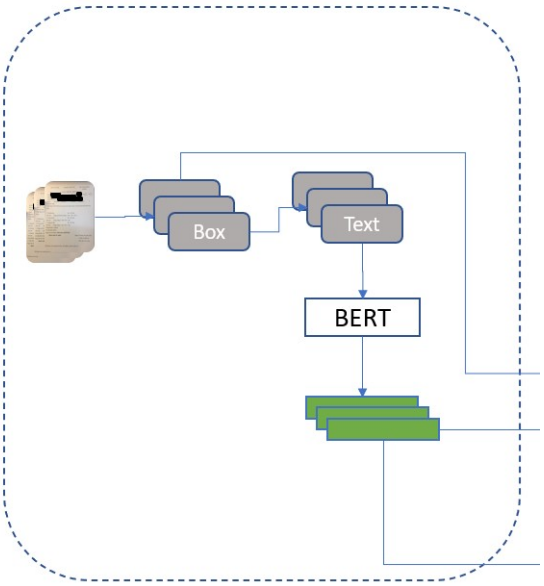
3.2.2 Chi tiết các thành phần của mô hình đề xuất

Nghiên cứu ngày càng tăng về học sâu đã dẫn đến việc sử dụng các phương pháp dựa trên mạng học sâu được áp dụng cho đồ thị. Với mạng học sâu, việc mô hình hóa các cấu trúc phi tuyến tính trở nên dễ dàng hơn, vì vậy deep-autoencoders được sử dụng để giảm số chiều.

Dựa trên những thông tin tương quan giữa những hộp văn bản trong tài liệu như hóa đơn, đơn thuốc, mô hình hóa chúng như một đồ thị đã cho phép nhà nghiên cứu hiểu về dữ liệu một cách hệ thống.

Mô hình đề xuất trong đồ án này tiếp cận theo hướng phân loại nút mạng trong đồ thị (node classification). Mô hình mạng nơ-ron đồ thị giải quyết bài toán trích xuất thông tin trong tài liệu gồm 3 mô-đun chính như sau:

Mô-đun tiền xử lý:



Hình 3.5: Hình vẽ mô tả mô-đun tiền xử lý

Ở module này, từ thông tin văn bản của hộp văn bản, thu được vec-tơ nhúng khi đưa qua mô hình huấn luyện sẵn BERT. Tiếp đó, đưa qua các lớp mạng Linear và Activation để giảm chiều của vec-tơ nhúng. Cho một câu văn bản $s_i = \left(w_1^{(i)}, w_2^{(i)}, ..., w_T^{(i)}\right)$, vec-tơ nhúng của câu văn được định nghĩa như sau:

$$te^{(i)} = BERTEncoder(w_{1:T}^{(i)}, \Theta_{benc}) \tag{3.1}$$

Trong đó:

$w_{1:T}^{(i)}$ là T từ đầu vào của câu văn thứ i; $te^{(i)}$ là đầu ra của mô hình BERT; Θ_{benc} là tham số của mô hình BERT

Trong ảnh tài liệu, N câu văn được mã hóa một cách độc lập nhau thu được N vec-tơ nhúng.

$$TE = \left[te^{(1)}, te^{(2)}, ..., te^{(N)}\right] \tag{3.2}$$

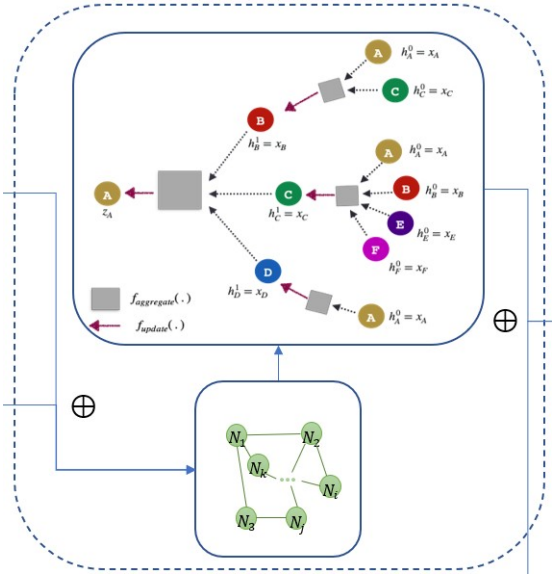
Bên cạnh đó, có những đặc trưng như độ dài của văn bản, số lượng ký tự là chữ số trong văn bản, tọa độ của hộp văn bản là những đặc trưng được em xem xét cho mô hình mạng nơ-ron đồ thị. Cụ thể, trong những bộ dữ liệu hiện tại, có những trường thông tin chứa chữ số nên đặc trưng mới cũng sẽ được mô hình đề xuất học trong quá trình huấn luyện.

Cách xây dựng đồ thị $G = (V, E)$ như sau:

- Nút của đồ thị được xây dựng dựa trên đặc trưng đã thu được ở trên.
- Cạnh của đồ thị được xây dựng dựa trên 2 yếu tố:

1. Thông tin về mối quan hệ giữa các hộp văn bản của bộ dữ liệu (nếu có)
2. Tìm kiếm 4 hộp văn bản xung quanh (trên, dưới, trái, phải) của hộp văn bản đang xét

Mô-đun mạng nơ-ron đồ thị:



Hình 3.6: Hình vẽ mô tả mô-đun mạng đồ thị

Trong mô-đun trước, những đặc trưng trên được kết hợp lại, làm đầu vào cho mô-đun mạng đồ thị này. Trong mô-đun này, chứa L lớp mạng GraphSage Conv để học được vec-tơ biểu diễn của nút mạng chứa thông tin của hộp văn bản bằng việc tổng hợp thông tin từ các nút hàng xóm. Việc tổng hợp những thông tin từ các nút hàng xóm cũng rất quan trọng bởi vì giúp mô hình tổng quát hóa được đối với dữ liệu chưa từng nhìn thấy. Em sử dụng hàm tổng hợp trung bình cộng để huấn luyện mô hình GraphSAGE.

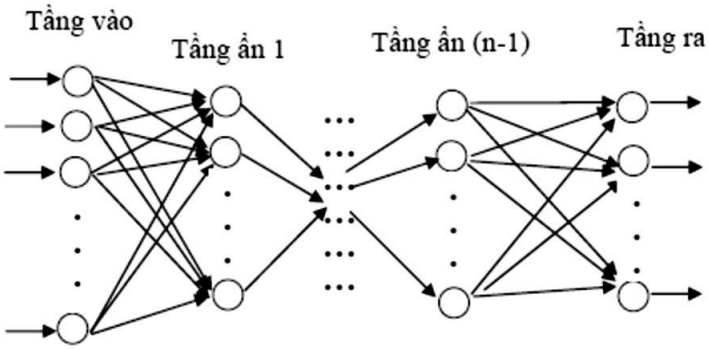
$$x_i^{'} = W_1 \cdot x_i + W_2 \cdot mean_{j \in \mathcal{N}(i)} x_j \tag{3.3}$$

Nhận thấy sự quan trọng từ thông tin của vec-tơ nhúng văn bản mang lại, em thực hiện nối vec-tơ nhúng văn bản và vec-tơ đặc trưng đồ thị tạo thành một vec-tơ tổng hợp, làm đầu vào cho mô-đun tiếp theo.

$$\hat{x_i} = x_i^{'} \oplus te^{(i)} \tag{3.4}$$

Mô-đun hậu xử lý: Trong mô-đun này, em sử dụng lớp mạng tuyến tính (Linear), Batch Normalization liên tiếp nhau để có thể giúp phân loại được nhãn của hộp văn bản trong tài liệu. Em sử dụng lớp mạng Batch Normalization ở giữa hai lớp

mạng tuyến tính (Linear) để tránh hiện tượng quá khớp, tăng tốc độ học khi huấn luyện mô hình mạng nơ-ron đề xuất. Mô hình của mô-đun hậu xử lý này giống với Perceptron Đa tầng (MLP), đầu vào của mạng MLP là vec-tơ $\hat{x_i}$ do mô-đun trước truyền tới, đầu ra của mạng MLP là vec-tơ chứa q chiều với q là số lớp cần phân loại của bài toán. Ở lớp mạng cuối cùng, đi qua hàm softmax thu được xác xuất của những lớp của mẫu t, từ đó, lấy ra lớp có xác xuất lớn nhất.



Hình 3.7: Kiến trúc của mạng MLP

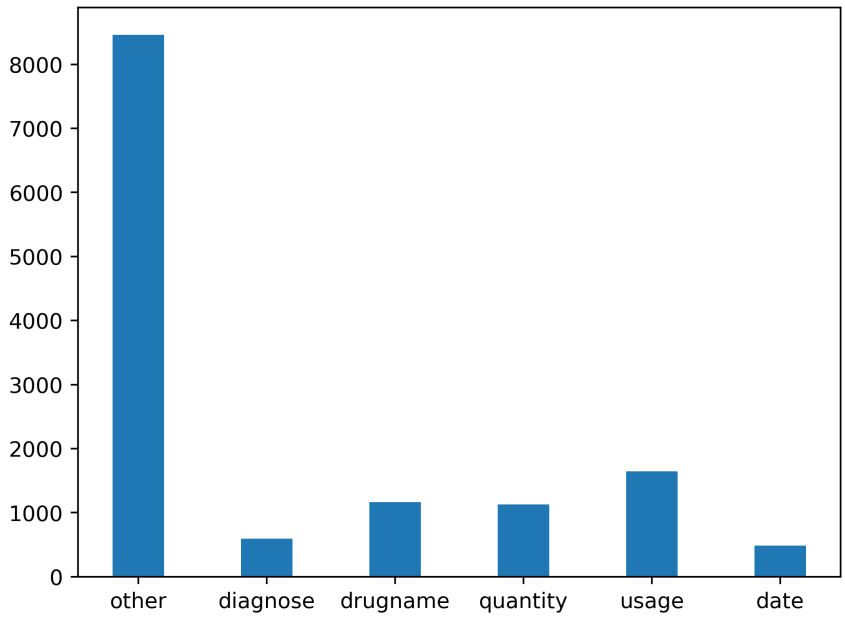
3.2.3 Hàm mất mát tiêu điểm

Hình 3.8 cho thấy các lớp trong bộ dữ liệu huấn luyện có số mẫu không cân bằng (lớp other có số mẫu lớn nhất khoảng hơn 8000 mẫu). Sự mất cân bằng này sẽ dẫn đến hiện tượng học quá khớp đối với những lớp xuất hiện nhiều trong ảnh tài liệu và không khớp đối với những lớp xuất hiện ít trong ảnh tài liệu. Vì vậy, em đề xuất sử dụng hàm mất mát tiêu điểm [26] trong mô hình đề cuất nhằm giảm thiểu sự mất cân bằng.Tác dụng của hàm mất mát tiêu điểm là việc giảm sự ảnh hưởng của những lớp có số lượng mẫu lớn và tập trung đến việc huấn luyện những lớp có số lượng mẫu ít hơn. Công thức toán học:

$$FL(p_t) = \alpha_t(1-p_t)^{\gamma}log(p_t) \tag{3.5}$$

Trong đó: p_t là giá trị xác suất dự đoán của mẫu t; α, γ là hai siêu tham số của hàm mất mát. $\gamma \in [0, 5]$

Hàm entropy chéo cân bằng (balanced cross entropy) [27] là hàm số cân bằng được tỷ lệ phân phối của mẫu dựa vào siêu tham số α. Tuy nhiên, hàm số này chưa thay đổi được suy giảm độ dốc. Đối với bộ dữ liệu mất cân bằng như hình 3.8, suy giảm độ dốc trong hàm mất mát bị ảnh hưởng lớn bởi những lớp có tần xuất xuất hiện nhiều khi huấn luyện mô hình, nghĩa là đạo hàm do những lớp có tần xuất xuất hiện lớn sẽ đóng góp nhiều hơn vào việc thay đổi hàm mất mát so với những lớp có tần xuất xuất hiện ít hơn. Bên cạnh đó, trọng số α của hàm balanced cross entropy không giúp tập trung vào các mẫu khó dự đoán trong các lớp.



Hình 3.8: Tần xuất xuất hiện của các lớp trong tập huấn luyện. Trục hoành là các lớp, trục tung là số mẫu trong tập huấn luyện

Ví dụ, đối với lớp chẩn đoán (diagnose) của bộ dữ liệu đơn thuốc tiếng Việt có độ dài văn bản không ổn định, nếu sử dụng hàm balanced cross entropy, tất cả những mẫu trong lớp chẩn đoán có cùng một trọng số, điều này dẫn đến việc mô hình chưa tập trung đúng và đủ vào những mẫu khó học (độ dài văn bản lớn) của lớp dữ liệu này.

Hàm mất mát tiêu điểm là hàm số điều chỉnh một cách triệt để, gia tăng sự ảnh hưởng của mẫu khó dự đoán (thường nằm trong lớp dữ liệu xuất hiện ít) lên suy giảm độ dốc. Bên cạnh đó, hàm mất mát tiêu điểm cũng giúp mô hình tập trung trực tiếp vào những mẫu khó dự đoán. Nhân tử điều chỉnh $(1 - p_t)^\gamma$ tác động lên hàm mất mát và độ dốc đạo hàm vì $(1 - p_t)^\gamma$ tỷ lệ nghịch với p_t, khi p_t lớn (đối với mẫu dễ dự đoán), nhân tử trên nhỏ, dẫn đến mức độ đóng góp vào hàm mất mát không đáng kể và ngược lại khi p_t nhỏ (đối với mẫu khó dự đoán), đóng góp vào hàm mất mát nhiều hơn. Giả sử, trường hợp dễ dự đoán có xác xuất $p_t = 0.9$ và trường hợp khó dự đoán có xác xuất $p_t = 0.1$ thì tỷ lệ chênh lệch đóng góp vào hàm mất mát với siêu tham số $\gamma = 1$ là:

$$\frac{(1 - 0.1)^1}{(1 - 0.9)^1} = \frac{0.9}{0.1} = 9 \tag{3.6}$$

3.3 Kết chương

Trong chương này, em đã trình bày về những đóng góp chính của đồ án tốt nghiệp, cụ thể là phương pháp xử lý tài liệu nghiêng dựa trên khái niệm bao lồi và phép chuyển đổi góc nhìn. Phương pháp này giúp tiền xử lý những ảnh tài liệu nghiêng. Đóng góp thứ 2 là mạng nơ-ron đề xuất. Trong mô hình đề xuất này, mô hình BERT giúp lấy được vec-tơ nhúng ngữ nghĩa của văn bản đầu vào, kết hợp những thông tin về mặt không gian, tọa độ của những hộp văn bản, tạo ra những đặc trưng đồ thị, giúp mô hình không chỉ học được ý nghĩa về mặt văn bản mà còn học được tương quan giữa những hộp văn bản trong tài liệu. Ngoài ra, trong mô hình đề xuất sử hàm mất mát tiêu điểm thay vì hàm balanced cross entropy, giúp tốc độ hội tụ của mô hình nhanh hơn. Những thực nghiệm về mô hình mạng nơ-ron đồ thị đề xuất được trình bày trong chương tiếp theo.

CHƯƠNG 4. ĐÁNH GIÁ ĐỘ CHÍNH XÁC CỦA MÔ HÌNH ĐỀ XUẤT

Trong phần này, em sẽ trình bày về phương pháp thí nghiệm và các kết quả thực nghiệm có được khi so sánh mô hình đề xuất và mô hình cơ sở.

4.1 Phương pháp thí nghiệm

Em đã sử dụng mô hình mạng nơ-ron đồ thị PICK, mô hình huấn luyện sẵn BERT cho bài toán phân loại văn bản làm mô hình cơ sở để so sánh với mô hình đề xuất. Khi so sánh kết quả với mô hình PICK, có thể kiểm chứng được hiệu quả của mô hình đề xuất vì hai mô hình có cùng hướng tiếp cận với mạng nơ-ron đồ thị. Khi so sánh với mô hình huấn luyện sẵn BERT với tác vụ phân loại văn bản, có thể kiểm chứng được độ hiệu quả của mô-đun mạng nơ-ron đồ thị và hàm mất mát tiêu điểm. Bảy thí nghiệm em thực hiện bao gồm:

- 1 thí nghiệm chạy mô hình PICK với tập dữ liệu đơn thuốc tiếng Việt
- 3 thí nghiệm chạy mô hình đề xuất trên 3 bộ dữ liệu nêu trên
- 2 thí nghiệm chạy mô hình đề xuất thiếu mô-đun đồ thị (hay còn gọi là mô hình huấn luyện sẵn BERT cho tác vụ phân loại văn bản) trên bộ dữ liệu đơn thuốc tiếng Việt và bộ dữ liệu SROIE
- 1 thí nghiệm chạy mô hình đề xuất với hàm mất mát balanced cross entropy

4.1.1 Cài đặt hai đề xuất

Mô hình mạng nơ-ron đề xuất: Em cài đặt mô hình đề xuất bằng ngôn ngữ lập trình Python và thư viện Deep-learning Pytorch, thư viện hỗ trợ việc xây dựng mạng nơ-ron đồ thị Pytorch Geometric, thư viện Transformer chứa mô hình huấn luyện sẵn BERT. Em sử dụng card đồ họa NVIDIA RTX 2080 Ti với 12 Gb RAM để huấn luyện, kiểm thử mô hình đề xuất.

Em sử dụng mô hình RoBERTa Tokenizer để biến đổi các từ trong câu văn bản thành các id tương ứng lấy từ bảng nhúng (embedding table), sau đó dùng mô hình huấn luyện sẵn RoBERTa phiên bản BASE với cấu hình mặc định $L = 12, H = 768, A = 12$ để lấy vec-tơ nhúng của văn bản, kết hợp cùng thông tin tọa độ (sau khi chuẩn hóa) thành một đặc trưng tổng hợp , làm đầu vào của mạng nơ-ron đồ thị.

Phần mạng nơ-ron đồ thị, em sử dụng 2 lớp mạng GraphSage Conv nối tiếp nhau để học tương quan giữa những hộp văn bản trên tài liệu bán cấu trúc như hóa đơn, đơn thuốc, etc. Mô hình mạng nơ-ron đồ thị GraphSAGE có đề xuất 3 hàm tổng hợp như hàm trung bình cộng (mean), hàm cực đại (max) và lớp mạng LSTM. Em sử dụng hàm tổng hợp là hàm trung bình cộng, để lấy trung bình cộng của thông

tin giữa những nút mạng hàng xóm, giúp mô hình có thể học tốt hơn đối với dữ liệu chưa từng huấn luyện.

Sau khi vec-tơ tổng hợp đi qua mạng nơ-ron đồ thị thu được một vec-tơ đặc trưng đồ thị, em kết hợp cùng vec-tơ nhúng lúc đầu tạo thành một vec-tơ tổng hợp làm đầu vào của khối hậu xử lý của mạng đề xuất. Điều này lấy ý tưởng của mạng Resnet giúp cho lớp mạng không quên được đặc trưng của vec-tơ nhúng văn bản.

Đối với hàm mất mát tiêu điểm, em chọn siêu tham số $\gamma = 1$ giúp điều chỉnh sự cân bằng của các lớp trong bộ dữ liệu tiếng Việt. Thật vậy, cách sử dụng hàm mất mát tiêu điểm giúp mô hình có thể tập trung vào những mẫu khó dự đoán, cải thiện tham số đánh giá.

Mô hình được huấn luyện với phương pháp tối ưu AdamW để cập nhật trọng số với siêu tham số $lr = 0.0005$. Em cũng điều chỉnh siêu tham số lr trong quá trình huấn luyện sử dụng tham số $num - warmup - step = 1000$, giúp giảm tác động của việc đi lệch hướng mô hình khi tiếp xúc với những dữ liệu mới.

Cách tăng cường dữ liệu huấn luyện: em sử dụng những đơn thuốc đã được gán nhãn, với những trường thông tin quan trọng để thêm vào những template đơn thuốc bệnh viện thu thập từ internet. Em làm theo các bước: xây dựng bộ template word của đơn thuốc bệnh viện, thêm các trường thông tin quan trọng vào template, chuyển đổi thành file ảnh. Cuối cùng từ file ảnh sẽ gán nhãn tự động bằng code.

Phương pháp tiền xử lý tài liệu nghiêng: Em sử dụng mô hình phát hiện văn bản CRAFT nhằm phát hiện tọa độ của những hộp văn bản trong tài liệu. Tiếp theo, dựa trên tất cả những điểm phát hiện được từ những hộp văn bản trên, dùng thư viện scipy để tìm được bao lồi của chúng. Khi đã có một tập đỉnh của bao lồi, dùng thư viện MinimumBoundingBox được cung cấp bằng mã nguồn mở trên github để tìm được 4 điểm của hình chữ nhật có diện tích nhỏ nhất xung quanh bao lồi đó với ý tưởng được trình bày ở chương 2. Và tìm ra ma trận của phép chuyển đổi góc nhìn giúp biến đổi hình ảnh nghiêng thành một hình ảnh có độ nghiêng nhỏ hơn trước, giúp mô hình nhận diện văn bản trong ảnh tài liệu có thể nhận diện tốt hơn. Những tọa độ của hộp văn bản được thành tọa độ mới trên ảnh tài liệu được áp dụng bởi phép biến đổi góc nhìn.

Cụ thể, đối với hàm biến đổi ảnh với 4 điểm, các bước thực hiện như sau: sắp xếp 4 điểm của hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi theo chiều kim đồng hồ với thứ tự điểm góc phía trên bên trái, điểm góc phía trên bên phải, điểm góc phía dưới bên phải, điểm góc phía dưới bên trái. Tiếp theo, tính độ rộng và chiều cao lớn nhất của hình tạo bởi 4 điểm trên, và thiết lập tọa độ của 4 điểm mới trên ảnh tài liệu sau khi áp dụng phép chuyển đổi góc nhìn, đó là điểm góc

phía trên bên trái $(0, 0)$, điểm góc phía trên bên phải $(maxWidth - 1, 0)$, điểm góc phía dưới bên phải $(maxWidth - 1, maxHeight - 1)$, điểm góc phía dưới bên trái $(0, maxHeight - 1)$. Từ đó, em tìm được ma trận của phép chuyển đổi góc nhìn M và ảnh tài liệu sau khi áp dụng phép chuyển đổi góc nhìn.

Đối với hàm quay hộp văn bản có tọa độ $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ với ma trận M thu được từ hàm built-in *cv2.getPerspectiveTransform* bằng các bước như sau: Thay đổi hình dạng (reshape) các điểm thành ma trận có chiều $[4, 2]$. Sau đó dùng hàm hstack của thư viện numpy giúp thêm một cột có giá trị 1 vào bên phải của ma trận có chiều $[4, 2]$ trên tạo ra một ma trận có chiều $[4, 3]$. Và sử dụng phép nhân ma trận giữa ma trận M có chiều $[3, 3]$ và ma trận có chiều $[4, 3]$ và các phép chuyển vị ma trận, cuối cùng thu được một ma trận có chiều $[4, 3]$. Để thu được ma trận kết quả, em lấy 2 cột đầu tiên của ma trận trên thu được tọa độ 4 điểm của một hộp văn bản trên ảnh tài liệu mới $(x_1', y_1', x_2', y_2', x_3', y_3', x_4', y_4')$, đây là kết quả được ánh xạ duy nhất bởi 4 điểm của hộp văn bản trên ảnh tài liệu nghiêng .

4.1.2 Tham số đánh giá

Để đánh giá sự hiệu quả của mô hình trích xuất thông tin văn bản từ ảnh, tham số đánh giá F1-score, precision, recall phù hợp hơn so với tham số đánh giá độ chính xác. Do đó, trong phạm vi đồ án này, em xin được sử dụng độ đo F1-score để đánh giá mô hình đề xuất cũng như so sánh kết quả với mô hình khác trong bài toán.

4.2 Bộ dữ liệu sử dụng

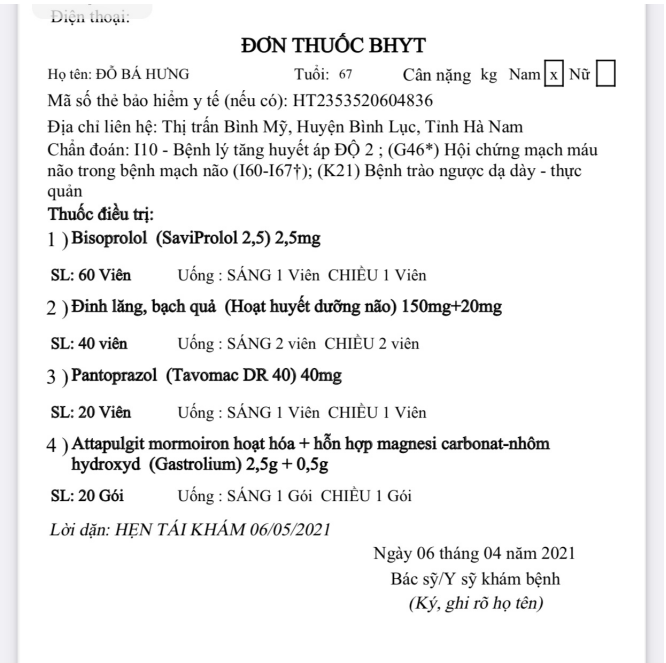
4.2.1 Bộ dữ liệu đơn thuốc tiếng Việt

Là bộ dữ liệu được thu thập tại bệnh viện ở Việt Nam, bao gồm 10 mẫu đơn thuốc khác nhau trong đó có 3 loại chứa bảng biểu và 7 loại không chứa bảng biểu. Trong bộ dữ liệu trên, có 609 ảnh đơn thuốc được gán nhãn. Ảnh đơn thuốc gồm 3 loại: file pdf, ảnh chụp màn hình, ảnh chụp từ điện thoại. Trong mỗi đơn thuốc, có 5 trường thông tin là chuẩn đoán bệnh, tên thuốc, số lượng, cách dùng, ngày tháng kê đơn. Mỗi đơn thuốc có 1 ảnh và 1 file json chứa các thông tin về hộp văn bản (id, tọa độ, văn bản, nhãn), những hộp văn bản được sắp xếp theo thứ tự từ trái sang phải, trên xuống dưới.

Phân chia	Chẩn đoán	Tên thuốc	Số lượng	Cách dùng	Ngày tháng	Khác
Train	592	1166	1129	1643	487	8462
Validation	108	260	254	366	105	1785
Test	149	261	261	341	123	1924

Bảng 4.1: Phân phối của các lớp trong bộ dữ liệu đơn thuốc tiếng Việt

Hình 4.1 là một đơn thuốc không chứa bảng biểu trong bộ dữ liệu đơn thuốc tiếng Việt, từ đơn thuốc trên, có thể thấy chẩn đoán bệnh nằm ở phía trên so với những trường thông tin quan trong khác. Tiếp đến là danh sách thuốc bao gồm tên thuốc, số lượng và cách dùng tương ứng. Thông tin ngày tháng kê đơn thuốc sẽ nằm gần cuối của đơn thuốc.



Điện thoại:

ĐƠN THUỐC BHYT

Họ tên: ĐỖ BÁ HƯNG Tuổi: 67 Cân nặng kg Nam [x] Nữ [ ]

Mã số thẻ bảo hiểm y tế (nếu có): HT2353520604836

Địa chỉ liên hệ: Thị trấn Bình Mỹ, Huyện Bình Lục, Tỉnh Hà Nam

Chẩn đoán: I10 - Bệnh lý tăng huyết áp ĐỘ 2 ; (G46*) Hội chứng mạch máu não trong bệnh mạch não (I60-I67†); (K21) Bệnh trào ngược dạ dày - thực quản

Thuốc điều trị:

1) **Bisoprolol (SaviProlol 2,5) 2,5mg**

SL: 60 Viên Uống : SÁNG 1 Viên CHIỀU 1 Viên

2) **Đinh lăng, bạch quả (Hoạt huyết dưỡng não) 150mg+20mg**

SL: 40 viên Uống : SÁNG 2 viên CHIỀU 2 viên

3) **Pantoprazol (Tavomac DR 40) 40mg**

SL: 20 Viên Uống : SÁNG 1 Viên CHIỀU 1 Viên

4) **Attapulgit mormoiron hoạt hóa + hỗn hợp magnesi carbonat-nhôm hydroxyd (Gastrolium) 2,5g + 0,5g**

SL: 20 Gói Uống : SÁNG 1 Gói CHIỀU 1 Gói

Lời dặn: HẸN TÁI KHÁM 06/05/2021

Ngày 06 tháng 04 năm 2021

Bác sỹ/Y sỹ khám bệnh

(Ký, ghi rõ họ tên)

Hình 4.1: Ảnh đơn thuốc trong bộ dữ liệu đơn thuốc Việt Nam

4.2.2 Bộ dữ liệu SROIE

Bộ dữ liệu bao gồm 626 ảnh huấn luyện và 347 ảnh kiểm thử. Mỗi hóa đơn bao gồm các dòng thông tin hộp văn bản và transcript tương ứng. Mỗi hóa đơn được gán nhãn với 4 kiểu thông tin là company, date, address, total và 1 nhãn other. Bộ dữ liệu này chủ yếu chứa chữ số và ký tự tiếng Anh. Bộ dữ liệu này có bố cục thay đổi với cấu trúc phức tạp.

Hình 4.2 là một ví dụ trong bộ dữ liệu SROIE. Điểm khó nhất của bộ dữ liệu này nằm ở hộp văn bản có nhãn là *total*. Trên mỗi tờ hóa đơn, có nhiều hộp văn bản có nhiều nội dung văn bản chứa chữ số, giống như hộp văn bản có nhãn *total* phân bố ở nhiều vị trí khác nhau.

Hình 4.2: Ảnh hóa đơn trong bộ dữ liệu SROIE

4.2.3 Bộ dữ liệu FUNSD

Bộ dữ liệu bao gồm 199 ảnh thật và các file json tương ứng. Trong mỗi file json chứa thông tin của từng hộp văn bản gồm id, box, text, label, words. Bộ dữ liệu bao gồm 3 nhãn question, answer, header và 1 nhãn other. Bộ dữ liệu được chia thành 149 ảnh huấn luyện và 50 ảnh kiểm thử.

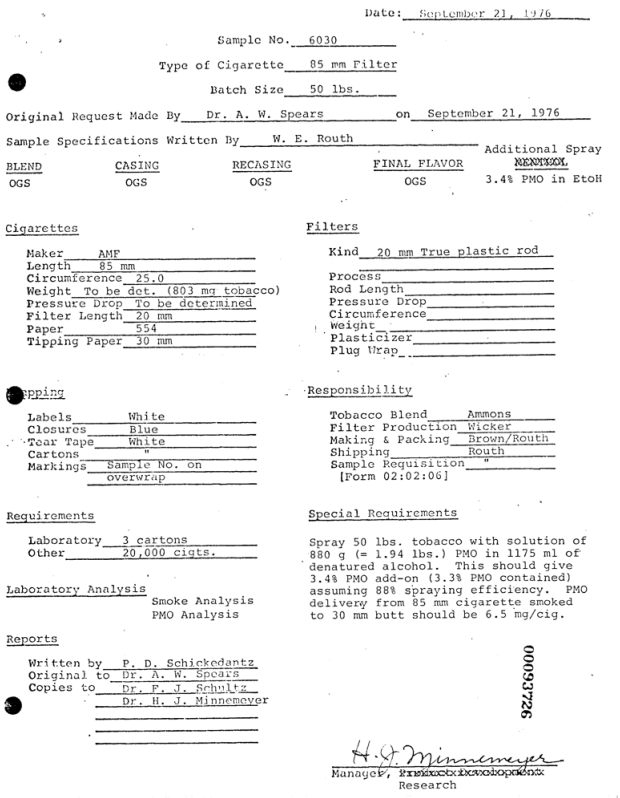
Hình 4.3 là một ví dụ trong bộ dữ liệu FUNSD.

Phân chia	Header	Question	Answer	Other	Total
Huấn luyện	441	3266	2802	902	7411
Kiểm thử	122	1077	821	312	2332

Bảng 4.2: Phân phối của các lớp trong bộ dữ liệu FUNSD

4.3 Kết quả thực nghiệm

So sánh với mô hình hiện hành: Trong phần này, em so sánh mô hình đề xuất với mô hình sử dụng mạng đồ thị trích xuất thông tin từ ảnh tài liệu PICK. Bảng 4.3 cho thấy kết quả so sánh. Trong mỗi cột, giá trị tốt nhất được in đậm. Mô hình đề xuất đạt được độ chính xác tốt nhất trên hai bộ dữ liệu (bộ dữ liệu đơn thuốc và



Date: September 21, 1976

Sample No. 6030

Type of Cigarette 85 mm Filter

Batch Size 50 lbs.

Original Request Made By Dr. A. W. Spears on September 21, 1976

Sample Specifications Written By W. E. Routh

BLEND	CASING	RECASING	FINAL FLAVOR	Additional Spray
OGS	OGS	OGS	OGS	3.4% PMO in EtOH

Cigarettes

Maker AMF
Length 85 mm
Circumference 25.0
Weight To be det. (803 mg tobacco)
Pressure Drop To be determined
Filter Length 20 mm
Paper 554
Tipping Paper 30 mm

Filters

Kind 20 mm True plastic rod
Process
Rod Length
Pressure Drop
Circumference
Weight
Plasticizer
Plug Wrap

Tipping

Labels White
Closures Blue
Tear Tape White
Cartons "
Markings Sample No. on overwrap

Responsibility

Tobacco Blend Ammons
Filter Production Wicker
Making & Packing Brown/Routh
Shipping Routh
Sample Requisition "
[Form 02:02:06]

Requirements

Laboratory 3 cartons
Other 20,000 cigts.

Special Requirements

Spray 50 lbs. tobacco with solution of 880 g (= 1.94 lbs.) PMO in 1175 ml of denatured alcohol. This should give 3.4% PMO add-on (3.3% PMO contained) assuming 88% spraying efficiency. PMO delivery from 85 mm cigarette smoked to 30 mm butt should be 6.5 mg/cig.

Laboratory Analysis

Smoke Analysis
PMO Analysis

Reports

Written by P. D. Schickedantz
Original to Dr. A. W. Spears
Copies to Dr. F. J. Schultz
Dr. H. J. Minnemeyer

00093726

H. J. Minnemeyer
Manager, Research

Hình 4.3: Ảnh tài liệu trong bộ dữ liệu FUNSD

Mô hình	Dữ liệu đơn thuốc	Dữ liệu SROIE	Dữ liệu FUNSD
Mô hình PICK	96	96.1	-
Mô hình đề xuất	**97**	**97**	**78**

Bảng 4.3: So sánh với mô hình PICK trên bộ dữ liệu đơn thuốc, bộ dữ liệu SROIE và bộ dữ liệu FUNSD với độ đo F1-score

SROIE). Cụ thể là, trên bộ đơn thuốc tiếng Việt ($+1\%$) và trên bộ dữ liệu SROIE ($+0.9\%$).

So sánh trên bộ dữ liệu đơn thuốc tiếng Việt: Trong phần này, em so sánh mô hình đề xuất với mô hình PICK. Bảng 4.4 cho thấy kết quả so sánh từng trường thông tin trong bộ dữ liệu. Trong mỗi hàng, giá trị tốt nhất được in đậm. Nhận thấy,những hộp văn bản mang nhiều thông tin về văn bản như *Chẩn đoán bệnh*, *Tên thuốc* được mô hình đề xuất học một cách khá hiệu quả.

Trường thuộc tính	Mô hình	Precision	Recall	F1-Score
Chuẩn đoán bệnh	PICK	**100**	70	82
	Mô hình đề xuất	94	**79**	**86**
Tên thuốc	PICK	97	82	89
	Mô hình đề xuất	**99**	**98**	**99**
Số lượng	PICK	**98**	**93**	**96**
	Mô hình đề xuất	95	90	92
Cách dùng	PICK	**100**	99	99
	Mô hình đề xuất	99	**99**	**99**
Ngày tháng	PICK	**99**	89	94
	Mô hình đề xuất	**99**	**98**	**98**

Bảng 4.4: So sánh kết quả trên từng trường thông tin giữa mô hình PICK và mô hình đề xuất với bộ dữ liệu đơn thuốc tiếng Việt

So sánh vai trò của các thành phần trong mô hình đề xuất: Mô hình đề xuất gồm 3 yếu tố quan trọng đó là mạng nơ-ron đồ thị, mô hình huấn luyện sẵn BERT và hàm mất mát tiêu điểm. Nếu thiếu mô-đun mạng nơ-ron đồ thị thì bảng 4.5 cho thấy kết quả trên 2 bộ dữ liệu đơn thuốc và SROIE giảm tương ứng 3% và 2%.

Mô hình	Dữ liệu đơn thuốc	Dữ liệu SROIE
Mô hình đề xuất (đầy đủ)	**97**	**97**
Thiếu mạng đồ thị	94	95

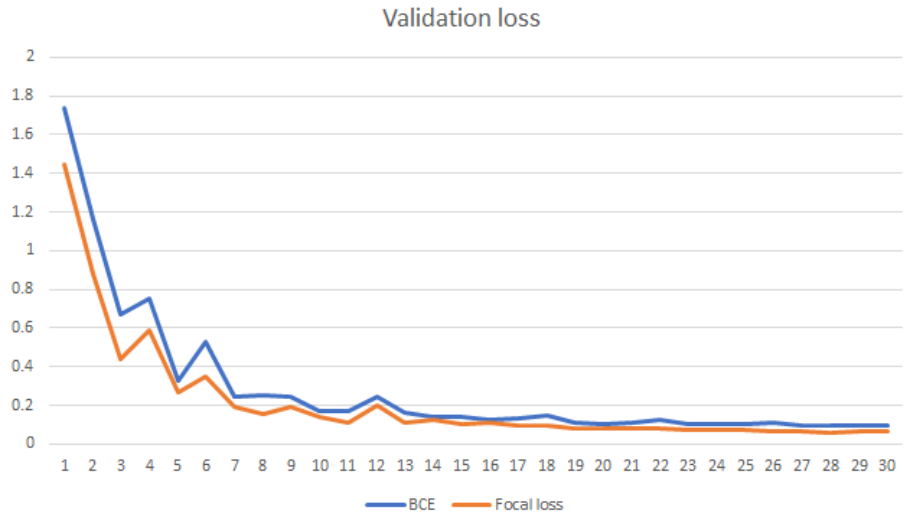
Bảng 4.5: Kết quả so sánh khi sử dụng mô hình đề xuất đầy đủ và mô hình đề xuất thiếu thành phần trên bộ dữ liệu đơn thuốc và bộ dữ liệu SROIE

So sánh độ hội tụ giữa hàm mất mát tiêu điểm và hàm mất mát balanced cross entropy khi huấn luyện mô hình trên bộ dữ liệu đơn thuốc tiếng Việt:

Hình 4.4 với đường màu cam biểu diễn giá trị của hàm balanced cross entropy và đường màu xanh biểu diễn giá trị của hàm mất mát tiêu điểm trong 30 epochs trên tập huấn luyện.

Hình 4.5 với đường màu cam biểu diễn giá trị của hàm balanced cross entropy và đường màu xanh biểu diễn giá trị của hàm mất mát tiêu điểm trong 30 epochs trên tập validation.

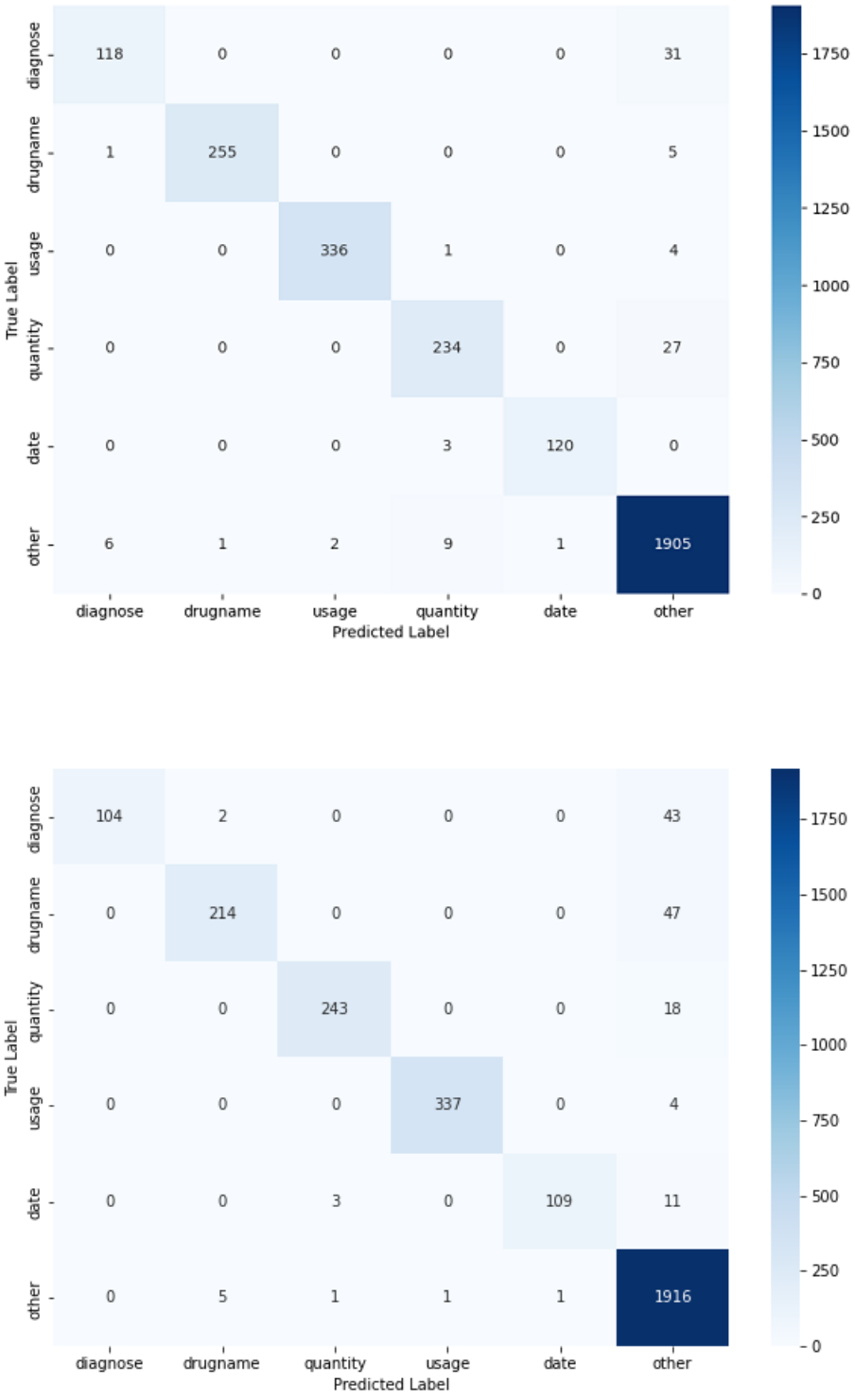
Hình 4.4: So sánh độ hội tụ giữa hai hàm mất mát trên tập huấn luyện



Hình 4.5: So sánh độ hội tụ giữa hai hàm mất mát trên tập validation

Trên hai đồ thị biểu diễn giá trị của hai hàm mất mát trên hai tập dữ liệu, có thể thấy hàm mất mát tiêu điểm luôn có giá trị nhỏ hơn so với hàm mất mát balanced cross entropy. Vì vậy, trên bộ dữ liệu đơn thuốc tiếng Việt, mô hình sử dụng hàm mất mát tiêu điểm với siêu tham số $\gamma = 1$ có tốc độ hội tụ nhanh hơn mô hình sử dụng hàm mất mát balanced cross entropy.

Trên bộ dữ liệu đơn thuốc tiếng Việt, hình 4.6.1 cho thấy kết quả confusion matrix của mô hình đề xuất trên tập kiểm thử. Hình 4.6.2 cho thấy kết quả confusion matrix của mô hình PICK trên tập kiểm thử. Em thấy kết quả của những trường *Chẩn đoán bệnh*, *Tên thuốc*, *Ngày tháng* của mô hình đề xuất khá hiệu quả.



Hình 4.6: Confusion Matrix của hai mô hình trên tập kiểm thử, hình trên là mô hình đề xuất, hình dưới là mô hình PICK

4.4 Kết chương

Trong chương này, em đã trình bày về phương pháp thí nghiệm, cách cài đặt đề xuất về xử lý tài liệu nghiêng, tham số đánh giá, bộ dữ liệu sử dụng cho bài toán trích xuất thông tin từ tài liệu (gồm bộ dữ liệu SROIE, bộ dữ liệu FUNSD, bộ dữ liệu tiếng Việt) và kết quả thực nghiệm đối với các bộ dữ liệu đó.

CHƯƠNG 5. TRANG WEB THỬ NGHIỆM

Trong chương này, em trình bày về trang web trích xuất thông tin từ đơn thuốc tiếng Việt. Trang web này cho phép người dùng tải lên một hình ảnh đơn thuốc và kết quả trả về là các trường thông tin liên quan đến đơn thuốc đó bao gồm chẩn đoán bệnh, tên thuốc, số lượng, cách dùng và ngày tháng.

5.1 Trang web trích xuất thông tin từ đơn thuốc tiếng Việt

Em viết một trang web giúp trích xuất thông tin từ đơn thuốc tiếng Việt. Trang web sử dụng một API có chức năng nhận ảnh đầu vào là một đơn thuốc tiếng Việt và trả về đầu ra là thông tin liên quan đến chẩn đoán bệnh và danh sách các thuốc trong đơn thuốc đó.

5.1.1 Giao diện chính của trang web

Hình 5.1 cho biết giao diện chính của trang web bao gồm 1 nút tải ảnh lên, kết quả của mô hình đề xuất trả về bao gồm thông tin về chẩn đoán bệnh, và danh sách các thuốc (tên thuốc, số lượng, cách dùng). Nhìn trên hình này, nút mũi tên màu trắng xanh cho phép người dùng tải ảnh trên máy tính cá nhân lên, sau đó, nếu ảnh đơn thuốc của người dùng đã thẳng và đẹp thì có thể sử dụng tùy chọn *None* hoặc nếu ảnh đơn thuốc của người dùng bị nghiêng thì có thể sử dụng tùy chọn *Convex Hull* và ấn nút **TRY DEMO** để hoàn thành. Sau đó, dữ liệu hình ảnh đơn thuốc sẽ được truyền về server để xử lý và trả về kết quả.

Và hình 5.2 cho biết giao diện trang web bao gồm 1 hình ảnh kết quả do API trả về. Trong hình ảnh kết quả do API trả về chứa những hộp văn bản của đơn thuốc tiếng Việt.

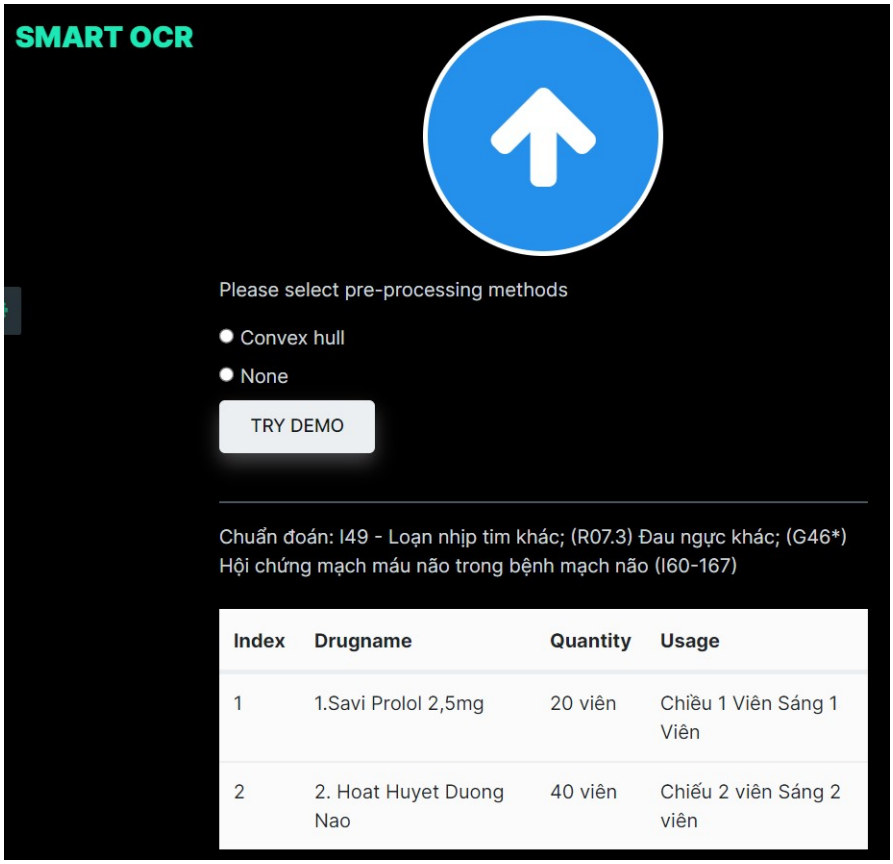
5.1.2 Công nghệ sử dụng

Các web framework đại diện cho một tập hợp các thư viện và mô-đun cho phép các nhà phát triển web viết mã, không cần lo lắng về các chi tiết cấp thấp như giao thức, quản lý luồng.

WSGI là kỹ thuật của giao diện chung giữa máy chủ web và trang web. Giao diện cổng máy chủ web WSGI được sử dụng như tiêu chuẩn trong việc phát triển trang web bằng ngôn ngữ lập trình Python.

Werkzeug là bộ công cụ WSGI, thực hiện yêu cầu, phản hồi và các tiện ích, cho phép xây dựng một web framework. Flask sử dụng Werkzeug làm một trong những cơ sở của framework.

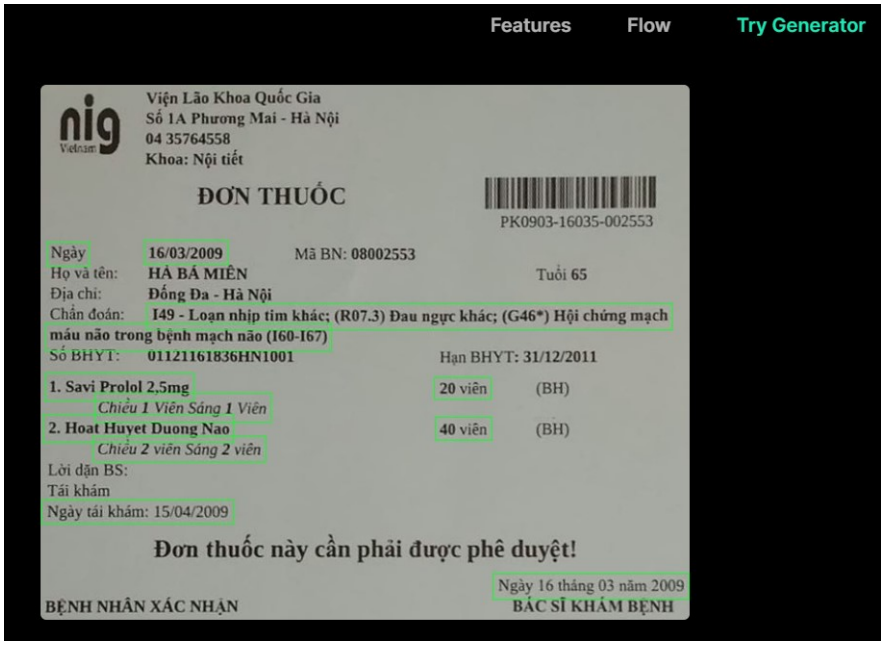
Jinja 2 là một template engine phổ biến cho ngôn ngữ lập trình Python. Hệ



Hình 5.1: Ảnh giao diện 1

thống web template kết hợp một giao diện mẫu với một nguồn dữ liệu cụ thể để hiển thị một trang web động.

Flask là một web framework được viết bằng ngôn ngữ lập trình Python. Flask dựa trên bộ công cụ Werkzeg WSGI và template engine (công cụ tách mã HTML) Jinja2. Flask là một micro web framework của ngôn ngữ lập trình Python không cần bất kỳ thư viện hoặc công cụ cụ thể. Flask cũng không có lớp trừu tượng hóa cơ sở dữ liệu, các thư viện xây dựng sẵn dựa trên bên thứ ba có sẵn và các hàm phổ biến hoặc các phương thức xác thực mẫu. Fask là một bộ lưu trữ giúp lập trình viên tạo ra các trang web dễ dàng hơn, có thể mở rộng, hiệu quả và có thể bảo trì bằng cách cung cấp code hoặc tiện ích mở rộng có thể sử dụng lại cho các nhiệm vụ phổ biến.



Hình 5.2: Ảnh giao diện 2

Đoạn code mẫu của chương trình Hello World được viết bằng ngôn ngữ Flask:

Listing 5.1: Đoạn code viết bằng Flask

```
from flask import Flask
app = Flask(__name__)

@app.route('/')
def hello_world():
    return 'Hello World!'

if __name__ == '__main__':
    app.run()
```

Khi chạy file *app.py*, chỉ cần chạy câu lệnh *flask run* hoặc *python app.py*, Flask mở cổng (default port) 5000, và nội dung trang web sẽ nằm ở đường dẫn **http://localhost:5000**.

5.1.3 Xây dựng trang web

API của hệ thống trích xuất thông tin từ đơn thuốc được thực hiện theo cách bước sau:

1. Phát hiện và nhận diện văn bản trong ảnh đơn thuốc
2. Chuyển đổi góc nhìn để thu được ảnh có độ nghiêng nhỏ nhờ bao lồi và phép chuyển đổi góc nhìn

3. Đưa qua mô hình mạng nơ-ron đề xuất thu được nhãn của hộp văn bản trong tập hợp nhãn được định nghĩa sẵn
4. Sử dụng tập luật dựa vào vị trí của các hộp văn bản để trả về chẩn đoán bệnh và danh sách các thuốc (tên thuốc, số lượng, cách dùng) kèm theo ảnh visulize kết quả

Trang web trích xuất thông tin từ đơn thuốc tiếng Việt được viết theo framework Flask. Em sử dụng Flask vì dễ dàng ghép phần mô hình đề xuất vào giao diện mẫu tạo ra một trang web có thể thực hành với chức năng cần thiết.

5.2 Kết chương

Trong chương này, em đã trình bày về chức năng, giao diện của trang web cũng như công nghệ được sử dụng khi xây dựng trang web này. Trong chương tiếp theo, em sẽ trình bày về kết luận và hướng phát triển.

CHƯƠNG 6. KẾT LUẬN

6.1 Kết luận

Trong phạm vi của đồ án tốt nghiệp, em đã thực hiện được ba đóng góp như sau:

1. Đề xuất phương pháp xử lý tài liệu nghiêng dựa trên thuật toán tìm bao lồi và hình chữ nhật có diện tích nhỏ nhất bao quanh bao lồi.
2. Đề xuất mô hình mạng nơ-ron đồ thị kết hợp với mô hình huấn luyện sẵn BERT, hàm mất mát tiêu điểm cho bài toán trích xuất thông tin trong tài liệu như hóa đơn, đơn thuốc
3. Trang web trích xuất thông tin từ đơn thuốc tiếng Việt

Đối với đóng góp thứ nhất, việc xử lý tài liệu nghiêng là cần thiết trong bài toán này. Bởi vì nó không chỉ giúp cải thiện độ chính xác trong khi phát hiện văn bản và nhận diện văn bản mà còn giúp tọa độ của những hộp văn bản trở nên dễ dàng học hơn trong mạng đồ thị. Thật vậy, cách thực hiện của phương pháp này bao gồm: tính toán những đỉnh thuộc bao lồi của những điểm góc của n hộp giới hạn được phát hiện bởi mô hình phát hiện văn bản CRAFT, sau đấy từ những điểm thuộc bao lồi thì tìm được hình chữ nhật có diện tích nhỏ nhất xung quanh bao lồi. Tiếp theo, sử dụng thuật toán chuyển đổi góc nhìn để biến đổi hình ảnh tài liệu nghiêng thành hình ảnh thẳng và tập trung vào những thông tin cần thiết (text).

Đối với đóng góp thứ hai, việc xây dựng mô hình trích xuất thông tin từ tài liệu cấu trúc như đơn thuốc và hóa đơn, bao gồm ba mô-đun chính: mô-đun tiền xử lý, mô-đun mạng nơ-ron đồ thị, mô-đun hậu xử lý. Trong mô-đun đầu tiên, một đồ thị được xây dựng bởi thông tin của những hộp văn bản. Nút mạng của đồ thị là thông tin văn bản trong hộp văn bản được cho qua mô hình huấn luyện sẵn RoBERTa để thu được vec-tơ ngữ nghĩa kết hợp với thông tin độ dài, thông tin về chữ số. Cạnh của đồ thị đó là việc kết nối nút hiện tại với 4 nút gần nhất nằm trên, dưới, trái, phải. Tiếp theo, là mô-đun mạng nơ-ron đồ thị, em sử dụng mô hình mạng GraphSAGE với nguyên lý tổng hợp thông tin của nút hiện tại với thông tin của các nút lân cận. Và sau đó những thông tin về đặc trưng đồ thị kết hợp với vec-tơ ngữ nghĩa do mô hình RoBERTa tạo ra trước đó, tạo ra vec-tơ tổng hợp, đưa qua mô-đun mạng MLP để phân loại ra những lớp tương ứng. Ngoài ra, em nhận thấy việc sử dụng hàm mất mát tiêu điểm có tốc độ hội tụ nhanh hơn so với hàm mất mát entropy chéo cân bằng. Bởi vì thành phần $(1 - p_t)^\gamma$ tỷ lệ nghịch với p_t, nên đối với những mẫu khó dự đoán có xác xuất p_t nhỏ, giúp tăng sự ảnh hưởng của những mẫu này vào hàm mất mát, ngược lại, những mẫu dễ dự đoán có xác xuất p_t lớn thì sự ảnh hưởng của

những mẫu này vào hàm mất mát là không đáng kể.

Đối với đóng góp thứ ba, việc trích xuất thông tin từ đơn thuốc tiếng Việt được thực hiện từ việc tải lên hình ảnh của đơn thuốc, chọn các tùy chọn tương ứng (*None* và *Convex hull*) và xác nhận gửi hình ảnh về server để xử lý, và kết quả cuối cùng là thông tin về chẩn đoán bệnh, danh sách các thuốc trong đơn và hình ảnh của đơn thuốc sau khi visualize các hộp văn bản chứa nội dung quan trọng.

Ngoài những thông tin liên quan đến văn bản, chữ số, độ dài của văn bản, cách xây dựng cạnh của đồ thị thì đặc trưng về hình ảnh đơn thuốc cũng là một đặc trưng đáng để xem xét, thêm vào mô hình. Tuy nhiên, trong phạm vị đồ án này, đặc trưng về hình ảnh của đơn thuốc đã chưa được thêm vào.

6.2 Hướng phát triển trong tương lai

Từ những hạn chế nêu trên, trong tương lai, em cũng sẽ thử nghiệm và đánh giá độ quan trọng của đặc trưng về mặt hình ảnh để thêm vào mô hình đề xuất nếu được.

CÁC CÔNG BỐ KHOA HỌC

1. Tran Bao Hieu, Hoang Duc Viet, **Nguyen Manh Hiep**, Pham Ngoc Bao Anh, Nguyen Duc Anh, Hoang Gia Bao, Hai-Phong Bui, Thanh Hung Nguyen, Phi Le Nguyen, Thi-Lan Le, "MC-OCR Challenge 2021: A Multi-modal Approach for Mobile-Captured Vietnamese Receipts Recognition", The 15th IEEE-RIVF International Conference on Computing and Communication Technologies, RIVF 2021 (Accepted).

TÀI LIỆU THAM KHẢO

[1] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "PICK: Processing key information extraction from documents using improved graph learning-convolutional networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2020.

[2] J. Ha, R. Haralick, and I. Phillips, "Recursive x-y cut using bounding boxes of connected components," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 2, 1995, 952–955 vol.2. DOI: 10.1109/ICDAR.1995.602059.

[3] F. Lebourgeois, Z. Bublinski, and H. Emptoz, "A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents," in *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, 1992, pp. 272–276. DOI: 10.1109/ICPR.1992.201771.

[4] A. Simon, J.-C. Pret, and A. Johnson, "A fast algorithm for bottom-up document layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 273–277, 1997. DOI: 10.1109/34.584106.

[5] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial naïve bayes classifier to text classification," in *Advanced multimedia and ubiquitous engineering*, Springer, 2017, pp. 347–352.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[8] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019.

[9] Y. He, C. Chen, J. Zhang, *et al.*, "Visual semantics allow for textual reasoning better in scene text recognition," *arXiv preprint arXiv:2112.12916*, 2021.

[10] T. Zheng, Z. Chen, S. Fang, H. Xie, and Y.-G. Jiang, "Cdistnet: Perceiving multi-domain character distance for robust text recognition," *ArXiv*, vol. abs/2111.11011, 2021.

[11] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *International Conference on Document Analysis and Recognition*, Springer, 2021, pp. 319–334.

[12] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.

[13] J. Ye, Z. Chen, J. Liu, and B. Du, "Textfusenet: Scene text detection with richer fused features," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 516–522.

[14] L. Xing, Z. Tian, W. Huang, and M. R. Scott, "Convolutional character networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[15] W. Hwang, J. Yim, S. Park, S. Yang, and M. Seo, "Spatial dependency parsing for semi-structured document information extraction," *arXiv preprint arXiv:2005.00642*, 2020.

[16] X. Liu, F. Gao, Q. Zhang, and H. Zhao, "Graph convolution for multimodal information extraction from visually rich documents," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 32–39. DOI: 10.18653/v1/N19-2005. [Online]. Available: https://aclanthology.org/N19-2005.

[17] C. Lockard, P. Shiralkar, X. L. Dong, and H. Hajishirzi, "ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 8105–8117. DOI: 10.18653/v1/2020.acl-main.721. [Online]. Available: https://aclanthology.org/2020.acl-main.721.

[18] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork, "Representation learning for information extraction from form-like documents," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. [Online]. Available: https://aclanthology.org/2020.acl-main.580.

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[20] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354*, 2016.

[21] A. R. Katti, C. Reisswig, C. Guder, *et al.*, "Chargrid: Towards understanding 2d documents," *arXiv preprint arXiv:1809.08799*, 2018.

[22] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[23] K. Javed and F. Shafait, "Real-time document localization in natural images by recursive application of a cnn," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, IEEE, vol. 1, 2017, pp. 105–110.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[25] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 2017, pp. 1597–1600.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[27] Q. Jodelet, X. Liu, and T. Murata, "Balanced softmax cross-entropy for incremental learning," in *International Conference on Artificial Neural Networks*, Springer, 2021, pp. 385–396.