

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH



ĐỒ ÁN CUỐI KÌ

TRÍCH XUẤT THÔNG TIN CHỨNG CHỈ TOEIC

Sinh viên thực hiện:

Nguyễn Phú Trung - 21521592

Trương Khánh Long – 21521750

Hoàng Hải Anh - 21521819

Giảng viên:

PGS.TS. Lê Đình Duy

ThS. Phạm Nguyễn Trường An

Thành phố Hồ Chí Minh, 18 tháng 01 năm 2024

LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn sâu sắc đến quý Thầy Cô trường Đại học Công nghệ thông tin, đặc biệt là quý thầy cô thuộc Khoa Khoa học máy tính, đã tạo cơ hội cho chúng em tiếp thu những kiến thức quý báu trong khoảng thời gian học tập môn Máy học. Những kiến thức được quý Thầy cô truyền đạt trong suốt thời gian qua là một bước đệm quan trọng để giúp chúng em hoàn thành đồ án môn học tốt hơn.

Chúng em cũng xin gửi lời cảm ơn sâu sắc tới PGS.TS Lê Đình Duy và ThS. Phạm Nguyễn Trường An - giảng viên đứng lớp môn Máy học đã tận tình giúp đỡ, hướng dẫn nhóm trong suốt quá trình thực hiện đồ án môn học. Qua những lời chỉ dẫn nhiệt tình của hai thầy, chúng em đã tiếp thu được nhiều kiến thức bổ ích cũng như trau dồi những kỹ năng cần thiết để hoàn thành đồ án một cách chẵn chu nhất.

Không chỉ vận dụng những kiến thức được giảng dạy trên lớp, nhóm cũng đã tự tìm hiểu, trau dồi thêm các kiến thức cần thiết thông qua tài liệu trên internet, giáo trình để có thể hoàn thiện đầy đủ yêu cầu đồ án môn học đưa ra. Tuy nhiên, trong quá trình thực hiện đồ án, nhóm có thể vẫn còn một vài sai sót không tránh khỏi. Do đó, nhóm rất mong nhận được những ý kiến đóng góp chân thành từ các thầy để có thể rút ra những bài học quý giá và biến nó trở thành hành trang, kinh nghiệm quý báu để nhóm thực hiện các đề tài tiếp theo trong tương lai.

Một lần nữa xin gửi đến hai thầy lời cảm ơn chân thành và tốt đẹp nhất.

MỤC LỤC

0. Giải trình sau khi vấn đáp.....	4
1. Tổng quan về đề án	5
1.1. Lý do chọn đề tài.....	5
1.2. Mô tả bài toán và phạm vi đề tài.....	5
1.3. Các công cụ hỗ trợ	6
1.3.1. Label Studio.....	6
1.3.2. Google Colaboratory.....	7
1.3.3. Github	8
2. Các nghiên cứu liên quan	8
2.1. Trích xuất thông tin từ chứng minh thư: [1] [2]	8
2.2. Nhận diện biển số xe ở Việt Nam: [3]	9
3. Mô tả bộ dữ liệu.....	10
4. Mô tả đặc trưng.....	11
4.1. Data Pipeline Processing.....	11
4.2. Tiền xử lý dữ liệu.....	11
4.3. Feature Engineering	12
5. Mô tả thuật toán máy học	13
6. Đánh giá, kết luận và hướng phát triển	14
6.1. Đánh giá	14
6.2. Kết luận.....	14
6.3. Hướng phát triển đề tài	15
TÀI LIỆU THAM KHẢO	16

TÓM TẮT

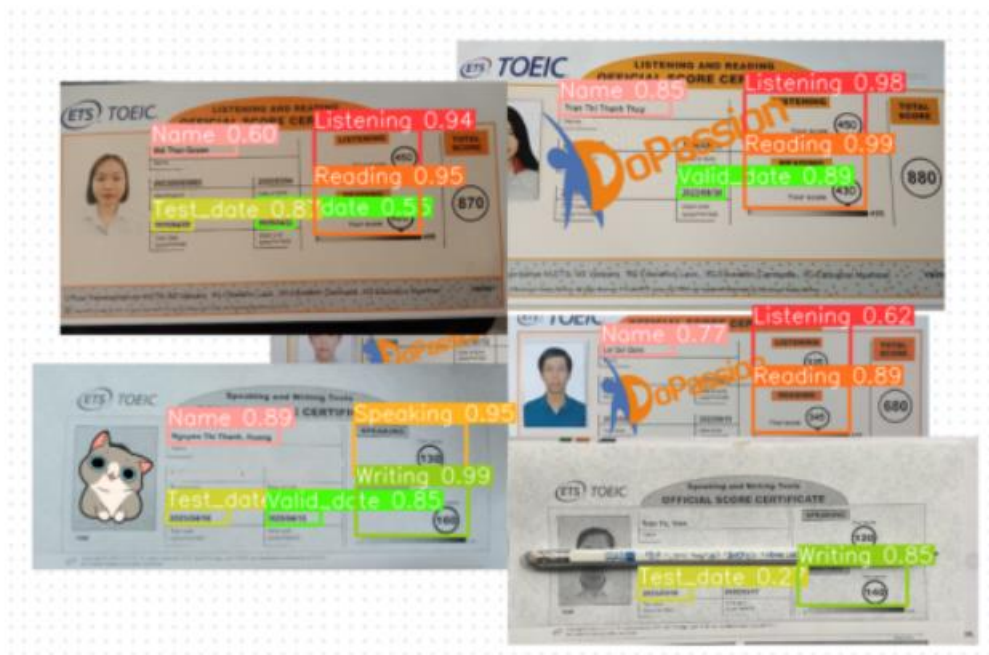
Đồ án Trích xuất thông tin trên chứng chỉ Toeic thực hiện việc trích xuất dữ liệu trên hình ảnh chứng chỉ Toeic. Bộ dữ liệu được nhóm thu thập thông qua liên hệ với trung tâm anh ngữ bao gồm hơn 150 ảnh chụp các chứng chỉ nghe-đọc, nói-viết. Nhóm đã hoàn thành việc trích xuất thông qua các bước tiền xử lý dữ liệu, gán nhãn dữ liệu sau đó sử dụng machine learning để chọn các vùng ảnh chứa thông tin từ đó có thể trích xuất dữ liệu thích hợp. Cuối cùng, nhóm đã thành công trong việc thực hiện các bước xử lý, trích xuất trên và hoàn tất việc đánh giá mô hình với các chỉ số tích cực. Kết quả cho thấy mô hình nhận diện với tỉ lệ chính xác cao.

Link github repository: <https://github.com/boo283/CS114.O11>

0. Giải trình sau khi vấn đáp

Thông qua buổi vấn đáp, nhóm đã trình bày các nghiên cứu liên quan đến đề tài của nhóm chi tiết trong phần 2 của bài báo cáo này. Bên cạnh đó, nhóm có nhận được yêu cầu của thầy trong việc phát triển đề án lên bằng cách nhận diện nhiều chứng chỉ toeic trong cùng một tấm ảnh.

Hiện tại, nhóm đã thành công trong việc nhận diện nhiều chứng chỉ Toeic trong một ảnh như hình:



Như hình ảnh trên, nhóm đã có thể trích xuất được các chứng chỉ riêng biệt, sau đó thực hiện lấy các thông tin trên từng chứng chỉ. Tuy nhiên, do chất lượng ảnh ban đầu kém nên một số thông tin trên chứng chỉ vẫn chưa thể trích xuất hoặc trích xuất sai, với các ảnh bị xoay nhóm vẫn chưa trích xuất được.

1. Tổng quan về đề án

1.1. Lý do chọn đề tài

Trong bối cảnh công nghệ số và xã hội ngày càng phát triển cùng với xu hướng hội nhập quốc tế hiện nay, ngoại ngữ đang dần khẳng định vai trò, vị thế của mình trong nhiều lĩnh vực của đời sống. Việc sở hữu cho riêng một từ một đến hai ngôn ngữ khác tiếng mẹ đẻ sẽ là một ưu thế vượt trội trong việc tìm kiếm cơ hội để thăng tiến, thành công. Chúng ta có thể đọc hiểu được tài liệu bằng tiếng nước ngoài, có thể giao tiếp với bạn bè quốc tế, làm việc trong các công ty đa quốc gia. Điều đó giúp cải thiện chất lượng cuộc sống của chính bản thân mình.

Để xác định được trình độ ngoại ngữ của một người, chúng ta có thể giao tiếp trực tiếp với họ hoặc cách nhanh nhất và phổ biến nhất chính là sử dụng kết quả các bài test của các tổ chức ngoại ngữ uy tín được in trên các loại chứng chỉ khác nhau. Điển hình như trong kỳ thi THPT Quốc Gia những năm gần đây, các loại chứng chỉ như TOEIC và IELTS có thể được sử dụng để thay thế toàn bộ bài thi môn tiếng Anh với số điểm được quy đổi theo quy định. Việc này giúp giảm bớt áp lực cho thí sinh cũng như đơn vị tổ chức kỳ thi. Tuy nhiên, với một số lượng lớn thí sinh như vậy, việc trích xuất thông tin và xác thực tính minh bạch các chứng chỉ bằng thủ công là một vấn đề cần được cân nhắc cả về thời gian, nhân lực và chi phí phát sinh. Thấu hiểu được những khó khăn trên, cùng với việc ứng dụng các tính năng mạnh mẽ của các mô hình machine learning hiện nay kết hợp với những kiến thức được học tại môn máy học, nhóm chúng em quyết định cùng nhau xây dựng hệ thống tự động trích xuất thông tin trên chứng chỉ TOEIC để tiết kiệm thời gian. Từ việc tiếp thu các kiến thức được truyền tải trên lớp và trong tài liệu kết hợp với việc trực tiếp thực hành thông qua các notebook được thầy cung cấp, chúng em hi vọng sẽ tạo ra được một hệ thống hoạt động hiệu quả với độ chính xác khi trích xuất cao nhất có thể.

1.2. Mô tả bài toán và phạm vi đề tài

Đề án nghiên cứu tập trung vào việc trích xuất thông tin từ hình ảnh chứng chỉ TOEIC được người dùng chụp và tải lên trang web. Đầu vào của bài toán là hình ảnh chụp chứng chỉ mặt chính diện, rõ nét, thể hiện đầy đủ các đường viền của chứng

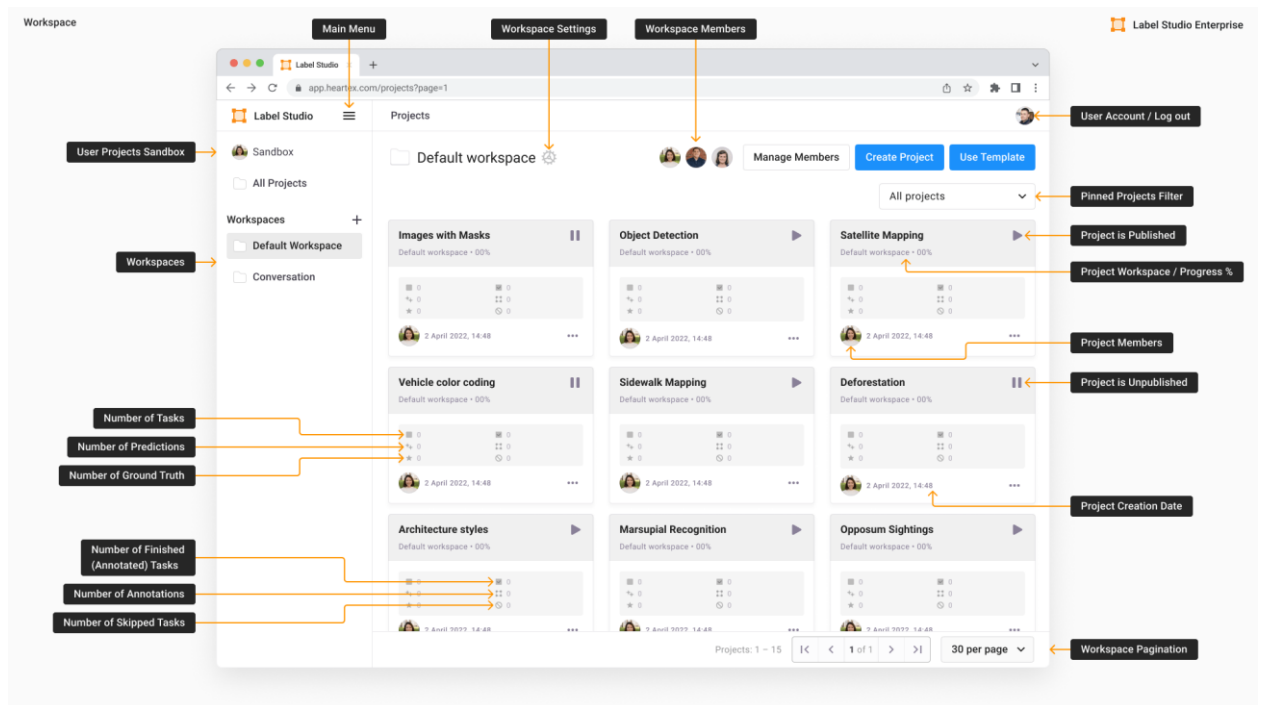
chỉ. Đầu ra của bài toán là các thông tin của thí sinh trên chứng chỉ (tên đầy đủ (không dấu), điểm phần thi Listening, điểm phần thi Reading, điểm phần thi Speaking, điểm phần thi Writing, Ngày thi, Ngày hết hạn chứng chỉ). Thông tin nào không trích xuất được do hình ảnh quá mờ hoặc bị che khuất sẽ được để trống hoặc dùng mô hình để dự đoán thông tin bị thiếu đó và lưu vào cơ sở dữ liệu.

Phạm vi đề tài chỉ gói gọn trong hai loại chứng chỉ TOEIC Listening – Reading và TOEIC Speaking – Writing. Do thời gian gấp rút cũng như nhóm mới tiếp cận đến các công nghệ máy học trong thời gian ngắn nên chỉ có thể trích xuất được trên hai loại chứng chỉ này. Đối tượng sử dụng hệ thống của chúng em là một số cơ quan quản lý giáo dục, các công ty, cơ sở giáo dục cũng như những nơi có sử dụng hai loại chứng chỉ trên để đánh giá năng lực ngoại ngữ đầu vào của từng cá nhân.

1.3. Các công cụ hỗ trợ

1.3.1. Label Studio

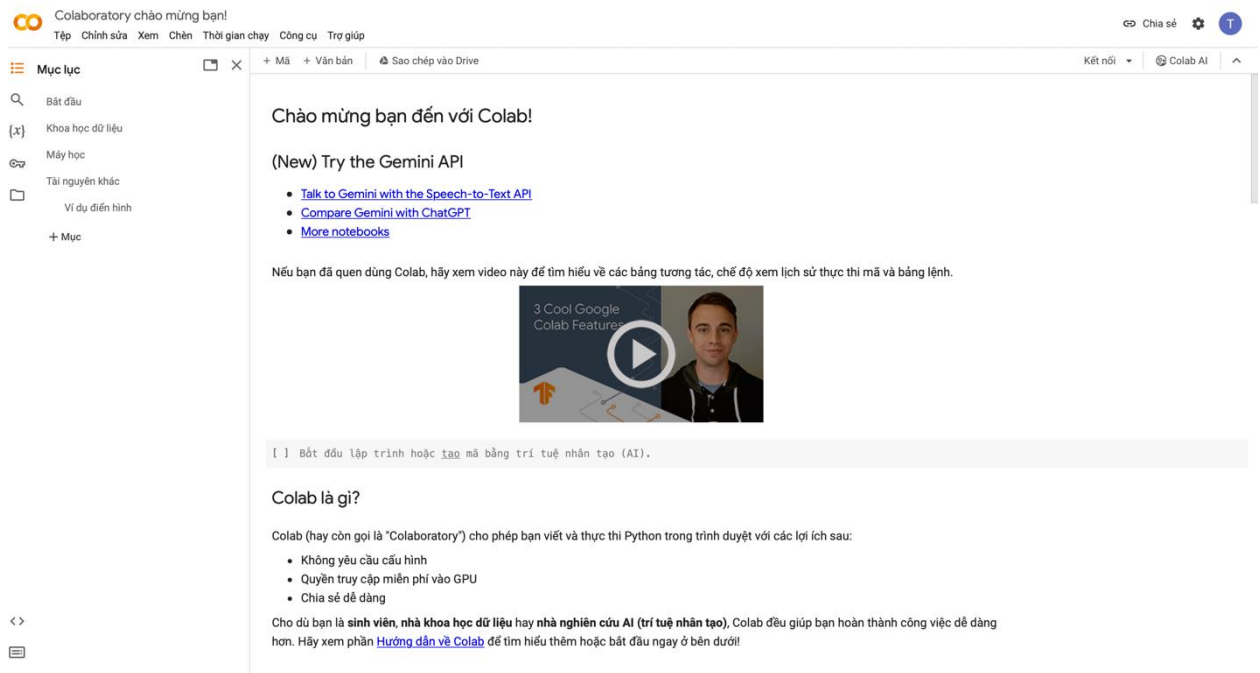
Công cụ hỗ trợ đánh nhãn dữ liệu mã nguồn mở, cho phép đánh nhãn nhiều định dạng dữ liệu như hình ảnh, âm thanh, văn bản, video và đây thời gian thông qua một giao diện người dùng thân thiện rõ ràng, sau đó xuất dữ liệu sang các định dạng mô hình khác nhau. Nó có thể được sử dụng để chuẩn bị nguồn dữ liệu thô hoặc cải thiện dữ liệu huấn luyện hiện tại để có được mô hình máy học chính xác.



Hình 1: Giao diện quản lý dự án của Label Studio

1.3.2. Google Colaboratory

Phần mềm giúp lưu trữ mã nguồn và thực thi chương trình trực tiếp trên trình duyệt thông qua máy ảo GPU, TPU được Google cung cấp miễn phí. Cho phép kết nối trực tiếp với Google Drive tạo thuận lợi trong việc trích xuất, chỉnh sửa, lưu trữ dữ liệu trực tiếp sau mỗi thao tác. Các thành viên có thể làm việc cùng nhau trên cùng một notebook tạo thuận lợi trong việc trao đổi thông tin.



Hình 2: Giao diện khởi đầu của Google Collab

1.3.3. Github

Nền tảng trực tuyến giúp lưu trữ toàn bộ source code cũng như dữ liệu của nhóm sau khi đã hoàn thành mọi thao tác chỉnh sửa. Phần mềm giúp người dùng bên ngoài có thể tham khảo cũng như nhận xét và đánh giá về toàn bộ dự án của nhóm. Link github chung của nhóm: <https://github.com/boo283/CS114.O11>

2. Các nghiên cứu liên quan

2.1. Trích xuất thông tin từ chứng minh thư: [1] [2]

- Phương pháp thực hiện:

+ Detect 4 góc của chứng minh thư: xem bốn góc của chứng minh thư như một object cần detect, sau đó xoay thẳng tọa độ của bốn góc này (bottom left, bottom right, top left, top right).

+ Xoay chứng minh thư: Tính tọa độ trung tâm của mỗi bounded box hay tọa độ của bốn góc để tìm đầy đủ tọa độ 4 góc (trong trường hợp bị thiếu). Sau đó, xoay thẳng ảnh lại dựa vào kích thước thực tế.

+ Detect vùng chữ có trong chứng minh thư: Sử dụng các mô hình có sẵn trong thư viện Tensorflow API để lấy ra các chữ có trong ảnh được crop ở bước trên.

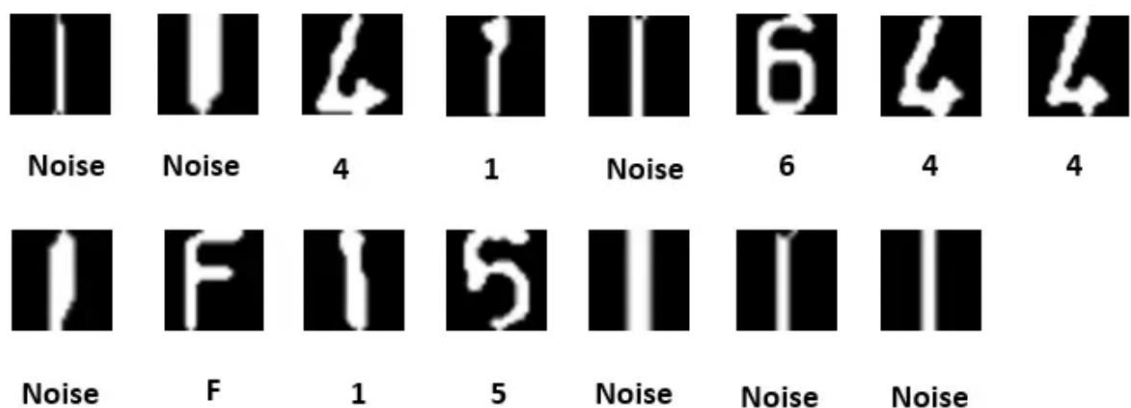
+ OCR: Với các box thu được từ bước trên, ta crop và đưa qua mô hình OCR để đọc.

2.2. Nhận diện biển số xe ở Việt Nam: [3]

- Phương pháp thực hiện:

+ Xác định vị trí chứa biển số xe trong ảnh chụp: Sử dụng YoloV3 để crop vùng chứa biển số xe ra khỏi ảnh gốc.

+ Tách từng kí tự trên biển số xe: Chuyển màu ảnh từ BGR sang HSV, sử dụng adaptive threshold để làm nổi bật những phần muốn lấy. Sử dụng thuật toán Connected components analysis – kết nối tất cả pixel có cùng giá trị thành một khối và gán nhãn. Nhờ đó, tất cả các pixel của từng kí tự sẽ được kết nối và tách ra khỏi biển số xe. Tuy nhiên trong đó cũng có vài kí tự nhiễu như khung biển số xe, dấu gạch ngang, và background. Vì vậy, cần loại bỏ các label của background trước sau đó thiết lập các giá trị ngưỡng để loại bỏ nhiễu.



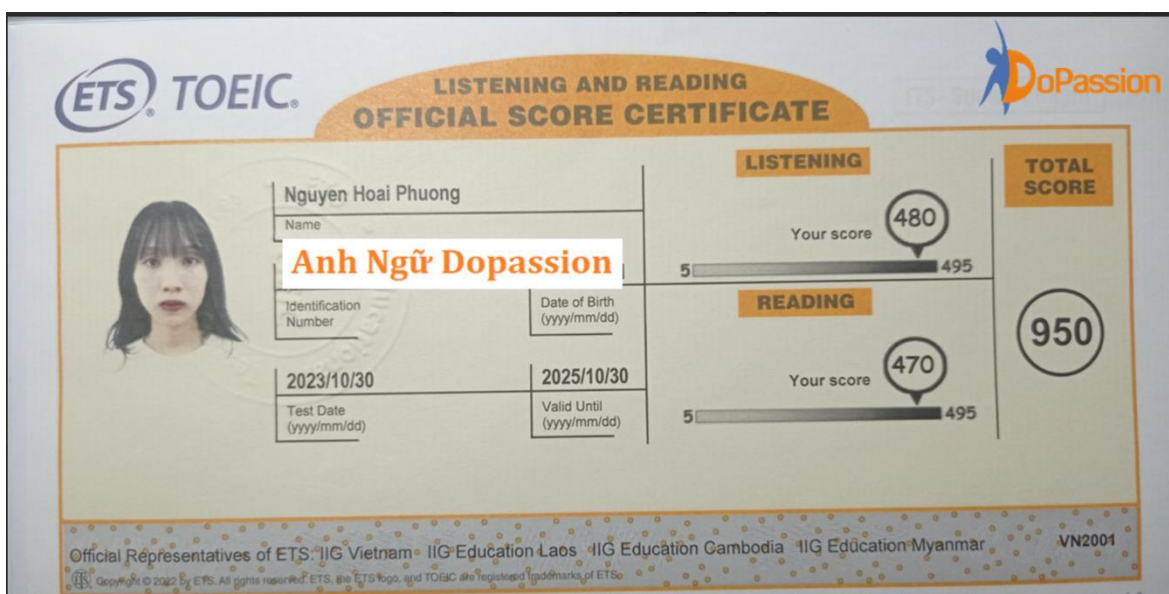
Hình 3: Kết quả thực hiện lấy kí tự trên biển số xe

+ Phân loại các kí tự thu được (sử dụng CNN): số lượng class trong bài toán phân loại là 32 bao gồm: 10 ký tự số từ (0-9) và 21 ký tự chữ cái (trừ I, J, O, Q) và một class Background dùng để lưu các giá trị nhiễu.

+ Xác định biển số xe gồm một hay hai dòng: ta sử dụng tọa độ (x,y) cho mỗi ký tự. Nếu tọa độ y của mỗi ký tự nhỏ hơn ngưỡng đề ra thì biển số xe gồm một dòng. Ngược lại thì biển số có hai dòng. Ta sử dụng giá trị x để xác định thứ tự của các ký tự từ trái qua phải và từ trên xuống dưới.

3. Mô tả bộ dữ liệu

Nhóm chủ động liên hệ trung tâm Anh ngữ Dopassion để thu thập các chứng chỉ. Hình ảnh chứng chỉ có được sau khi các học viên đi thi ở IIG hoặc các tổ chức khác sau đó đăng tải lên website của trung tâm hoặc nhắn tin trực tiếp đến cho giáo viên thông qua các ứng dụng trực tuyến như Messenger, Zalo hoặc Gmail. Số lượng chứng chỉ thu thập được đến thời điểm hiện tại là 161 bao gồm 84 chứng chỉ Listening – Reading và 77 chứng chỉ Speaking – Writing.

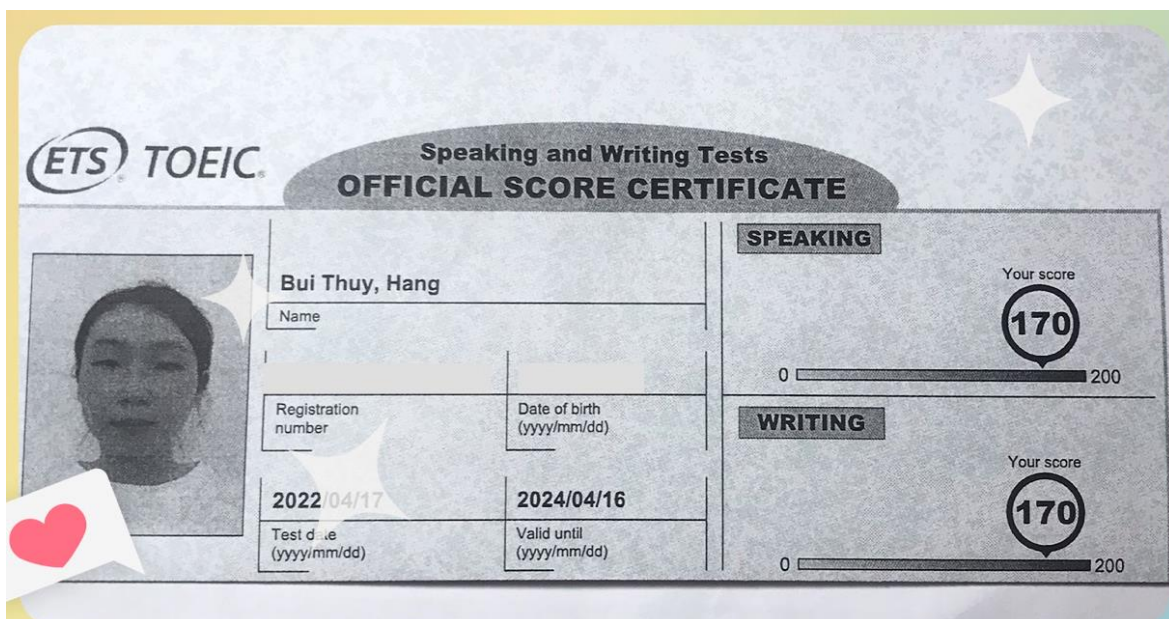


This is a TOEIC Listening and Reading Official Score Certificate for Nguyen Hoai Phuong. The certificate includes a photo of the test taker, a watermark of the Dopassion logo, and a table of scores. The Listening score is 480 and the Reading score is 470, both on a scale of 5 to 495. The total score is 950. The test date is 2023/10/30 and the valid until date is 2025/10/30. The certificate is issued by ETS TOEIC and is an official representative of ETS in Vietnam, Laos, Cambodia, and Myanmar.

TEST TAKER INFORMATION		LISTENING	READING	TOTAL SCORE
Name	Nguyen Hoai Phuong	Your score	Your score	950
Identification Number		480	470	
Test Date (yyyy/mm/dd)	2023/10/30	5 495		
Valid Until (yyyy/mm/dd)	2025/10/30	5 495		

Official Representatives of ETS: IIG Vietnam, IIG Education Laos, IIG Education Cambodia, IIG Education Myanmar. VN2001

Hình 4: Chứng chỉ TOEIC Listening - Reading



This is a TOEIC Speaking and Writing Tests Official Score Certificate for Bui Thuy, Hang. The certificate includes a photo of the test taker, a watermark of the Dopassion logo, and a table of scores. The Speaking score is 170 and the Writing score is 170, both on a scale of 0 to 200. The test date is 2022/04/17 and the valid until date is 2024/04/16. The certificate is issued by ETS TOEIC.

TEST TAKER INFORMATION		SPEAKING	WRITING
Name	Bui Thuy, Hang	Your score	Your score
Registration number		170	170
Test date (yyyy/mm/dd)	2022/04/17	0 200	
Valid until (yyyy/mm/dd)	2024/04/16	0 200	

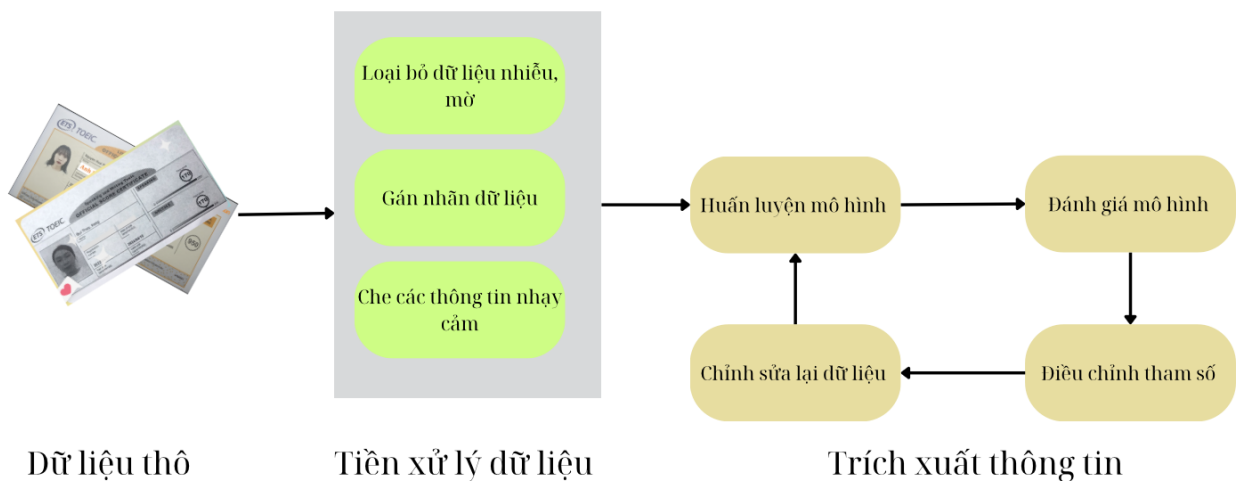
Hình 5: Chứng chỉ TOEIC Speaking – Writing

Ngoài ra, nhóm còn thực hiện tạo một bộ dữ liệu với tập train: 70 ảnh, với sự kết hợp từ hình ảnh các chứng chỉ như sau:



4. Mô tả đặc trưng

4.1. Data Pipeline Processing



Hình 6: Sơ đồ quy trình thực hiện

4.2. Tiền xử lý dữ liệu

Toàn bộ dữ liệu và quy trình được lưu trữ trong một folder trên Google Drive. Do hình ảnh cần đảm bảo được tính bảo mật thông tin cá nhân của từng học viên nên sẽ có một vài vị trí bị che khuất. Không chỉ có vậy, hình ảnh được thu thập

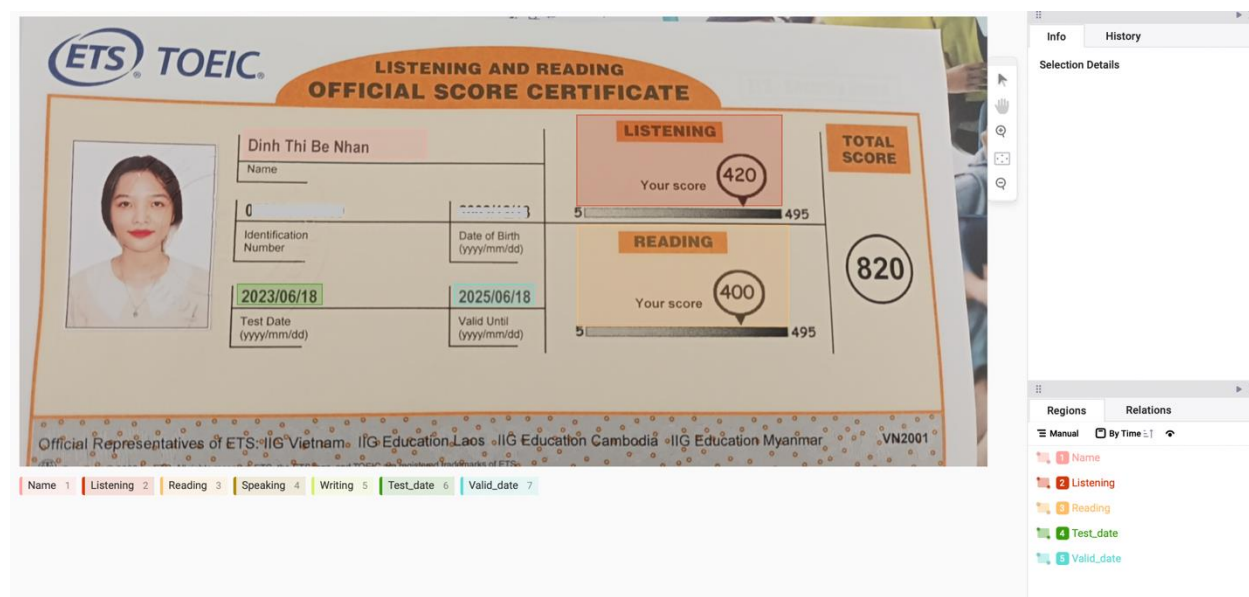
thông qua nền tảng internet nên có thể bị mờ, bị bể, bị nghiêng, Do đó, nhóm quyết định loại bỏ một số trong các hình ảnh chứng chỉ đó.

4.3. Feature Engineering

Nhóm sử dụng công cụ Label Studio để tiến hành gán nhãn cho từng ảnh. Quá trình label sẽ giúp xác định các bounded box tại vị trí chứa thông tin cần trích xuất tương ứng. Các đặc trưng thu được sẽ được lưu trữ lại và phân tích trong phần mô tả đặc trưng. Sau khi toàn bộ hình ảnh đã được gán nhãn, nhóm sẽ xuất kết quả dưới định dạng YOLO. Kết quả thu được bao gồm một thư mục chứa 2 thư mục, trong đó chứa các tệp tin khác nhau và khác ở phần định dạng:

- + Thư mục chứa các tệp tin txt bao gồm các thông tin về ID và tọa độ của box tương ứng với định dạng $x\ y\ x\ y$ với giá trị dạng thập phân từ 0 đến 1. Trong đó $x\ y$ đầu tiên tương ứng với tọa độ góc trên bên trái và $x\ y$ phía sau tương ứng với tọa độ góc dưới bên phải của bounded box.











- + Thư mục chứa các hình ảnh đã được gán nhãn với các bounded box.



Hình 7: Kết quả gán nhãn cho một mẫu dữ liệu sử dụng Label Studio

Shared with me > ... > Test_set > L&R ▾

Type ▾ People ▾ Modified ▾

Name ▾	Owner	Last modified ▾	File size
 labels	 me	Jan 17, 2024 me	—
 images	 me	Jan 17, 2024 me	—
 notes.json 	 me	Jan 17, 2024 me	484 bytes
 classes.txt 	 me	Jan 17, 2024 me	61 bytes

Hình 8: Kết quả sau khi xuất file YOLO

Sau đó, nhóm tạo thêm tập tin mydata.yaml để lưu trữ các thông tin bao gồm: đường dẫn đến thư mục chứa dataset (path), đường dẫn đến thư mục chứa train/val/test dataset (train/val/test) và thông tin về ID number và tên tương ứng với định dạng “ID: name” (names).

Cuối cùng, nhóm xử lý các hình ảnh bên trong bounded box giúp nâng cao khả năng nhận diện và trích xuất thông tin của mô hình bao gồm các công việc như xác định vùng chứa dữ liệu cần lấy trong bounding box, chuyển đổi vùng quan trọng từ dạng mảng các pixel sang ảnh dạng grayscale bằng hàm convert(“L”), áp dụng các bộ lọc giảm nhiễu, làm rõ ảnh, tăng cường độ tương phản để làm nổi bật phần văn bản trong ảnh.

5. Mô tả thuật toán máy học

Đầu tiên, nhóm phân chia bộ dữ liệu thành ba tập: tập validation (24 samples), tập training (19 samples) và tập train (137 sample) ở cả hai thư mục Images và Labels sau khi hoàn thành quá trình tiền xử lý dữ liệu.

Tiếp theo, nhóm sử dụng Model YOLOv8 để phát hiện đối tượng dựa trên convolutional neural network (CNN) với đầu vào là một hình ảnh. Đầu tiên, chia hình ảnh thành một lưới các ô. Đối với mỗi ô, mô hình sẽ tìm kiếm loại đối tượng có thể có trong ô đó, tọa độ của bounding box bao quanh đối tượng đó và độ tự tin cho bounding box. Các bounding box có độ tự tin cao hơn một ngưỡng nhất định

có khả năng cao chứa đối tượng thật sự. Độ tự tin được tính dựa trên kết quả của hàm logistic regression với hàm kích hoạt là hàm sigmoid.

Cuối cùng, nhóm sử dụng hai mô hình Pytesseract và EasyOCR để tiến hành đọc các ký tự xuất hiện trong hình ảnh đưa vào. Đầu tiên, Pytesseract nhận diện văn bản trong ảnh bằng cách chuyển đổi toàn bộ hình ảnh thành ảnh chỉ gồm hai màu trắng và đen để tăng cường độ tương phản. Sau đó, sử dụng thư viện Tesseract để phân tích các thành phần của văn bản đó và nhận diện ký tự trong ảnh. Kết quả của mô hình là thông tin nằm trên ảnh kèm theo tọa độ của vùng chứa thông tin đó. Bên cạnh đó, EasyOCR chia ảnh đầu vào thành nhiều vùng sau đó đưa chúng vào mô hình máy học để phân tích ảnh và nhận diện ký tự trong vùng văn bản.

6. Đánh giá, kết luận và hướng phát triển

6.1. Đánh giá

Hiệu suất được đánh giá dựa trên độ chính xác của thông tin được trích xuất từ các mô hình. Ở đây, nhóm tính độ chính xác dựa trên từng bounding box (từng thành phần của thông tin) trên tổng số bounding box có được khi tiến hành gán nhãn. Tỷ lệ đọc chính xác của một hình ảnh chứng chỉ được xác định khi toàn bộ thông tin trên hình ảnh chứng chỉ khớp hoàn toàn với thông tin mà mô hình đã phát hiện và gửi trả kết quả ngược trở lại. Nhóm tiến hành tính toán độ chính xác của từng bounding box và thu được kết quả. Dưới đây là kết quả thu được ở dạng nhận diện chứng chỉ trên ảnh riêng biệt (1 chứng chỉ/ ảnh):

	Name	Listening	Reading	Speaking	Writing	Test_Date	Valid_Date
Accuracy_PyTesseract	0.842105	0.947368	0.842105	0.736842	0.789474	0.210526	0.947368
Accuracy_Easy_OCR	1.000000	0.947368	1.000000	1.000000	0.894737	0.473684	0.894737

Hình 9: Kết quả thu được của mô hình nhận diện thông tin

Hiệu suất trong việc nhận diện nhiều chứng chỉ trong một ảnh còn kém chủ yếu do hình ảnh khi được trích xuất có chất lượng kém.

6.2. Kết luận

Qua quá trình thực hiện đồ án, nhóm đã thành công trong việc đọc được cơ bản các thông tin trên chứng chỉ TOEIC về cả thông tin chữ lẫn thông tin số. Các thành

viên cũng đã tìm hiểu được cách ứng dụng các mô hình máy học vào thực tế cũng như các bước cơ bản để giải một bài toán machine learning. Toàn bộ quy trình từ thu thập dữ liệu đến xử lý cũng như gán nhãn, nhóm đều dành thời gian để tự tay làm. Không chỉ có vậy, nhóm còn bước đầu tìm hiểu về object detection cũng như text recognition và cách cài đặt chúng. Tuy nhiên, với kiến thức và kinh nghiệm còn non nớt cũng như thời gian có hạn, đề án của nhóm vẫn còn nhiều vấn đề chưa thể giải quyết được. Nhóm chúng em hi vọng thầy có thể đưa ra những lời góp ý chân thành nhất để đề án chúng em có thể hoàn thiện hơn.

6.3. Hướng phát triển đề tài

Nhóm sẽ tăng cường độ chính xác của mô hình khi thực hiện trích xuất thông tin với ảnh được chụp ở nhiều điều kiện khác nhau với chất lượng tốt hơn. Tìm kiếm một model có thể nhận diện được một cách chính xác nhất các thông tin do mô hình được nhóm sử dụng vẫn còn nhận diện sai và thiếu một vài trường hợp. Sau khi hoàn thành tốt được đề tài, nhóm sẽ phát triển hệ thống tự động nhận diện các loại chứng chỉ khác như IELTS, TOEFL, và nhiều loại chứng chỉ hơn nữa.

TÀI LIỆU THAM KHẢO

- [1]: <https://viblo.asia/p/trich-xuat-thong-tin-tu-chung-minh-thu-bJzKmaRwK9N>
- [2]: <https://jst-hau1.vn/media/30/uffile-upload-no-title30803.pdf>
- [3]: <https://viblo.asia/p/nhan-dien-bien-so-xe-viet-nam-Do754P9L5M6>