

An OCR Toolbox for Vietnamese Documents

Phạm Minh Khôi
University of Science
Ho Chi Minh city, Vietnam
Email: pmkhoi@selab.hcmus.edu.vn
Student ID: 18120043

Nguyễn Hồ Thăng Long
University of Science
Ho Chi Minh city, Vietnam
Email: nhtlong@selab.hcmus.edu.vn
Student ID: 18120134

Tóm lược—Các hệ thống Nhận dạng ký tự quang học (OCR) đang được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau chẳng hạn hệ thống tự động trong văn phòng, nhà máy, dạy học trực tuyến, đặc biệt là trong các hệ thống trích xuất thông tin từ hóa đơn. Chúng tôi đề xuất một quy trình nhận diện văn bản từ ảnh chụp hóa đơn trong cuộc sống bằng thiết bị di động đồng thời chuẩn hóa, xác định vùng tiềm năng, trích xuất thông tin và tự sửa lỗi. Sử dụng những từ khóa này để trả lời những truy vấn về tên cửa hàng, địa chỉ, thời gian thực hiện giao dịch, tổng giá tiền. Trong quá trình phát triển, chúng tôi đã xây dựng phương án để nhận diện chứng từ liên quan tới tiếng Việt, giải pháp nhận diện dc khá tốt chữ in trên hóa đơn, chứng từ, tài liệu, sửa lỗi các từ trong tập dữ liệu. Hơn nữa hệ thống hỗ trợ hóa đơn không có cấu trúc, có thể sử dụng trong nhiều loại chứng từ khác nhau.

Ngoài ra chúng tôi còn chứng minh hệ thống còn cung cấp cho người dùng khả năng huân luyện, tùy chỉnh và áp dụng vào bài toán khác như số hóa ảnh chụp chứng minh nhân dân, căn cước công dân, là một phần quan trọng trong hệ thống eKYC, giải pháp đã sử dụng trên một số mẫu từ internet, là tiên đề cho quy trình hoàn thiện để có thể sử dụng trong thương mại điện tử và an toàn thông tin. Mã nguồn có thể tìm thấy ở [đây](#).

I. GIỚI THIỆU

Các hóa đơn thường mang lại các thông tin cần thiết cho các công ty, và đa số chúng thuộc dạng giấy hoặc là các văn bản dạng điện tử như tệp tin PDF hoặc dạng ảnh. Để quản lý những thông tin này một cách hiệu quả, các công ty trích xuất và lưu trữ chúng vào cơ sở dữ liệu của họ. Optical Character Recognition (OCR) là một công nghệ phục vụ cho việc nhận diện chữ viết trong ảnh một cách tự động và có thể giải quyết bài toán này một cách hiệu quả. Ngày nay OCR được áp dụng đa số vào các bài toán như điện tử hóa văn bản, xác thực danh tính, hệ thống thương mại điện tử, và nhận diện biển số xe,... OCR truyền thống chủ yếu tập trung vào trích xuất văn bản từ tài liệu in. Tuy nhiên văn bản trong tự nhiên là một bài toán nan giải hơn bởi vì các tác động trực quan như chữ viết bị méo mó, mờ nhạt, che lấp, chìm trong nền hoặc thậm chí là ảnh hướng bởi góc nhìn khác nhau. Mặc dù những khó khăn này, các thuật toán deep learning ngày nay hoạt động tương đối tốt và càng phát triển.

Trong những năm gần đây, việc phát hiện và nhận diện văn bản trong thực tiễn tự nhiên đã được giải quyết phổ biến bởi các cộng đồng sử dụng thị giác máy tính. Chính vì lẽ đó, chúng tôi đã không chọn bài toán OCR trên ảnh hóa đơn scan có chất lượng cao như tập dataset để đánh giá. Thay vào đó, để tiếp cận bài toán này, nhóm tác giả đã sử dụng bộ dữ liệu do người Việt xây dựng mang tên MC-OCR, bộ dữ liệu này

có đầy đủ những thách thức gấp phải bao gồm các ngữ cảnh trong thực tế. Kết quả thu được chúng tôi đã đề xuất hệ thống bao gồm các bước xác định vùng chứa hóa đơn, chuẩn hóa, dự đoán vùng chứa nội dung, xác định văn bản chính sửa lỗi chính tả, phân loại nội dung văn bản, truy vấn thông tin hóa đơn.

Ở phần nhận diện và chuẩn hóa hóa đơn, chúng tôi sử dụng các phương pháp xử lý ảnh thay vì mô hình chọn ra 4 góc vì khả năng diễn giải của các phương pháp xử lý ảnh cổ điển mang tính thuyết phục hơn, mặc dù theo số liệu kết quả chúng đem lại thấp hơn so với các phương pháp học sâu, tuy nhiên để mô hình tổng quan nhất và có thể áp dụng vào các bài toán khác, chúng tôi dùng các phương pháp truyền thống để có thể tách được hóa đơn ra khỏi nền.

Ở phần ước tính các vùng văn bản, đã có nhiều mô hình đề xuất cho bài toán object detection trong những năm gần đây như yolo, efficientnet ..vv nhưng ở đây chúng tôi sử dụng mô hình chính là PAN. Chúng tôi cũng sử dụng phép tăng cường xoay để mô hình có thể bắt biến và có khả năng nhận diện trước các ảnh đầu vào khó học. Điểm đặc biệt của PAN khi chúng tôi thử nghiệm chính là đầu ra của dữ liệu là 4 điểm chứ không phải là một hộp chữ nhật chứa văn bản, và dựa theo khảo sát thì 4 điểm này tạo thành một hình tứ giác có độ nghiêng là độ nghiêng của chữ, quả thật sau khi biến đổi về hình chữ nhật, mô hình phần nào đã giải quyết bài toán deslant trong thực tế, từ đó khiến mô hình OCR hoạt động tốt hơn rất nhiều.

Ở phần nhận diện văn bản, chúng tôi sử dụng mô hình transformer để giải quyết bài toán này. Transformer đang là mô hình SOTA hiện tại trong hầu hết các bài toán từ thi giác cho đến xử lý ngôn ngữ. Mô hình này giúp việc trích xuất văn bản từ ảnh một cách hiệu quả.

Ở phần chỉnh sửa lỗi sai chính tả, chúng tôi áp dụng các thuật toán so khớp bao gồm trie và edit distance. Bằng cách thống kê và lọc bỏ một số từ trong quá trình thu thập, chúng tôi đã xây dựng một bộ từ điển từ dữ liệu hóa đơn bao gồm cá cụm từ chỉ từ loại {đơn vị cung cấp, thời gian, địa điểm, tổng tiền}. Quá trình này khiến cho việc truy vấn thông tin nhanh chóng và dễ dàng hơn.

Ở phần truy vấn thông tin, chúng tôi sử dụng các phương pháp làm sạch từ và so khớp để có thể trả về loại đúng nhất của đầu vào, hơn nữa chúng tôi còn sử dụng thêm mô hình PhoBERT để có thể rút trích được ngữ nghĩa của các cụm từ và phân loại. Từ đó chúng tôi gán nhãn và dựng được các mẫu thông tin câu trả lời cho các câu hỏi bao gồm {đơn vị cung

cấp, thời gian, địa điểm, tổng tiền}

Kết quả chúng tôi thu được cho thấy hệ thống hoạt động rất tốt trên các hóa đơn ở các góc khác nhau, hình dạng khác nhau, tuy nhiên vẫn còn hạn chế ở phương pháp xử lý ảnh cổ điển (ảnh 9). Hệ thống cũng đã khẳng định khả năng mở rộng và ứng dụng của giải pháp trên bài toán tương tự, chẳng hạn như chứng minh nhân dân cũng đã được thử nghiệm và ghi nhận một số kết quả ở ảnh 8.

II. CÁC CÔNG VIỆC LIÊN QUAN

Đa số các hệ thống và nghiên cứu đều chia bài toán này thành 2 bước để giải quyết. Bước đầu tiên là phát hiện vị trí các vùng có thể chứa văn bản và trích xuất những vùng này từ ảnh gốc. Sau đó tiến hành nhận dạng nội dung văn bản từ những vùng này và tạo thành các chuỗi kết quả.

Tiền xử lý (preprocessing) là một công đoạn không thể thiếu trong các bài toán OCR để chống lại sự nhạy cảm của mô hình. bao gồm một số phương pháp như deskew là cân bằng lai các chữ nghiêng so với trục X, deslant là cân bằng lai độ nghiêng của chữ so với trục Y ... Bởi vì kì vọng đầu vào của các bài toán OCR thường sẽ là những tấm ảnh scan trắng và đen. Tuy nhiên trong thực tế, các giá trị trắng và đen có thể có mức độ màu khác nhau do những yếu tố như ánh sáng, màu mực. Cho nên khử noise, contrast là một bước cần thiết. Một số giải pháp có thể kể đến như lọc ngưỡng trên ảnh xám (grayscale), adaptive threshold. Hoặc có thể kể đến một số mạng MLP để làm sạch đầu vào của dữ liệu.

Phương pháp dựa trên thuộc tính thủ công. Các phương pháp truyền thống sử dụng thuộc tính thủ công để thực hiện việc phát hiện văn bản. Những hệ thống này sử dụng đặc trưng như là MSER [1], Stroke Width [2], hoặc HOG [3] để phát hiện các vùng chứa văn bản. Ở bước phân loại, các mô hình như SVM, hoặc KNN [4] được sử dụng trên những vùng đó.

Một ví dụ điển hình là HMM(hidden markov model), sử dụng đặc trưng thủ công. Một cách thường thấy là sử dụng một sliding window chạy dọc qua câu văn bản. Vector đặc trưng đầu vào có thể được biểu diễn như sau: trọng lượng (weight) của cửa sổ dựa trên: số điểm đen, trọng tâm cửa sổ, moment bậc 2 của cửa sổ. Đặc trưng này có thể mô tả được số lượng pixel của văn bản trên từng cửa sổ và phân phôi của chúng, vị trí của upper và lower contour trong cửa sổ, hướng (đạo hàm) của contour. Sử dụng đặc trưng để đưa qua một mô hình dạng chuỗi như HMM.

Phương pháp dựa trên Deep Learning có thể kể đến các nhóm sau:

Đầu tiên là *temporal classification*. 2 step chính bao gồm từ ảnh 2D thành vector 1D, từ đó cân chỉnh (align) so với nhãn đầu vào của từ đó

Image to sequence. Hướng trích xuất đặc trưng của deep learning. Trích xuất đặc trưng từ ảnh (C,W,H) thành (C,W) (sẽ lấy pooling chiều H) lúc này W (chiều rộng của ảnh có thể coi như timestep T. Qua một số lớp hidden và cuối cùng là linear để đưa về không gian độ lớn của vocab

Sequence alignment. Lúc huấn luyện thì (Connectionist Temporal Classification) CTCLoss [5] sẽ tìm cách align sequence input từ model và sequence target từ groundtruth để

model học. Giới hạn là sẽ không khai thác được meaning mà chỉ dựa hoàn toàn vào đặc trưng hình ảnh

Sq2Seq Lấy cảm hứng từ NLP(Neural machine translation) như sau Mô hình sẽ thực hiện một bài toán gần như dự đoán trạng thái tiếp theo là gì, vừa tối ưu encoder (bắt kè mô hình CNN nào) và decoder giống như cơ chế của NMT

Một số phương pháp trích xuất đặc trưng hiện tại cũng là sự kế thừa và kết hợp của các phương pháp trên. Gupta et al. [6] đề xuất một phương pháp sử dụng cấu trúc YOLO để tìm vùng chứa văn bản. Bartz et al. [7] và Wojna et al. [8] áp dụng cơ chế Spatial attention để xây dựng mạng có thể huấn luyện end-to-end và thu được một số kết quả tốt. STN-OCR [7] hoạt động giống với ý tưởng của mô hình RCNN [9], trong đó họ tạo ra nhiều lối để tìm vị trí văn bản, sau đó trích những vùng này ra rồi phân loại. Wojna et al. [8] thực hiện việc phát hiện văn bản bằng nhiều góc nhìn của một ảnh chứ không dựa vào bounding box. Họ trích xuất đặc trưng của 4 góc nhìn khác nhau của cùng một bức ảnh sau đó kết hợp sử dụng spatial attention và RNN để dự đoán các đoạn câu trong ảnh. Ngoài ra, PP-OCR [10] cũng là một hệ thống OCR siêu nhẹ với độ chính xác cao được áp dụng phổ biến hiện nay, Du[10] đề xuất sử dụng một mạng segmentation để phân đoạn văn bản, sau đó với các phép xử lý ảnh cơ bản để tìm ra tọa độ. Sau đó đưa qua một mạng CRNN [11] là mô hình trích xuất đặc trưng ảnh theo cơ chế chuỗi thời gian. Mô hình kế thừa hàm lỗi Connectionist Temporal Classification(CTC) [5] để tránh sự bất ổn định giữa dự đoán và nhãn thật. Tuy nhiên, CTC có một số hạn chế khi số lượng kí tự tối đa có thể dự đoán bằng với kích thước của feature map. Do đó, cần phải điều chỉnh kiến trúc mô hình để phù hợp với từng bộ dữ liệu khác nhau.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Dữ liệu được sử dụng

Chúng tôi sử dụng 2 bộ dữ liệu chính để huấn luyện và đánh giá các mô hình của các giai đoạn:

- Bộ dữ liệu MC-OCR từ cuộc thi RIVF2021¹. Bộ dữ liệu chứa 1000 ảnh huấn luyện và 400 ảnh kiểm thử, là các hình ảnh hóa đơn Việt Nam được chụp bởi thiết bị di động, xem Ảnh 1.
- Bộ dữ liệu SROIE19 của cuộc thi ICDAR 2019², đây là một tập chứa các hình ảnh hóa đơn bằng tiếng Anh. Bộ dữ liệu gồm 700 ảnh huấn luyện và 400 ảnh để đánh giá, xem Ảnh 2.

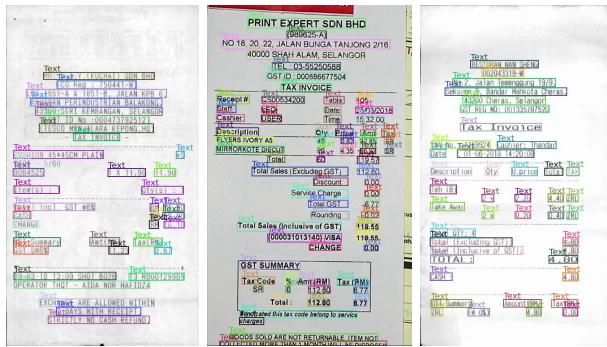
Trong đó, tập SROIE19 được dùng để huấn luyện mô hình PAN; và tập MC-OCR được dùng để huấn luyện mô hình TransformerOCR và PhoBERT.

¹<https://rivf2021-mc-ocr.vietnlp.com/>

²<https://rrc.cvc.uab.es/?ch=13>



Ảnh 1: Một vài ảnh dữ liệu từ tập MC-OCR. Từ trái qua phải: Hóa đơn bình thường, hóa đơn xoay ngang, hóa đơn lật ngược 180 độ



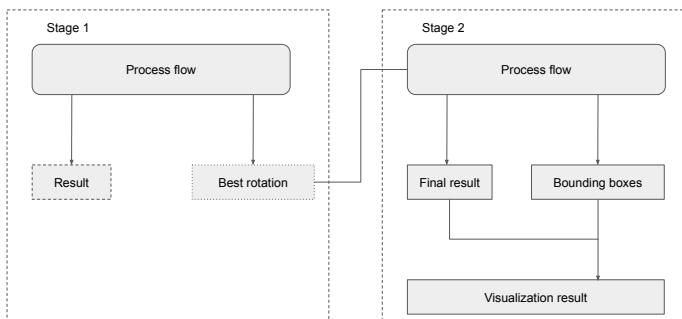
Ảnh 2: Một vài ảnh dữ liệu từ tập SROIE19. Khác với tập MC-OCR, các hóa đơn trong bộ dữ liệu này ở góc nhìn trực diện

B. Sơ lược

Chúng tôi đề xuất một phương pháp hoàn chỉnh để trích xuất văn bản từ hóa đơn, bao gồm hai giai đoạn chính:

- Giai đoạn đầu tiên: Tạm thời ước lượng vùng có chữ, xác định góc tốt nhất và ra kết quả.
- Giai đoạn trực quan hóa: Sau khi có góc tốt nhất, để trực quan hóa, chúng tôi thực hiện thuật toán lại một lần nữa và thu về được kết quả hiển thị.

Sơ đồ thực hiện được mô tả ở ảnh 3



Ảnh 3: Quy trình tổng quan

Cả hai giai đoạn đều sử dụng các module tuần tự như sau: Tiền xử lý ảnh, phát hiện văn bản trong ảnh, sau đó là nhận dạng văn bản và cuối cùng là trích xuất thông tin từ ảnh. Từ ảnh dữ liệu gốc, một số phương pháp tiền xử lý sẽ được áp

dụng để chuẩn hóa, sau đó mô hình sẽ đề xuất các vùng có thể chứa văn bản trong ảnh và đưa những vùng này vào nhận dạng để cho ra kết quả cuối cùng. Mô tả rõ hơn ở ảnh 4

C. Các bước thực hiện chính

1) Tiền xử lý dữ liệu:

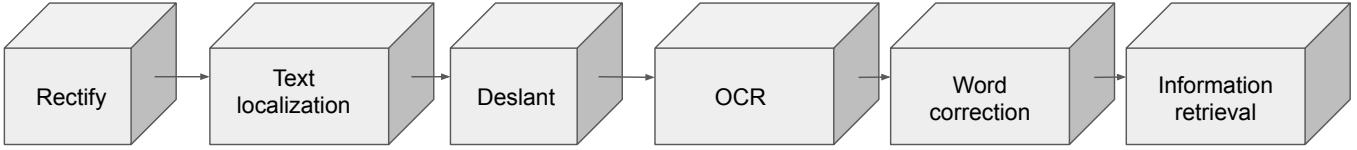
Việc tiền xử lý dữ liệu là một công việc quan trọng trong mọi bài toán. Nhận thấy rằng trong bộ dữ liệu MC-OCR, các tờ hóa đơn không nằm ở một vị trí nhất định trong ảnh và không phải lúc nào cũng thẳng, đều và dễ đọc, kể cả việc mỗi hóa đơn có kích thước khác nhau. Có thể nhìn thấy trong Fig 1. Các hóa đơn có nhiều kích thước khác nhau, hoặc bị xoay nghiêng, thậm chí là lật ngược 180 độ. Điều này có thể gây khó khăn cho các mô hình phát hiện và nhận dạng, do đó việc tiền xử lý ảnh là cần thiết. Chúng tôi áp dụng phương pháp xử lý ảnh đơn giản để phát hiện hóa đơn trong ảnh, sau đó áp dụng các kỹ thuật cổ điển như xoay ảnh, cắt ảnh để có thể trích xuất những vùng phát hiện ra và chuẩn hóa về cùng một góc nhìn, sau đó mới tiến hành tìm kiếm văn bản.

Phát hiện hóa đơn trong ảnh. Để tìm ra vị trí của hóa đơn, chúng tôi sử dụng thuật toán phát hiện biên cạnh Canny, kết quả trả về cuộc thuần toán là các đường biên cạnh trong ảnh, sau đó chúng tôi tiến hành tìm các đường bao (contour) có cường độ lớn nhất dựa trên các đường biên cạnh đã phát hiện. Đường bao có thể được giải thích đơn giản là một đường cong nối tất cả các điểm liên tục (đọc theo đường biên), có cùng màu hoặc cường độ. Chúng tôi giả định rằng trong một bức hình, biên cạnh của hóa đơn sẽ là đường bao lớn nhất. Sau khi tìm được đường bao lớn nhất, chúng tôi thực hiện phép cắt vùng này ra khỏi ảnh gốc, sau đó sử dụng phép biến đổi hình học để xoay ảnh về vị trí trực diện. Tuy nhiên, các trường hợp ảnh nằm quá nghiêng, nghiêng 90 độ, hoặc 180 độ sẽ không chính xác. Do đó chúng tôi sẽ tìm một phép xoay là bội số của 90 độ để trả về kết quả có độ nghiêng nhỏ nhất. Một số ví dụ mẫu có thể thấy ở Bảng 4.

2) Phát hiện văn bản:

Để phát hiện văn bản có trong hóa đơn và phân tách các từ, mô hình Pixel Aggregation Network (PAN) được giới thiệu bởi Wang et al. [12] vì mô hình này đạt được sự cân bằng tốt giữa độ chính xác và thời gian chạy. Ngoài ra mô hình còn cho thấy đã thể hiện tốt độ đo về mặt ngữ nghĩa của các cụm từ, chính vì vậy nhóm tác giả đánh giá PAN là một mô hình đáng thử nghiệm,

Các mô hình huấn luyện sẵn của PAN đã học rất tốt trên bộ dữ liệu chữ trong cảnh (như là CTW1500, SynthText) do đó có thể sẽ không hoạt động tốt với dữ liệu dạng hóa đơn. Do đó cần phải huấn luyện lại trên bộ dữ liệu phù hợp. Đầu ra của PAN là các vùng đề xuất chứa văn bản, những vùng này có thể có các hình dạng bất kỳ. Do khi huấn luyện mô hình đã áp dụng các kỹ thuật xoay ảnh nên thuật toán vẫn hoạt động tốt trên hóa đơn bị nghiêng. Bảng 5 mô tả một số ví dụ mẫu cho việc phát hiện văn bản và so sánh.



Ảnh 4: Quy trình thực hiện chính

3) Nhận dạng văn bản:

Sau khi phát hiện được các vùng chứa văn bản, công đoạn tiếp theo sẽ là nhận diện văn bản. Ở đây VietOCR³ - một framework nhận dạng văn bản tiếng Việt hoạt động ổn định và hiệu quả. Hướng phát triển là kế thừa mô hình AttentionOCR từ framework này. Mô hình này là sự kết hợp giữa mô hình CNN và mô hình Attention. Cách hoạt động của mô hình tương tự như kiến trúc của mô hình Seq2Seq, ảnh đầu vào sẽ được nén thành các vector đặc trưng, sau đó đưa qua mô hình LSTM, sau đó tại mỗi thời điểm mô hình sẽ dự đoán từ tiếp theo là gì. Xem ví dụ ở Bảng 5

4) Chính sửa lỗi chính tả:

Bằng các phương pháp thống kê thông thường, chúng tôi sử dụng các thuật toán cổ điển so khớp chuỗi để nhận dạng các từ đã gấp trong từ điển xây dựng bao gồm cây tìm kiếm Trie [13] và Edit Distance [14]. Ngoài ra, chúng tôi sử dụng thêm Biểu thức chính quy (Regular Expression hay RegEx) để lọc thời gian từ văn bản một cách hiệu quả. Edit Distance được mô tả như sau: Độ tương đồng của hai chuỗi S_1 và S_2 được xác định bởi công thức sau, bằng cách tính 2 lần số lượng kí tự giống nhau K_m chia cho tổng số kí tự của 2 chuỗi. Các kí tự giống nhau K_m gồm chuỗi con giống nhau dài nhất và các kí tự giống nhau ở các vùng rỗi rạc.

$$D_{ro} = \frac{2K_m}{|S_1| + |S_2|} \quad (1)$$

trong đó điểm đánh giá độ tương đồng là D_{ro} có giá trị trong khoảng 0 và 1: $0 \leq D_{ro} \leq 1$

Sử dụng nguõng để thay thế một số từ đã gấp trong văn bản.

5) Trích xuất thông tin từ văn bản:

Để trích xuất ra các thông tin trong hóa đơn, chúng tôi sử dụng 2 loại phương pháp chính để gán nhãn cho từng loại

Phân loại văn bản dựa theo deep learning Sau khi lấy được nội dung của từng đoạn text trong ảnh đã phát hiện, chúng tôi sử dụng mô hình PhoBERT là state-of-the-art của VinAI[15] áp dụng cho các bài toán chữ viết Tiếng Việt để huấn luyện phân loại thành các lớp được định danh sẵn trong tập MC-OCR.

³<https://pbccquoc.github.io/vietocr/>

Phân loại văn bản dựa theo rule based Sau quá trình chỉnh sửa, do một số từ đã xuất hiện có cùng thể loại với từ trong tập dữ liệu, ta hoàn toàn có thể sử dụng ánh xạ 1-1 để thực hiện gán nhãn cho một số từ cụ thể. Do trường hợp ở đây là hóa đơn, các trường thông tin sẽ dễ dàng lặp lại, do vậy rule based rất hữu dụng trong giai đoạn trích xuất thông tin.

Kết quả nhãn dự đoán của một đoạn văn bản cuối cùng sẽ là sự kết hợp tuyến tính của hai phương pháp trên. Xem kết quả mẫu ở ảnh 7

IV. THỰC NGHIỆM VÀ KẾT LUẬN

A. Tham số và quy trình huấn luyện

Trước tiên, chúng tôi thử nghiệm việc phát hiện văn bản trên mô hình PAN, chúng tôi cũng so sánh trước và sau khi áp dụng kỹ thuật tiền xử lý ảnh. Chúng tôi tiền huấn luyện mô hình PAN trên bộ dữ liệu SROIE19. Đầu tiên ảnh được thay đổi về kích thước 640×640 , áp dụng qua một số kỹ thuật tăng cường ảnh như xoay 90 độ, lật ngang lật dọc, mà không có một phương pháp nào thay đổi màu của ảnh. Sau đó, ảnh huấn luyện sẽ đi qua mô hình, và học để tối ưu hàm loss. Chúng tôi sử dụng thuật tối ưu Adam, tốc độ học 0.001 và huấn luyện suốt 300 epochs trên card đồ họa Tesla T4 của Google Colab. Độ hiệu quả của mô hình được đánh giá bằng công thức Mean Average Accuracy (MAP), Pixel Accuracy và IOU.

Model	Image Size	MAP@0.5	Pixel Accuracy	IOU
PAN (baseline)	640 x 640	0.708	0.9517	0.9117
PAN (rotation)	640 x 640	0.669	0.9321	0.8898

Bảng 1: Đánh giá điểm của PAN trên tập SROIE19_val, rotation là có sử dụng phép xoay để tăng cường dữ liệu, baseline là không sử dụng phép nào.

Nhận thấy dữ liệu trong tập MCOCR bao gồm các vùng chứa văn bản, chúng tôi tiến hành cắt vùng có chữ và phân chia tập dữ liệu theo tỉ lệ 8:2 để tiến hành huấn luyện. Chúng tôi sử dụng pretrained từ framework VietOCR và tiến hành finetuning.

Model	train loss	val loss	acc full seq	acc per char
Transformer OCR	0.551	0.476	0.8906	0.9815

Bảng 2: Đánh giá mô hình Transformer trên tập dữ liệu MC-OCR.

Ở giai đoạn cuối cùng là dự đoán phân loại các bounding box dựa trên nội dung. Chúng tôi sử dụng pretrained của

VinAI, framework NLP của Huggingface để tiến hành huấn luyện. Chúng tôi loại bỏ các từ được lặp lại để tránh data leakage.

Backbone	train loss	val loss	acc train set	acc val set
PhoBERT-base	0.105	0.509	0.978	0.924

Bảng 3: Đánh giá mô hình phân loại trên tập dữ liệu MC-OCR sạch.

B. Chi tiết cài đặt và suy diễn

1) Hóa đơn:

Sau khi hoàn tất huấn luyện, chúng tôi sử dụng tập private MCOCR để tiến hành thử nghiệm.



Ảnh 5: Mô tả toàn bộ quá trình trích xuất thông tin hóa đơn:

1. Phát hiện hóa đơn, cắt phần hóa đơn ra và chuẩn hóa.
2. Phát hiện văn bản.
3. Cắt các văn bản ra và nhận diện.
4. Trích xuất thông tin.

Chúng tôi sử dụng mô hình PAN để phát hiện các vùng chứa văn bản trên tập dữ liệu MC-OCR. Mặc dù, dữ liệu huấn luyện cho mô hình ở ngôn ngữ Anh, nhưng khi áp dụng vào bộ dữ liệu này, mô hình vẫn hoạt động tương đối tốt. Từ kết quả quan sát được từ bảng 1, thấy rằng kết quả sau khi thêm phép xoay thấp hơn một chút, tuy nhiên đó là do bộ dữ liệu SROIE19 không có dữ liệu hóa đơn nào bị xoay. Chúng tôi vẫn sử dụng mô hình rotation cho những thí nghiệm sau vì nó hỗ trợ tính bất biến với phép xoay. Chúng tôi tiến hành chạy mô hình trên toàn bộ tập dữ liệu sau đó cắt xén và lưu lại các vùng mô hình phát hiện được từ ảnh, nói rõ hơn, chúng tôi sắp xếp thứ tự của những vùng này theo tọa độ từ trái sang phải, từ trên xuống để thuận tiện trong việc nối kết quả lúc sau.

Sau khi trích xuất các đoạn văn bản, chúng tôi sử dụng mô hình Transformer OCR trên từng đoạn văn bản để trích xuất ra chữ. Khi sử dụng mô hình cho việc dự đoán, để thực sự xử lý vấn đề hóa đơn bị xoay, chúng tôi thực hiện việc này ở bước này. Chúng tôi kiểm tra hướng của văn bản dựa trên dòng đầu tiên của hóa đơn, bằng việc xoay dòng này theo 4 hướng và lần lượt cho qua mô hình Transformer OCR và chọn hướng có điểm xác suất cao nhất. Sau đó áp dụng hướng xoay này cho tất cả các đoạn còn lại của hóa đơn.

Các đoạn văn bản trích xuất được sửa lỗi chính tả bằng cách áp dụng phương pháp Edit Distance và cây Trie để thay thế đoạn câu bằng câu có khoảng cách lớn hơn 0.85 theo 2 phương pháp này.

Sau đó các đoạn này được phân loại thành các nhãn "SELLER", "TIMESTAMP", "TOTAL_COST", "ADDRESS" hoặc "NONE". Đối với nhãn "TIMESTAMP", chúng tôi sử dụng các câu RegEx để lọc các đoạn thuộc dạng thời gian như "XX:XX:XX", "XX-XX-XX", "XX.XX.XX" và "XX/XX/XXXX" (trong đó X là số tự nhiên 0 đến 9). Đối với những đoạn còn lại, chúng tôi kết hợp 3 bộ phân loại PhoBERT, Edit Distance và tổng hợp chúng theo các bước bên dưới để cho ra kết quả cuối cùng.

Gọi, L_{bert} , L_{trie} , L_{ed} lần lượt là nhãn dự đoán của các phương pháp PhoBERT, Trie, Edit Distance. Tương tự có S_{bert} , S_{trie} , S_{ed} là xác suất dự đoán của 3 phương pháp này. Đồng thời, gọi nhãn dự đoán cuối cùng và xác suất của nó là L , S . Nếu đoạn đang xét không thuộc các dạng của câu RegEx thì:

- 1) Nếu có 2 trên 3 (hoặc cả 3) phương pháp đều dự đoán cùng một nhãn, thì phân loại theo nhãn đó, và xác suất bằng tổng của 2 (hoặc 3) xác suất của phương pháp đó.
- 2) Nếu 3 phương pháp dự đoán 3 nhãn khác nhau:
 - a) Nếu $S_{ed} \geq 0.4$, thì $L = L_{ed}$, $S = S_{ed}$
 - b) Nếu $S_{trie} \geq 0.25$, thì $L = L_{trie}$, $S = S_{trie}$
 - c) Ngược lại $L = L_{bert}$, $S = \frac{S_{bert}}{3}$

2) Chứng minh nhân dân:

Nhận thấy rất nhiều chứng từ cũng có thể giải quyết được bằng phương án chúng tôi đề ra, chúng tôi áp dụng vào bài toán khác chính là số hóa ảnh chụp chứng minh nhân dân, căn cước công dân đây là một bài toán thực tế có thể sử dụng trong thương mại điện tử và an toàn thông tin. Trực quan hóa quá trình dự đoán của hệ thống có thể xem ở ảnh 6. Điều đặc biệt ở đây khi chúng tôi sử dụng cùng một mô hình đã huấn luyện trên tập MCOCR, kết quả trả ra tốt hơn so với bộ tham số gốc. Điều này chứng tỏ rằng các nguồn tài nguyên học của các mô hình có thể chuyển đổi cho các bài toán khác nhau trong ngôn ngữ Tiếng Việt.



Ảnh 6: Mô tả toàn bộ quá trình trích xuất thông tin Chứng minh nhân dân Việt Nam ⁴

Áp dụng cùng một quy trình với ảnh chứng minh nhân dân Việt Nam (nguồn dữ liệu từ internet). Hệ thống của chúng tôi đã trả ra những kết quả khả quan và chứng minh tính tổng quan của hệ thống. Mô tả ở ảnh 8. Có thể quan sát rằng kết quả của mô hình dự đoán vị trí văn bản lúc này đã liền mạch với nhau mặc dù chưa được huấn luyện qua bộ dữ liệu chứng minh nhân dân.

Ở hệ thống đề xuất, chúng tôi chưa có cách gán nhãn cho các kết quả đề ra để phục vụ bài toán truy vấn khi nguồn dữ

liệu là chứng minh thư. Tuy nhiên ở phạm vi bài báo, chúng tôi đã chứng minh được sự tổng quát của mô hình nhận diện vùng tiềm năng chứa văn bản và nhận dạng nội dung văn bản, cũng như tính linh hoạt của hệ thống đã đề xuất.

REFERENCES

- [1] L. Neumann and J. Matas, “A method for text localization and recognition in real-world images,” in *Asian conference on computer vision*, pp. 770–783, Springer, 2010.
- [2] B. Epshtain, E. Ofek, and Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2963–2970, IEEE, 2010.
- [3] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in *2011 International Conference on Computer Vision*, pp. 1457–1464, IEEE, 2011.
- [4] K. Wang and S. Belongie, “Word spotting in the wild,” in *European conference on computer vision*, pp. 591–604, Springer, 2010.
- [5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.
- [6] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2315–2324, 2016.
- [7] C. Bartz, H. Yang, and C. Meinel, “Stn-ocr: A single neural network for text detection and text recognition,” 2017.
- [8] Z. Wojna, A. Gorban, D.-S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, “Attention-based extraction of structured information from street view imagery,” 2017.
- [9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” 2014.
- [10] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, “Pp-ocr: A practical ultra lightweight ocr system,” 2020.
- [11] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.
- [12] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network,” 2020.
- [13] A. Thue, “Über die gegenseitige lage gleicher teile gewisser zeichenreihen,” 1912.
- [14] J. W. Ratcliff and D. E. Metzener, “Pattern matching: the gestalt approach,” in *Dr. Dobb's Journal*, p. 46, July 1988.
- [15] D. Q. Nguyen and A. T. Nguyen, “Phobert: Pre-trained language models for vietnamese,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037–1042, 2020.

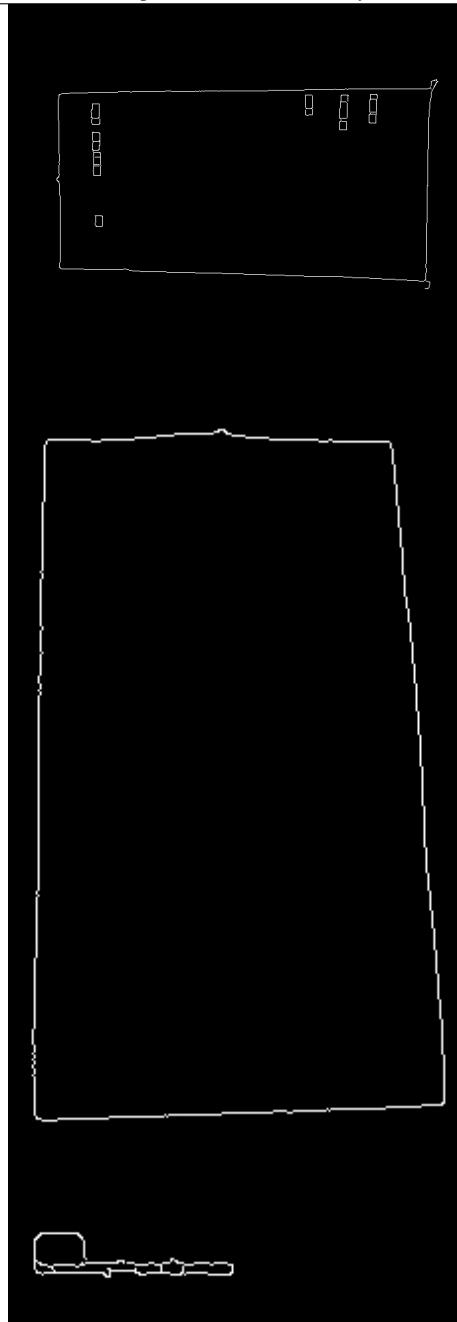
⁴Ảnh dữ liệu từ internet

V. PHỤ LỤC

Ảnh gốc



Các biên cạnh phát hiện bởi thuật Canny



Hóa đơn được trích xuất ra

PHÓ MÔ
Tổ 7 khu Tân Lập 4 - P.Cẩm Thủy -
Tp.Cẩm Phả - Quảng Ninh
ĐT: 0858.931.931

HÓA ĐƠN BÁN HÀNG
Số: HD130820-0039 - Bản: B. 1[A]
13/08/2020 - 21:52

Khách hàng:
SĐT:
Địa chỉ:

Đơn giá	SL	Thành tiền
Huống dương óc chó	12,000	12,000
Trà chanh	10,000	20,000
Cộng tiền hàng:	32,000	
Chiết khấu:	0	
Tổng cộng:	32,000	
Tiền khách đưa:	32,000	
Tiền thừa:	0	

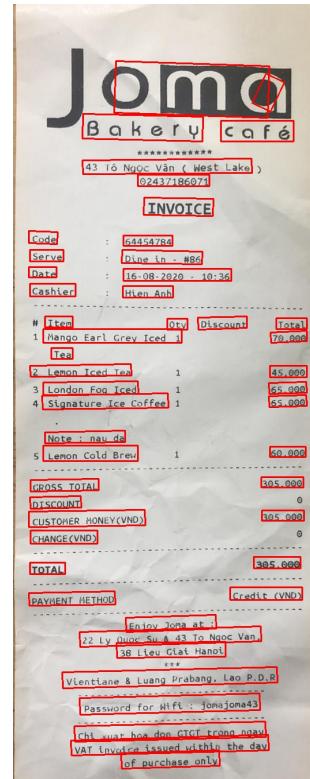
Ba mươi hai nghìn đồng chẵn
Cám ơn và hẹn gặp lại!
Powered by POS365 VN

Bảng 4: Sử dụng Canny để phát hiện và trích vùng hóa đơn ra

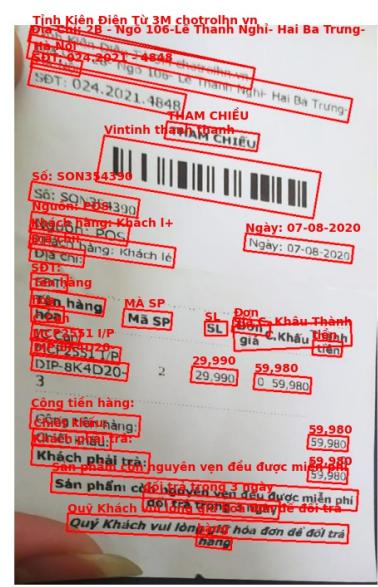
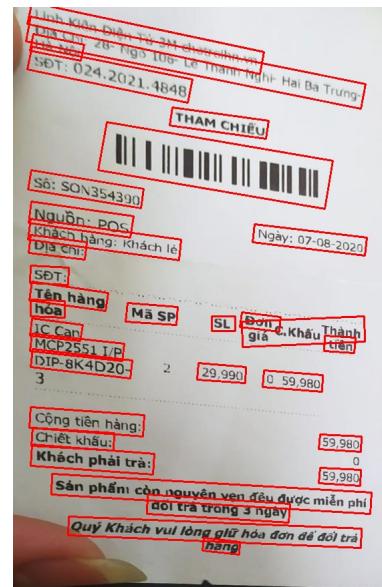
Ảnh gốc



Kết quả của tiền xử lý + PAN



Kết quả của Transformer OCR



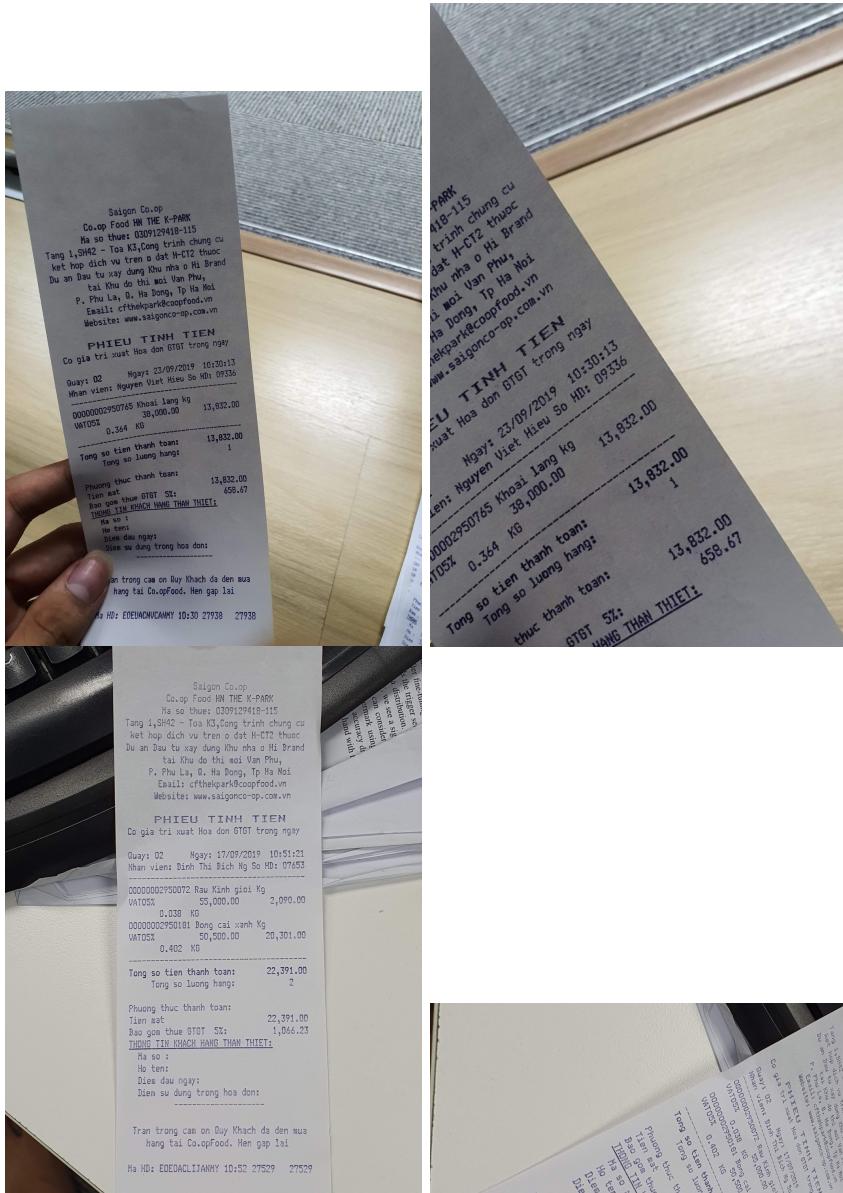
Bảng 5: Dự đoán của mô hình PAN và Transformer OCR trên một vài mẫu dữ liệu. Về trực quan, PAN cho kết quả tốt.

Nhân	Giá trị	Nhân	Giá trị	Nhân	Giá trị
SELLER	bakery	SELLER	THÚC COFFEE	SELLER	Tỉnh Kiên Điện Tử 3M chotrolhn vn
ADDRESS	43 Tô Ngọc Vân (Nest Lake	ADDRESS	22 Quang Trung, P10, Gò Vấp	ADDRESS	Địa Chỉ: 2B – Ngõ 106-Lê Thanh Nghị- Hai Ba Trưng-
TIMESTAMP	16-08-20 10:36	TIMESTAMP	30.03.2019	TIMESTAMP	07-08-2020
TOTAL_COST	TOTAL 305.000	TOTAL_COST	Tiền Thanh Toán 35 000	TOTAL_COST	Khách phải trả 59,980

Ảnh 7: Kết quả truy vấn thông tin sử dụng kết hợp kết quả từ các phương pháp truy vấn từ các ảnh trong Bảng 5



Ảnh 8: Kết quả sử dụng qui trình trên đối với dữ liệu Chứng minh nhân dân Việt Nam ⁴. Theo từ trên xuống: ảnh gốc, mô hình gốc, mô hình sau khi finetune. Có thể thấy kết quả sau khi huấn luyện mô hình cho các box liền mạch hơn.



Ảnh 9: Một số kết quả tệp ở khâu phát hiện hóa đơn, đa số các trường hợp tệp do phần nền quá sáng so với hóa đơn. Từ trái sang phải: Ảnh gốc, phần hóa đơn được phát hiện