# TRÍCH XUẤT THÔNG TIN TỪ CHỨNG CHỈ TIẾNG ANH TOEIC

Đặng Thị Thúy Hồng[1,2], Trần Huỳnh Quốc An[1,2], Phạm Tiến Dương[1,2], Đỗ Trọng Hợp[1,2]

[1] Đại học Công Nghệ Thông Tin, Đại học Quốc gia Thành phố Hồ Chí Minh
[2] 20520523.gm.uit.edu.vn
20520955@gm.uit.edu.vn
20520122@gm.uit.edu.vn
hopdt@.uit.edu.vn

**Tóm tắt nội dung** Key Information Extraction là lĩnh vực đóng vai trò quan trọng trong các hệ thống phân tích và trích xuất các trường thông tin từ văn bản, các dạng tài liệu có thể là hóa đơn, văn bản hành chính, các loài giấy tờ tùy thân quan trọng... Nhờ sự phát triển của Deep Learning đặc biệt là mô hình Grap Neural Network (GNN) được ứng dụng rộng rãi trong khai thác thông tin, các nghiên cứu và ứng dụng trong những năm gần đây đã đạt được nhiều thành tựu trong việc trích xuất những trường thông tin chính (KIE) từ tài liệu. Trong bài báo này, chúng tôi tiến hành hệ thống trích xuất thông tin chính (KIE) từ hình ảnh chụp các Chứng chỉ tiếng anh TOEIC 2 kỹ năng Listening (nghe) và Reading (đọc), bộ dữ liệu được chúng tôi tự thu thập, áp dụng phương pháp Data Augmentation để tăng sinh dữ liệu và tiến hành gán nhãn. Chúng tôi sử dụng mô hình YOLOv5 để phát hiện các văn bản có trong hình ảnh, sau đó tiến hành nhận diện các ký tự xuất hiện trong văn bản với mô hình EasyOCR và trích xuất các trường thông tin chính. Hệ thống bước đầu đạt được kết quả tốt khi nhận diện và trích xuất được các trường thông tin chính như mong muốn.

**Keywords:** Text detection, Optical Character Recognition, Key Information Extraction

## 1 Giới thiệu

Trong thế giới hiện nay, chúng ta đang chìm ngập trong một lượng thông tin khổng lồ ảnh hưởng lớn đến cuộc sống của con người theo nhiều cách khác nhau. Công nghệ ngày càng phát triển, chúng ta có xu hướng chuyển đổi phần lớn các tài liệu, văn bản bằng giấy sang tài liệu, văn bản dạng số như các chứng từ, giấy tờ điện tử để dễ dàng trong việc lưu trữ và tìm kiếm. Bên cạnh đó, năm 2020 - 2021, thế giới nói chung và Việt Nam nói riêng đã trải qua khoảng thời gian đại dịch hoành hành dữ dội, khiến cho việc di chuyển, gặp gỡ nhau trở nên khó khăn hơn. Vì vậy, "Work from home" (làm việc tại nhà) trở thành giải pháp tối ưu cho các doanh nghiệp, trường học... và việc chuyển đổi các tài

liệu thành dạng tài liệu số cũng là một phần giúp cho các công việc được giải quyết trơn tru và đạt hiệu quả cao trong khoảng thời gian giãn cách xã hội. Với khối lượng thông tin số khổng lồ như hiện nay thì lượng hình ảnh tài liệu, văn bản chứa những thông tin quan trọng cũng chiếm phần lớn. Gần đây, việc nhận dạng và trích xuất các thông tin từ hình ảnh thu hút được sự quan tâm lớn từ cộng đồng nghiên cứu và đây cũng là bài toán đầy tính thách thức bởi vì ảnh có thể bị mờ, không chất lượng do thiếu ánh sáng gây khó khăn trong quá trình xử lý. Tuy nhiên, với sự phát triển và tốc độ xử lý của các mô hình học máy, vấn đề nhận dạng văn bản đã được giải quyết bằng nhiều phương pháp. Một trong những phương pháp này là phát hiện các ký tự quang học riêng lẻ (Optical Character Recognition – OCR) và sau đó kết hợp chúng thành các từ có nghĩa phục vụ cho bài toán trích xuất thông tin (Key Information Extraction – KIE). Trong bài báo này, chúng tôi tiến hành xây dựng một hệ thống nhận dạng và trích xuất các thông tin chính quan trọng trên bộ dữ liệu bao gồm hình ảnh chụp Chứng chỉ tiếng anh TOEIC mà chúng tôi tự thu thập. Đầu tiên để phát hiện và định vị văn bản có trong hình ảnh chúng tôi đã áp dụng mô hình YOLOv5, sau đó tiến hành nhận diện các ký tự xuất hiện trong văn bản với mô hình EasyOCR và cuối cùng trích xuất tám trường thông tin chính bao gồm NAME, ID (Identification Number), DOB (Date of Birth), TEST DATE, VALID UNTIL, LISTENING SCORE, READING SCORE, TOTAL SCORE. Tiếp theo của bài báo này, chúng tôi sẽ đề cập đến các công trình liên quan ở phần 2, giới thiệu bộ dữ liệu ở phần 3, tiếp theo là trình bày phương pháp nghiên cứu ở phần 4. Phân tích kết quả đạt được và đánh giá ở phần 5 kế tiếp, cuối cùng kết luận và hướng phát triển ở phần 6.

## 2 Công trình liên quan

Cho đến nay, đã có nhiều phương pháp đột phá được công bố phục vụ các công trình nghiên cứu phát hiện, nhận dạng và trích xuất văn bản trong hình ảnh các tài liệu cấu trúc hay bán cấu trúc chẳng hạn như biên lai, thẻ ngân hàng, danh thiếp, hóa đơn, v.v. và đạt được độ chính xác cao. Đối với bước tiền xử lý ảnh, người ta thường áp dụng các kỹ thuật xử lý ảnh truyền thống và kỹ thuật xử lý ảnh sâu mạng lưới thần kinh như canny, thuật toán phát hiện cạnh, phương pháp Otsu [1], U-Net [2] và VGG [3]. Bên cạnh đó, việc áp dụng các mô hình học sâu để phát hiện ký tự cũng đạt được hiệu suất cao hơn và tiêu tốn thời gian xử lý thấp hơn so với học máy truyền thống. Ví dụ, để phát hiện văn bản, mô hình Character Region Awareness for Text Detection (CRAFT) [4] giúp phát hiện hiệu quả các vùng văn bản bằng cách phát hiện từng ký tự và liên kết chúng lại với nhau. Hay mô hình YOLO [5] đã đề xuất một phương pháp khác có thể phát hiện các đối tượng trong thời gian thực một cách hiệu quả. Để nhận diện văn bản, các mô hình CRNN [6], Tesseract [7], TRBA [8]... được áp dụng có thể phân loại và nhận diện ở cấp độ ký tự đến cấp độ từ một cách hiệu quả. Các mô hình này có thể xác định vị trí của từng ký tự, sau đó nhận diện chúng và nhóm các ký tự thành từ ngữ có nghĩa. Đối với bài báo này, để đạt được độ chính xác cao, chúng tôi đã áp dụng các mô hình theo các bước như một qui

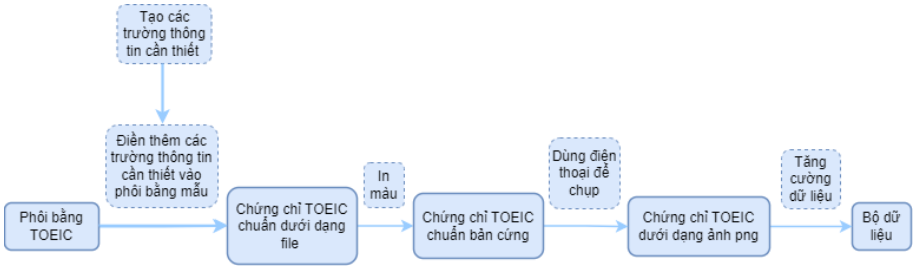
trình end-to-end để trích xuất thông tin từ hình ảnh của Chứng chỉ tiếng anh TOEIC. Đầu ra của bước trước được sử dụng làm đầu vào để huấn luyện các mô hình cho bước tiếp theo.

## 3 Bộ dữ liệu

Trong bài báo này, chúng tôi sử dụng bộ dữ liệu TOEICText bao gồm khoảng 600 bức ảnh chụp các chứng chỉ tiếng anh TOEIC do chúng tôi tự thiết kế dựa trên chứng chỉ mẫu và một phần trong đó được thu thập từ các trang mạng xã hội. Quá trình xây dựng bộ dữ liệu gồm có các bước: Thu thập và tiền xử lý dữ liệu, Gán nhãn dữ liệu.

### 3.1 Thu thập và tiền xử lý dữ liệu

Đầu tiên, chúng tôi thiết kế một chứng chỉ trống chưa chứa bất kỳ thông tin nào của người nhận dựa trên chứng chỉ mẫu gọi là phôi bằng; sau đó chúng tôi tạo các trường thông tin của người nhận chứng chỉ này bao gồm Name, Identification Number, Date of Birth, Test Date, Valid Until, Listening Score, Reading Score, Total Score; cuối cùng chúng tôi tiến hành đưa các trường thông tin vào phôi bằng và căn chỉnh để giống với chứng chỉ mẫu nhất. Để xây dựng được bộ dữ liệu chất lượng nhất, chúng tôi tiến hành in các chứng chỉ và dùng điện thoại để chụp lại. Vì số lượng tương đối ít, chúng tôi đã áp dụng một số phương pháp tăng cường dữ liệu như thay đổi kích thước, xoay hình, tăng giảm độ sáng,... để đa dạng bộ dữ liệu. Bên cạnh đó, chúng tôi cũng kết hợp tìm và thu thập các chứng chỉ từ các nền tảng mạng xã hội và chúng tôi xây dựng được bộ dữ liệu khoảng 600 bức ảnh chụp các chứng chỉ TOEIC. Quá trình xây dựng bộ dữ liệu được tóm tắt ở sơ đồ bên dưới:




**Hình 1.** Quá trình xây dựng bộ dữ liệu

### 3.2 Gán nhãn dữ liệu

Chúng tôi đã sử dụng tool OpenLabeling để gán nhãn cho các hình ảnh với 8 nhãn là 8 trường thông tin cần trích xuất. Bộ nhãn được mô tả ở bảng sau:

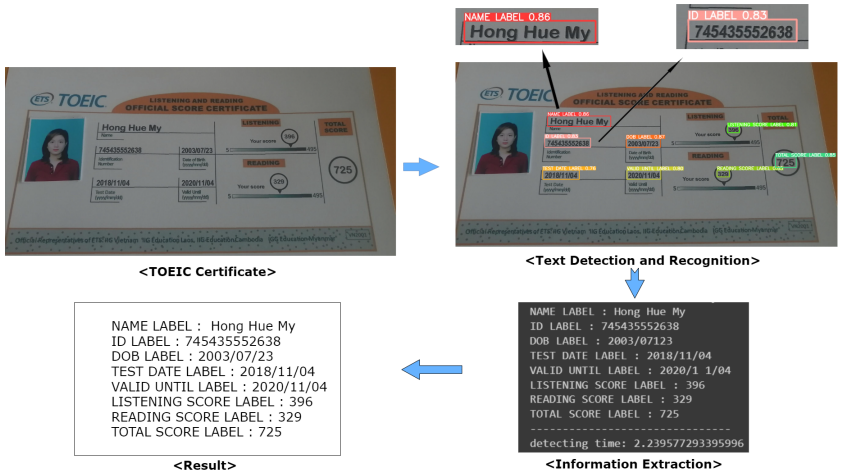
**Bảng 1.** Bảng chi tiết các nhãn của bộ dữ liệu

| **Tên nhãn** | **Mô tả** |
|---|---|
| *NAME* | *Họ tên* |
| *ID (Identification Number)* | *Số định danh* |
| *DOB (Date of Birth)* | *Ngày sinh (yyyy/mm/dd)* |
| *TEST DATE* | *Ngày thi (yyyy/mm/dd)* |
| *VALID UNTIL* | *Chứng chỉ có hiệu lực đến ngày (yyyy/mm/dd)* |
| *LISTENING SCORE* | *Điểm thi kỹ năng nghe* |
| *READING SCORE* | *Điểm thi kỹ năng đọc* |
| *TOTAL SCORE* | *Tổng điểm 2 kỹ năng* |

Trong quá trình gán nhãn, chúng tôi tiến hành tạo các bounding box là các tứ giác bao quanh các vùng chứa chữ cần trích xuất, mỗi bounding box sẽ có màu sắc khác nhau để dễ phân biệt. Sau đó sẽ xuất ra các file label chứa thông tin là tọa độ tâm, chiều dài, chiều rộng của các bounding box đã gán.

## 4 Phương pháp thực hiện

Hệ thống trích xuất thông tin bao gồm các bước được mô tả ở hình sau:

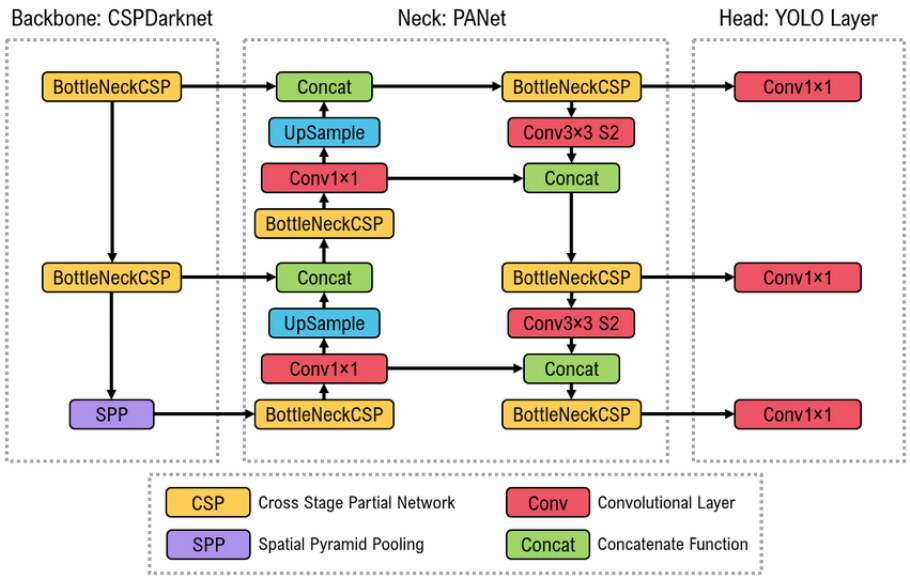



**Hình 2.** Phương pháp thực hiện trích xuất các thông tin từ Chứng chỉ TOEIC

### 4.1 Text Detection

YOLO (You only look once) là một thuật toán phát hiện đối tượng theo thời gian thực do Joseph Redmon, Santosh Divvala, Ross Girshick và Ali Farhadi tạo ra vào năm 2015 và được huấn luyện trước trên tập dữ liệu COCO. Từ đó đến nay, nhóm đã cải thiện mô hình này và nhiều phiên bản YOLO đã được phát hành. YOLO nổi tiếng về tốc độ và độ chính xác, được sử dụng trong nhiều ứng dụng như: chăm sóc sức khỏe, giám sát an ninh và xe tự lái.

YOLO là mạng thần kinh tích chập được tối ưu hóa để phát hiện đối tượng. Chúng tôi sử dụng YOLOv5 để phát hiện văn bản trong bài viết này. YOLOv5 [9] là phiên bản hiện đại của YOLO được công bố vào năm 2020 bởi Glenn Jocher (Người sáng lập và Giám đốc điều hành của Utralytics). Kiến trúc mạng YOLOv5 được chia thành ba phần, bao gồm Backbone: CSPDarknet, Neck: PANET và Header: YOLO Layer. Đầu tiên, ảnh sẽ đi vào lớp CSPDarknet để trích xuất các đặc trưng của các đối tượng trong ảnh, sau đó chuyển tiếp tới lớp PANet để tổng hợp các đặc trưng. Cuối cùng, đầu ra tại lớp YOLO sẽ dựa vào các đặc trưng để xác định vị trí tọa độ điểm và kích thước của đối tượng cần nhận diện.

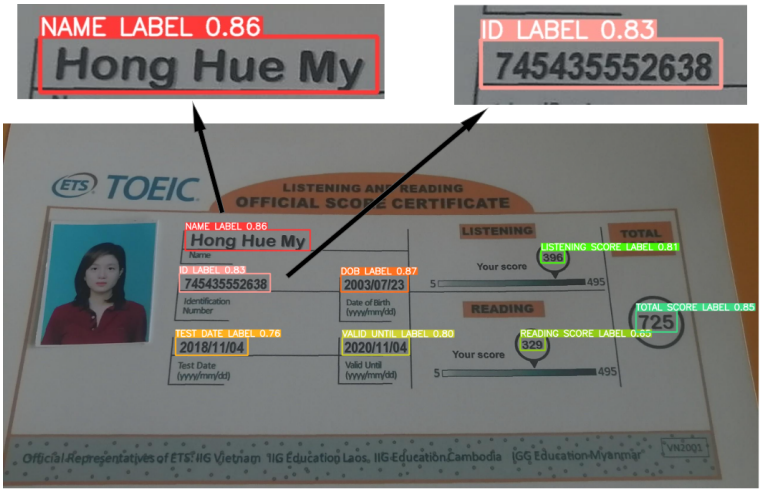



**Hình 3.** Kiến trúc của YOLOv5

Với tốc độ phát hiện các đối tượng và độ chính xác cao, chúng tôi đã nhận được kết quả khả quan khi tiến hành huấn luyện YOLOv5 để phát hiện văn bản trên bộ dữ liệu.

**Hình 4.** Trước và sau khi phát hiện văn bản bằng YOLOv5

Sau khi phát hiện văn bản trên ảnh của các Chứng chỉ, chúng tôi cắt các vùng văn bản từ hình ảnh, tiếp đó phương pháp phát hiện ký tự quang học được áp dụng để nhận dạng ký tự.



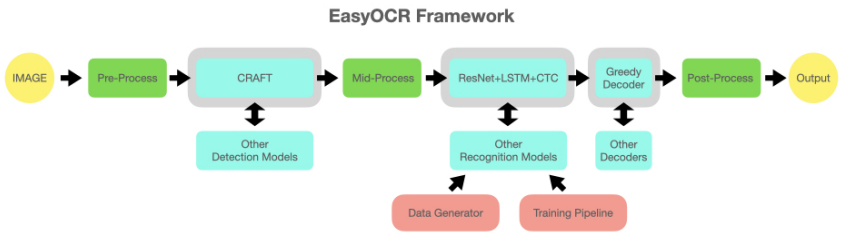

**Hình 5.** Trước và sau khi cắt các vùng văn bản từ các Chứng chỉ

### 4.2 Text Extraction

Với số lượng dữ liệu bị hạn chế, chúng tôi đã sử dụng mô hình EasyOCR được huấn luyện trước để có thể nhận dạng các ký tự và trích xuất thông tin.

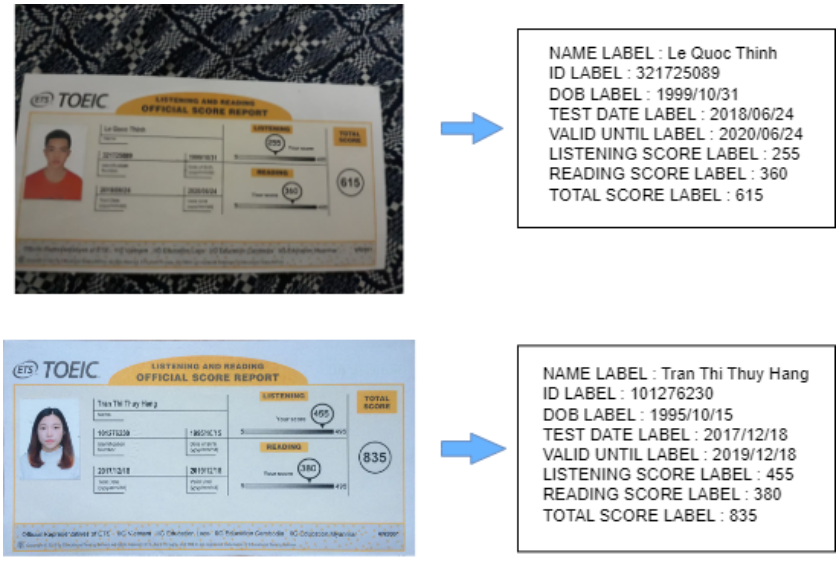
EasyOCR [10] là một mô hình nhận dạng ký tự quang học với tỷ lệ nhận dạng cao. Mô hình này dựa trên PyTorch và sử dụng thuật toán CRAFT để phát hiện các vùng ký tự, CRAFT là mô hình phát hiện văn bản cảnh dựa trên mạng neural network. Để nhận diện, EasyOCR sử dụng kỹ thuật CRNN, bao gồm ba thành phần chính: feature extraction (ResNet), sequence labeling (LSTM) và

decoding (CTC). EasyOCR pipeline cũng bao gồm một số bước tiền xử lý đầu vào nên đây một trong những mô hình xử lý rất tốt cho tác vụ trích xuất thông tin.




**Hình 6.** EasyOCR pipeline

Đầu vào của EasyOCR là tập các ảnh chứa vùng văn bản được cắt ra sau khi đã phát hiện bằng YOLOv5 và mô hình này sẽ huấn luyện để nhận diện các ký tự có trong ảnh. Cuối cùng, chúng tôi tiến hành thử nghiệm bằng cách đưa hình ảnh hoàn chỉnh của các Chứng chỉ vào mô hình và nhận được một số kết quả.




**Hình 7.** Trích xuất thông tin từ Chứng chỉ tiếng anh TOEIC

# 5 Kết quả thử nghiệm

## 5.1 Phương pháp đánh giá

Hai phương pháp được áp dụng để đánh giá hiệu suất là Accuracy và Character Error Rate (CER).

**Accuracy** là tỉ lệ phần trăm dự đoán đúng giá trị của từng nhãn của bộ dữ liệu. Độ đo Accuracy được định nghĩa như sau:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

trong đó TP là True Positive , TN là True Negative, FP là False Positieve, FN là False Negative.

**Character Error Rate** là tỷ lệ lỗi ký tự, đánh giá mức độ hiệu quả của phương pháp OCR phát hiện ký tự quang học. CER được định nghĩa như sau:

$$CER = \frac{i + s + d}{n},$$

trong đó n là tổng số ký tự trong văn bản tham chiếu; i, s, d là số lần chèn, thay thế và xóa các ký tự cần thiết để chuyển đổi đầu ra OCR từ văn bản tham chiếu đầu vào.

## 5.2 Kết quả thu được

**Bảng 2.** Đánh giá kết quả với Accuracy và CER

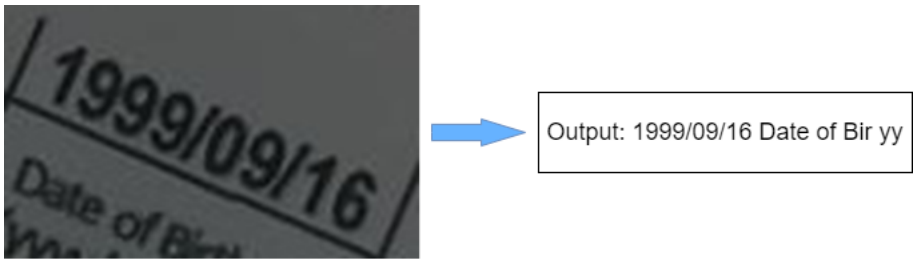
| **LABEL** | **ACCURACY** | **CER** |
|---|---|---|
| *NAME* | *0.44* | *0.31* |
| *ID* | *0.67* | *0.16* |
| *DOB* | *0.05* | *0.17* |
| *TEST DATE* | *0.27* | *0.16* |
| *VALID UNTIL* | *0.25* | *0.17* |
| *LISTENING SCORE* | *0.84* | *0.12* |
| *READING SCORE* | *0.81* | *0.14* |
| *TOTAL SCORE* | *0.85* | *0.07* |

**Nhận xét:** Nhìn chung, mô hình nhận diện ký tự và trích xuất thông tin khá chính xác với các nhãn chứa thông tin điểm thi như: nhãn "TOTAL SCORE" (Độ chính xác: 0.85, Tỷ lệ sai ký tự: 0.07), nhãn "LISTENING SCORE" (Độ chính xác: 0.84, Tỷ lệ sai ký tự: 0.12) và nhãn "READING SCORE" (Độ chính xác: 0.81, Tỷ lệ sai ký tự: 0.14). Tuy nhiên đối với các nhãn chứa thông tin ngày

thì độ chính xác đạt được khá thấp: nhãn "DOB" (Độ chính xác: 0.05, Tỷ lệ sai ký tự: 0.17), nhãn "TEST DATE" (Độ chính xác: 0.27, Tỷ lệ sai ký tự: 0.16), nhãn "VALID UNTIL" (Độ chính xác: 0.25, Tỷ lệ sai ký tự: 0.17).

### 5.3 Phân tích lỗi

**Lỗi 1:** Trích xuất thừa ký tự:




**Hình 8.** Trích xuất thừa ký tự

Trong quá trình gán nhãn và huấn luyện, đa số dữ liệu ảnh trong bộ dữ liệu ở dạng thẳng nên khi mô hình trích xuất thông tin của những ảnh dạng nghiêng sẽ trích xuất thừa sai khác so với văn bản gốc.

**Lỗi 2:** Trích xuất sai ký tự:


**Hình 9.** Trích xuất sai ký tự

Với một số mẫu dữ liệu chứa ký tự "/", mô hình trích xuất sai thành ký tự "1" bởi vì mẫu dữ liệu khá mờ và tối, nên mô hình đã "hiểu nhầm" và nhận diện sai ký tự đó.

## 6 Kết luận và hướng phát triển

Trong bài báo này, chúng tôi đã xây dựng bộ dữ liệu bao gồm khoảng 600 hình ảnh các Chứng chỉ tiếng Anh TOEIC do chúng tôi tự thu thập và gán nhãn.

Bên cạnh đó, chúng tôi cũng xây dựng hệ thống end-to-end để trích xuất thông tin từ bộ dữ liệu bao gồm hai bước liên tiếp: Text Detection, Text Extraction. Mô hình YOLOv5 đã được được sử dụng để phát hiện các vùng ký tự và để nhận diện, trích xuất thông tin từ hình anh Chứng chỉ, chúng tôi đã sử dụng mô hình EasyOCR. Sau khi huấn luyện mô hình đã trích xuất được các thông tin mà chúng tôi mong muốn.

Hiện tại, bộ dữ liệu của chúng tôi vẫn còn nhiều hạn chế về số lượng và chất lượng như ảnh bị thiếu sáng, bị mờ, bị nghiêng, ảnh quá nhỏ... khiến cho quá trình huấn luyện của chúng tôi gặp khó khăn và trích xuất ra những thông tin bị sai lệch, chưa chính xác. Vì vậy, chúng tôi sẽ tiếp tục phát triển bộ dữ liệu, xây dựng thêm mô hình tiền xử lý hình ảnh khi trước khi đưa vào huấn luyện và mở rộng thêm bộ dữ liệu để phục vụ cho bài toán trích xuất thông tin có thể đạt được kết quả tốt hơn, chuẩn xác hơn.

## Tài liệu


1. Tan Nguyen Thi Thanh and Khanh Nguyen Trong, “A Method for Segmentation of Vietnamese Identification Card Text Fields” International Journal of Advanced Computer Science and Applications(IJACSA), 10(10), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0101057
2. Zeng, Zitao et al. “RIC-Unet: An Improved Neural Network Based on Unet for Nuclei Segmentation in Histology Images.” IEEE Access 7 (2019): 21420-21428.
3. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
4. Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection,” CoRR, vol. abs/1904.01941, 2019.
5. Jocher, G.; Stoken, A.; Borovec, J.; Christopher, S.T.; Laughing, L.C. Ultralytics/yolov5: v4.0-nn.SiLU() activations; Weights  Biases logging, PyTorch Hub integration. Zenodo 2021.
6. Chen, Lei, and Shaobin Li. "Improvement research and application of text recognition algorithm based on CRNN." Proceedings of the 2018 International Conference on Signal Processing and Machine Learning. 2018.
7. Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." International Journal of Computer Applications 55.10 (2012): 50-56.
8. J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, “What is wrong with scene text recognition model comparisons? dataset and model analysis,” CoRR, vol. abs/1904.01906, 2019.
9. C. W. Chuang and C. P. Fan, “Deep-Learning Based Joint Iris and Sclera Recognition with YOLO Network for Identity Identification,” Journal of Advances in Information Technology, vol. 12, pp. 60-65, 2021.
10. J. Memon, M. Sami, R. A. Khan and M. Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR),” IEEE Access, vol. 8, pp. 142642-142668, 2020.
11. T. et al., “End-to-end text recognition with convolution neural networks,” IEEE Int. Conf. Pattern Recognit, 2012.
12. El Abbadi, Nidhal Khdhair. "Scene Text detection and Recognition by Using Multi-Level Features Extractions Based on You Only Once Version Five (YOLOv5) and

Maximally Stable Extremal Regions (MSERs) with Optical Character Recognition (OCR)." Al-Salam Journal for Engineering and Technology 2.1 (2023): 13-27.
13. X. -S. Vu, Q. -A. Bui, N. -V. Nguyen, T. T. Hai Nguyen and T. Vu, "MC-OCR Challenge: Mobile-Captured Image Document Recognition for Vietnamese Receipts," 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), Hanoi, Vietnam, 2021, pp. 1-6, doi: 10.1109/RIVF51545.2021.9642077.
14. YOLOv5, `https://github.com/ultralytics/yolov5`. Last accessed 18 Feb 2023
15. EasyOCR, `https://github.com/JaidedAI/EasyOCR`. Last accessed 6 Jan 2023