**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**---🙠🕮🙢---**

**Logo

Description automatically generated**

**XỬ LÝ ẢNH VÀ ỨNG DỤNG**

**Đề tài:**

**TRÍCH XUẤT THÔNG TIN HÓA ĐƠN TIẾNG VIỆT**

Giảng viên hướng dẫn: Cáp Phạm Đình Thăng

Thành viên:

* **Nguyễn Đức Thắng - 19522206**
* **Nguyễn Xuân Minh – 19521848**
* **Hà Văn Thanh - 19522224**
* **Hoàng Quân - 19522077**

Lớp: **CS406.N11**

**THÀNH PHỐ HỒ CHÍ MINH – NĂM 2023**

1. **Tổng quan đề tài**
2. **Giới thiệu**

Trích xuất thông tin từ hóa đơn tự động là phương pháp sử dụng các kiến thức, công nghệ trong thị giác máy tính , xử lý ngôn ngữ tự nhiên … để phát hiện, nhận diện và trích xuất các thông tin có trong hóa đơn.

Các dữ liệu được trích xuất có thể là các thông tin quan trọng như: người mua, người bán, ngày lập hóa đơn, giá trị tổng hóa đơn, v…v

1. **Mô tả bài toán**

Bài toán trích xuất thông tin hóa đơn sử dụng ảnh chụp từ các hóa đơn như một đầu vào (input) qua một mô hình xử lý và đưa ra thông tin kết quả (output) của hóa đơn đó. Dữ liệu đầu ra của nhóm trong bài toán này gồm các trường dữ liệu: người bán (SELLER), địa chỉ điểm bán (ADDRESS), thời gian mua - lập hóa đơn (TIMESTAMP) và tổng chi phí của hóa đơn (TOTAL COST)



1. **Dữ liệu sử dụng**

Bộ dữ liệu dùng trong bài toán này là ảnh chụp hóa đơn được xuất từ các siêu thị, cửa hàng ở Việt Nam như: Bách Hóa Xanh, Circle K, Vin Mart, Ministop,...

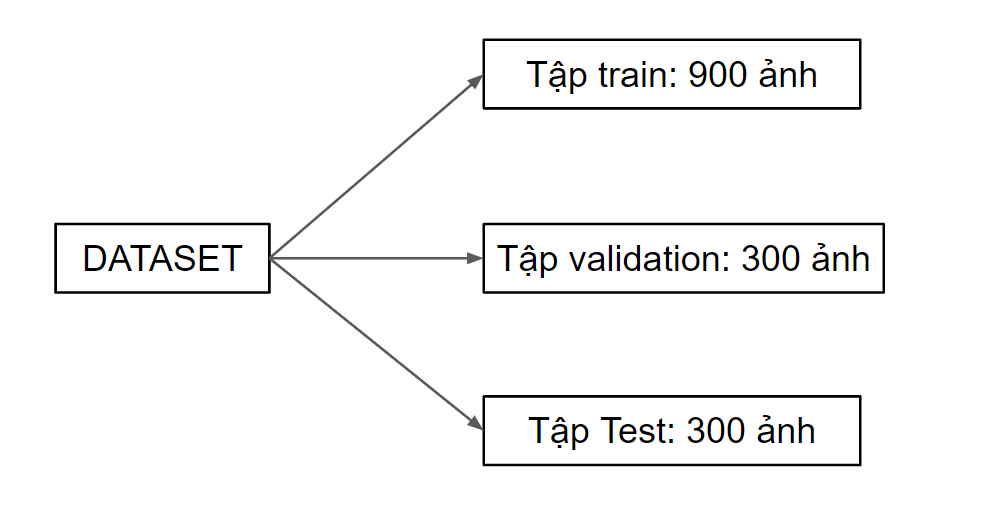
Bộ dữ liệu gồm khoảng 1500 ảnh được sưu tầm từ Internet và các ảnh được nhóm chụp thêm để bổ sung vào bộ dữ liệu



*Một số ví dụ về ảnh hóa đơn trong bộ dataset*

Bộ dữ liệu của nhóm nhìn chung vẫn chưa đủ tốt để thực hiện huấn luyện ngay, trong bộ dataset vẫn còn chứa nhiều các ảnh cần phải xử lý như: ảnh quá sáng hay quá tối, ảnh nghiêng ( 1 góc hoặc nằm ngang). Những ảnh này ảnh hưởng tới kết quả của mô hình nhóm sử dụng do đó cần phải thông qua các bước tiền xử lý để mang lại kết quả tốt hơn.

Nhóm đã thực hiện chia bộ dataset như sau:

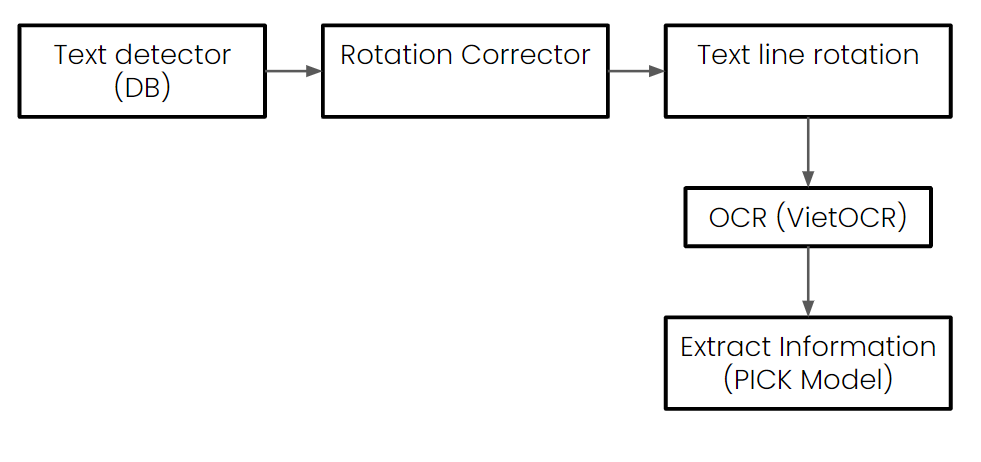


1. **Ý tưởng & Phương pháp thực hiện**

Ý tưởng để thực hiện bài toán này nhóm đã làm như sau:

1. Tiền xử lý: từ ảnh đầu vào, thực hiện tách nền (chỉ quan tâm tới những vùng chứa thông tin (text) của ảnh hóa đơn. Nhóm sử dụng các phương pháp xử lý ảnh cơ bản (threshold, contour, .. ). Đầu ra của quá trình này là ảnh đã được xử lý tách nền.
2. Text detector: xác định vùng chứa thông tin (text) trong ảnh hóa đơn, những thông tin này sẽ được bọc bởi các bounding box . Giai đoạn này sử dụng phương pháp Differentiable Binarization (DB). Đầu ra của quá trình này là ảnh với các bounding box chứa thông tin của ảnh
3. Rotation corrector: thực hiện xoay, căn chỉnh ảnh hóa đơn cho thẳng. Phương pháp này sử dụng các bounding box có được ở bước 2 để tính toán góc xoay và xoay ảnh để thu được ảnh thẳng nhất có thể.
4. TextLine rotation: Sau khi xoay hóa đơn cho thẳng ở bước 3, thì vẫn có thể tồn tại những vùng chứa text bị nghiêng, nên sẽ thực hiện căn chỉnh cho thẳng đối với các vùng chứa này. Ở bước này nhóm sử dụng xử lý ảnh cơ bản.
5. OCR: bước này là bước đọc các thông tin (Text) từ ảnh đã được xoay ở trên. Mô hình sử dụng ở bước này là Viet OCR với pre-trained của mô hình vgg19\_seq2seq. Cuối bước này, nhóm thu được thông tin text có trong mỗi bức hình, đi kèm với mỗi thông tin là tọa độ bounding box của text có trong bức hình.
6. Extract Information: Đây là bước để chắt lọc ra những thông tin cần khai thác từ hóa đơn. Bằng thông tin thu được ở bước 5 ( text + bounding box ) và ảnh thu được ở bước 4, nhóm tiến hành training model để trích xuất thông tin theo 4 trường như đã nêu ở trên. Ở bước này nhóm sử dụng một mô hình có tên là PICK model của tác wenwenyu.
7. **Các bước thực hiện**

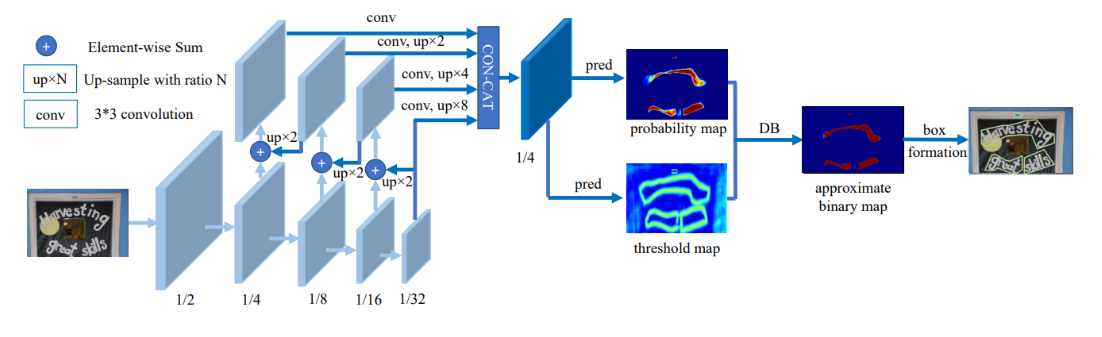
Như đã nói ở trên, thì dưới đây là tổng quan các bước được nhóm tiến hành cho quy trình từ xử lý input, đào tạo mô hình để trích xuất thông tin hóa đơn Tiếng Việt:



# **TEXT DETECTOR**

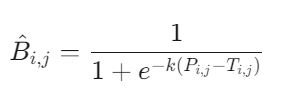
Phương pháp: Nhị phân hóa khác biệt - Differentiable Binarization (DB)

Kiến trúc mô hình DB:



*Nguồn:* [*Real-time Scene Text Detection with Differentiable Binarization (arxiv.org)*](https://arxiv.org/pdf/1911.08947v2.pdf)

Khác với những phương pháp xử lý trước đây, DB có cơ chế đặt ngưỡng tự động để phân biệt đâu là điểm ảnh, đâu là điểm ảnh đối tượng bằng cách sử dụng hàm để tính giá trị cho ma trận bản đồ xác suất là:



Trong đó B là bản đồ nhị phân, T là ngưỡng động được học từ mô hình, k là hệ số khuếch đại, P là bản đồ xác suất. Đây là một hàm khả vi trên tập xác định, do đó có thể đạo hàm và do đó có thể tối ưu so với mô hình.

Kết thúc bước này, nhóm thu được kết quả là các vùng chứa text (dưới dạng tọa độ bounding box có trong ảnh).

1. **ROTATION CORRECTOR & TEXT LINE ROTATION**

Phương pháp: chỉ sử dụng các phương pháp xử lý ảnh cơ bản cho phần này.

Cách triển khai: Sử dụng kết quả của bước trước để tìm ra box có độ dài lớn nhất, sau đó tiến hành xoay ảnh theo box đó.



1. **OPTICAL CHARACTER RECOGNITION - OCR**

Phương pháp: sử dụng pre-train từ open source VietOCR.

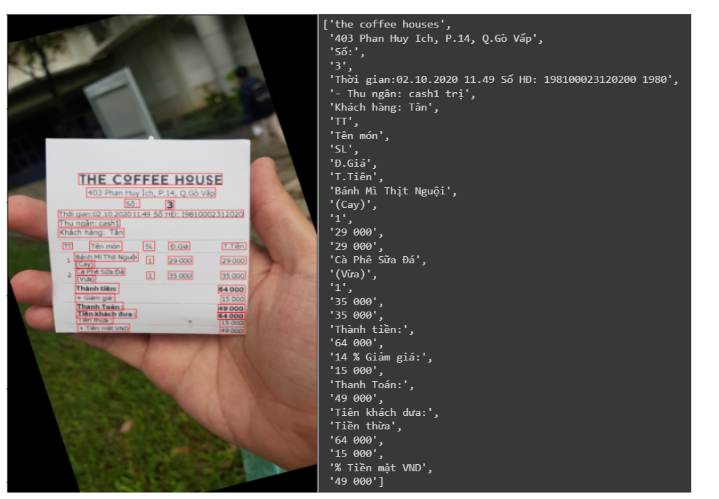
**VietOCR** là một thư viện mã nguồn mở được public bởi anh Phạm Quốc. Cung cấp 2 kiến trúc trong việc nhận dạng tiếng việt: VGG19 kết hợp transformer và VGG19 kết hợp seq2seq đã được huấn luyện trên tập dữ liệu 10m ảnh.

Lý do sử dụng VietOCR

* Mô hình đã được huấn luyện trên tập dữ liệu 10 triệu ảnh nên độ chính xác đạt được ở mức cao mà không cần phải xây dựng hoặc huấn luyện lại từ đầu.
* Là một trong những mô hình nhận dạng tiếng việt tốt nhất hiện nay.
* Tốc độ mô hình nhanh, dễ sử dụng

Sau khi có được kết quả (ảnh) từ bước 3 (rotation corrector & text line rotation), ảnh sẽ được dùng làm input của mô hình OCR.

Kết quả sau bước OCR là thông tin dạng văn bản được in trên hóa đơn



1. **KEY INFORMATION EXTRACTION - KIE**

Phương pháp: Sử dụng PICK (Graph Learning + Convolutional Network)

Mục đích: trích xuất văn bản của một số trường chính từ dữ liệu đã cho và lưu chúng vào các tài liệu có cấu trúc.

PICK : **P**rocessing Key **I**nformation Extraction from Documents using Improved Graph Learning-**C**onvolutional Networ**K**s là một mô hình kết hợp Graph Learning module tinh chỉnh hiệu quả, giúp mô hình có thể nắm được quan hệ giữa các vùng cần trích xuất, từ đó giúp mô hình phân loại hiệu quả và chính xác hơn. PICK còn sử dụng rất nhiều feature của đối tượng bao gồm text, image, position để làm tăng khả năng biểu diễn của chúng, giúp PICK có hiệu năng tốt hơn.

Ở bước này, sau khi thu được thông tin các trường text ( bao gồm text + bounding box) ở bước 5, và các ảnh sau khi đã được tiền xử lý ( ở bước 4), nhóm tiến hành training và đánh giá mô hình dựa trên bộ dataset đã chia.

# **III. Kết quả**

Kết quả của bài toán được nhận xét dựa trên 3 tiêu chí:

* mean entity recall (mER) = True Positives / (True Positives + False Negatives)
* mean entity precision (mEP) = True Positives / (True Positives + False Positives)
* mean entity F\_1 (mEF) = 2 \* (Precision \* Recall) / (Precision + Recall)

Kết quả đánh giá trên tập test:

|  | mER | mEP | mEF |
| --- | --- | --- | --- |
| Tổng | 0.881204 | 0.781904 | 0.828589 |

Một số ví dụ về kết quả thu được:



# **IV. Kết Luận (Nhận xét + hướng phát triển )**

Nhìn chung, bài toán trích xuất thông tin từ ảnh hóa đơn là bài toán sử dụng các mô hình kết hợp với nhau, ở đó kết quả của mô hình trước sẽ được dùng làm đầu vào cho mô hình sau. Do đó kết quả của một quá trình đều ảnh hưởng tới đầu ra cuối cùng của bài toán.

Trong phạm vi đề tài, vì nhóm sử dụng bộ dataset có thể còn nhiều nhiễu và không đa dạng các loại hóa đơn từ nhiều cửa hàng (đa số là hóa đơn của cửa hàng Vinmart) nên các mô hình thu được vẫn còn nhiều hạn chế.

Nhận ra những khuyết điểm đó, nhóm có đề xuất để phát triển tiếp bài toán mang lại kết quả tốt hơn, nhóm cần thu thập đa dạng hơn các loại hóa đơn từ nhiều nguồn, ngoài ra, cần phải tiếp cận, thử thêm nhiều phương pháp để chọn ra phương pháp tốt nhất có thể.

Qua bài toán này, nhóm đã học thêm được nhiều các mô hình xử lý và kết hợp chúng thành một quy trình để giải quyết các bài toán cụ thể, một việc có thể được ứng dụng rất nhiều trong tương lai.

Nguồn tài liệu tham khảo:

<https://viblo.asia/p/bai-toan-phat-hien-chu-text-detection-va-mo-hinh-db-phan-1-3Q75wNm9lWb>

<https://viblo.asia/p/information-extraction-trong-ocr-la-gi-phuong-phap-nao-de-giai-quyet-bai-toan-yMnKMjzmZ7P>

<https://sci-hub.se/10.1109/ICDAR.2019.00244>

<https://learnopencv.com/optical-character-recognition-using-paddleocr/#PP-OCR->

<https://arxiv.org/pdf/2004.07464.pdf>