ĐẠI HỌC QUỐC GIA TP.HCM
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

----------□□□----------

BÁO CÁO MÔN HỌC:

CSC15006 – NHẬP MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

OCR CHỮ NÔM VỚI YOLO

Lớp: 22CNTThuc

Giảng viên hướng dẫn:

- PGS.TS. Đinh Điền
- TS. Nguyễn Hồng Bửu Long
- TS. Lương An Vinh

Sinh viên thực hiện:

Nguyễn Thị Ngọc Trang	22127421
Phan Lê Đức Anh	22127020
Trần Hoàng Linh	22127233
Nguyễn Gia Phúc	22127482

TP. Hồ Chí Minh, ngày 4 tháng 1 năm 2025

LỜI CẢM ƠN

Nhóm chúng em xin gửi lời chân thành cảm ơn nhất đến các thầy và các anh chị đã hướng dẫn chúng em trong quá trình thực hiện đồ án lần này.

Chúng em trân trọng sự hướng dẫn tận tình của thầy Đinh Điền trong cả một quá trình vừa qua. Mặc dù thầy rất bận bịu với công việc nhưng thầy vẫn cố gắng sắp xếp thời gian để hướng dẫn cho nhóm.

Chúng em cám ơn sự hỗ trợ và tiếp sức thầm lặng từ thầy Lương An Vinh trong quá trình thực hiện.

Chúng em cũng trân quý những kiến thức quý báu thầy Nguyễn Hồng Bửu Long đã truyền đạt trên lớp để giúp chúng em có một nền tảng vững chắc thực hiện đồ án lần này.

Những sự hỗ trợ chu đáo từ các anh chị hướng dẫn cũng là một yếu tố không thể thiếu giúp cho chúng em hoàn thành đồ án này một cách tốt đẹp.

Mặc dù quá trình thực hiện là rất gian nan và nhiều chông gai nhưng nhóm đã vượt qua và đi đến được kết quả cuối cùng. Tất cả cũng một phần không nhỏ nhờ vào sự hỗ trợ của các thầy và anh chị.

MỤC LỤC

DANH SÁCH HÌNH

I. TỔNG QUAN

1. Giới thiệu đề tài

- OCR (hay Optical Character Recognition) là công nghệ chuyển đổi hình ảnh văn bản thành văn bản số hóa, và công nghệ OCR được ứng dụng rất nhiều trong đời sống của chúng ta từ những tiện ích như google dịch hay sao chép văn bản nhanh chóng.
- Nhận thấy được tính ứng dụng cao của công nghệ OCR để giúp con người chúng ta tìm hiểu, học tập được sâu sắc và nhanh chóng hơn về văn hóa, lịch sử của những thế hệ cha ông, tổ tiên đi trước; người Việt Nam chúng ta đã ứng dụng và tìm hiểu công nghệ này nhanh chóng. Từ đó, việc tìm hiểu các công cụ có sẵn giúp hiện thực hóa việc OCR chữ Nôm là một việc thiết yếu. Đề tài OCR chữ Nôm với các mô hình hay công cụ khác nhau là một để tài thú vị để chúng ta tìm hiểu kỹ hơn về các công cụ tiềm năng có thể sử dụng để giúp quá trình tìm hiểu về văn hóa, văn học chữ Nôm ngày xưa của chúng ta được thuận lợi hơn, giúp số hóa các tài liệu cổ, tạo điều kiện cho việc nghiên cứu ngôn ngữ học và văn học.
- Vậy, tại sao chúng ta lại chọn YOLO?
    - Tốc độ xử lý của mô hình nhanh
    - Hiệu quả cao để nhận diện từng chữ một thay vì cả một câu như các mô hình để nhận diện chữ Nôm khác
- Mục tiêu:
    - Giúp xây dựng hệ thống nhận diện chữ Nôm
    - Các dữ liệu còn có thể được dùng để huấn luyện các mô hình khác

2. Mức độ và phạm vi thực hiện

- Mức độ: tập trung vào xây dựng dữ liệu và tìm hiểu các mô hình huấn luyện bằng hình ảnh. Đặc biệt, khác với các đề tài OCR chữ Nôm mà những người đi trước đã thực hiện, nhóm chọn thực hiện OCR hình ảnh ở mức ký tự trong khi các đề tài trước thực hiện ở mức câu.
- Phạm vi dữ liệu: các tài liệu cổ bằng chữ Nôm như:
    - Sách Nôm công giáo 1995 – 078 – Thiên Chúa Thánh Giáo Khải Mông
    - Sách Nôm công giáo 1995 – 090 – Thiên Chúa Thánh Giáo Hối Tội Kinh
    - Sách Nôm công giáo 1995 – 081 – Dọn mình trước chịu Cô-mô-nhong
    - Sách Nôm công giáo 1995 – 080 – Những điều ngắm trong các ngày lễ trọng
- Ngoài ra, do trong quá trình thực hiện gán nhãn, bộ dữ liệu trên có kích thước nhỏ và có độ chính xác không cao nên nhóm bổ sung bộ ngữ liệu được lấy từ đề tài NomNaOCR [1]:
    - Lục Vân Tiên
    - Truyện Kiều bản 1866
    - Truyện Kiều bản 1871
    - Truyện Kiều bản 1872
    - Đại Việt Sử Ký Toàn Thư – Quyển Thủ
    - Đại Việt Sử Ký Toàn Thư – Ngoại kỷ toàn thư

- Đại Việt Sử Ký Toàn Thư – Bản kỷ toàn thư
- Đại Việt Sử Ký Toàn Thư – Bản kỷ thực lục
- Đại Việt Sử Ký Toàn Thư – Bản kỷ tục biên

II. BỘ DỮ LIỆU

1. Khái quát chung

- Các dữ liệu được cung cấp đóng vai trò rất quan trọng trong việc huấn luyện mô hình OCR chữ Nôm, vì chữ Nôm có nhiều đặc điểm phức tạp về mặt cấu trúc cũng như hình dạng ký tự nên bộ dữ liệu được dùng để huấn luyện cần phải phong phú và được xử lý cẩn thận
- Lĩnh vực dữ liệu được chọn cho đồ án này là về mặt tôn giáo, cụ thể là các sách Nôm công giáo, để có thể huấn luyện cho mô hình nhận dạng được các ký tự ở trong lĩnh vực tương tự. Hơn nữa, bộ dữ liệu được bổ sung thêm các tác phẩm chữ Nôm khác trong lịch sử làm cho vốn từ của bộ dữ liệu được phong phú hơn:
    - Sách Nôm công giáo 1995 – 078 – Thiên Chúa Thánh Giáo Khải Mông
    - Sách Nôm công giáo 1995 – 090 – Thiên Chúa Thánh Giáo Hối Tội Kinh
    - Sách Nôm công giáo 1995 – 081 – Dọn mình trước chịu Cô-mô-nhong
    - Sách Nôm công giáo 1995 – 080 – Những điều ngắm trong các ngày lễ trọng
    - Lục Vân Tiên
    - Truyện Kiều bản 1866
    - Truyện Kiều bản 1871
    - Truyện Kiều bản 1872
    - Đại Việt Sử Ký Toàn Thư – Quyển Thủ
    - Đại Việt Sử Ký Toàn Thư – Ngoại kỷ toàn thư
    - Đại Việt Sử Ký Toàn Thư – Bản kỷ toàn thư
    - Đại Việt Sử Ký Toàn Thư – Bản kỷ thực lục
    - Đại Việt Sử Ký Toàn Thư – Bản kỷ tục biên

2. Thu thập dữ liệu

- Nguồn dữ liệu: các văn bản Nôm cổ được thu thập từ nhiều nguồn khác nhau như Thư viện Quốc gia Việt Nam, các công trình nghiên cứu chữ Nôm
- Từ các dữ liệu đã được cung cấp, tiến hành xử lý sơ bộ để tăng cường chất lượng hình ảnh như cân bằng sáng, tăng độ tương phản, giảm nhiễu và tiến hành sắp xếp các dữ liệu và gán nhãn dữ liệu lần lượt ở các mức độ câu và ký tự

3. Công cụ hỗ trợ gán nhãn

- Roboflow: giúp trực quan hóa và hiển thị các nhãn dữ liệu được gán, phát hiện và kiểm tra sai sót trong nhãn dữ liệu.
- PPOCRLabel (custom model): OCR các ký tự.
- YOLOv5 (custom model): phát hiện ký tự, lấy bounding box của từng ký tự.

4. Các quy trình gán nhãn:

Do tính chất của các file ngữ liệu thô khác nhau nên nhóm có các quy trình khác nhau để gán nhãn dữ liệu.

Cách thứ nhất:

- Trích xuất hình ảnh từ các file pdf, tách các trang chữ Nôm và chữ Quốc Ngữ thành các folder riêng.
- Từ hình ảnh trích xuất được, tiến hành phát hiện từng ký tự chữ Nôm bằng mô hình YOLOv5, thu được bounding box của từng ký tự. Dữ liệu được lưu dưới dạng file txt và có định dạng YOLOv5 PyTorch TXT (nhưng các thông số không được chuẩn hóa để giảm tính sai số khi chuyển đổi định dạng dữ liệu):

  class_id x_center y_center width height
- Từ các bounding box thu được, tiến hành sắp xếp dựa trên tọa độ theo thứ tự từ trên xuống dưới, từ phải sang trái và xử lý chuyển đổi bounding box này về dạng tọa độ 4 điểm để có thể áp dụng OCR được bằng ứng dụng PPOCRlabel.
- Tiến hành ghi tất cả các bounding box ra một file xlsx với định dạng file như sau:
  - ID: ID của các hình trong tập dữ liệu, có định dạng <Name>.XXX.YYY, trong đó:
    - Name: tên của file dữ liệu thô.
    - XXX: số trang của hình đó trong file dữ liệu thô.
    - YYY: thứ tự của bounding box trong hình.
  - Image_name: Tên file hình được lưu trong máy tương ứng với từng hình của tập dữ liệu.
  - Image_Box: Tọa độ 4 điểm của bounding box.
  - SinoNom_OCR: Được để trống do ta chỉ có thông tin bounding box của từng ký tự chứ chưa nhận dạng được ký tự.
- Dùng công cụ check_label được các thầy hướng dẫn cung cấp để chuyển đổi file xlsx về file Label.txt và fileState.txt để dùng cho công cụ PPOCRLabel.
- Sử dụng công cụ PPOCRLabel để hiển thị bounding box phát hiện được tương ứng với từng hình, xóa các bounding box nhận dạng sai ký tự nếu có (các bounding box nhận dạng dấu mộc của sách, số trang sách)
- Tiến hành OCR từng ký tự chữ Nôm và xuất kết quả OCR ra file Label.txt mới.
- Trích xuất các dữ liệu cần thiết từ file Label.txt, lọc các bounding box nhận dạng nhiều hơn 1 ký tự (do model YOLOv5 có thể phát hiện các bounding box chồng lấn lên nhau làm PPOCRLabel nhận dạng ra 2 ký tự).
- Gióng từng ký tự chữ Nôm với từng chữ Quốc ngữ (thu được bằng cách OCR các trang chữ Quốc ngữ của file sách) bằng thuật toán M.E.D Leveinshtein [2].
- Sau khi có được các cặp chữ Nôm và Quốc ngữ tương ứng cùng với thông số của bounding box, chuyển đổi dữ liệu về dạng YOLOv5 PyTorch TXT với định dạng như sau:

  class_id x_center y_center width height
  - class_id: chữ Quốc ngữ tương ứng với bounding box.
  - x_center, y_center, width, height: các thông số của bounding box theo định dạng YOLOv5 PyTorch TXT (được chuẩn hóa).

- Như thế, ta có được các file dữ liệu được gán nhãn của từng hình tương ứng.

Cách thứ hai:

Dữ liệu được xây dựng dựa trên model phát hiện chữ Hán Nôm và tập dữ liệu của đồ án giữa kỳ vì dữ liệu giữa kì đã được gán nhãn cho từng câu tương đối chính xác.

- Thực hiện dò tất cả tọa độ mà model phát hiện chữ Nôm đã được huấn luyện trước đó.
- Song song với đó lấy toàn bộ nhãn đã được gán ở giữa kì như một khung để đảm bảo chỉ lấy những chữ có tọa độ nằm trong khung này.
- Đảm bảo chữ luôn nằm trong khung giúp giảm thiểu những chữ bị thừa bên ngoài khung như dấu chấm, chữ ghi chú, v.v.
- Nhưng có những chữ được cho là nằm trong khung nhưng bị lệch ra bên ngoài tương đối không nhiều. Nên từ đây có thể lấy những chữ có diện tích của box chữ giao với diện tích của box câu sao cho phần giao này không vượt quá 50% diện tích của box chữ.
- Sắp xếp lại tọa độ của các box chữ từ trên xuống dưới.
- Từ file *result.xlsx* giữa kì lấy từng câu Quốc ngữ tương ứng với từng box câu.
- Lấy ra từng chữ trong câu Quốc ngữ tương ứng với box câu để gán nhãn cho box chữ.
- Chuyển đổi tất cả các tọa độ sang format của YOLOv5 PyTorch TXT.

Cách thứ ba:

- Đầu tiên, thực hiện trích xuất hình ảnh từ các file pdf, tách các trang chữ Nôm và chữ Quốc Ngữ thành các folder riêng.
- Từ hình ảnh trích xuất được, tiến hành phát hiện từng ký tự chữ Nôm bằng mô hình YOLOv5, thu được bounding box của từng ký tự. Dữ liệu được lưu dưới dạng file txt và có định dạng YOLOv5 PyTorch TXT (nhưng các thông số không được chuẩn hóa để giảm tính sai số khi chuyển đổi định dạng dữ liệu):

  class_id x_center y_center width height
- Tạo project trên Roboflow, tiến hành upload data lên để kiểm tra khả năng phát hiện kí tự của mô hình YOLOv5. Ta nhận thấy với ngữ liệu hiện tại mô hình nhận diện chồng box, thiếu box ở khu vực tối và nhận diện cả những kí tự dư thừa, hoặc bị xoay và mất trang.
- Sau đó, chia ngữ liệu Nôm ra các folder khác nhau để xử lí theo tính chất của từng ảnh, thực hiện các kỹ thuật tiền xử lí ngữ liệu như: xoay trang theo đúng chiều, cắt ảnh để giảm thiểu độ nhiễu do các nội dung khác không liên quan, xử lý độ sáng và độ tương phản theo thông số phù hợp với ảnh để lấy được cả vùng tối và vùng sáng của ảnh, tăng giảm kích thước ảnh.
- Dùng mô hình YOLOv5 đã được train sẵn để lấy bounding box của từng kí tự.
- Tải lại bộ ngữ liệu lên Roboflow để kiểm tra các bounding box và nhận thấy đã xử lí được khá tốt vấn đề chồng box, hạn chế các kí tự không liên quan và nhận diện được tốt hơn các vùng dữ liệu.

- Với các bounding box tốt hơn thu được, tiến hành sắp xếp dựa trên tọa độ theo thứ tự từ trên xuống dưới, từ phải sang trái. Tuy nhiên, do ngữ liệu có số lượng các box nghiêng khá lớn nên chọn một thông số "tolerance" phù hợp với bộ ngữ liệu để lấy được nhiều box trong một câu, giúp tăng độ chính xác khi gióng hàng các chữ. Tiếp tục dùng Roboflow để kiểm tra ngữ liệu đến khi đạt yêu cầu tốt nhất có thể.
- Tiếp tục gán các class cho các box chữ theo thứ tự đi từ mức trang, mức câu rồi đến mức chữ như cách gióng hàng để tô màu chữ cho ngữ liệu giữa kì vì ngữ liệu giữa kì đã được gán nhãn khá chính xác (được tiền xử lí nhiễu hiệu quả nên chỉ trừ những trường hợp đặt biệt hoặc những trường hợp do công cụ OCR chữ Quốc ngữ không chính xác dẫn đến việc không khớp ở một số dòng). Cụ thể, bộ ngữ liệu giữa kì được xử lí gióng hàng như sau :
    - Gióng hàng mức trang: Chia các box trong một ảnh và đánh số trang Nôm ngược chiều với số trang Quốc ngữ. Đồng thời trong quá trình OCR chữ Quốc ngữ, việc cắt ảnh có để lại số trang (được đưa vào đánh dấu header cho nội dung thuộc một trang), ta gióng hàng với trang được đánh số tương ứng trong các trang chữ Nôm được OCR.
    - Gióng hàng mức dòng: Gióng dựa trên độ dài các box nếu dữ liệu bị mất box câu. Nếu dữ liệu không bị mất box câu thì có thể gióng trực tiếp để đảm bảo câu được khớp hơn.
    - Gióng hàng mức chữ: dùng thuật toán M.E.D Leveinshtein để gióng hàng chữ Nôm trong câu với chữ Quốc ngữ trong câu dựa vào hai từ điển SinoNom_similar_Dic và QuocNgu_SinoNom_Dic được cung cấp.
- Sau khi gióng hàng chữ, ta dùng các box chữ đã được xếp theo câu gióng hàng theo thứ tự với cặp Quốc ngữ và Nôm.
- Cuối cùng ta được nhãn là chữ Quốc ngữ tương ứng với box chữ Nôm
- Tạo file yaml tổng hợp các class chữ Quốc ngữ rồi đánh số thứ tự để gán đúng class id cho dataset theo cấu trúc của YOLOv5 PyTorch TXT.
- Kiểm lại bộ ngữ liệu một lần nữa và có thể điều chỉnh trong Roboflow nếu cần thiết.

5. Các khó khăn và cách xử lý

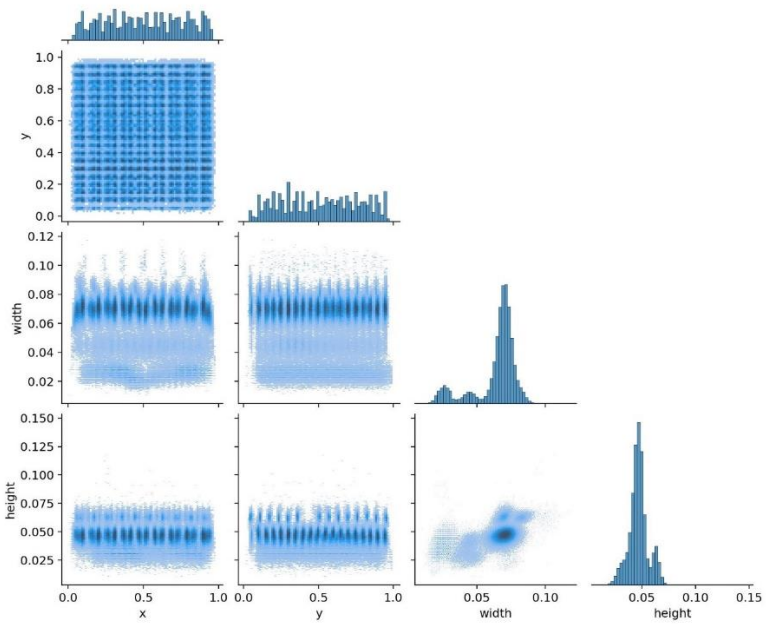
- Các khó khăn:
    - Hình ảnh nhiễu, mờ.
    - Ánh sáng không đều dẫn đến các khu vực tối màu không thể OCR hoặc OCR thiếu sót.
    - Nhiều ký tự có cấu trúc tương đồng làm ảnh hưởng đến quá trình xác định.
    - Ngữ liệu thô là các tài liệu chữ Nôm viết tay nên dẫn đến các trường hợp như sau:
        - Các ký tự Nôm bị viết dính lại với nhau làm cho model phát hiện chữ hoạt động với độ chính xác thấp.
        - Các ký tự Nôm viết tay xấu dẫn đến việc nhận dạng chữ với độ chính xác thấp.
        - Phông chữ khác nhau làm tăng độ khó của việc nhận diện.

        - Chữ Quốc ngữ được trích xuất bằng công cụ OCR nên độ chính xác bị giảm đi đáng kể.
    - Kỹ thuật gióng hàng dữ liệu còn chưa hiệu quả dẫn đến độ chính xác của bộ ngữ liệu không cao.
    - Chữ Nôm là một ngôn ngữ cổ dẫn đến hạn chế về số lượng tài liệu cũng như các công cụ để giúp kiểm tra độ chính xác

  Ta có thể thấy rằng độ chính xác của bộ ngữ liệu ở mỗi bước trong quy trình gán nhãn bị giảm đi đáng kể. Do đó, bộ ngữ liệu được gán nhãn cuối cùng có độ chính xác không cao và bị mất dữ liệu khá nhiều.
- Cách xử lý:
    - Áp dụng các kỹ thuật giúp tăng cường độ chính xác khi nhận diện như xoay, phóng to, cắt ảnh.
    - Sử dụng các kỹ thuật xử lý ảnh như cân bằng sáng, giảm nhiễu và làm mịn hình ảnh.
    - Chuẩn bị bộ ngữ liệu chữ Quốc ngữ với độ chính xác cao nhất có thể để bù cho độ chính xác thấp của bộ ngữ liệu chữ Nôm:
        - Thử các công cụ OCR hiệu quả hơn.
        - Sau khi OCR, bộ ngữ liệu được tinh chỉnh bằng công cụ sửa lỗi chính tả để cho ra bộ ngữ liệu đúng chính tả.
        - Kiểm tra thủ công bộ ngữ liệu.

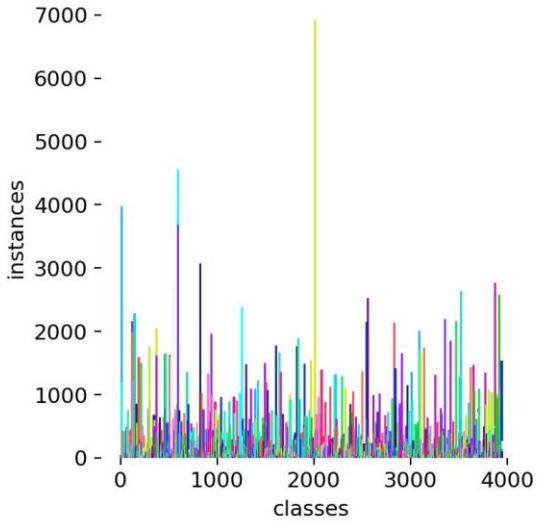
6. Tổng quan bộ dữ liệu gán nhãn.

- Tổng số hình ảnh: 3294
- Tổng số chữ Quốc ngữ được gán nhãn (Vocabularies): 3945
- Tương quan dữ liệu:



Hình 1. Biểu đồ thể hiện sự tương quan các giá trị trong bộ dữ liệu

- Thống kê các class_id:



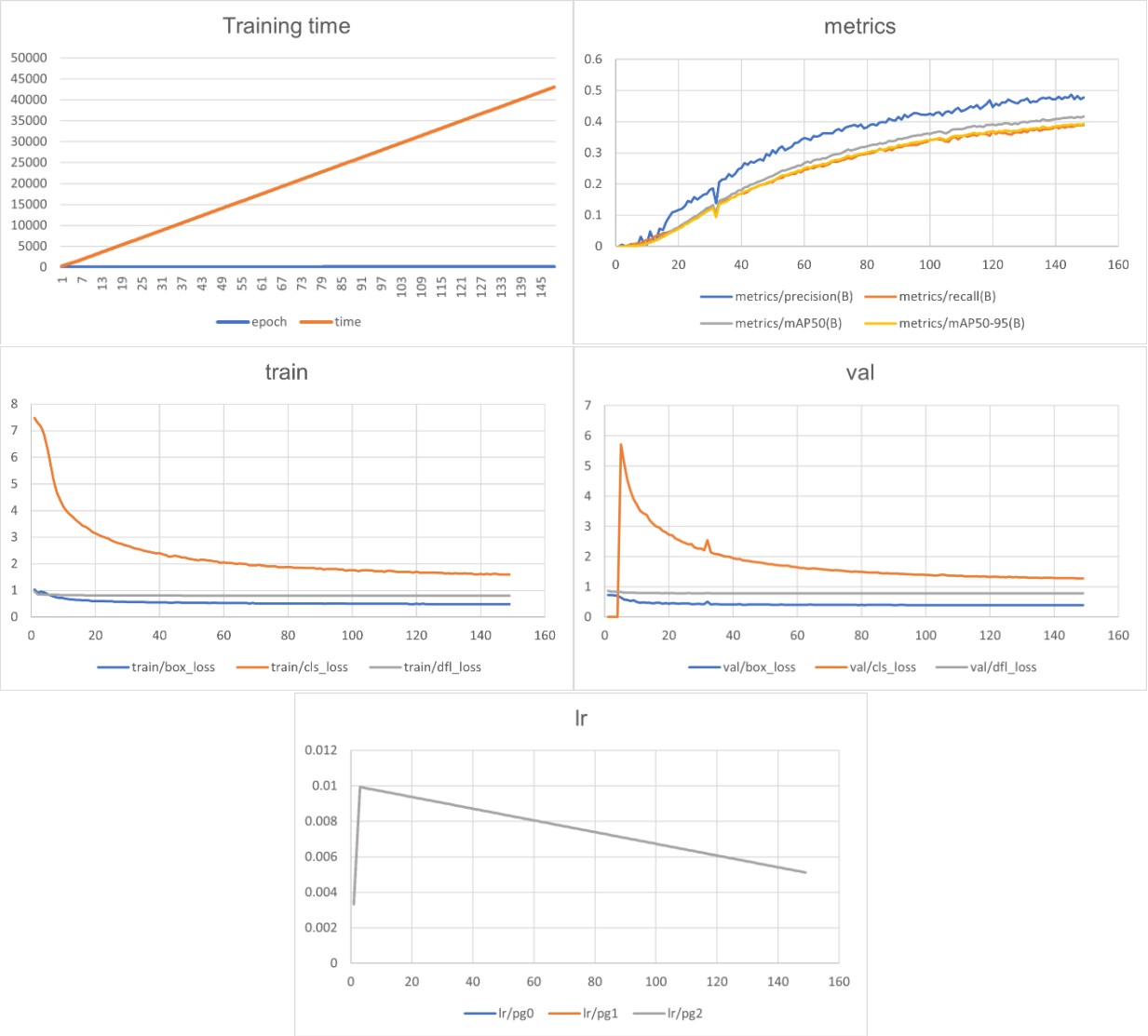
Hình 2. Biểu đồ thể hiện số lượng của từng class_id

7. Khái quát mô hình

- YOLOv11 [3] là một phiên bản cải tiến của dòng mô hình YOLO, được thiết kế để thực hiện tác vụ phát hiện đối tượng trong thời gian thực với độ chính xác và tốc độ vượt trội.
- Những điểm nổi bật chính của YOLOv11 bao gồm:
    - **Kiến trúc cải tiến**: Tận dụng các khối convolutional hiện đại và cơ chế attention tiên tiến để cải thiện khả năng trích xuất đặc trưng từ hình ảnh.
    - **Tăng độ chính xác**: Sử dụng các phương pháp tối ưu hóa mới, chẳng hạn như phân tầng feature maps hoặc loss function tinh chỉnh, giúp cải thiện độ chính xác trong việc dự đoán vị trí và phân loại đối tượng.
    - **Hiệu năng thời gian thực**: Mặc dù được cải tiến phức tạp hơn, YOLOv11 vẫn giữ được tốc độ xử lý nhanh, đáp ứng tốt các yêu cầu ứng dụng trong thời gian thực.
    - **Hỗ trợ nhiều kích thước đối tượng**: Với các cải tiến trong kiến trúc backbone và head, YOLOv11 hoạt động tốt trên cả các đối tượng nhỏ, trung bình và lớn trong cùng một hình ảnh.

8. Huấn luyện mô hình

- Mô hình YOLOv11 được nhóm huấn luyện trên nền tảng Kaggle.
- Kết quả huấn luyện mô hình:



Hình 3. Kết quả huấn luyện mô hình.

Learning rate (lr)

- Learning rate ban đầu cao (khoảng 0.01) giúp mô hình học nhanh trong giai đoạn đầu.
- Sau khoảng 20 epoch, learning rate giảm dần tuyến tính, xuống mức khoảng 0.004 ở epoch cuối. Đây là một chiến lược scheduler giảm tuyến tính (linear decay), phổ biến trong việc tối ưu hóa mô hình.
- Các nhóm tham số khác nhau (pg0, pg1, pg2) được áp dụng learning rate đồng bộ, cho thấy không có sự ưu tiên đặc biệt nào giữa các nhóm tham số.

Giai đoạn đầu, learning rate cao giúp mô hình nhanh chóng học được các thông số cơ bản. Khi loss giảm dần, việc giảm learning rate giúp mô hình tránh overshooting (nhảy qua giá trị tối ưu) và tinh chỉnh các thông số tốt hơn.

Loss

- Các đồ thị train/val loss (box_loss, cls_loss, dfl_loss) cho thấy quá trình hội tụ ổn định của mô hình.

- Loss (train/val): Loss bắt đầu ở mức khoảng 1.0, giảm nhanh chóng trong 20 epoch đầu, sau đó duy trì ở mức thấp dưới 0.5. Đây là loss dùng để tối ưu hóa vị trí và kích thước bounding box, phản ánh khả năng mô hình học cách định vị đối tượng tốt.
- Classification Loss (train/val): Ban đầu rất cao (khoảng 7.0), sau đó giảm đều và đạt mức 2.0 sau khoảng 80 epoch, ổn định dần về sau. Đây là chỉ số quan trọng vì nó đánh giá khả năng phân loại chính xác của mô hình. Việc loss giảm chứng tỏ mô hình dần hiểu được cách phân loại các đối tượng.
- DFL Loss (train/val): DFL Loss bắt đầu ở mức thấp (khoảng 0.5) và hầu như không thay đổi nhiều trong toàn bộ quá trình huấn luyện. Điều này cho thấy mô hình đã tối ưu tốt khía cạnh này ngay từ đầu, nhấn mạnh hiệu quả của mô hình với loss này.
- Sự đồng nhất giữa train/val loss: Train và val loss gần như đồng nhất ở mọi giai đoạn, cho thấy mô hình không bị overfitting hay underfitting.

Metrics

- Precision: Precision tăng ổn định, đạt mức 0.55 ở cuối quá trình huấn luyện. Mô hình có khả năng dự đoán chính xác các bounding box thuộc về đối tượng thực tế, tức là mô hình ít đưa ra dự đoán sai (false positives thấp).
- Recall: Recall tăng dần, đạt mức 0.45 ở epoch cuối. Mô hình phát hiện được 45% số đối tượng thực tế có mặt trên ảnh. Tuy nhiên, Recall thấp hơn Precision, điều này nghĩa là mô hình bỏ sót một số đối tượng (false negatives).
- mAP50: mAP50 (Mean Average Precision tại IoU = 0.50) tăng đều, đạt 0.50 ở cuối. Mô hình dự đoán chính xác bounding boxes với mức độ chồng lấp (IoU) tối thiểu là 50%.
- mAP50-95: mAP50-95 tăng ổn định, đạt 0.40 ở epoch cuối. Đây là chỉ số khó đạt được hơn vì nó đòi hỏi mô hình phải đạt hiệu quả trên nhiều mức IoU khác nhau (từ 0.50 đến 0.95). Chỉ số này đạt 0.40 cho thấy mô hình có khả năng tổng quát hóa tốt.

Thời gian huấn luyện

- Thời gian huấn luyện trung bình cho mỗi epoch ổn định, chứng tỏ tài nguyên phần cứng và quá trình huấn luyện không gặp vấn đề về bottleneck.
- Biểu đồ thời gian phản ánh tính đồng nhất và ổn định của môi trường huấn luyện.

Tổng hợp đánh giá

Ưu điểm:

- Hội tụ tốt: Loss giảm đều, ổn định và không có dấu hiệu overfitting.
- Hiệu suất cao: Precision (0.55) và mAP50 (0.50) cao, phù hợp với nhiều ứng dụng thực tế.
- Chiến lược learning rate hợp lý: Điều chỉnh giảm dần learning rate giúp mô hình hội tụ ổn định.

Hạn chế:

- Recall thấp hơn Precision: Một số đối tượng bị bỏ sót, gây ảnh hưởng đến khả năng bao quát của mô hình.
- Thời gian hội tụ lâu: Cần tối ưu thêm để rút ngắn thời gian huấn luyện.

Cải thiện

Tăng Recall:

- Sử dụng dữ liệu đa dạng hơn, tập trung vào các mẫu khó nhận diện (đối tượng nhỏ, bị che khuất).
- Điều chỉnh trọng số của classification loss (cls_loss) để tăng ưu tiên phát hiện các đối tượng.

Tối ưu thời gian huấn luyện:

- Áp dụng mixed-precision training để giảm thời gian tính toán.
- Sử dụng distributed training nếu có nhiều GPU.

Tăng dữ liệu:

- Bổ sung các tập dữ liệu từ nhiều nguồn khác nhau để cải thiện khả năng tổng quát hóa.

Hyperparameter Tuning:

- Kiểm tra các giá trị khác nhau cho learning rate, batch size, và các tham số khác để tìm cấu hình tốt nhất.

III. ĐÁNH GIÁ VÀ KẾT QUẢ

1. Phương pháp đánh giá

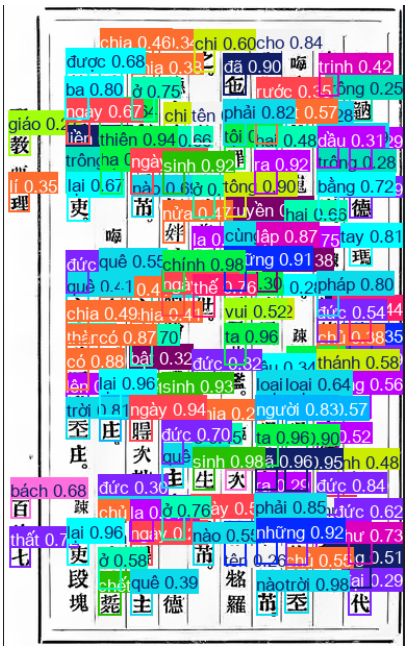
- Dùng model huấn luyện để dự đoán kết quả trên một ảnh bất kỳ có cùng tính chất với các ảnh trong dataset.
- Kết quả cho ra class và bounding box của các ký tự. Ta trích xuất các thông số ra để xử lý.
- Tiến hành sắp xếp các bounding box dựa trên tọa độ theo thứ tự từ trên xuống dưới, từ phải sang trái.
- Sau khi sắp xếp các bounding box, ta lấy ra class là các chữ Quốc ngữ dự đoán được để kiểm tra.
- Tiến hành dùng thuật toán M.E.D Leveinshtein để gióng hàng các chữ Quốc ngữ dự đoán với các chữ Quốc ngữ trong văn bản gốc của ảnh đánh giá.
- Đếm số cặp mà 2 chữ đều giống nhau và đếm tổng số chữ trong văn bản Quốc ngữ gốc, ta được tỉ lệ chính xác của kết quả dự đoán.

2. Kết quả thử nghiệm

- Dùng một bộ ngữ liệu bên ngoài để test mô hình, ta được kết quả như các hình bên dưới:

Kết quả phát hiện được tổng 81/168 kí tự được gán chính xác cho hình trên.

Kết quả phát hiện được tổng 67/160 kí tự được gán chính xác cho hình trên.



Kết quả phát hiện được tổng 111/224 kí tự được gán chính xác cho hình trên.

IV. THAM KHẢO

[1] N. Đ. D. A. Đặng Hoàng Quân. [Trực tuyến]. Available: https://github.com/ds4v/NomNaOCR.

[2] “Introduction to Levenshtein distance,” 31 1 2024. [Trực tuyến]. Available: https://www.geeksforgeeks.org/introduction-to-levenshtein-distance/.

[3] “ultralytics,” [Trực tuyến]. Available: https://github.com/ultralytics/ultralytics.