FPT University HCMC

Hỗ trợ robot hiểu ngôn ngữ con người và thực hiện các tác vụ đa dạng, phức tạp

Mentor: Anh Lương Hữu Dũng

Sinh viên thực hiện:

Đào Khang— SE183427

Huỳnh Khánh Quang— SE183777

—

MUC LUC

DANH SÁCH HÌNH ẢNH

1. Lời nói đầu

Trong bối cảnh cách mạng và AI phát triển nhanh chóng, robot không chỉ có khả năng thực thi nhiệm vụ vật lý, mà ngày càng được trêng bị thêm khả năng nhận diện ngôn ngữ tự nhiên và tương tác linh hoạt với con người. Mục tiêu của dự án này là xây dựng một hệ thống robot có thể nhận lệnh ngôn ngữ tiếng Việt tự nhiên, dịch sang tiếng Anh, trích xuất thông tin về đối tượng và vị trí, sau đó nhận dạng và segment đối tượng trong ảnh dựa theo YOLOv8-seg để rồi đưa qua hệ thống Camera 3D và robot cụ thể ở đây là AUBO để thực hiện mệnh lệnh đã được đề ra

2. Mục tiêu và Vấn đề

2.1. Mục tiêu

- Xây dựng hệ thống nhận lệnh ngôn ngữ tiếng Việt tự nhiên.
- Dịch sang tiếng Anh để phù hợp với các mô hình xử lý NLP.
- Trích xuất đối tượng và vị trí mong muốn.
- Áp dụng YOLOv8 segmentation để nhận dạng và tô màu phân đoạn vật thể tương ứng trong ảnh và xác định trọng tâm vật ấy
- Kết hợp cùng với hệ thống Camera 3D và Robot để thực hiện nhiệm vụ

2.2. Vấn đề

- Tiếng Việt có ngữ pháp phức tạp đối với xử lý NLP.
- Trình phiên dịch Tiếng Anh còn nhiều cấu trúc chưa thể giống Tiếng Việt
- Kết hợp nhiều module: nhận diện giọng nói, dịch thuật, NLP, và Computer Vision.
- Dữ liệu ảnh thực tế có nhiều nhiễu tạp, vị trí đối tượng phức tạp.
- Chưa xử lý được những hình ảnh 3D

3. Phương pháp

3.1. Model

3.1.1. YOLOv8x-seg

Dự án sử dụng mô hình YOLOv8x-seg, là phiên bản lớn nhất của dòng mô hình YOLOv8 chuyên cho nhiệm vụ segmentation (phân vùng đối tượng trong ảnh). YOLOv8x-seg có khả năng phát hiện vật thể với độ chính xác cao, hỗ trợ segment chính xác từng pixel và hoạt động tốt trong thời gian thực.

3.1.2. spacy

Ngoài ra, dự án cũng sử dụng mô hình spaCy, là một mô hình NLP nhẹ nhưng hiệu quả, được huấn luyện để nhận biết cấu trúc ngữ pháp tiếng Anh như thực thể, cụm danh từ, động từ, giới từ, vị trí,...

3.2. Danh sách các hàm chính

3.2.1. def speech_to_text()

- **Mục đích**: Ghi âm giọng nói người dùng bằng micro, nhận diện nội dung tiếng Việt, sau đó dịch sang tiếng Anh.
- **Thư viện**: speech_recognition, deep_translator
  - **recognizer.listen()**: Dùng để ghi âm
  - **recognizer.recognize_google()**: Dùng để chuyển giọng nói thành văn bản tiếng Việt
  - **GoogleTranslator.translate()**: Dùng để dịch văn bản tiếng Việt vừa nhận sang tiếng Anh

3.2.2. def extract_object_and_location(user_input)

- **Mục đích**: Dùng thư viện NLP spaCy để trích xuất thông tin từ câu văn: đối tượng (object) và vị trí (location).
- **Thư viện**: spacy
  - **nlp(user_input)**: Dùng để phân tích cú pháp câu
  - **token.text.lower() in valid_locations**: Dùng để xác định từ chỉ vị trí
  - **doc.noun_chunks**: Dùng để tìm cụm danh từ để xác định object

3.2.3. def detect_segment_object(image_path, target_object, target_location)

- **Mục đích**: Nhận diện và phân đoạn đối tượng theo label và vị trí chỉ định trong ảnh bằng YOLOv8 segmentation.
- **Thư viện**: cv2, numpy, ultralytics, matplotlib
  - **YOLO(image)**: phát hiện và phân đoạn đối tượng
  - **label == target_object**: Lọc ra đối tượng cần được phân đoạn
  - **cv2.fillPoly(), cv2.rectangle(), cv2.circle()**: Dùng để hiển thị mask, bounding box, trọng tâm
  - **matplotlib.pyplot.imshow()**: Hiển thị hình ảnh kết quả

3.2.4. Main

- **user_input = speech_to_text()**: Ghi âm, dịch, và trả về câu tiếng Anh.
- **extract_object_and_location(user_input)**: Trích object và location từ câu lệnh.
- **detect_segment_object(image_path, object_name, target_location)**: Dựa vào object + location để nhận diện đúng vật thể.

4. Kết quả

4.1. Model gốc không finetune

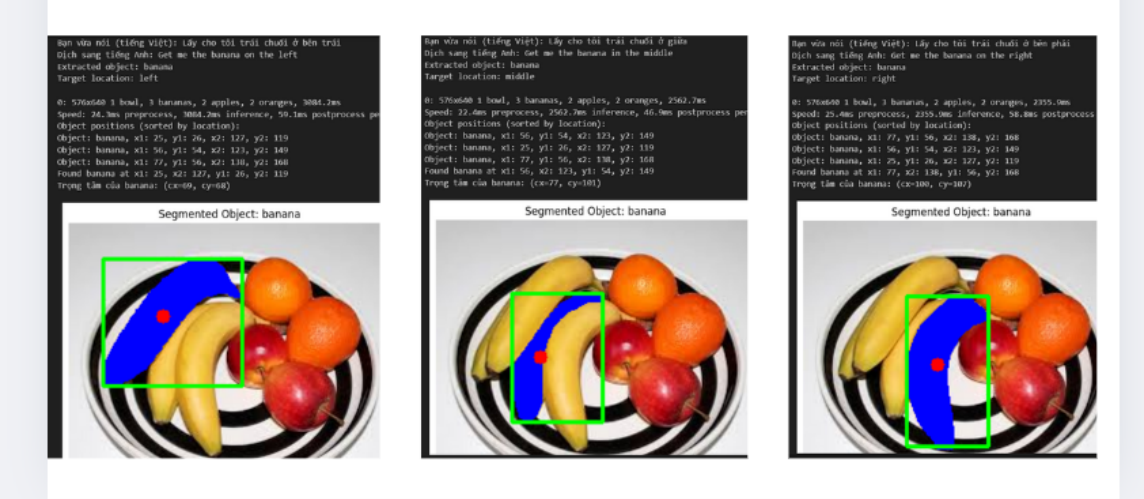

Figura 1: Kết quả của Model YOLOv8x-seg gốc dùng để nhận diện và phân đoạn vị trí trái cây (ở đây là chuối) theo vị trí trái, phải, giữa

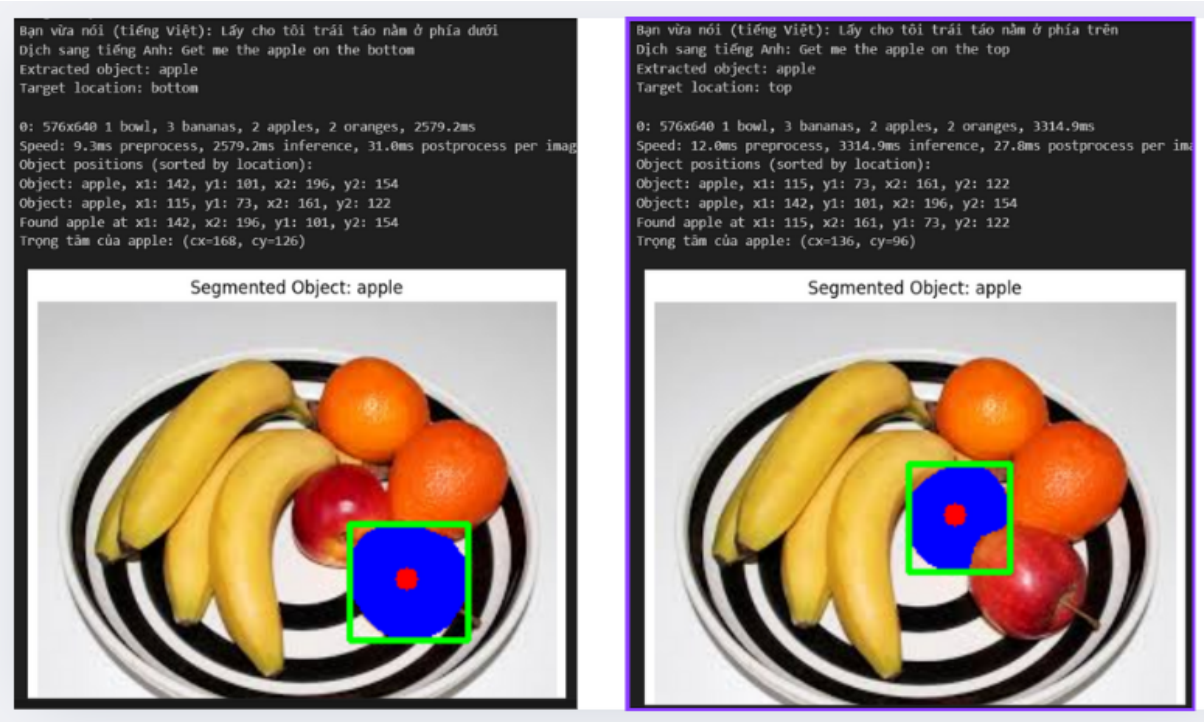

Figura 2: Kết quả của Model YOLOv8x-seg gốc dùng để nhận diện và phân đoạn vị trí trái cây (ở đây là táo) theo vị trí trên, dưới

4.2. Model đã finetune

4.2.1. Finetune bu lông

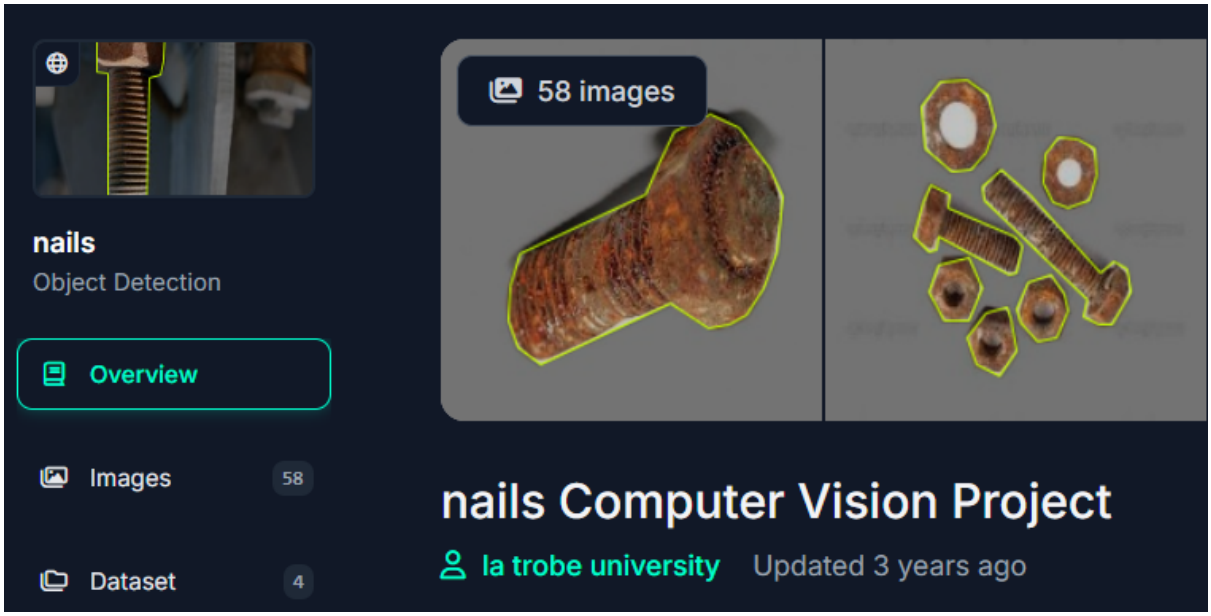

Figura 3: Dataset

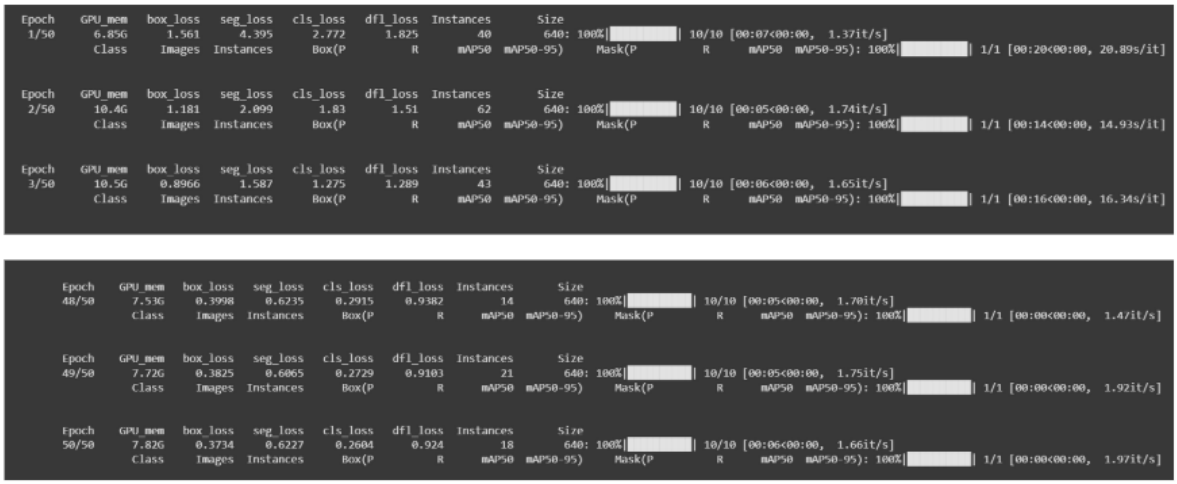

Figura 4: Thông số trên từng epoch

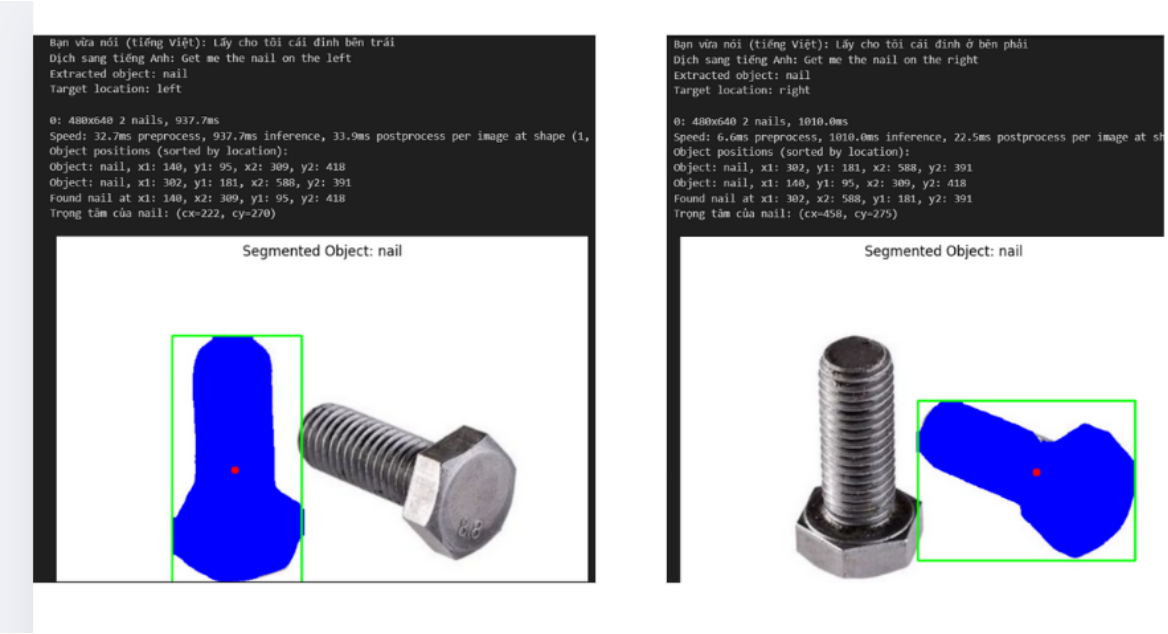

Figura 5: Kết quả

5. Kết luận và Hướng phát triển

5.1. Kết luận

Hệ thống đã có thể:

- Hiểu được câu lệnh bằng tiếng Việt.
- Trích xuất chính xác đối tượng (label) và vị trí (right, left, middle,...)
- Nhận dạng đúng đối tượng trong ảnh bằng YOLOv8-seg
- Phân đoạn đối tượng mong muốn

5.2. Hướng phát triển

5.2.1. Về hướng phát triển:

Về sự phát triển mà bọn em hướng tới, chúng em mong muốn tiếp tục cải tiến hệ thống bằng cách tích hợp một mô hình NLP mạnh mẽ hơn như các biến thể Transformer chuyên biệt để nâng cao khả năng hiểu ngữ nghĩa và xử lý câu lệnh phức tạp hơn. Đồng thời, chúng em cũng mong muốn được mở rộng dự án để có cơ hội được tiếp cận với hệ thống camera 3D và robot thực tế để có thể thử nghiệm mô hình trong môi trường thực, từ đó đưa giải pháp tiến gần hơn tới khả năng ứng dụng trong các bài toán robot gắp thả thông minh.

5.2.2. Về các nội dung cần nghiên cứu:

- Nghiên cứu về cách sử dụng và vận hành của Camera 3D
- Tìm hiểu thêm về lý thuyết cũng như cách kết nối và sử dụng Robot để có thể liên kết với Camera 3D và code Python

5.2.3. Về các kết quả mong đợi:

- Có thể kết nối hệ thống đã được lập trình trên Python với cái thiết bị tân tiến như Camera 3D để có thể xử lý hình ảnh, nhận diện và phân đoạn đối tượng cũng như kết nối với hệ thống Robot để có thể thực hiện những mệnh lệnh đã được đề ra bằng giọng nói trong thời gian thực