HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Thị Thu Thủy

PHÁT HIỆN ĐỐI TƯỢNG TỪ VỰNG MỞ CÓ KÍCH THƯỚC NHỎ TRONG ẢNH CHỤP TỪ DRONE SỬ DỤNG OWL-VIT KẾT HỢP SAHI

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Hà Nội - NĂM 2024

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS Phạm Văn Cường

Phản biện 1: TS. Lê Quốc Hưng

Phản biện 2: PGS.TS. Đỗ Trung Tuấn

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 08 giờ 30 phút, ngày 20 tháng 03 năm 2024

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại công nghệ thông tin phát triển như hiện nay, đặc biệt là lĩnh vực AI với các hệ thống nhận diện thông qua camera ngày càng nhiều. Trong đó phát hiện đối tượng là một tính năng được được nhiều hệ thống sử dụng, nhất là trong các hệ thống phân tích khách hàng của cửa hàng, hệ thống giám sát an ninh. Hơn thế nữa, tính năng này còn được áp dụng vào để phân tích ảnh từ các thiết bị bay không người lái được lắp camera như drone. Ảnh chụp từ drone chụp được rất nhiều vật thể, góc camera chụp rộng. Tuy nhiên một trong những vấn đề gặp phải của nó khi sử dụng những phương pháp phát hiện đối tượng truyền thống hiện nay đó là những đối tượng loại nhỏ thường hay bị bỏ sót.

Những năm gần đây, phát hiện đối tượng (object detection) theo hướng từ vựng mở (open-vocabulary - OV) đã thu hút sự quan tâm nghiên cứu ngày càng nhiều. Khác với phát hiện đối tượng truyền thống chỉ nhận dạng các đối tượng thuộc các danh mục cố định, phát hiện đối tượng từ vựng mở nhằm mục tiêu phát hiện các đối tượng trong một tập hợp danh mục mở. Các mô hình huấn luyện cả hình ảnh-ngôn ngữ (vision-language) để phát hiện các đối tượng từ vựng mở. Trong đó, Vision Transformer for Open-World Localization (OWL-ViT) là một trong những mô hình phát hiện tốt nhất hiện nay. Tuy nhiên thực tiễn cho thấy mô hình này phát hiện những đối tượng có kích thước nhỏ không tốt.

Để giải quyết các vấn đề này, đề án nghiên cứu cải tiến mô hình Vision Transformer for Open-World Localization (OWL-ViT) kết hợp với kỹ thuật Slicing Aided Hyper Inference (SAHI) để phát hiện đối tượng từ vựng mở cho các đối tượng kích thước nhỏ trong ảnh chụp từ drone.

2. Tổng quan về vấn đề nghiên cứu

Với sự phát triển của công nghệ kèm theo sự phát triển của dữ liệu, những ý tưởng mới kết hợp giữa văn bản, để phát hiện đối tượng trong ảnh tốt hơn. Và một hướng đi mới cho bài toán này chính là phát hiện đối tượng từ vựng mở (open-vocabulary - OV). Trong phát hiện đối tượng truyền thống, mô hình chỉ phát hiện được những đối tượng cụ thể đã được đào tạo (tập đối tượng cố định). Ngược lại, nhờ sự kết hợp đào tạo giữa cả dữ liệu văn và hình ảnh, phát hiện đối tượng từ vựng mở, mô hình có thể phát hiện được cả những đối tượng chưa được huấn luyện. Nhận đầu vào là một cặp hình ảnh – văn bản (image-text), văn bản gồm những danh từ cần phát hiện trong bức ảnh, sau đó mô hình phát hiện đối tượng từ vựng mở sẽ cho ra kết quả phát hiện gồm các hộp bao vật thể (bounding boxes) và tên ứng với các danh từ của đối tượng. Như vậy, phát hiện đối tượng từ vựng mở đã khắc phục được vấn đề tập đối tượng phát hiện bị hạn chế, sự kết giữa văn bản và hình ảnh sẽ giúp tăng độ chính xác cho phát hiện đối tượng trong ảnh. Một trong những mô hình cho kết quả tốt nhất với phát hiện đối tượng từ vựng mở hiện nay đó là Vision Transformer for Open-World Localization (OWL- ViT). Mô hình OWL-ViT được thiết kế dựa trên kiến trúc cơ bản của Vision Transformer và Text Transformer, huấn luyện trước nó với một tập dữ liệu lớn gồm các cặp hình ảnh-văn bản. Để phát hiện đối tượng từ vựng mở, loại bỏ token pooling và thêm vào hai đầu phân loại và xác định vị trí đối tượng (object classification head và object localization head) làm đầu ra của Vision Transformer encoder. Đầu xác định vị trí đối tượng là một mạng nơ ron truyền thẳng nhiều lớp-Multi-Layer Perceptron (MLP), cho ra tọa độ các đối tượng là

các hộp bao (bounding boxes), số hộp bao bằng số lượng từ của đầu vào văn bản (mỗi từ ứng với một danh từ là một lớp đối tượng cần được phát hiện). Phần văn bản đầu vào được xử lý qua Text Transformer encoder sẽ được sử dụng kết hợp với đầu phần loại để phân loại đối tượng, gán nhãn cho đối tượng đã được xác định với bounding box. Đây là một mô hình có kiến trúc đơn giản, rất dễ để mở rộng phát triển. Tuy nhiên thực nghiệm cho thấy OWL-ViT phát hiện các đối tượng kích thước nhỏ không tốt, cụ thể là rất nhiều đối tượng nhỏ trong bước ảnh không được phát hiện. Và đây là một vấn đề của mô hình cần được cải thiện. Có rất nhiều giải pháp giúp mô hình phát hiện đối tượng kích thước nhỏ trong ảnh được tốt hơn. Đề án này sử dụng một trong những kỹ thuật mới và tốt nhất hiện nay đó là Slicing Aided Hyper Inference (SAHI) để giải quyết vấn đề này cho OWL-ViT. Kỹ thuật SAHI có thể áp dụng với bất kì mô hình phát hiện đối tượng nào. Trong quá trình tinh chỉnh, phương pháp này chia bức ảnh thành các nhiều phần chồng lấn nhau (overlapping patches). Các patches này được thay đổi kích thước, tuy nhiên vẫn giữ nguyên tỷ lệ khung hình, tạo ra các bức ảnh tăng cường, nhằm mục đích tăng kích thước của đối tượng so với trong hình ảnh gốc.

Từ trên, có thể thấy rằng OWL-ViT đã sử dụng kết hợp giữa thông tin giữa văn bản và hình ảnh để phát hiện đối tượng từ vựng mở. Điều này giúp cho việc phát hiện đối tượng chính xác hơn, phát hiện được các lớp đối tượng không qua huấn luyện. Đây cũng là môt mô hình với kiến trúc cơ bản, rất dễ để mở rộng và phát triển thêm. Trong khi đó, kỹ thuật SAHI với phương pháp chia cắt hình ảnh giúp cho việc phát hiện đối tượng kích thước nhỏ hiệu quả hơn, đã khắc phục một vấn đề thường xuyên gặp phải trong các mô hình phát hiện đối tượng. Bằng việc tận dụng ưu điểm của mô hình OWL-ViT và kỹ thuật SAHI, đề án sẽ trình bày phương pháp kết hợp OWL-ViT để phát hiện đối tượng từ vựng mở có kích thước nhỏ trong ảnh chụp từ drone, loại ảnh có góc chụp từ trên cao, kích thước ảnh lớn, các đối tượng trong ảnh kích cỡ rất nhỏ.

3. Mục đích nghiên cứu

Đề án này nghiên cứu phát hiện đối tượng từ vựng mở có kích thước nhỏ trong ảnh chụp từ drone. Giải pháp sử dụng mô hình OWL-ViT để phát hiện đối tượng từ vựng mở, kết hợp với kỹ thuật SAHI giúp mô hình phát hiện đối tượng tốt hơn. Với việc phát hiện đối tượng nhỏ trong ảnh tốt hơn sẽ giúp cho các hệ thống AI giám sát, phân tích thông tin qua camera nhận phát hiện nhiều đối tượng hơn, giúp cho việc phân tích hình ảnh được chi tiết và chính xác hơn.

Khía cạnh lý thuyết:

- Nghiên cứu: Hiểu sâu hơn về hướng đi mới trong phát hiện đối tượng là phát hiện đối tượng từ vựng mở. Nghiên cứu mô hình OWL-ViT và kỹ thuật SAHI, khả năng kết hợp áp dụng vào bài toán phát hiện đối tượng kích thước nhỏ trong ảnh chụp từ drone.
- Phân tích so sánh: Để thực hiện nghiên cứu đánh giá phương pháp kết hợp giữa OWL-ViT + SAHI sẽ giúp phát hiện đối tượng kích thước nhỏ tốt hơn so với chỉ sử dụng OWL-ViT hoặc SAHI kết hợp với phương pháp khác bằng cách lập bảng so sánh kết quả trên tập dữ liệu VisDrone2019-Detection.
- Hiểu vấn đề: Để hiểu được những vấn đề, thách thức trong bài toán phát hiện đối tượng kích thước nhỏ trong ảnh chụp từ drone. Từ đó có những ý tưởng để nghiên cứu giải pháp.

Khía cạnh thực tiễn:

- Cài đặt mô hình: Lập trình mô hình kết hợp giữa OWL-ViT + SAHI để nhận diện những đối tượng kích thước nhỏ trong ảnh chụp từ drone.
- Ứng dụng trong tạo bộ dữ liệu: Phương pháp OWL-ViT + SAHI sẽ giúp thực hiện đánh nhãn dữ liệu tự động hiệu quả hơn khi có thể phát hiện các đối tượng nhỏ tốt hơn. Từ đó, có thể ứng dụng xây dựng công cụ đánh nhãn dữ liệu tự động với phương pháp này.
- Ứng dụng trong sản phẩm: Phát hiện đối tượng trong ảnh chụp từ drone có tính ứng dụng cao trong các hệ thống giám sát, phân tích ở không gian rộng phát hiện vật thể từ trên cao.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Đối tượng: Đối tượng chính của nghiên cứu này là dữ liệu ảnh, cụ thể là các đối tượng có kích thước nhỏ được chụp từ drone.
- Công nghệ: Các công nghệ được nghiên cứu là mô hình phát hiện đối OWL-ViT và kỹ thuật SAHI.

Phạm vi nghiên cứu:

- Phạm vi: Ảnh chụp từ drone. Cụ thể là bộ dữ liệu VisDrone2019- Detection.
- Khung thời gian: Dự án dự kiến sẽ được hoàn thành trong khoảng thời gian bốn tháng. Nghiên cứu sẽ bao gồm các bài báo và bộ dữ liệu tính đến cuối năm 2023.

5. Phương pháp nghiên cứu

Đề án được nghiên cứu dựa trên cả lý thuyết và thực nghiệm. Xây dựng mô hình dựa trên các phương pháp được nghiên cứu từ các bài báo khoa học về phát hiện đối tượng trong ảnh.

Thu thập dữ liệu:

- Bộ dữ liệu drone: Bộ dữ liệu được chụp từ drone VisDrone2019-Detection, gồm các hình ảnh chụp từ trên cao, góc rộng với rất nhiều đối tượng kích thước nhỏ trong ảnh. Bao gồm 10,209 ảnh với 2,6 triệu đối tượng như: người đi bộ, đám đông người, xe đạp, ô tô, xe bán tải, xe tải hạng nặng, xe ba bánh, xe ba gác, xe bus.
- Tiền xử lý dữ liệu: Lọc bỏ ảnh mờ, nhiễu. Chỉnh sửa lại nhãn đánh sai. Đưa về dạng nhãn dữ liệu phù hợp cho mô hình.

Thiết lập thử ngiệm:

- Môi trường: Thực nghiệm sẽ được tiến hành trên một máy tính được kiểm soát để đảm bảo khả năng thử nghiệm nhiều lần.
- Công cụ và thư viện: Ngôn ngữ lập trình Python sẽ được sử dụng cùng với các thư viện hỗ trợ ngôn ngữ này.

Phương pháp:

- Nghiên cứu tài liệu: nghiên cứu các bài báo khoa học về OWL-ViT và SAHI để hiểu sâu hơn về mô hình và kỹ thuật trên.
- Xây dựng phương pháp: kết hợp giữa mô hình phát hiện đối tượng từ vựng mở OWL-ViT và kỹ thuật hỗ trợ phát thiện đối tượng kích thước nhỏ SAHI.

- Đào tạo mô hình: Xử lý dữ liệu, tiến hành tinh chỉnh (fine-tuning) với các bộ dữ liệu VisDrone2019-Detection.
- Số liệu đánh giá: Mô hình sẽ được đánh giá dựa trên điểm AP thu được trên các bộ dữ liệu.
- Đánh giá: Đánh giá kết quả mô hình dự đoán trên các tập dữ liệu. So sánh OWL-ViT với trước và sau khi sử dụng thêm SAHI. Ngoài ra cũng so sánh thêm với các kết quả trên các tập dữ liệu đã có của những mô hình phát hiện đối tượng khác như TOOD, FCOS, VFNet.
- Xây dựng demo cho phương pháp OWL-ViT+SAHI để thấy rõ kết quả phát hiện đối tượng từ vựng mở có kích thước nhỏ trong ảnh.

Từ mục tiêu, nhiệm vụ nghiên cứu, đề án sẽ được cấu trúc với ba chương nội dung chính như sau:

Chương 1:Tổng quan vấn đề nghiên cứu

Chương 2: Phương pháp OWL-ViT kết hợp SAHI

Chương 3: Thực nghiệm và đánh giá

Chương I - TỔNG QUAN VẤN ĐỀ NGHIÊN CỨU

1.1 Bài toán phát hiện đối tượng

1.1.1 Tổng quan phát hiện đối tượng

Phát hiện đối tượng (object detection): Là nhiệm vụ khó khăn hơn và là sự kết hợp của cả hai nhiệm vụ trên: Vẽ một bounding box xung quanh từng đối tượng quan tâm trong ảnh và gán cho chúng một nhãn. Kết hợp cùng nhau, tất cả các vấn đề này được gọi là object recognition hoặc object detection. Đầu vào: một hình ảnh có một hoặc nhiều đối tượng. Đầu ra: một hoặc nhiều bounding box tương ứng với nhãn đối tượng. Ví dụ như mô hình nhận đầu vào là một bức ảnh chứa cả hai con vật chó và mèo, kết quả đầu ra sẽ là hai hộp bao hình chữ nhật bao quanh mỗi con vật và tên con vật tương ứng cạnh mỗi hộp.

Bài toán phát hiện đối tượng đề cập đến khả năng của hệ thống máy tính và phần mềm để định vị các đối tượng trong một hình ảnh và xác định từng đối tượng. Object Detection đã được sử dụng rộng rãi để phát hiện khuôn mặt, phát hiện xe, đếm số người đi bộ, hệ thống bảo mật và xe không người lái,…

1.1.2 Phát hiện đối tượng trong ảnh chụp từ Drone

Trong thời đại công nghệ thông tin phát triển như hiện nay, đặc biệt là lĩnh vực AI với các hệ thống nhận diện thông qua camera ngày càng nhiều. Trong đó phát hiện đối tượng là một tính năng được được nhiều hệ thống sử dụng, nhất là trong các hệ thống phân tích khách hàng của cửa hàng, hệ thống giám sát an ninh. Hơn thế nữa, tính năng này còn được áp dụng vào để phân tích ảnh từ các thiết bị bay không người lái được lắp camera như drone. Ảnh chụp từ drone chụp được rất nhiều vật thể, góc camera chụp rộng. Phát hiện đối tượng từ ảnh chụp của drone có nhiều ứng dụng quan trọng, bao gồm:

- Giám sát và An ninh: Các mô hình phát hiện đối tượng có thể được sử dụng để phát hiện các đối tượng xâm nhập lạ tại các khu vực quan trọng như biên giới, các cơ sở quân sự.
- Quản lý môi trường và tài nguyên: Phát hiện đối tượng từ ảnh chụp của drone có thể hỗ trợ trong việc giám sát môi trường tự nhiên như phát hiện vùng đất rừng bị tàn phá, theo dõi sự biến đổi của các khu vực đất và đánh giá tình trạng đại dương; theo dõi tài nguyên nông nghiệp, như đánh giá mật độ cây trồng, giám sát sự phát triển và phát hiện bất thường trong vườn trồng.
- Quản lý thiên tai và khắc phục hậu quả: Drone có thể được sử dụng mô hình AI có thể giúp phát hiện các vùng bị tổn thương, đánh giá mức độ thiệt hại và hỗ trợ quyết định khắc phục hậu quả.
- Quản lý công trình và xây dựng: Bằng ảnh chụp từ Drone, các mô hình phát hiện đối tượng sẽ phát hiện các công trình xây dựng trái phép trái với quy hoạch.

1.2 Các nghiên cứu liên quan

1.2.1 Một số phương pháp phát hiện đối tượng từ vựng đóng

Phát hiện đối tượng từ vựng đóng là các mô hình đã được cố định hóa tên các đối tượng là các từ vựng nhất định và từ ngữ, tên loại đối tượng không được sử dụng trong quá trình học tập của mô hình.

Ban đầu, các mô hình "one-stage" và "two stage" như SSD và Faster-RCNN rất được ưa chuộng. Đây đều là các mô hình thuần CNN, kiến trúc mô hình phần lớn xây dựng đều dựa trên các lớp CNN. Vấn đề của phương pháp two stage" Faster-RCNN có ốc độ phát hiện chậm do phụ thuộc vào số vùng đề xuất từ RPN, còn "one-stage" SSD thì các đặc trưng học được từ mô hình này từ các lớp trước đó không đủ phức tạp, điều này dẫn đến hiệu suất kém hơn trên các đối tượng nhỏ hơn.

Gần đây, mô hình DETR (Detection Transformer) được phát triển, sử dụng Transformer để phát hiện đối tượng. Khác với những phương pháp truyền thống ở trên dựa trên việc hiệu chỉnh phân loại các loại đối tượng và độ tin cậy của hộp bao vật thể trên anchor boxes được định nghĩa từ trước. Vì Transformer thực chất biến đổi chuỗi nên DETR có thể coi như là quá trình biến đổi từ chuỗi hình ảnh đến đối tượng truy vấn. Mô hình DETR là một hướng tiếp cận mới so với hướng tiếp cận CNN truyền thống. Vấn đề của DETR cũng như các mô hình trước đó là khá tệ trong việc phát hiện đối tượng nhỏ.

1.2.2 Phát hiện đối tượng từ vựng mở

Với sự phát triển của công nghệ kèm theo sự phát triển của dữ liệu, những ý tưởng mới kết hợp giữa văn bản, để phát hiện đối tượng trong ảnh tốt hơn. Và một hướng đi mới cho bài toán này chính là phát hiện đối tượng từ vựng mở (open-vocabulary- OV). Nhờ sự kết hợp đào tạo giữa cả dữ liệu văn và hình ảnh, phát hiện đối tượng từ vựng mở, mô hình có thể phát hiện được cả những đối tượng chưa được huấn luyện. Nhận đầu vào là một cặp hình ảnh – văn bản (image-text), văn bản gồm những danh từ cần phát hiện trong bức ảnh, sau đó mô hình phát hiện đối tượng từ vựng mở sẽ cho ra kết quả phát hiện gồm các hộp bao vật thể (bounding boxes) và tên ứng với các danh từ của đối tượng. Như vậy, phát hiện đối tượng từ vựng mở đã khắc phục được vấn đề tập đối tượng phát hiện bị hạn chế, sự kết giữa văn bản và hình ảnh sẽ giúp tăng độ chính xác cho phát hiện đối tượng trong ảnh.

1.2.3 Một số kỹ thuật hỗ trợ phát hiện đối tượng nhỏ.

Có một số kỹ thuật hỗ trợ phát hiện đối tượng nhỏ trong thị giác máy tính. Các phương pháp này sử dụng các chiến lược và thuật toán khác nhau để cải thiện hiệu suất phát hiện, đặc biệt là cho các đối tượng có kích thước nhỏ. Dưới đây là một số kỹ thuật phổ biến hiện tại.

Kim tự tháp ảnh (Image Pyramid): Phương pháp này tạo ra nhiều phiên bản với tỷ lệ khác nhau của ảnh đầu vào bằng cách thực hiện giảm mẫu hoặc tăng mẫu. Các phiên bản này, gọi là các cấp độ kim tự tháp, cung cấp các độ phân giải khác nhau của ảnh. Các mô hình phát hiện đối tượng có thể áp dụng thuật toán phát hiện trên mỗi cấp độ pyramid để xử lý các đối tượng ở các tỷ lệ khác nhau.

Cửa sổ trượt (Sliding Window): Phương pháp này bao gồm việc trượt một cửa sổ có kích thước cố định trên ảnh ở các vị trí và tỷ lệ khác nhau. Tại mỗi vị trí cửa sổ, bộ phát hiện đối tượng áp dụng một mô hình phân loại để xác định xem có đối tượng nào xuất hiện hay không.

Kỹ thuật trích xuất đặc trưng đa tỷ lệ (Multiple Scale Feature Extraction): Phương pháp này xử lý ảnh ở nhiều độ phân giải khác nhau hoặc áp dụng các lớp tích chập với các lĩnh vực nhận thức khác nhau.

Tăng cường dữ liệu (Data augmentation): Các phương pháp tăng cường như cắt ngẫu nhiên, thay đổi kích thước, xoay, hoặc thêm nhiễu nhân tạo có thể giúp tạo ra các biến thể

trong tập dữ liệu, cho phép mô hình học các đặc trưng mạnh mẽ cho các đối tượng nhỏ. Các kỹ thuật tăng cường cũng có thể mô phỏng các tỷ lệ, quan điểm và che phủ khác nhau của đối tượng, giúp mô hình phát hiện tổng quát tốt hơn với các tình huống thực tế.

Học chuyển giao (Transfer learning): Các mô hình được tiền huấn luyện, đặc biệt là những mô hình với kiến trúc mạng CNN sâu, ghi lại các đặc trưng phân cấp phong phú có lợi cho việc phát hiện đối tượng nhỏ. Bằng cách điều chỉnh mô hình được tiền huấn luyện trên các bộ dữ liệu mục tiêu, mô hình hiện đối tượng có thể nhanh chóng thích nghi với các nhiệm vụ mới, sử dụng các biểu diễn đã học và cung cấp khả năng phát hiện tốt hơn cho các đối tượng nhỏ.

1.3 Vấn đề phát hiện đối tượng kích thước nhỏ trong ảnh

1.3.1 Nguyên nhân phát hiện đối tượng kích thước nhỏ không tốt trong ảnh Drone

Ảnh chụp từ drone chụp được rất nhiều vật thể, góc camera chụp rộng. Nhiệm vụ phát hiện đối tượng gặp nhiều khó khăn do kích thước nhỏ và độ phân giải thấp của các đối tượng, cũng như các yếu tố khác như che khuất, nhiễu nền và biến đổi trong điều kiện ánh sáng. Ngoài ra còn rất nhiều lý do khiến các phương pháp phát hiện đối tượng truyền thống phát hiện đối tượng kích thước nhỏ kém được nêu ra ở dưới đây.

Đầu tiên là do giới hạn vùng tiếp nhận (Limited Receptive Field). Khái niệm này dùng để chỉ phạm vi không gian của ảnh đầu vào (trường nhìn) có tác động đến đầu ra của một nơron hoặc bộ lọc cụ thể trong mạng nơron tích chập (CNN). Mỗi nơron trong một lớp tích chập sẽ có một trường nhìn, tức là vùng của ảnh đầu vào mà nó nhận thức. Kích thước của trường nhìn thường nhỏ hơn kích thước toàn bộ ảnh đầu vào. Khi đi sâu vào mạng, trường nhìn của các nơron sẽ càng nhỏ đi do các phép tích chập và lớp gộp. Trong các mô hình phát hiện đối tượng thông thường, vùng tiếp nhận có thể bị hạn chế, điều này có nghĩa là mạng không có đủ thông tin ngữ cảnh xung quanh các đối tượng nhỏ. Kết quả là, mô hình có thể gặp khó khăn trong việc phát hiện và xác định vị trí chính xác các đối tượng này do vùng tiếp nhận không đủ.

Nguyên nhân thứ hai là do biến đổi tỷ lệ (Scale Variation). Các đối tượng nhỏ thể hiện sự biến đổi tỷ lệ đáng kể so với các đối tượng lớn trong một hình ảnh. Các bộ phát hiện đối tượng được huấn luyện trên các bộ dữ liệu chủ yếu gồm các đối tượng lớn, chẳng hạn như ImageNet hoặc COCO, có thể gặp khó khăn trong việc tổng quát hóa cho các đối tượng nhỏ do sự chênh lệch về tỷ lệ.

Lý do thứ ba, thiên hướng dữ liệu huấn luyện (Training Data Bias). Các mô hình phát hiện đối tượng thường được huấn luyện trên các bộ dữ liệu quy mô lớn, có thể chứa các thiên hướng về các đối tượng lớn do sự phổ biến của chúng. Thiên hướng này có thể không cố ý ảnh hưởng đến hiệu suất của bộ phát hiện đối tượng khi xử lý các đối tượng nhỏ. Kết quả là, mô hình có thể chưa được tiếp xúc đủ với các ví dụ huấn luyện đa dạng về các đối tượng nhỏ. Điều này dẫn đến sự thiếu ổn định và độ chính xác phát hiện giảm đi đối với các trường hợp đối tượng nhỏ.

Cuối cùng là việc xác định vị trí chính xác của các đối tượng nhỏ có thể gặp khó khăn do độ phân giải không gian hạn chế của các bản đồ đặc trưng trong kiến trúc mạng nơron tích chập. Các chi tiết tinh tế cần thiết để xác định vị trí chính xác có thể bị mất hoặc trở nên không thể phân biệt ở độ phân giải thấp hơn. Các đối tượng nhỏ có thể bị che khuất bởi các đối tượng

lớn khác hoặc các nền nhiễu, làm khó khăn thêm cho việc xác định vị trí. Những yếu tố này có thể góp phần làm cho các bộ phát hiện đối tượng thông thường không thể xác định và phát hiện các đối tượng nhỏ một cách chính xác.

1.3.2 Phương pháp phát hiện đối tượng đề xuất

Các phương pháp được nêu ở mục 1.2.2 giúp mô hình cải thiện độ chính xác khi phát hiện các đối tượng kích thước nhỏ. Tuy nhiên những kỹ thuật đó có các nhược điểm như: Pyramid Image và Sliding Window gây tốn kém chi phí tính toán và bộ nhớ. Multiple Scale Feature Extraction cần phải sửa lại kiến trúc mạng khi áp dụng, khó khăn thay đổi các kiến trúc mạng phức tạp. Data augmentation có thể gây mất thông tin khi áp dụng các biến đổi không phù hợp, giảm chất lượng và độ tin cậy của dữ liệu. Transfer learning không cải thiện nhiều kết quả cho các đối tượng nhỏ hiếm gặp. Bên cạnh đó các phương pháp phát hiện đối tượng từ vựng đóng cũng còn nhiều vấn đề đối với các đối tượng kích thước nhỏ. Nhất là các mô hình dựa trên kiến trúc CNN thường bị Limited Receptive Field như đã nêu ở trên.

Để khắc phục những vấn đề được nêu trên, đề án đề xuất sử dụng phương pháp kết hợp giữa mô hình phát hiện đối tượng từ vựng mở Vision Transformer for Open-World Localization (OWL-ViT) và kỹ thuật Slicing Aided Hyper Inference (SAHI) giúp phát hiện các đối tượng kích thước nhỏ tốt hơn. Mô hình OWL-ViT được thiết kế dựa trên kiến trúc cơ bản của Vision Transformer và Text Transformer chứ không sử dụng kiến trúc CNN. Mô hình sẽ được huấn luyện trước nó với một tập dữ liệu lớn gồm các cặp hình ảnh-văn bản. Nhờ sự kết hợp đào tạo giữa cả dữ liệu văn và hình ảnh, phát hiện đối tượng từ vựng mở, mô hình có thể phát hiện được cả những đối tượng chưa được huấn luyện. Như vậy, phát hiện đối tượng từ vựng mở OWL-ViT đã khắc phục được vấn đề tập đối tượng phát hiện bị hạn chế, sự kết giữa văn bản và hình ảnh sẽ giúp tăng độ chính xác cho phát hiện đối tượng trong ảnh. Bên cạnh đó, sử dụng thêm kỹ thuật SAHI, một trong những kỹ thuật mới và tốt nhất hiện nay. Với phương pháp này chia cắt hình ảnh giúp cho việc phát hiện đối tượng kích thước nhỏ hiệu quả hơn, đã khắc phục một vấn đề thường xuyên gặp phải trong các mô hình phát hiện đối tượng.

1.4 Kết luận chương

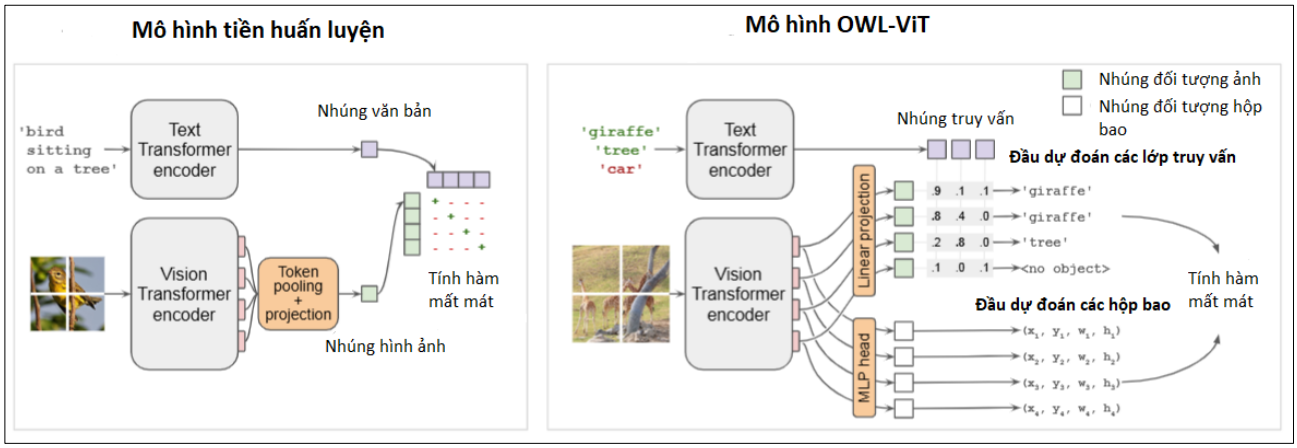
Tại chương này, đề án cung cấp một cái nhìn tổng quan về bài toán phát hiện đối tượng. Phân biệt các nhiệm vụ phân loại hình ảnh, định vị đối tượng với phát hiện đối tượng. Trình bày bài toán phát hiện đối tượng trong ảnh chụp từ drone và các ứng dụng thực tế như giám sát an ninh, quản lý tài nguyên, giám sát xây dựng,... Đề án cũng trình bày các nghiên cứu liên quan đến vấn đề phát hiện đối tượng kích thước thước nhỏ trong ảnh chụp Drone. Cụ thể là phát hiện đối tượng từ vựng đóng như SSD, Faster-RCNN, DETR và các vấn đề của chúng. Bên cạnh đó là trình bày các kỹ thuật phổ biến hỗ trợ phát hiện đối tượng nhỏ trong ảnh: kim tự tháp ảnh, cửa sổ trượt, trích xuất đặc trưng đa tỷ lệ, tăng cường dữ liệu, học chuyển giao . Đề án chỉ ra vấn đề hiện tại của phát hiện đối tượng kích thước nhỏ trong ảnh Drone. Nêu ra nguyên nhân của vấn đề, nhược điểm của các phương pháp trước. Từ đó đề xuất giải pháp đề xuất kết hợp giữa OWL-ViT và SAHI cho bài toán. Các chương sau sẽ đi vào khía cạnh kỹ thuật của phương pháp này.

Chương 2 - PHƯƠNG PHÁP OWL-VIT KẾT HỢP SAHI

2.1 Mô hình Vision Transformer cho Open-World Localization (OWL-ViT)

2.1.1 Tổng quan mô hình OWL-ViT

OWL-ViT bắt đầu với kiến trúc Vision Transformer, đã được chứng minh là có khả năng mở rộng cao, và tiến hành tiền huấn luyện đối ngẫu trên một tập dữ liệu hình ảnh-văn bản lớn. Sau đó để chuyển giao mô hình sang nhiệm vụ phát hiện đối tượng, thực hiện một số thay đổi. Đầu tiên, loại bỏ lớp pooling token cuối cùng và thay vào đó gắn một đầu phân loại nhãn và một đầu dự đoán bounding box cho mỗi token đầu ra của Transformer Encoder. Phân loại từ vựng mở mở được kích hoạt bằng cách thay thế trọng số của lớp phân loại cố định bằng class-name embeddings được thu được từ mô hình văn bản. Tiếp theo, điều chỉnh lại mô hình đã được tiền huấn luyện trên các tập dữ liệu phát hiện tiêu chuẩn bằng cách sử dụng hàm mất mát bipartite matching. Như vậy, cả mô hình hình ảnh và văn bản đều được điều chỉnh lại từ đầu đến cuối giống như hình 2.1. Bên trái là mô hình tiền huấn luyện, bên phải là mô hình OWL-ViT sau khi được điều chỉnh lại.



Hình 2.1: Kiến trúc mô hình OWL-ViT

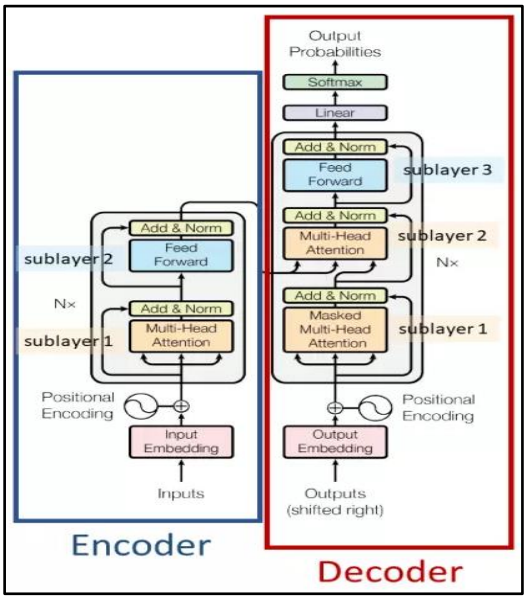
Để phân loại các đối tượng đã phát hiện với từ vựng mở, mô hình sử dụng text embeddings, thay vì class embeddings đã học, trong lớp đầu ra của đầu phân loại. Các text embeddings, được gọi là truy vấn (queries), được tạo ra bằng cách đưa tên đối tượng hoặc các mô tả đối tượng văn bản khác qua text encoder. Nhiệm vụ của mô hình sau đó là dự đoán, đối với mỗi đối tượng, một bounding box và xác suất mà mỗi truy vấn áp dụng cho đối tượng đó. Các truy vấn có thể khác nhau cho mỗi hình ảnh. Kết quả là, mỗi hình ảnh sẽ có không gian nhãn riêng biệt, được xác định bởi một tập hợp các chuỗi văn bản. Phương pháp này bao gồm cả việc phát hiện đối tượng với từ vựng đóng, trong đó toàn bộ tập hợp tên danh mục đối tượng được sử dụng làm tập truy vấn cho mỗi hình ảnh.

Khác với một số phương pháp khác, mô hình không kết hợp tất cả các truy vấn cho một hình ảnh thành một chuỗi token (token sequence) duy nhất . Thay vào đó, mỗi truy vấn bao gồm một token sequence riêng biệt đại diện cho một mô tả đối tượng riêng biệt và được xử lý riêng biệt bởi bộ mã hóa văn bản. Ngoài ra, kiến trúc của mô hình không bao gồm việc kết hợp giữa image encoder và text encoder. Mặc dù việc kết hợp sớm có vẻ có lợi theo nhiều suy đoán nhưng thực tế nó làm giảm hiệu suất suy luận đáng kể vì việc mã hóa một truy vấn yêu cầu một quá trình chuyển tiếp qua toàn bộ mô hình hình ảnh và cần được lặp lại cho mỗi kết hợp hình ảnh/truy vấn. Trong OWL-ViT có thể tính toán các nhúng truy vấn độc lập với

hình ảnh, cho phép sử dụng hàng ngàn truy vấn cho mỗi hình ảnh, nhiều hơn nhiều so với việc kết hợp sớm.

2.1.2 Text Encoder của mô hình

Trong mô hình OWL-ViT sử dụng Encoder của Transformer tiêu chuẩn để mã hóa phần văn bản. Nhìn vào sơ đồ kiến trúc của Transformer ở dưới có thể thấy rằng nó được chia thành hai phần rõ ràng đó là encoder và decoder. Do mô hình OWL-ViT chỉ sử dụng phần encoder nên đề án sẽ tập trung phân tích các phần của Transformer encoder và bỏ qua phần decoder.



Hình 2.2: Kiến trúc Transformer

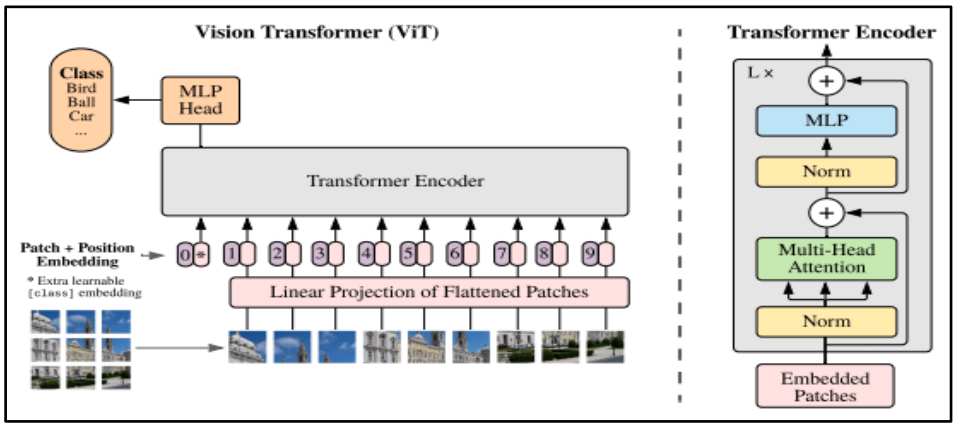
Đầu tiên của phần encoder là Input Embeddings, máy tính không hiểu câu chữ mà chỉ đọc được số, vectơ, ma trận; vì vậy ta phải biểu diễn câu chữ dưới dạng vectơ, gọi là input embedding. Điều này đảm bảo các từ gần nghĩa có vectơ gần giống nhau. Hiện nay đã có khá nhiều pretrained word embeddings như GloVe, Fasttext, gensim Word2Vec,... cho chúng ta lựa chọn. Word embeddings phần nào cho giúp ta biểu diễn ngữ nghĩa của một từ, tuy nhiên cùng một từ ở vị trí khác nhau của câu lại mang ý nghĩa khác nhau. Đó là lý do Transformer có thêm một phần Positional Encoding để thêm thông tin về vị trí của một từ. Tiếp đến là Self-Attention-cơ chế giúp Transformer "hiểu" được sự liên quan giữa các từ trong một câu. Vấn đề của Self-attention là attention của một từ sẽ luôn "chú ý" vào chính nó. Để tránh xảy ra điều này vì cái ta mong muốn muốn là đặc trưng liên quan giữa các từ khác nhau trong câu. Tác giả đã giới thiệu một phiên bản nâng cấp hơn của Self-attention là Multi-head attention. Ý tưởng rất đơn giản là thay vì sử dụng 1 Self-attention (1 head) thì ta sử dụng nhiều Attention khác nhau (multi-head) và biết đâu mỗi Attention sẽ chú ý đến một phần khác nhau trong câu. Vì mỗi "head" sẽ cho ra một ma trận attention riêng nên ta phải concat các ma trận này và nhân với ma trận trọng số W_O để ra một ma trận attention duy nhất (weighted sum). Và tất nhiên, ma trận trọng số này cũng được điều chỉnh trong khi quá trình huấn luyện.

Sau khi đi qua Mutil Multi-head Attention cho ra các vectơ attention Z, chúng sẽ đi qua phần Add & Normalize tạo thành sub-layer 1 như trong mô hình tổng quan về Transformer ở trên. Mỗi sub-layer đều là một khối dư (residual block). Cũng giống như residual blocks trong Computer Vision, các kết nối tắt (skip connections) trong Transformer cho phép thông tin đi

qua sub-layer trực tiếp. Thông tin này (x) được cộng với attention (z) của nó và thực hiện chuẩn hóa (normalization) với Layer Normalization. Cuối cùng là Feed Forward, sau khi được chuẩn hóa, các vectơ z được đưa qua mạng kết nối đầy đủ (fully connected) và cho ra các text/query embeddings. Vì các vectơ này không phụ thuộc vào nhau nên ta có thể tận dụng được tính toán song song cho cả câu.

2.1.3 Vision Encoder của mô hình

Mô hình Vision Transformer chia hình ảnh thành các mảnh nhỏ hơn gọi là "điểm chú ý" (patches) và biến đổi chúng thành các vectơ. Các vectơ này sau đó được đưa vào một mạng Transformer, cho phép mô hình học được sự tương tác giữa các điểm chú ý và xử lý thông tin không gian rộng hơn



Hình 2.3: Kiến trúc Vision Transformer

Trong phần đầu tiên Linear Projection of Flattened Patches, khác với các mô hình CNN cho bài toán image classification, ảnh input đầu vào cho mô hình CNN đó là toàn bộ ảnh với kích thước cố định. Tuy nhiên Vision Trans có một cách xử lý khác. Với mỗi ảnh đầu vào, ViT xử lý bằng cách chia ảnh ra thành các phần có kích thước bằng nhau (patch) và sau đó thêm các thông tin cần thiết, quá trình này gọi là Patch Embedding Bước tiếp theo, đưa các patches này về dạng vector bằng cách duỗi thẳng (flatten) các patches này ra. Hình 2.3 trên mô tả phần Linear Projection. Thực chất Linear Projection là một lớp Dense với đầu vào là flattened vector của các patches, đầu ra sẽ là vectơ embedding tương ứng với từng patch. Tương tự như với mô hình Transformer gốc. Positional embedding trong mô hình ViT sẽ chứa thông tin về vị trí của patch trong ảnh (spatial information). Nếu như chỉ embedding các patch và đưa vào mô hình Transformer thì với 2 ảnh ở bên trên sẽ hoàn toàn không có sự khác biệt. Do đó ta cần thêm thông tin về vị trí cho mỗi patch. Sau khi có vectơ positional embedding cho mỗi patch ta sẽ cộng các vectơ này tương ứng với embedding vector của từng patch đã tính ở trên và thu được các vector embedding vừa chứa thông tin của vùng ảnh vừa chứa thông tin về vị trí của nó trong ảnh. Vector vị trí này có kích thước 1D giúp giảm kích thước lưu trữ so với vector 2D. Các position embedding được cộng vào các vectơ patch của ảnh tương ứng. Giống như minh họa trong hình 29, với 9 patch ảnh được chia tương ứng với 9 position embedding từ 1 đến 9. Tuy nhiên lại còn thừa vị trí 0*. Phần * ở đây chính là Class Embedding. Trong quá trình tiền huấn luyện bằng bộ mã hóa Transformer, ta luôn cần một nhãn lớp ở vị trí 0. Khi chúng ta truyền các hình ảnh patch làm đầu vào, luôn cần thêm một

token phân loại ở vị trí đầu tiên như được thể hiện trong hình 29. Kiến trúc Vision Transformer dùng trong phân loại ở hình 27 với các vectơ sau khi đi qua Transformer encoder sẽ đi đến đầu classification là một khối Multilayer perceptron đưa ra kết quả cuối cùng là xác suất tương ứng với các class. Trong hình 30 bên trái là kiến trúc tiền huấn luyện (pre-trained) của Vision Transformer encoder, các vectơ đầu ra của encoder cũng được đưa qua token pooling + projection để tạo ra một image embedding tổng quát cho toàn bộ hình ảnh sau đó kết hợp với phần text encoder để phân loại ảnh. Khác với những thứ trên, OWL-ViT sử dụng để phát hiện đối tượng nên các vectơ sau khi đi qua vision Vision Transformer encoder sẽ được đồng thời đi qua hai đầu MLP và đầu Linear projection. Đầu Linear projection dùng để dự đoán ra nhãn của đối tượng. Đầu MLP (Multilayer perceptron) để dự đoán ra các bounding box của đối tượng.

2.1.4 Hàm mất mát

Mô hình OWL-ViT sử dụng hàm mất mát khớp hai phía (bipartite matching loss) được giới thiệu trong mô hình DETR. Hàm mất mát khớp hai phía thực hiện thực chất là so sánh các lớp và hộp giới hạn được dự đoán của mỗi truy vấn đối tượng N = 100 với các nhãn đúng, được đệm đến cùng độ dài N (vì vậy, nếu một hình ảnh chỉ chứa 4 đối tượng, 96 nhãn sẽ chỉ có "không có đối tượng" làm lớp và "không có hộp giới hạn" làm hộp giới hạn). Thuật toán Hungarian matching được sử dụng để tìm một ánh xạ một-đến-một tối ưu cho mỗi truy vấn N tới mỗi nhãn N. Tiếp theo, ta sử dụng hàm mất mát chéo entropy chuẩn (cho các lớp) và tổ hợp tuyến tính của hàm mất mát L1 và hàm mất mát generalized IoU (đối với hộp giới hạn) để tối ưu các tham số của mô hình.

Mô hình OWL-ViT điều chỉnh nó cho việc phát hiện đối tượng vựng mở như sau. Do công sức cần thiết để gán nhãn cho các tập dữ liệu phát hiện một cách toàn diện, các tập dữ liệu với số lượng lớn danh mục được gán nhãn theo cách liên minh. Các tập dữ liệu như vậy có không gian nhãn không phân biệt, có nghĩa là mỗi đối tượng có thể có nhiều nhãn. Do đó, mô hình sử dụng hàm mất mát focal sigmoid cross-entropy thay vì softmax cross-entropy làm hàm mất mát phân loại. Hơn nữa, vì không phải tất cả các đối tượng đều được gán nhãn trong mỗi hình ảnh, các tập dữ liệu liên minh cung cấp cả những nhận dạng tích cực (xuất hiện) và tiêu cực (không xuất hiện) của các đối tượng truy vấn cho mỗi hình ảnh. Trong quá trình huấn luyện, đối với một hình ảnh cụ thể, mô hình sử dụng tất cả các nhận dạng tích cực và tiêu cực của nó như các truy vấn. Hơn nữa, mô hình còn ngẫu nhiên chọn mẫu các đối tượng theo tỷ lệ tần suất của chúng trong dữ liệu và thêm chúng như các "giả-tiêu-cực" để có ít nhất 50 tiêu cực cho mỗi hình ảnh.

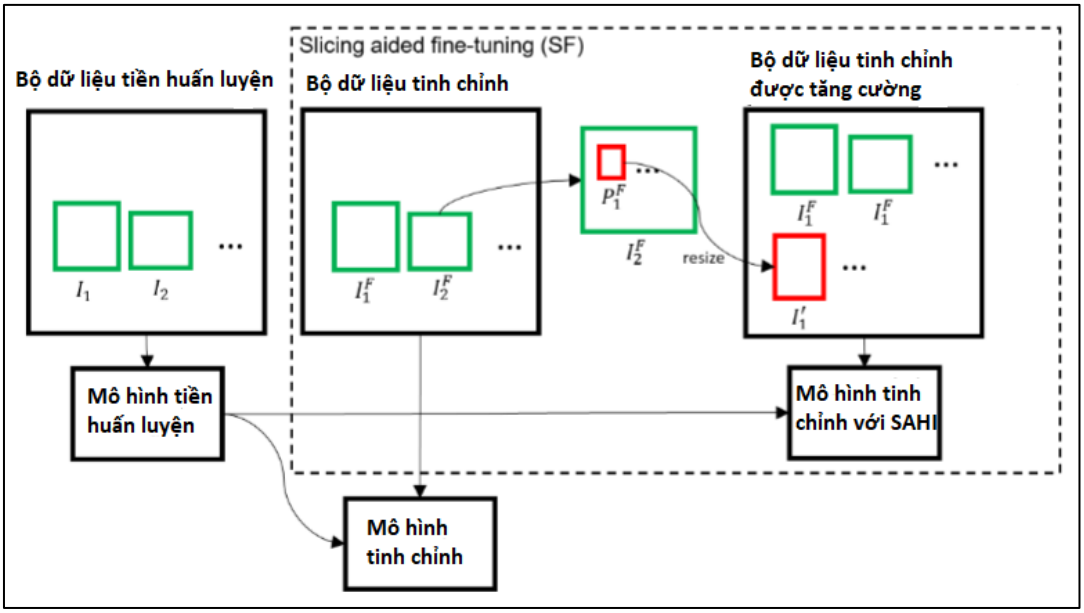
2.2 Kỹ thuật Slicing Aided Hyper Inference (SAHI)

Phát hiện các đối tượng nhỏ và các đối tượng ở xa trong cảnh quan sát là một thách thức lớn trong các ứng dụng giám sát. Những đối tượng như vậy được đại diện bởi một số lượng nhỏ pixel trong hình ảnh và thiếu chi tiết đủ, làm cho việc phát hiện chúng bằng các bộ phát hiện thông thường trở nên khó khăn. Kỹ thuật Slicing Aided Hyper Inference (SAHI) được đề xuất, hỗ trợ các mô hình phát hiện đối tượng nhỏ bằng việc chia nhỏ hình ảnh và fine-tuning. Phương pháp này chia hình ảnh đầu vào thành các phần chồng lấn nhau (overlap

patches), gây ra khu vực pixel tương đối lớn hơn cho các đối tượng nhỏ so với hình ảnh được đưa vào mạng.

2.2.1 Phương pháp SAHI cho tinh chỉnh mô hình

Các mô hình phát hiện đối tượng phổ biến thường được cung cấp trọng số được huấn luyện trước trên các tập dữ liệu như ImageNet và MS COCO. Điều này cho phép chúng ta tinh chỉnh mô hình bằng cách sử dụng các tập dữ liệu nhỏ hơn và trong khoảng thời gian đào tạo ngắn hơn so với việc huấn luyện từ đầu với các tập dữ liệu lớn. Tuy nhiên, các tập dữ liệu thông thường này thường chứa hình ảnh có độ phân giải thấp (640 x 480) với các đối tượng lớn chiếm phần lớn khu vực pixel (trung bình chiếm 60% chiều cao của hình ảnh). Mô hình được huấn luyện trước trên các tập dữ liệu này cho kết quả phát hiện tốt cho các hình ảnh tương tự. Tuy nhiên, hiệu suất phát hiện đối tượng nhỏ trên các hình ảnh có độ phân giải cao được tạo bởi drone và camera giám sát chất lượng cao thường thấp hơn đáng kể.

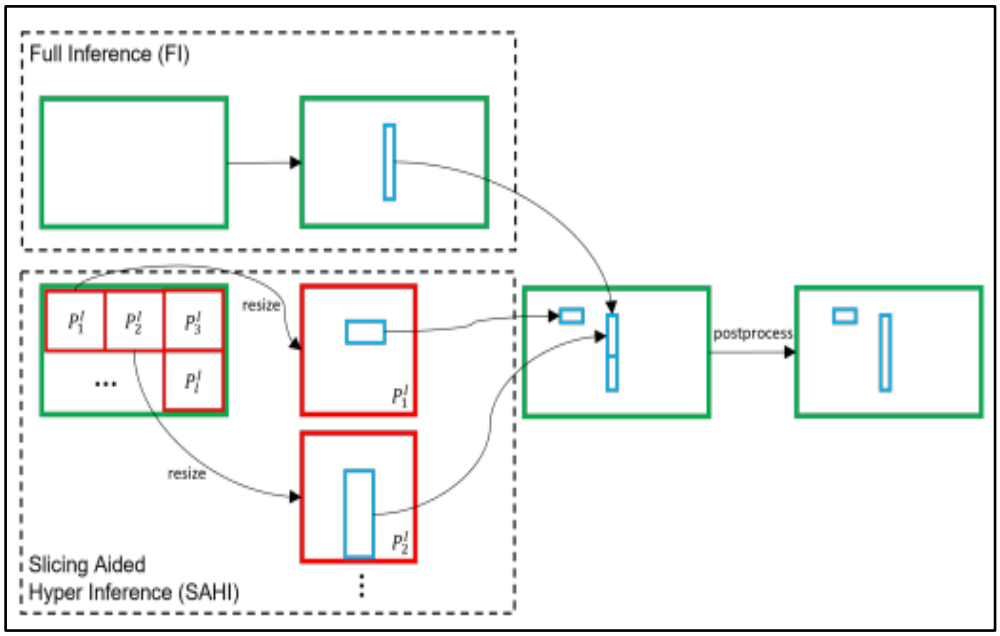


Hình 2.4: Phương pháp SAHI cho tinh chỉnh mô hình (Slicing aided fine-tuning)

Để khắc phục vấn đề trên, SAHI hỗ trợ quá trình tinh chỉnh mô hình phát hiện đối tượng như sau. Đầu tiên, mở rộng tập dữ liệu bằng cách trích xuất các patch từ tập dữ liệu fine-tuning như được thể hiện trong Hình 31. Mỗi hình ảnh I_1^F, I_2^F, ..., I_j^F được chia thành các patch chồng lấn nhau P_1^F, P_2^F, ..., P_k^F với kích thước hai chiều M và N được lựa chọn trong khoảng $[M_{min}, M_{max}]$ và $[N_{min}, N_{max}]$ là các siêu tham số. Trong quá trình fine-tuning, các patch được điều chỉnh kích thước để giữ nguyên tỷ lệ khung hình sao cho chiều rộng hình ảnh nằm trong khoảng từ 800 đến 1333 pixel, từ đó tạo ra các hình ảnh mở rộng I_1', I_2', ..., I_k', trong đó kích thước đối tượng tương đối lớn hơn so với hình ảnh gốc. Những hình ảnh này I_1', I_2', ..., I_k', cùng với hình ảnh gốc I_1^F, I_2^F, ..., I_j^F (để hỗ trợ phát hiện các đối tượng lớn), được sử dụng trong quá trình fine-tuning. Bởi vì, khi kích thước mảnh giảm, các đối tượng lớn có thể không vừa vào một mảnh và các khu vực giao nhau, điều này có thể dẫn đến hiệu suất phát hiện kém cho các đối tượng lớn.

2.2.2 Phương pháp SAHI cho suy luận mô hình

Trong giai đoạn suy luận, ảnh truy vấn gốc I được chia thành l phần chồng lên nhau M × N: P_1^I, P_2^I, ... P_l^I. Sau đó, mỗi phần cắt được điều chỉnh kích thước mà vẫn giữ tỷ lệ khung hình ban đầu. Tiếp theo, quá trình phát hiện đối tượng được thực hiện độc lập trên từng phần chồng lên nhau. Một quá trình suy luận đầy đủ (có thể dùng thêm tinh chỉnh với SAHI) sử dụng ảnh gốc có thể được áp dụng để phát hiện các đối tượng lớn hơn. Cuối cùng, kết quả dự đoán của các phần chồng lên nhau và (nếu được sử dụng) kết quả tính chỉnh được hợp nhất trở lại kích thước ban đầu bằng cách sử dụng thuật toán Non-maximum Suppression (NMS). Trong quá trình NMS, các hộp có tỷ lệ giao nhau (Intersection over Union - IoU) cao hơn ngưỡng khớp đã được xác định trước Tm được khớp và đối với mỗi khớp, các kết quả phát hiện có xác suất phát hiện thấp hơn ngưỡng đã xác định T_d sẽ bị loại bỏ.



Hình 2.5: Phương pháp SAHI cho suy luận mô hình (Slicing aided hyper inference)

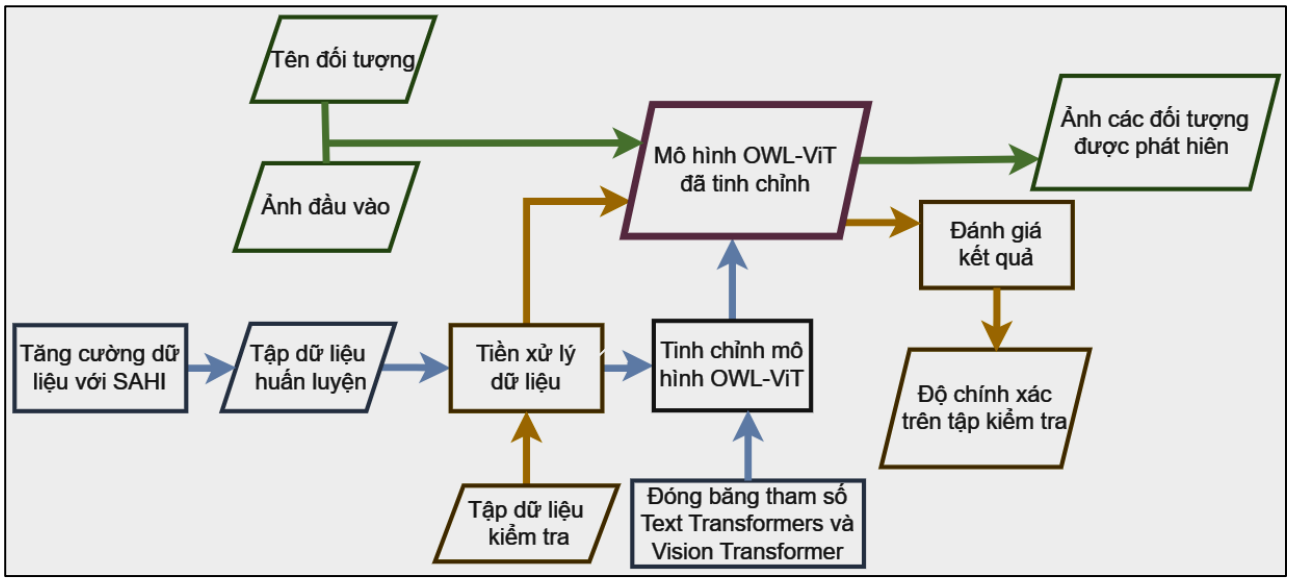
2.3 Phương pháp kết hợp giữa OWL-ViT và SAHI cho phát hiện đối tượng

2.3.1 Tổng quan phương pháp

Mô hình OWL-ViT tuy được đào tạo với tập dữ liệu lớn với nhiều loại vật thể nhưng đối với những vật thể có kích thước nhỏ trong ảnh mô hình phát hiện vẫn còn kém. Điều này có thể do nhiều nguyên nhân như đã nêu ở chương I, tuy nhiên nguyên nhân chủ yếu là do mô hình chưa được đào tạo nhiều với vật thể nhỏ như trong tập dữ liệu Visdrone2019. Vậy nên đề án sử dụng kỹ thuật SAHI để tăng cường dữ liệu cho bộ dữ liệu đào tạo. Tận dụng sức mạnh đã được đào tạo với một nguồn dữ liệu cực lớn từ OWL-ViT, tiến hành tinh chỉnh mô hình này với bộ dữ liệu đã được tăng cường. Trong quá trình tinh chỉnh này có thay đổi hàm mất mát để đạt kết quả tốt hơn.

Sau khi mô hình được tinh chỉnh, đánh giá mô hình trên tập kiểm tra của bộ dữ liệu Visdrone2019 với độ đo mAP. Mô hình này cũng nhận đầu vào là ảnh và tên đối tượng dưới dạng văn bản, cho ra kết quả suy luận là ảnh đầu ra với xác suất dự đoán có đối tượng, tên đối tượng và bounding box tương ứng. Bounding box có dạng (cx, cy, w, h) với (cx, cy) là tọa độ tâm và (w, h) là chiều rộng và chiều cao của bounding box.

Từng bước của phương pháp phát hiện đối tượng kết hợp OWL-ViT và SAHI được miêu tả chi tiết trong sơ đồ hình 2.6 dưới đây:



Hình 2.6: Phương pháp phát hiện đối tượng kết hợp OWL-ViT và SAHI

2.3.2 Xử lý trong tinh chỉnh mô hình

Trong phần tinh chỉnh mô hình, bước đầu tiên đó là chuẩn bị dữ liệu. Đối với phần chuẩn bị dữ liệu, đề án sử dụng kỹ thuật SAHI để tăng cường dữ liệu. SAHI cắt những ảnh trong tập đào tạo của tập Visdrone2019 ra thành các phần nhỏ hơn. Trộn lẫn chúng với tập đào tạo ban đầu, ta được bộ dữ liệu đào tạo mới với kích thước ảnh đa dạng hơn. Điều này sẽ giúp cho những vật thể nhỏ của bộ dữ liệu được đào tạo nhiều hơn với các ảnh kích thước khác nhau. Ảnh gốc ban đầu được chia thành các ảnh nhỏ, và giữ nguyên được nhãn ban đầu của vật thể. Những vật thể nhỏ trong những ảnh mới này được phóng to ra so với kích cỡ của bức ảnh mới. Nhờ vậy khi cho vào đào tạo, mô hình sẽ học được những đặc rõ ràng của các đối tượng cần phát hiện trong bức ảnh. Và việc gia tăng số lượng ảnh sẽ giúp mô hình được học đi học lại các đối tượng, giúp phát hiện đối tượng tốt hơn. Trong tập dữ liệu Visdrone2019 các ảnh có kích thước lớn, đồng thời độ phân giải cao nên việc cắt nhỏ ảnh hoàn toàn phù hợp vì không làm ảnh hưởng đến chất lượng ảnh. Sau khi cắt ra bằng SAHI, các bức hình nhỏ không hề bị mờ và đảm bảo chất lượng cho mô hình khi đào tạo.

Sau khi xử lý xong dữ liệu, tiếp đến là tinh chỉnh mô hình OWL-ViT. Vì OWL-ViT là một mô hình có kích cỡ mạng rất lớn, để đào tạo cả mô hình cần một nguồn tài nguyên lớn cả về thiết bị máy móc lẫn dữ liệu. Vậy nên đề án thực hiện tinh chỉnh mô hình với tập dữ liệu đã tăng cường bằng SAHI ở trên. Tận dụng sức mạnh của mô hình đã được đào tạo với các tập dữ liệu lớn, ta đóng băng phần Text Encoder và Vision Encoder, tiến hành tinh chỉnh trên phần còn lại của mạng, bao gồm hai phần quan trọng là đầu dự đoán bounding box và dự đoán tên lớp cho đối tượng. Trong quá trình tinh chỉnh sử dụng hàm mất mát theo OWL-ViT Adaptation. Thuật toán tối ưu tham số được sử dụng là AdamW, một phiên bản cải tiến hơn so với Adam.

2.3.3 Thuật toán tối ưu

Đề án sử dụng thuật toán tối ưu AdamW là một phiên bản nâng cấp của thuật toán Adam vốn rất phổ biến và nổi tiếng thường được sử dụng mặc định. Khi được kiểm tra trên

một loạt các nhiệm vụ học sâu khác nhau như phân loại hình ảnh, mô hình ngôn ngữ cấp ký tự và phân tích cú pháp, Adam không có hiệu suất tổng quát tốt như SGD với động lượng. Sự khác biệt của Adam nằm ở cách thức triển khai không hiệu quả của trọng số tiêu biến (weight decay). AdamW sửa đổi đơn giản một chút so với Adam bằng cách tách riêng trọng số tiêu biến (weight decay) khỏi các bước tối ưu hóa được thực hiện trong hàm mất mát. Điểm khác biệt lớn nhất của AdamW và Adam được thể hiện rõ ràng qua phần weight decay trong hình so sánh hai thuật toán ở dưới đây:

Algorithm 2 Adam with L_2 regularization and Adam with decoupled weight decay (AdamW)

1: **given** $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda \in \mathbb{R}$
2: **initialize** time step $t \leftarrow 0$, parameter vector $\boldsymbol{\theta}_{t=0} \in \mathbb{R}^n$, first moment vector $\boldsymbol{m}_{t=0} \leftarrow \boldsymbol{0}$, second moment vector $\boldsymbol{v}_{t=0} \leftarrow \boldsymbol{0}$, schedule multiplier $\eta_{t=0} \in \mathbb{R}$
3: **repeat**
4: $t \leftarrow t + 1$
5: $\nabla f_t(\boldsymbol{\theta}_{t-1}) \leftarrow \text{SelectBatch}(\boldsymbol{\theta}_{t-1})$ ▷ select batch and return the corresponding gradient
6: $\boldsymbol{g}_t \leftarrow \nabla f_t(\boldsymbol{\theta}_{t-1}) + \lambda \boldsymbol{\theta}_{t-1}$
7: $\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\boldsymbol{g}_t$ ▷ here and below all operations are element-wise
8: $\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$
9: $\hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t / (1 - \beta_1^t)$ ▷ β_1 is taken to the power of t
10: $\hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t / (1 - \beta_2^t)$ ▷ β_2 is taken to the power of t
11: $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$ ▷ can be fixed, decay, or also be used for warm restarts
12: $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta_t \left(\alpha \hat{\boldsymbol{m}}_t / (\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon) + \lambda \boldsymbol{\theta}_{t-1} \right)$
13: **until** *stopping criterion is met*
14: **return** optimized parameters $\boldsymbol{\theta}_t$

Hình 2.7: Mã giả thuật toán tối ưu của Adam và AdamW

Nhìn vào mã giả thuật toán ở trên, khi Adam chạy trên hàm mất mát f, và cộng với regularization, những trọng số có xu hướng có gradient lớn trong f sẽ không được điều chỉnh nhiều. Bởi vì trong L2 regularization, cả hai gradient của hai số hạng (hàm mất mát và regularization) đều được điều chỉnh theo độ lớn của chúng. Do đó, trọng số x với độ lớn gradient lớn s được điều chỉnh theo một lượng tương đối nhỏ hơn so với các trọng số khác. Ngược lại, việc tách riêng weight decay của AdamW sẽ điều chỉnh tất cả các trọng số với cùng tốc độ λ, điều chỉnh hiệu quả các trọng số x với s lớn hơn so với chuẩn hóa L2 tiêu chuẩn. Như trong hình trên, AdamW tách riêng weight decay khỏi bước tối ưu hóa được thực hiện trong hàm mất mát (dòng 6) bằng cách chuyển nó xuống phần cập nhật trọng số (dòng 12). AdamW đem lại hiệu suất tổng quát tốt hơn đáng kể với việc tách riêng weight decay so với Adam L2 regularization. Điều này đã được chứng minh qua thực nghiệm, việc tách riêng weight decay làm cho các cài đặt tối ưu của tốc độ học tập và weight decay độc lập hơn nhiều, nhờ đó giảm bớt tối ưu hóa siêu tham số.

2.3.4 Hàm mất mát

Hàm mất mát (loss function) là một hàm số được sử dụng để đo lường mức độ sai lệch giữa đầu ra dự đoán của mô hình và giá trị thực tế (ground truth). Mục tiêu của hàm mất mát là tối thiểu hóa sai lệch này, từ đó cung cấp một đánh giá về hiệu suất của mô hình. Hàm mất mát được sử dụng cho OWL-ViT trong đề án như sau:

$$L_{total} = L_{class} + L_{Unmatched} + L_{GIoU} + L_{Box}$$

Trong đó L_{class} là hàm mất mát phân loại tương phản (contrast classification loss) được tính dựa trên những truy vấn đúng (positive queries) và truy vấn sai (negative queries) cho từng bounding box được dự đoán trong ảnh. Ngoài ra cũng có một hàm mất mát

$L_{Unmatched}$ để loại bỏ các bounding box giả được phát hiện với xác suất cao cho truy vấn đúng. Nghĩa là bounding box được dự đoán đúng nhãn phân loại nhưng kích thước bounding box này không khớp với ground truth box. Hàm mất mát của mô hình sử dụng hàm mất mát L1 (L1 loss) và hàm mất mát GIoU (GIoU loss) cho phần hàm mất mát định vị vật thể (localization loss). Hàm mất mát L1 còn được gọi là tổn thất tuyệt đối trung bình (mean absolute loss) hoặc sai số tuyệt đối trung bình (mean absolute error). Nó chỉ đơn giản là tổng của sự khác biệt tuyệt đối giữa giá trị thực tế y_{true} và giá trị dự đoán $y_{predict}$. GIoU (Generalized IoU) là một phiên bản cải tiến của IoU trong việc đánh giá độ tương đồng giữa hai bounding box trong phát hiện đối tượng. GIoU được đưa ra để giải quyết nhược điểm của IoU khi không có sự giao nhau giữa các hộp giới hạn. GIoU tính toán không chỉ tỷ lệ chồng lấn mà còn đánh giá sự gần nhau của hai bounding box. Nó sẽ tính toán tỷ lệ chồng lấn (overlap ratio) cùng với một thành phần khoảng cách (distance term) giữa hai hộp giới hạn. Kết quả là, GIoU có thể phân biệt được giữa các hộp không giao nhau và cho kết quả tốt hơn so với IoU trong các trường hợp này.

2.4 Kết luận chương

Trong chương 2, đề án trình bày từ tổng quan đến chi tiết các nghiên cứu liên quan của phương pháp. Đầu tiên là trình bày về mô hình phát hiện đối tượng từ vựng mở OWL-ViT với các phần Text Encoder, Vision Encoder và hàm mất mát của mô hình. Tiếp đến là mô tả chi tiết về kỹ thuật SAHI với hai phần sử dụng trong tinh chỉnh và sử dụng trong suy luận. Đề án cũng trình bày chi tiết ý tưởng kết hợp hai phần lại và đưa ra giải pháp cho vấn đề phát hiện đối tượng kích thước nhỏ trong ảnh chụp từ Drone. Phương pháp sử dụng SAHI để tăng cường dữ liệu cho tập dữ liệu đào tạo. Sau đó tinh chỉnh mô hình OWL-ViT với tập dữ liệu đào tạo đã được tăng cường. Đề án mô tả thuật toán tối ưu AdamW được cải tiến từ Adam và hàm mất mát sử dụng trong quá trình tinh chỉnh. Tiếp đến chương 3 viết về việc triển khai và đánh giá hiệu quả của phương pháp.

Chương 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1 Mô tả các bộ dữ liệu

3.1.1 Nguồn dữ liệu

Các thiết bị bay không người lái (UAVs), hay còn gọi là drone, được trang bị máy ảnh và đã được triển khai nhanh chóng trong một loạt các ứng dụng, bao gồm nông nghiệp, nhiếp ảnh từ không gian, giao hàng nhanh và giám sát. Do đó, việc hiểu tự động dữ liệu hình ảnh thu thập từ các nền tảng này trở nên rất đòi hỏi, đồng thời đưa thị giác máy tính gần gũi hơn với các drone. Vậy nên một bộ dữ liệu chuẩn đoán quy mô lớn với đánh dấu chính xác cho các nhiệm vụ quan trọng của thị giác máy tính, mang tên đã ra đời, nhằm kết hợp thị giác và drone lại với nhau. Bộ dữ liệu VisDrone2019 được thu thập bởi nhóm AISKYEYE tại Laboratoire of Machine Learning and Data Mining, Đại học Thiên Tân, Trung Quốc. Bộ dữ liệu gồm 288 đoạn video được tạo thành từ 261,908 khung hình và 10,209 ảnh tĩnh, được chụp bởi các máy ảnh gắn trên drone khác nhau, bao gồm nhiều khía cạnh khác nhau bao gồm vị trí (chụp từ 14 thành phố khác nhau cách nhau hàng nghìn km tại Trung Quốc), môi trường (đô thị và nông thôn), đối tượng (người đi bộ, xe cộ, xe đạp, v.v.), và mật độ (cảnh vắng và tắc nghẽn).

3.1.2 Chi tiết dữ liệu

Bộ dữ liệu được sử dụng trong đề án là bộ dữ liệu ảnh tĩnh từ chụp từ drone sử dụng để phát hiện đối tượng VisDrone2019-Detection, gồm các hình ảnh chụp từ trên cao, góc rộng với rất nhiều đối tượng kích thước nhỏ trong ảnh. Bao gồm 10,209 ảnh 3 tập đào tạo-xác thực-kiểm thử được với số lượng ảnh lần lượt là 6471, 548, 1610. Số lượng đối tượng là 2,6 triệu đối tượng bao gồm các loại sau: người đi bộ, đám đông người, xe đạp, ô tô, xe bán tải, xe tải hạng nặng, xe ba bánh, xe ba gác, xe bus

3.2 Quá trình tinh chỉnh mô hình OWL-ViT kết hợp SAHI

3.2.1 Thông số, độ đo

Đề án sử dụng độ đo trong phát hiện đối tượng là mAP viết tắt của mean Average Precision (độ chính xác trung bình). Một AP (Average Precision) được tính toán cho mỗi lớp đối tượng riêng lẻ trong bài toán object detection. mAP là giá trị trung bình của tất cả các AP này. Thông số mAP là một phép đo quan trọng để đánh giá tổng thể hiệu suất của mô hình object detection. Nó là một chỉ số phổ biến và được sử dụng rộng rãi trong cả nghiên cứu và ứng dụng thực tế.

3.2.2 Tăng cường dữ liệu với SAHI

Ở chương II, đề án đã việc tăng cường dữ liệu đào tạo cho bộ dữ liệu bằng kỹ thuật SAHI. Tại phần thực nghiệm này, ta chia các ảnh trong tập đào tạo thành các ảnh có kích thước nhỏ hơn (ảnh slice) là 840x840, 960x960 với tỷ lệ chồng lấn là 0 và 0.25. Từ đó ta được 4 tập dữ liệu nhỏ hơn là 840x840 tỷ lệ chồng lấn 1, 840x840 tỷ lệ chồng lấn 0.25, 960x960 tỷ lệ chồng lấn 0 và 960x960 tỷ lệ chồng lấn 0.25. Do số lượng đối tượng trong các lớp không được cân bằng, điều này dẫn tới việc mất cân bằng trong phát hiện đối tượng, nghĩa là mô hình sẽ dự đoán nghiên về lớp có số lượng đối tượng được học nhiều hơn, vì nó có nhiều hơn trong tập đào tạo. Để giải quyết vấn đề đề này ta áp dụng SAHI để tăng cường dữ liệu. Sau khi chia ảnh trong tập đào tạo thành các ảnh có kích thước nhỏ hơn, ta tăng cường

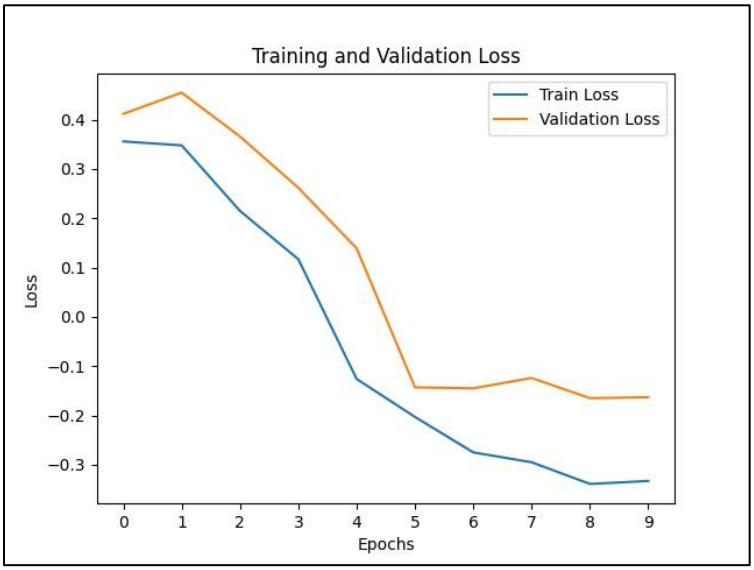
những lớp ít dữ liệu bằng cách lấy thêm các đối tượng trong ảnh slice trong quá trình tinh chỉnh.

Bảng 3.1: Mô tả chi tiết về số lượng đối tượng được tăng cường theo lớp

Lớp	**Tập đào tạo ban đầu**	**Slice 840-0**	**Slice 840-025**	**Slice 960-0**	**Slice 960-025**	**Tập đào tạo mới**
car (3)	**144.866**	280.028	299.263	328.313	359.478	***144.866***
pedestrian (0)	**79.337**	145.894	154.630	167.470	179.868	***79.337***
motor (9)	**29.647**	59.054	62.814	65.710	75.213	***29.647***
people (1)	**27.059**	52.687	57.093	57.047	66.631	***27.059***
van (4)	**24.956**	**45.081**	48.579	52.629	**57.374**	***127.411***
truck (5)	**12.875**	**25.088**	**26.231**	**28.988**	**31.362**	***124.544***
bicycle (2)	**10.480**	**20.641**	23.538	23.077	26.387	***31.121***
bus (8)	**5.926**	**10.087**	**10.335**	**11.750**	**12.538**	***50.656***
tricycle (6)	**4.812**	**9.612**	10.331	9.290	**11.597**	***25.921***
awning-tricycle (7)	**3.246**	**6.965**	**7.470**	6.997	**9.259**	***26.940***
Tổng						***667.502***

3.2.3 Tinh chỉnh mô hình với bộ dữ liệu Visdrone

Sau khi tăng cường dữ liệu của tập đào tạo ở trên bằng SAHI, tinh chỉnh mô hình OwL-ViT với 10 epochs, tốc độ học là 3e-7. Mô hình được đánh giá độ mất mát trên tập val của VisDrone2019-Detection, kết quả đào tạo mô hình được trình bày ở dưới đây.



Hình 3.1. Kết quả train loss và val loss trong quá trình huấn luyện

Nhận xét:
-Giá trị giữa train loss và val loss đều có xu hướng giảm dần trong quá trình huấn luyện. Từ những epoch 7 về sau, val loss và train loss không giảm nhiều.
-Mô hình đạt kết quả tốt nhất trên tập val với val loss = -0.165 khi train loss = -0.339 tại epoch thứ 8

3.3 Đánh giá kết quả

3.3.1 Kết quả Owl-ViT kết hợp SAHI

Từ kết quả của huấn luyện trên, chọn mô hình ở epoch thứ 8 có kết quả tốt nhất để đánh giá trên tập kiểm thử của bộ dữ liệu Visdrone2019. Sử dụng độ đo mAP với ngưỡng IoU = 0.5 theo 3 loại đối tượng lớn, vừa và nhỏ ở bảng 3.2 và 3.3 dưới đây:

Bảng 3.2: Kết quả đánh giá mô hình OwL-ViT + SAHI trên tập kiểm tra

STT	Độ đo	Kết quả	Ghi chú
1	mAP@0.5	28.5	Cho toàn bộ đối tượng
2	mAP@0.5s	17.5	Đối tượng kích thước nhỏ
3	mAP@0.5m	43.2	Đối tượng kích thước vừa
4	mAP@0.5l	48.7	Đối tượng kích thước lớn

Bảng 3.3: So sánh OwL-ViT + SAHI và OwL-ViT ban đầu trên tập kiểm tra

	mAP@0.5	mAP@0.5s	mAP@0.5m	mAP@0.5l
OwL-ViT ban đầu	21.3	10.7	38.5	45.2
OwL-ViT + SAHI	28.5	17.5	43.2	48.7

Nhận xét:
- Nhìn chung OwL+ SAHI tăng kết quả lên đáng kể so với OwL ban đầu chưa được fine-tune. Tăng mAP@0.5 trên toàn tập kiểm tra là 7.2%.
- Đối với đối tượng loại nhỏ mAP@0.5s tăng 6.8%. So với mô hình gốc ban đầu, đây là một kết quả khá tốt.

3.2.2 So sánh với các mô hình khác kết hợp SAHI

Với kết quả trên, ta so sánh mô hình OwL-ViT kết hợp SAHI với một số mô hình khác cũng kết hợp với SAHI được đánh giá cùng trên tập kiểm tra của bộ dữ liệu VisDrone2019.

Bảng 3.4: So sánh OwL-ViT kết hợp SAHI với các mô hình khác

	mAP@0.5	mAP@0.5s	mAP@0.5m	mAP@0.5l
FCOS	25.8	14.2	39.6	45.1
VFNet	28.8	16.8	44.0	47.5
TOOD	29.4	18.1	44.1	50.0
OwL-ViT + SAHI	28.5	17.5	43.2	48.7

Nhận xét:

- So sánh 4 mô hình trên, có thể thấy độ chính xác mAP của mô hình OwL-ViT kết hợp SAHI đứng thứ 3, chỉ hơn FCOS. Tuy nhiên so về loại đối tượng nhỏ mAP@0.5 thì mô hình lại đứng thứ 2, chỉ sau TOOD 0.6%.

-Với loại đối tượng lớn mAP@0.5l, mô hình cũng đứng thứ 2. Loại đối tượng vừa mAP@0.5m tuy cũng đứng thứ 3 nhưng chỉ kém TOOD là 0.9% và hơn khá nhiều so với FCOS là 3.6%.

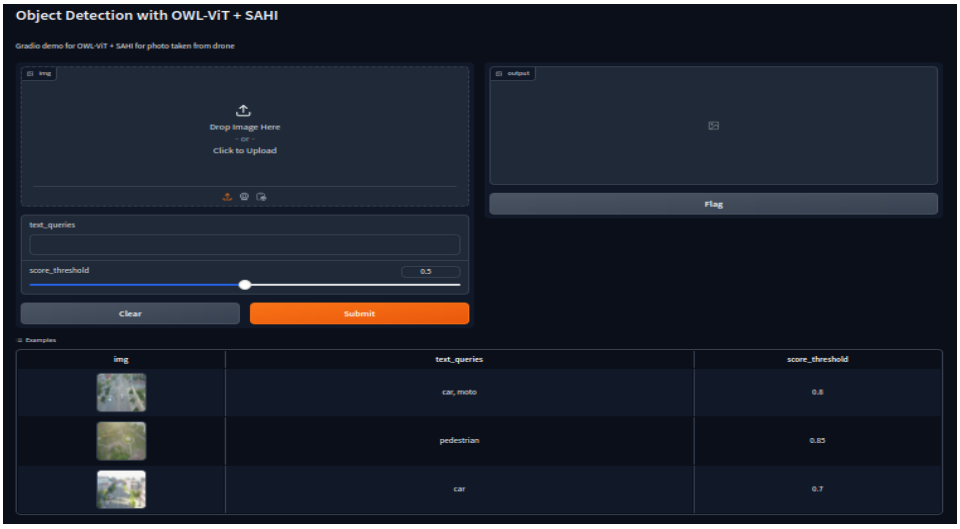
3.4 Demo

3.4.1. Giới thiệu về Hugging Face Gradio

Hugging Face Gradio là một công cụ mã nguồn mở được phát triển bởi Hugging Face, nhằm giúp xây dựng giao diện người dùng tương tác cho mô hình Trí tuệ Nhân tạo (AI) một cách dễ dàng. Gradio kết hợp giữa khả năng xử lý ngôn ngữ tự nhiên và hình ảnh của Hugging Face với Gradio, một thư viện giao diện người dùng tương tác mã nguồn mở. Gradio cung cấp các công cụ để tạo ra các giao diện tương tác cho mô hình AI một cách nhanh chóng, mà không yêu cầu nhiều kiến thức về lập trình hoặc giao diện người dùng. Với Gradio, bạn có thể xây dựng giao diện người dùng cho mô hình AI trong vài dòng mã, cho phép người dùng tương tác với mô hình và xem kết quả trực tiếp.

3.4.2 Xây dựng giao diện demo

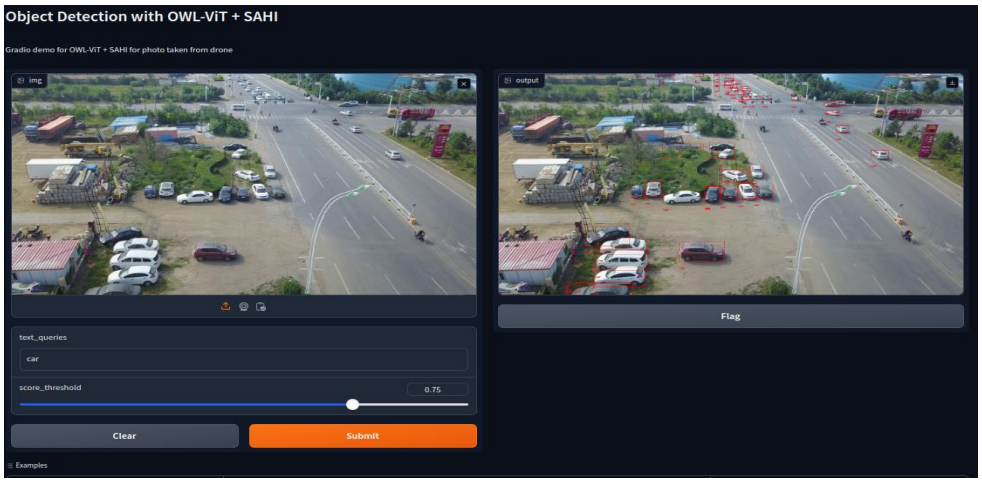
Xây dựng demo trên Hugging Face với mô hình OWL-ViT kết hợp SAHI đã được huấn luyện ở trên để phát hiện các đối tượng trong ảnh chụp từ Drone. Giao diện có hai phần chính là Input (bên trái) và Output(bên phải). Examples: Các ảnh ví dụ kèm ngưỡng có sẵn, người dùng có thể kéo vào để demo nhanh kết quả.



Hình 3.2: Giao diện demo của mô hình OWL-ViT kết hợp SAHI

3.4.3 Kết quả demo

Sau khi xây dựng giao diện demo xong, dưới đây là kết quả sử dụng các chức năng trên giao diện.



Hình 3.3. Kết quả phát hiện đối tượng trên giao diện Demo

Ngoài ra, có thể sử dụng phần example, phát hiện đối tượng cùng ảnh ngưỡng có sẵn để demo kết quả. Chỉ cần click vào ảnh muốn demo trong danh sách và ấn Submit để xem kết quả.

3.5 Kết luận chương

Với chương 3, đề án trình bày cụ thể về quá trình thực nghiệm và đánh giá hiệu quả của phương pháp được đề xuất. Mô tả chi tiết về các bộ dữ liệu được sử dụng trong quá trình tinh chỉnh và đánh giá. Trình bày cách thức tăng cường bộ dữ liệu đào tạo của tập dữ liệu Visdrone2019 bằng SAHI. Đồng thời, đề án cũng trình bày về quá trình tinh chỉnh mô hình OWL-ViT với các bước cụ thể. Đánh giá các kết quả của mô hình OWL+SAHI với mô hình gốc OWL-ViT để thấy được sự hiệu quả. So sánh thêm với các kết quả đã có với các phương pháp khác trên các bộ dữ liệu sử dụng. Xây dựng demo cho phương pháp để dễ dàng sử dụng mô hình phát hiện đối tượng cho ảnh, trực quan hóa kết quả.

KẾT LUẬN

Các kết quả đạt được của đề án tốt nghiệp:

- Nghiên cứu các bài toán phát hiện đối tượng, phát hiện đối tượng từ vựng mở và phát hiện đối tượng có kích thước nhỏ trong ảnh chụp từ drone. Trình bày kiến trúc mô hình OWL-ViT, kỹ thuật SAHI.
- Nghiên cứu phương pháp kết hợp giữa mô hình phát hiện đối tượng từ vựng mở OWL-ViT và kỹ thuật SAHI giúp phát hiện đối tượng kích thước nhỏ trong ảnh hiệu quả hơn. Từ đó tận dụng được kết hợp giữa văn bản và hình ảnh để phát hiện được nhiều đối tượng kích cỡ nhỏ trong ảnh chụp từ drone.
- Tiến hành thực nghiệm và đánh giá độ hiệu quả của phương pháp OWL-ViT cho việc phát hiện đối tượng kích thước nhỏ trong ảnh chụp từ drone. Xây dựng demo cho phương pháp để dễ dàng sử dụng mô hình phát hiện đối tượng và trực quan hóa kết quả.

Hướng nghiên cứu tiếp theo:

- Mô hình: nghiên cứu mô hình giúp phát hiện đối tượng với tốc độ nhanh và chính xác hơn cho video. Nghiên cứu thêm các kỹ thuật có thể áp dụng để giải quyết những vấn đề khác trong phát hiện đối tượng.
- Ứng dụng: mở rộng nghiên cứu tính ứng dụng của phương pháp có thể áp dụng vào những lĩnh vực nào khác trong cuộc sống để xây dựng sản phẩm thực tế.