BÁO CÁO TUẦN 2

Dự án: AI Nhận diện Bối cảnh Trong Khung Hình
Thời gian: 12/4– 10/4

1. Mục tiêu trong tuần

- Tìm hiểu và thử nghiệm các mô hình mô tả ảnh (image captioning).
- Lựa chọn thuật toán nhận diện đối tượng phù hợp.
- Tìm và chọn bộ dữ liệu huấn luyện.
- Xây dựng pipeline xử lý ảnh đầu vào.
- Kết hợp nhận diện đối tượng với sinh mô tả ngữ cảnh.

2. Công việc đã thực hiện

- **a. Khảo sát & chọn mô hình sinh mô tả ảnh (Image Captioning)**

Mô hình đã khảo sát:

Mô hình	Kiến trúc chính	Ưu điểm	Nhược điểm
Show and Tell	CNN + LSTM	Cơ bản, dễ hiểu	Độ chính xác thấp
Show, Attend and Tell	CNN + Attention + LSTM	Có chú ý, mô tả sát hơn	Khó triển khai hơn
BLIP	Vision Transformer + LM	Pretrained mạnh, tự nhiên	Cần GPU, nặng
OFA	Unified Transformer	Đa nhiệm (caption, VQA...)	Cấu hình phức tạp
GIT	ViT + T5	Tự nhiên, mạnh mẽ	Cốn mới, ít hỗ trợ

Mô hình đã chọn: BLIP

- Pretrained mạnh, tự nhiên.
- Dễ triển khai với HuggingFace.
- Kết hợp tốt với YOLO và nhận diện vật thể.

Kết quả thử nghiệm:

- Tạo được mô tả khá tự nhiên và đúng ngữ cảnh.

b. Chọn thuật toán nhận diện đối tượng

Thuật toán đã chọn: YOLOv8

- Hiệu suất cao, nhẹ gọn.
- Hỗ trợ object detection, segmentation, pose.

- Dễ train và fine-tune theo nhu cầu dự án.

c. Chọn bộ dữ liệu huấn luyện

Đã chọn: MS COCO dataset

- 330k+ ảnh, 80 loại đối tượng, 5+ caption/1 ảnh.
- Annotation rõ ràng cho cả detection và captioning.
- Được dùng luyện các mô hình SOTA như YOLO, BLIP...

d.Xây dựng pipeline xử lý ảnh

- Ảnh đầu vào → Resize + Normalize → YOLOv8 nhận diện → BLIP sinh mô tả.
- Có thể kết hợp caption theo đối tượng (region-based captioning).

3.Kết quả đạt được

- Chạy thành công BLIP và sinh mô tả tự nhiên.
- Nhận diện đối tượng bằng YOLOv8.
- Hoàn thiện pipeline: ảnh → nhận diện → mô tả.
- Chọn xong dữ liệu COCO để huấn luyện/fine-tune.

4. Khó khăn gặp phải

- BLIP cần GPU mạnh, khá nặng.
- Dữ liệu thế loại đặc thù thiếu (nếu muốn đặc thù hoá theo ngành).
- Caption sinh ra đôi khi chung chung hoặc thừa tự nhiên.

5. Kế hoạch tuần sau

- So sánh thêm mô hình caption (OFA, Show and Tell).
- Tích hợp giao diện test pipeline (web/app).
- Fine-tune BLIP trên COCO hoặc bộ dữ liệu riêng.
- Bắt đầu tích hợp YOLO + captioning trong hệ thống.

Khả năng tự huấn luyện mô hình

1. Mô hình sinh mô tả ảnh (BLIP)

Tự train theo 2 cách:

- **Fine-tune BLIP pretrained**:
  - Cần GPU (>=8GB), dữ liệu đúng format.
  - Dễ dùng, nhanh và hiệu quả cao.
- **Train from scratch**:
  - Cần multi-GPU, dữ liệu lớn, thời gian luyện lâu.
  - Dẻ overfit nếu dữ liệu nhỏ.

2. Mô hình nhận diện đối tượng (YOLO)

Có thể tự train YOLOv5, YOLOv8, YOLOv7...

- Cần ảnh + annotation định dạng YOLO (.txt).
- Có thể dùng LabelImg, Roboflow, Makesense.ai để gán nhãn.
- Chạy dễ dàng bằng Python hoặc CLI (Ultralytics).

Thời gian train nhanh (1-2h), GPU trung bình có thể chạy được.

Chiến lược dự án:

Thành phần	Hành động	Mục tiêu
BLIP	Fine-tune từ pre-trained	Tăng tính tự nhiên và đúng ngữ cảnh
YOLOv8	Tự train từ đầu	Nhận diện sát các vật thể trong ảnh