TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI TP. HỒ CHÍ MINH

VIỆN CÔNG NGHỆ THÔNG TIN VÀ ĐIỆN, ĐIỆN TỬ

Deep Learning

010112710502 - Group 8 - Topic 2

Object Detection (YOLO-style Head on COCO)

Lecturer: PhD. Nguyễn Thị Khánh Tiên

No.	Name	ID
1	Lê Văn An	0560205001827
2	Phạm Gia Bảo	093205005065
3	Dương Hưng	086205003910
4	Huỳnh Hậu	066205010571
5	Huỳnh Thị Cẩm Giang	051305006893

HỒ CHÍ MINH, 2026

Mục Lục

Tóm Tắt

Đề tài Object Detection của học phần Deep Learning tập trung vào việc xây dựng một mô hình phát hiện đối tượng từ đầu trên bộ dữ liệu COCO 2017 đã được tùy chỉnh, trong đó số lượng lớp được giảm từ 80 xuống còn 10 nhằm phù hợp với mục tiêu nghiên cứu và điều kiện tính toán. Sau quá trình tách lọc, tập dữ liệu thu được gồm khoảng 30.000 ảnh cho tập huấn luyện và 1.303 ảnh cho tập validation. Mô hình được thiết kế gồm ba thành phần chính: backbone ResNet-50 pretrained để trích xuất đặc trưng, neck FPN nhằm khai thác thông tin đa tỉ lệ, và detection head theo YOLO-style để dự đoán bounding box, objectness và nhãn lớp. Toàn bộ pipeline được triển khai bằng PyTorch và huấn luyện trên nền tảng Kaggle, cho phép kiểm soát chi tiết quá trình huấn luyện và đánh giá mô hình.

Kết quả thực nghiệm cho thấy mô hình có khả năng hội tụ ổn định, tuy nhiên hiệu năng vẫn còn ở mức cơ bản, với chỉ số AP@0.5 cao nhất đạt khoảng 0.48. Kết quả này phản ánh tính chất baseline của mô hình trong phạm vi đề tài, đồng thời cho thấy tiềm năng cải thiện trong các nghiên cứu tiếp theo thông qua việc mở rộng dữ liệu, tăng cường augmentation và tối ưu kiến trúc mô hình.

Chương 1. Giới thiệu

1.1. Mục tiêu

Mục tiêu của đề tài là xây dựng và triển khai một mô hình object detection dựa trên mạng nơ-ron tích chập (Convolutional Neural Networks – CNNs) nhằm phát hiện và phân loại các đối tượng trong hình ảnh. Thông qua đề tài, sinh viên củng cố kiến thức về CNNs trong việc trích xuất đặc trưng không gian, đồng thời hiểu rõ cách mở rộng CNN cho bài toán phức tạp hơn như object detection thông qua các thành phần backbone, neck và detection head. Bên cạnh đó, đề tài hướng đến việc đánh giá mô hình bằng các chỉ số chuẩn của object detection như AP50 và mAP.

1.2. Phạm vi và đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là bài toán phát hiện đối tượng trong hình ảnh tĩnh dựa trên CNNs và học sâu. Phạm vi nghiên cứu được giới hạn trên bộ dữ liệu COCO 2017, với ~30,000 ảnh cho tập huấn luyện và ~1,000 ảnh cho tập đánh giá. Xây dựng lớp COCODataset để có thể đọc được ảnh đầu vào, và metadata chứa (height, width), bbox, ground truth của ảnh. Mô hình sử dụng ResNet-50 pretrained làm backbone để trích xuất đặc trưng, kết hợp Feature Pyramid Network (FPN) nhằm khai thác đặc trưng đa tỉ lệ và YOLO-style head để dự đoán bounding box và nhãn lớp. Quá trình huấn luyện và đánh giá được triển khai bằng PyTorch trên nền tảng Kaggle, không tập trung vào các kỹ thuật tối ưu nâng cao hay so sánh với các mô hình state-of-the-art.

Chương 2. Tổng quan lý thuyết

2.1. Bài toán Object Detection

Object Detection (phát hiện đối tượng) là bài toán trong thị giác máy tính nhằm xác định vị trí và phân loại các đối tượng xuất hiện trong ảnh. Đầu ra của mô hình bao gồm các bounding box biểu diễn vị trí đối tượng và nhãn lớp tương ứng. Khác với bài toán phân loại ảnh, Object Detection đồng thời giải quyết hai nhiệm vụ là định vị và phân loại, do đó có độ phức tạp cao hơn. Bài toán này đóng vai trò quan trọng trong nhiều ứng dụng thực tế như giám sát an ninh, xe tự hành và robot.

2.2. Các phương pháp phát hiện đối tượng phổ biến

Các phương pháp Object Detection hiện nay có thể chia thành hai nhóm chính: two-stage detector và one-stage detector. Two-stage detector, tiêu biểu là Faster R-CNN, thực hiện lần lượt hai bước gồm sinh vùng đề xuất và phân loại, hồi quy bounding box, thường đạt độ chính xác cao nhưng chi phí tính toán lớn. Ngược lại, one-stage detector như YOLO và SSD thực hiện trực tiếp việc dự đoán bounding box và nhãn lớp trong một bước duy nhất, cho tốc độ xử lý nhanh và kiến trúc đơn giản hơn, phù hợp với các ứng dụng yêu cầu thời gian thực.

2.3. Backbone và Feature Pyramid Network (FPN)

Trong các mô hình Object Detection hiện đại, backbone CNN đóng vai trò trích xuất đặc trưng từ ảnh đầu vào. Các mạng như ResNet thường được sử dụng nhờ khả năng học đặc trưng sâu và ổn định. Ngoài ra, do các đối tượng trong ảnh có kích thước khác nhau, việc khai thác đặc trưng đa tỉ lệ là rất quan trọng. Feature Pyramid Network (FPN) được đề xuất nhằm kết hợp thông tin từ các tầng có độ phân giải khác nhau, giúp cải thiện khả năng phát hiện cả đối tượng nhỏ và lớn.

Chương 3. Dataset và Tiền xử lý dữ liệu

3.1. Dataset

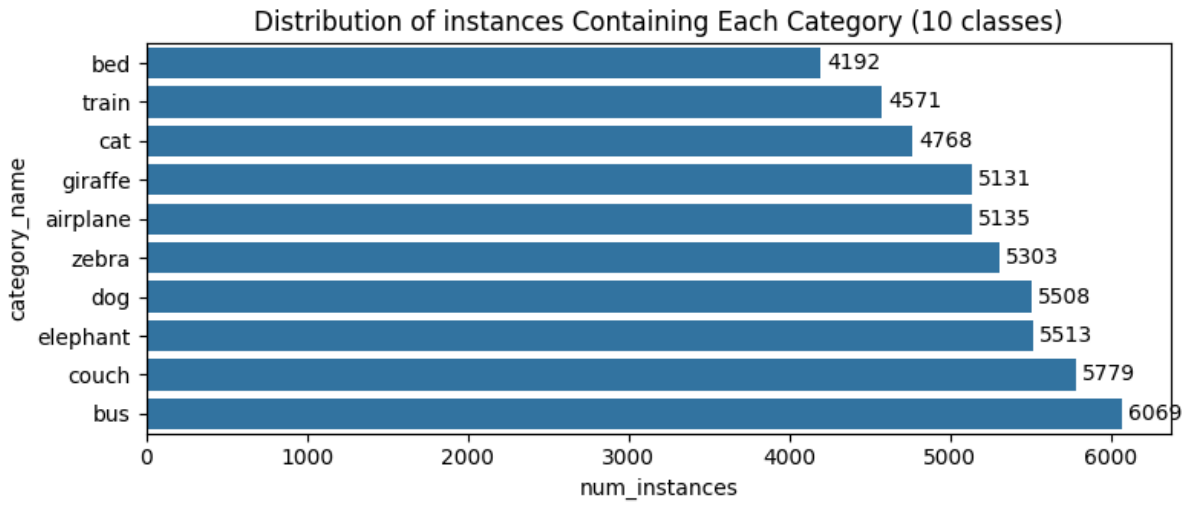
Bộ dữ liệu được sử dụng trong đề tài là COCO 2017, phù hợp với yêu cầu của bài toán phát hiện đối tượng (Object Detection). Phiên bản COCO 2017 bao gồm khoảng 118.000 ảnh trong tập train và 5.000 ảnh trong tập validation, với tổng cộng 80 lớp đối tượng khác nhau. Dataset gồm hai thành phần chính: thư mục ảnh chứa toàn bộ các ảnh đầu vào phục vụ cho quá trình huấn luyện và đánh giá mô hình, và tệp instances.json ở định dạng JSON, đóng vai trò là metadata mô tả chi tiết thông tin nhãn của từng ảnh. File này chứa các annotation về đối tượng, bounding box, phân loại lớp, mối quan hệ giữa ảnh và nhãn, đồng thời cung cấp thêm các thông tin liên quan đến giấy phép hình ảnh và dữ liệu cho bài toán segmentation. COCO 2017 là một bộ dữ liệu chuẩn, được cộng đồng nghiên cứu sử dụng rộng rãi trong huấn luyện và so sánh hiệu năng của các mô hình phát hiện đối tượng.

3.2. Tiền xử lý

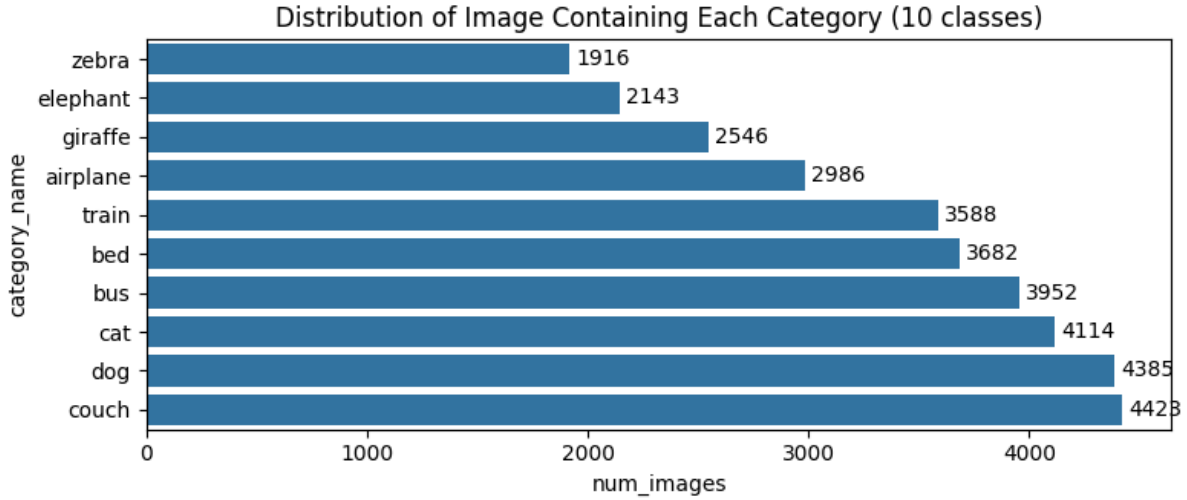
Do COCO 2017 có quy mô lớn với 80 lớp đối tượng, việc huấn luyện trực tiếp trên toàn bộ dataset đòi hỏi nhiều tài nguyên tính toán. Vì vậy, theo yêu cầu của đề tài và giảng viên hướng dẫn, số lượng lớp được giảm xuống còn 10 nhằm phù hợp với mục tiêu nghiên cứu và điều kiện phần cứng. Các lớp được lựa chọn để sử dụng trong đề tài bao gồm: airplane, bus, train, cat, dog, elephant, zebra, giraffe, couch và bed. Sau quá trình tách lọc, tập dữ liệu thu được gồm 30.919 ảnh cho tập train và 1.303 ảnh cho tập validation.

Quá trình tùy chỉnh dataset được thực hiện trực tiếp trên tệp instances.json bằng cách thống kê các lớp trong tập train, lựa chọn 10 lớp mục tiêu và lọc lại categories, annotations cũng như images tương ứng, chỉ giữ những ảnh có chứa bounding box của các lớp đã chọn. Cách làm này giúp giảm dung lượng dữ liệu và tối ưu quá trình huấn luyện, đồng thời vẫn đảm bảo đúng cấu trúc của bài toán Object Detection.

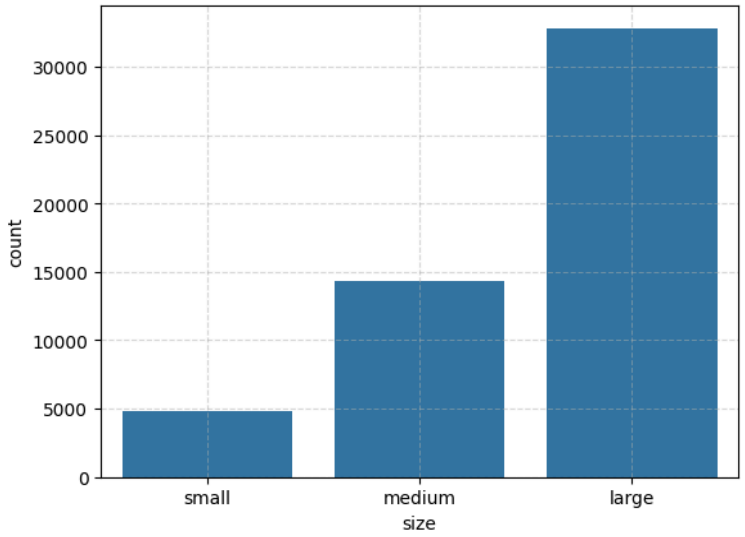
Bên cạnh đó, trong giai đoạn tiền xử lý hình ảnh, kỹ thuật letterbox được áp dụng để đưa các ảnh về cùng kích thước đầu vào của mô hình. Phương pháp này thực hiện thay đổi kích thước ảnh theo tỉ lệ ban đầu kết hợp với padding vào các vùng trống, nhằm tránh làm méo hình và giữ nguyên tỉ lệ đối tượng. Nhờ đó, chất lượng thông tin hình ảnh được bảo toàn tốt hơn, góp phần cải thiện hiệu quả huấn luyện và độ chính xác của mô hình.



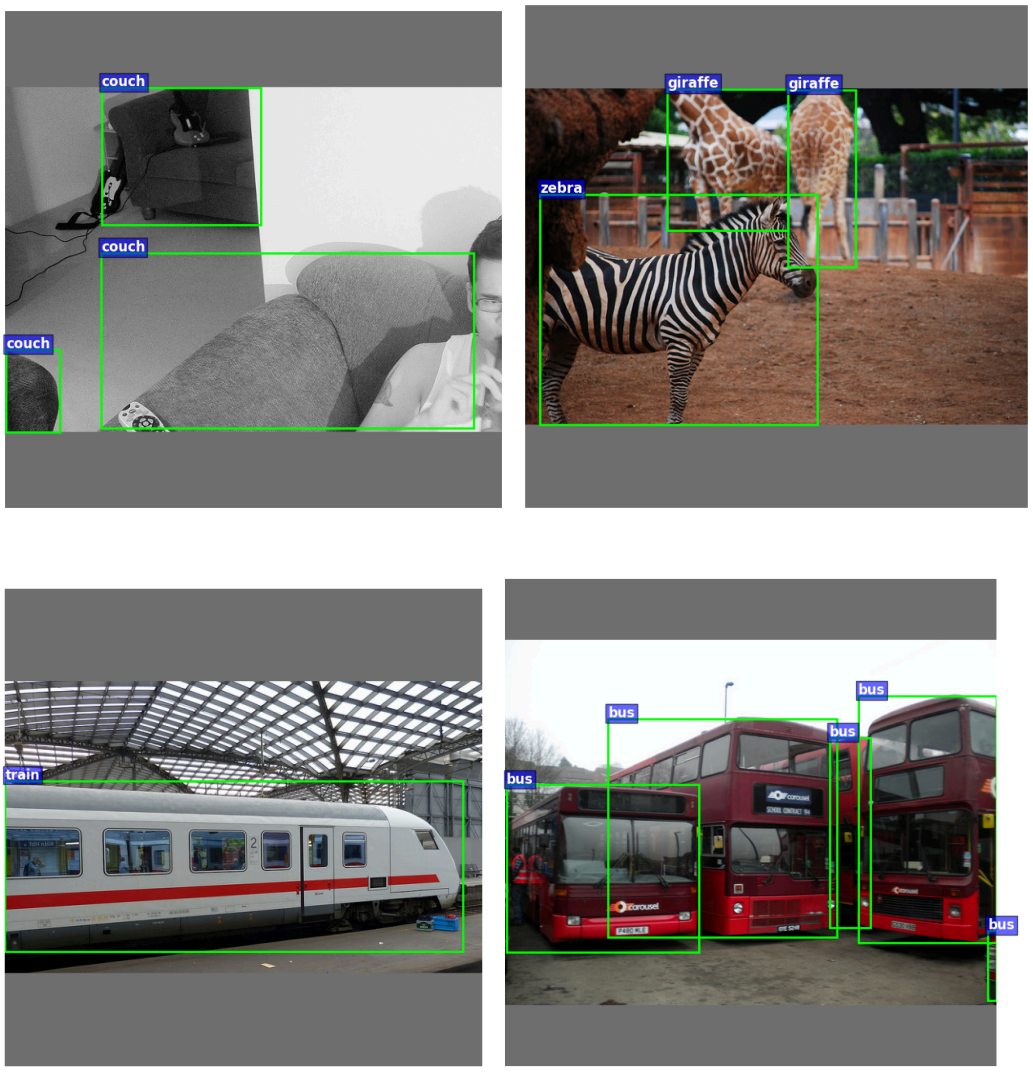
Hình 3.2.1. Phân phối số lượng instances theo từng category.



Hình 3.2.2. Phân phối số lượng image theo từng category.



Hình 3.2.3. Thống kê kích thước Bbox: nhỏ – trung bình – lớn (10 classes)



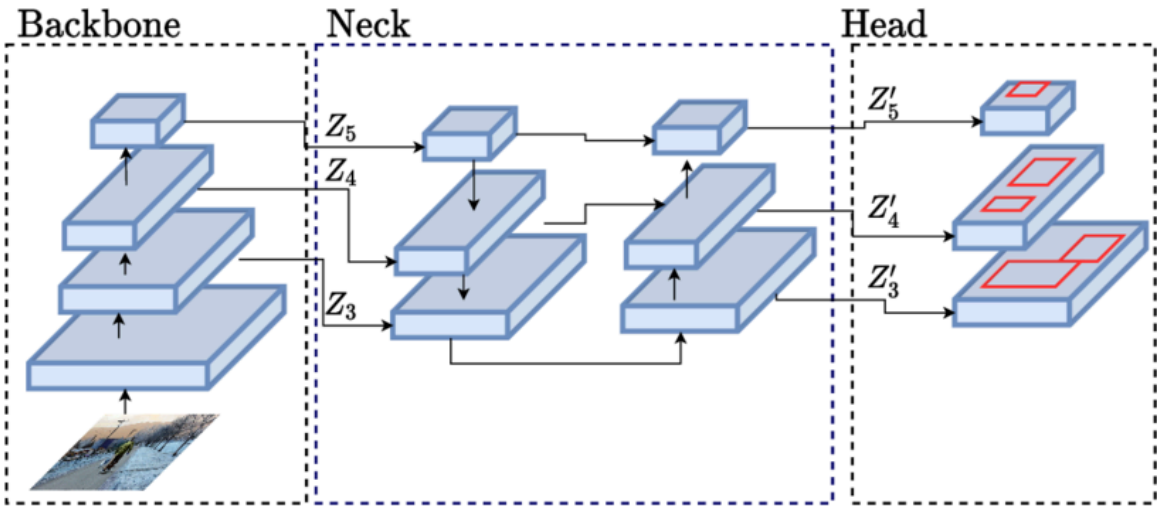
Hình 3.2.4. Ảnh và Bbox kèm theo labels của train dataset.

Chương 4. Phương pháp đề xuất

4.1. Kiến trúc mô hình

Mô hình được xây dựng theo hướng one-stage detector, gồm ba thành phần chính: backbone ResNet50, Feature Pyramid Network (FPN) và head dự đoán dạng decoupled. ResNet50 được sử dụng để trích xuất đặc trưng, khởi tạo với trọng số pretrained và cố định các tầng đầu nhằm giữ lại đặc trưng cơ bản và giảm chi phí huấn luyện. Từ backbone, ba mức đặc trưng C3, C4 và C5 được đưa vào FPN để tạo ra các feature map đa tỉ lệ P3, P4 và P5, giúp mô hình phát hiện hiệu quả các đối tượng có kích thước khác nhau.

Head dự đoán được thiết kế theo dạng decoupled, trong đó nhánh phân loại và nhánh hồi quy được tách riêng. Nhánh hồi quy dự đoán bounding box và objectness, trong khi nhánh phân loại dự đoán xác suất các lớp đối tượng. Kiến trúc này giúp tăng khả năng biểu diễn và ổn định hơn trong huấn luyện. Tổng thể mô hình đảm bảo cân bằng giữa độ chính xác và chi phí tính toán, phù hợp cho bài toán phát hiện 10 lớp đối tượng.



Hình 4.1.1. Sơ đồ tổng quát kiến trúc mô hình phát hiện đối tượng.

4.2. Hàm mất mát

Hàm mất mát của mô hình được xây dựng theo dạng tổng hợp, bao gồm ba thành phần chính: mất mát hồi quy bounding box, mất mát objectness và mất mát phân loại, với mục tiêu đồng thời tối ưu vị trí, độ tin cậy và nhãn lớp của đối tượng. Tổng hàm mất mát được biểu diễn:

$$L = \lambda_{box} L_{box} + \lambda_{obj} L_{obj} + \lambda_{cls} L_{cls}$$

Trong đó, L_{box} sử dụng CIoU loss để đánh giá mức độ trùng khớp giữa bounding box dự đoán và ground truth, giúp cải thiện cả vị trí, kích thước và hình dạng hộp bao. Hai thành phần L_{obj} và L_{cls} được tính bằng Binary Cross Entropy với logits, lần lượt cho bài toán xác định sự tồn tại của đối tượng và phân loại lớp.

Các trọng số $\lambda_{box} = 0.1$, $\lambda_{obj} = 0.7$, $\lambda_{cls} = 0.5$, được thiết lập nhằm cân bằng đóng góp của từng thành phần. Ngoài ra, mô hình áp dụng hệ số cân bằng theo từng mức đặc trưng P3, P4 và P5 để ổn định huấn luyện trên nhiều tỉ lệ. Thiết kế này giúp mô hình học hiệu quả cả về định vị và phân loại đối tượng trong một khuôn khổ thống nhất.

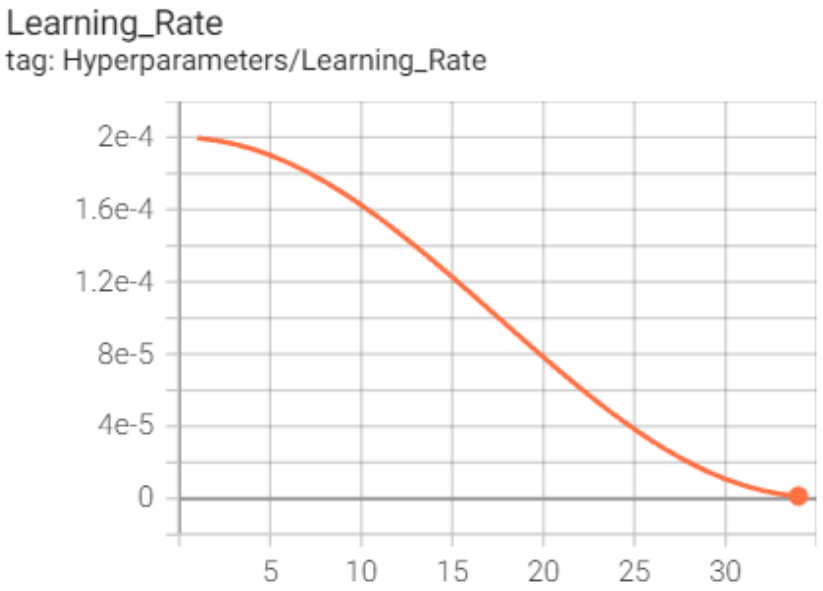
4.3. Thiết lập huấn luyện

Mô hình được huấn luyện bằng thuật toán tối ưu AdamW với learning rate ban đầu là 2×10^{-4} và hệ số weight decay 1×10^{-4}. AdamW giúp cải thiện khả năng tổng quát hóa so với Adam thông thường thông qua việc tách riêng cơ chế suy giảm trọng số khỏi cập nhật gradient.

Chiến lược điều chỉnh learning rate sử dụng Cosine Annealing, trong đó learning rate giảm dần theo dạng cosin từ giá trị ban đầu xuống giá trị tối thiểu 1×10^{-4} trong suốt quá trình huấn luyện. Cách tiếp cận này giúp mô hình hội tụ ổn định và hạn chế rơi vào điểm tối ưu cục bộ kém.

Quá trình huấn luyện được thực hiện trong 35 epoch, batch size là 32, với mỗi epoch bao gồm một vòng huấn luyện và một vòng đánh giá trên tập validation. Mixed Precision Training được áp dụng thông qua GradScaler nhằm giảm tiêu thụ bộ nhớ GPU và tăng tốc độ tính toán.

Hiệu năng mô hình được theo dõi bằng chỉ số mAP@0.5 trên tập validation. Mô hình có giá trị mAP@0.5 tốt nhất sẽ được lưu lại làm checkpoint tốt nhất. Đồng thời, các giá trị loss và metric của từng epoch được ghi lại bằng TensorBoard để phục vụ việc theo dõi và phân tích quá trình huấn luyện.



Hình 4.3.1. Đường cong điều chỉnh Learning Rate theo Cosine Annealing

Chương 5. Thực nghiệm và Kết quả

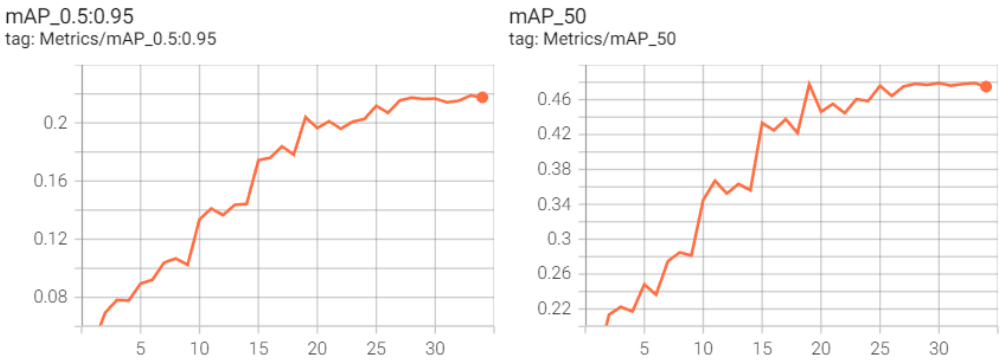
5.1. Metric đánh giá

Trong đề tài, hiệu năng của mô hình được đánh giá thông qua hai chỉ số chính là AP@0.5 (AP50) và mAP@0.5:0.95. AP@0.5 đo độ chính xác trung bình của mô hình tại ngưỡng IoU bằng 0.5, phản ánh khả năng phát hiện đối tượng trong điều kiện tương đối dễ. Trong khi đó, mAP@0.5:0.95 là giá trị trung bình AP trên nhiều ngưỡng IoU từ 0.5 đến 0.95 với bước 0.05, thể hiện khả năng tổng quát của mô hình trong các điều kiện đánh giá khắt khe hơn. Hai metric này được sử dụng phổ biến trong bài toán Object Detection và phù hợp để phản ánh toàn diện hiệu năng của mô hình.

Ngoài hai metric trên, các chỉ số như Precision, Recall, F1-score và Average Recall cũng thường được sử dụng trong đánh giá mô hình Object Detection. Tuy nhiên, trong phạm vi đề tài, AP@0.5 và mAP@0.5:0.95 được lựa chọn làm hai chỉ số chính để đánh giá hiệu năng mô hình.

5.2. Kết quả định lượng

Kết quả định lượng được thu thập trên tập validation trong suốt quá trình huấn luyện. Hình minh họa cho thấy cả hai chỉ số mAP@0.5:0.95 và AP@0.5 đều tăng dần theo số epoch và có xu hướng hội tụ ở các epoch cuối. Ở giai đoạn cuối quá trình huấn luyện, mô hình đạt mAP@0.5:0.95 xấp xỉ 0.22 và AP@0.5 đạt khoảng 0.48. Kết quả này cho thấy mô hình đã học được các đặc trưng cơ bản phục vụ cho bài toán phát hiện đối tượng trên tập dữ liệu đã được tùy chỉnh, đồng thời đóng vai trò như một baseline cho các hướng cải tiến trong tương lai.



Hình 5.2.1. Biểu đồ kết quả mAP và AP50.

Biểu đồ loss cho thấy box loss và classification loss trên tập train giảm rõ rệt theo epoch, trong khi các giá trị tương ứng trên tập validation giảm chậm hoặc có xu hướng tăng. Điều này cho thấy mô hình có dấu hiệu overfitting, đặc biệt ở nhánh phân loại, khi mô hình học tốt dữ liệu huấn luyện nhưng khả năng tổng quát hóa sang tập validation còn hạn chế. Tuy vậy, các metric mAP vẫn tăng dần, cho thấy mô hình vẫn cải thiện được khả năng phát hiện đối tượng và có thể xem như một baseline hợp lệ cho bài toán.

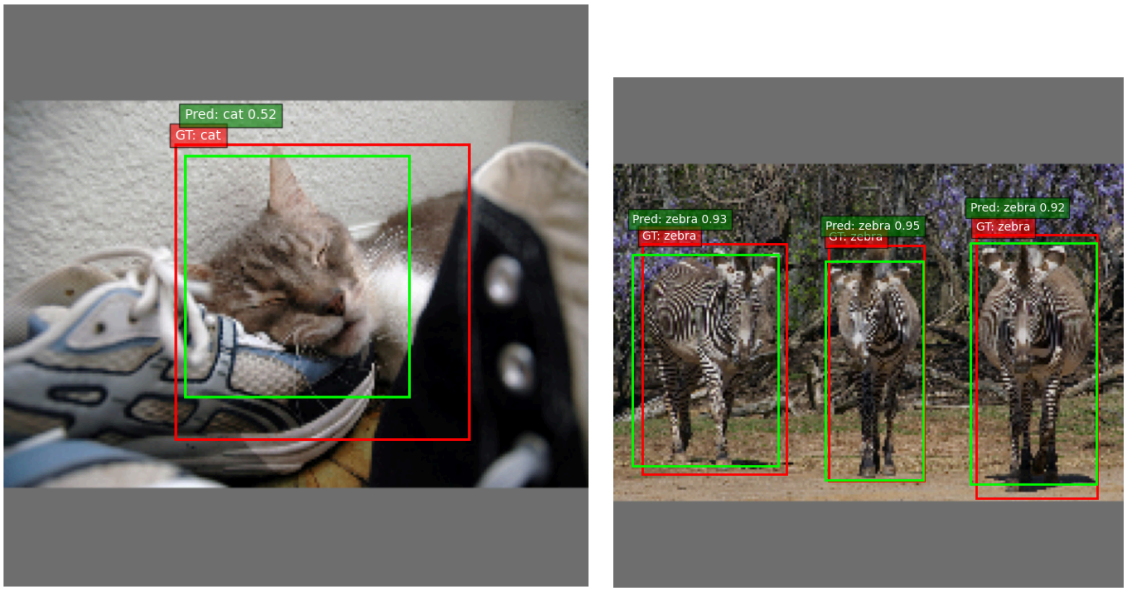
Hình 5.2.2. Đồ thị biểu diễn loss/box và loss/class

Hình 5.3.2. Đồ thị biểu diễn loss/objectness và loss/totalLoss

5.3. Kết quả định tính

Kết quả trực quan cho thấy mô hình có khả năng phát hiện và phân loại đúng các đối tượng mục tiêu trong nhiều bối cảnh. Ở ảnh thứ nhất, mô hình nhận diện đúng lớp cat với bounding box bao phủ tương đối tốt đối tượng, dù vẫn còn sai lệch nhẹ về vị trí. Ở ảnh thứ hai, mô hình phát hiện đồng thời nhiều đối tượng zebra với độ tin cậy cao, cho thấy khả năng xử lý tốt các trường hợp có nhiều đối tượng trong cùng một ảnh.

Tuy nhiên, vẫn tồn tại một số ảnh mà mô hình không nhận diện được hoặc bỏ sót đối tượng, đặc biệt trong các trường hợp đối tượng nhỏ, bị che khuất hoặc bối cảnh phức tạp. Điều này cho thấy mô hình hiện tại phù hợp ở mức baseline và còn nhiều dư địa để cải thiện trong các nghiên cứu tiếp theo.



Hình 5.3.1. So sánh bounding box ground truth và bounding box dự đoán của mô hình trên tập validation

Chương 6. Phân tích và Thảo luận

Kết quả thực nghiệm cho thấy mô hình có khả năng học và hội tụ ổn định, tuy nhiên xuất hiện hiện tượng overfitting, khi mô hình học rất tốt trên tập huấn luyện nhưng khả năng tổng quát hóa sang tập validation còn hạn chế. Mô hình vẫn gặp khó khăn trong việc phát hiện các đối tượng nhỏ, các trường hợp bị che khuất hoặc xuất hiện trong bối cảnh phức tạp, đồng thời tồn tại hiện tượng nhầm lẫn giữa các lớp có đặc trưng hình thái tương đồng, dẫn đến cả false positive và false negative.

Nguyên nhân chủ yếu đến từ việc tập dữ liệu huấn luyện đã được rút gọn, số lượng ảnh và mức độ đa dạng chưa cao, cũng như việc mô hình chưa áp dụng các kỹ thuật augmentation nâng cao hay các cơ chế regularization mạnh. Do đó, kết quả đạt được trong đề tài mang tính chất baseline cho bài toán phát hiện 10 lớp đối tượng trên COCO 2017 đã tùy chỉnh.

Mặc dù hiệu năng chưa cao, mô hình vẫn đảm bảo tính đúng đắn về mặt kiến trúc và quy trình huấn luyện, tạo nền tảng cho việc thử nghiệm, mở rộng và cải tiến các phương pháp nâng cao trong những nghiên cứu tiếp theo.

Chương 7. Kết luận và Hướng phát triển

Đề tài đã xây dựng thành công một mô hình phát hiện đối tượng theo hướng one-stage detector, áp dụng backbone ResNet50 kết hợp FPN và head dự đoán dạng decoupled, cùng với hàm mất mát kết hợp CIoU và BCE. Mô hình cho thấy khả năng hội tụ ổn định và đạt được các chỉ số đánh giá phù hợp với mục tiêu nghiên cứu trên tập dữ liệu COCO 2017 đã được tùy chỉnh còn 10 lớp.

Qua quá trình thực hiện, bài học rút ra là việc lựa chọn kiến trúc phù hợp, thiết lập hàm mất mát và chiến lược điều chỉnh learning rate đóng vai trò quan trọng đối với hiệu năng mô hình. Bên cạnh đó, chất lượng và sự đa dạng của dữ liệu huấn luyện có ảnh hưởng trực tiếp đến khả năng tổng quát hóa.

Trong tương lai, mô hình có thể được cải tiến bằng cách mở rộng tập dữ liệu, bổ sung các kỹ thuật tăng cường dữ liệu nâng cao, thử nghiệm các backbone mạnh hơn hoặc tích hợp các cơ chế attention và neck nâng cao như PANet. Ngoài ra, việc tinh chỉnh siêu tham số và tối ưu cho các ứng dụng thời gian thực cũng là những hướng phát triển tiềm năng.

Phụ lục

Phụ lục A. Link Repo Github

Toàn bộ mã nguồn của đề tài được công khai tại:
GitHub: *Repo-DL-ObjectDetetion-Group8*

Phụ lục B. Các thông số huấn luyện

Tham số	Giá trị
Train dataset size	~30k image
Val dataset size	~1k image
Số lớp (categories COCO2017)	10
Tên các lớp	Airplane, bus, train, cat, dog, elephant, zebra, giraffe, couch, bed
Kích thước ảnh	224x224
Optimizer	AdamW
Learning rate ban đầu	2e-4
Weight decay	1e-4
Scheduler	CosineAnnealingLR
Learning rate tối thiểu	1e-6
Batch size	32
Number epochs	35
Loss	CIou + BCE(obj, cls)

Tài liệu tham khảo:

1. Video - Introduction to Object Detection in Deep Learning
2. Slide - Deep learning - Chapter 5 - CNNs parts 1
3. Slide - Deep learning - Chapter 5 - CNNs parts 2 - Object detection survey