THÔNG TIN CHUNG

- Link YouTube video của báo cáo: https://youtu.be/qCSqRuyEheQ
- Link slides:

  Hoàn Nguyễn Huy - CS2205.SEP2025.DeCuong.FinalReport.Template.Slide

Họ và Tên: Nguyễn Huy Hoàn

MSHV: 250101022

- Lớp: CS2205.SEP2025
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 0
- Link Github:

  https://github.com/hoannh-uitgrad/Research-Methodology

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NGHIÊN CỨU VÀ ĐÁNH GIÁ HIỆU NĂNG MÔ HÌNH RT-DETR TRONG BÀI TOÁN PHÁT HIỆN ĐỐI TƯỢNG THỜI GIAN THỰC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

STUDY AND EVALUATION OF RT-DETR MODEL FOR REAL-TIME OBJECT DETECTION

TÓM TẮT *(Tối đa 400 từ)*

Trong thị giác máy tính, phát hiện đối tượng thời gian thực là bài toán nền tảng cho các ứng dụng trọng yếu như xe tự hành hay giám sát thông minh. Dù các mô hình YOLO hiện đang chiếm ưu thế nhờ cân bằng tốt giữa tốc độ và độ chính xác , chúng vẫn tồn tại điểm yếu cố hữu là sự phụ thuộc vào thuật toán hậu xử lý NMS (Non-Maximum Suppression). Cơ chế này gây ra độ trễ biến thiên, làm mất tính ổn định khi suy luận và cản trở quá trình tối ưu hóa đầu cuối. Nhằm khắc phục hạn chế đó, mô hình RT-DETR (công bố tại CVPR 2024) đã đề xuất kiến trúc đột phá giúp dung hòa hiệu quả tính toán của CNN với khả năng xử lý tập hợp của Transformer, loại bỏ hoàn toàn NMS.

Đề cương này trình bày kế hoạch tái hiện và đánh giá hiệu năng RT-DETR. Nghiên cứu tập trung phân tích sâu hai cải tiến cốt lõi: Bộ mã hóa lai hiệu quả (Efficient Hybrid Encoder) giúp giải quyết nút thắt tính toán bằng cách tách biệt xử lý đa tầng , và Cơ chế chọn lọc truy vấn (Uncertainty-minimal Query Selection) giúp tối ưu hóa việc khởi tạo dựa trên độ không chắc chắn.

Quá trình thực nghiệm sẽ được tiến hành trên tập dữ liệu chuẩn MS COCO val2017 , thực hiện so sánh đối chứng (benchmark) giữa RT-DETR (Backbone ResNet-50/101) và YOLOv8 (phiên bản L/X) trên cùng phần cứng GPU. Thông qua hai chỉ số Độ chính xác (AP) và Tốc độ (FPS), nghiên cứu kỳ vọng khẳng định vị thế của RT-DETR trong việc thiết lập chuẩn mực hiệu năng mới cho các tác vụ thời gian thực.

GIỚI THIỆU

1. Bối cảnh và Động lực nghiên cứu:

Phát hiện đối tượng thời gian thực (Real-time object detection) là bài toán then chốt trong thị giác máy tính, đóng vai trò nền tảng cho các hệ thống xe tự hành và giám sát thông minh. Trong thập kỷ qua, các mô hình CNN một giai đoạn như YOLO (You Only Look Once) đã trở thành chuẩn mực công nghiệp nhờ sự cân bằng tốt giữa tốc độ và độ chính xác. Tuy nhiên, sự phát triển của YOLO đang tiệm cận giới hạn do các ràng buộc về kiến trúc.

2. Vấn đề tồn tại Hạn chế lớn nhất của các dòng YOLO hiện nay (kể cả YOLOv8) là sự phụ thuộc vào thuật toán hậu xử lý Triệt tiêu phi cực đại (NMS). NMS gây ra hai vấn đề nghiêm trọng:

- Độ trễ không ổn định: Thời gian xử lý bị biến thiên tùy thuộc vào số lượng vật thể trong ảnh, gây khó khăn cho các ứng dụng yêu cầu thời gian thực khắt khe.
- Rào cản tối ưu hóa: NMS đòi hỏi các siêu tham số thủ công, ngăn cản mô hình đạt được khả năng học đầu cuối (end-to-end) thực sự.

3. Giải pháp đề xuất: RT-DETR Nhằm giải quyết nút thắt trên, mô hình RT-DETR (Real-Time DEtection TRansformer) được giới thiệu tại CVPR 2024 đã đề xuất một hướng tiếp cận đột phá: kết hợp ưu điểm của CNN và Transformer để loại bỏ hoàn toàn NMS.

Nghiên cứu này tập trung tái hiện và đánh giá RT-DETR dựa trên hai cải tiến cốt lõi:

- Efficient Hybrid Encoder: Tách biệt tương tác nội bộ và đa tầng để giảm chi phí tính toán.
- Uncertainty-minimal Query Selection: Tối ưu hóa việc khởi tạo truy vấn để nâng cao độ chính xác.

Mục tiêu cuối cùng là kiểm chứng định lượng xem liệu RT-DETR có thực sự thay thế được YOLOv8 trong các tác vụ thời gian thực hay không.

MỤC TIÊU

1. **Phân tích cơ sở lý thuyết:** Nghiên cứu sâu kiến trúc RT-DETR để làm rõ hạn chế của thuật toán NMS trong các mô hình YOLO hiện tại, đồng thời phân tích hiệu quả của hai cải tiến cốt lõi là Bộ mã hóa lai (Efficient Hybrid Encoder) và Cơ chế chọn lọc truy vấn (Uncertainty-minimal Query Selection) trong việc tối ưu hóa tốc độ và độ chính xác.
2. **Triển khai thực nghiệm:** Thiết lập môi trường và tái hiện thành công mô hình RT-DETR (phiên bản R50/R101) cùng mô hình đối chứng YOLOv8 trên tập dữ liệu chuẩn MS COCO val2017, đảm bảo các điều kiện phần cứng đồng nhất để phục vụ cho việc đánh giá công bằng.
3. **Đánh giá và So sánh định lượng:** Thực hiện đo đạc và so sánh chi tiết hiệu năng giữa RT-DETR và YOLOv8 dựa trên các chỉ số then chốt: Độ chính xác trung bình (AP), Tốc độ khung hình (FPS) và Độ trễ (Latency). Mục tiêu là kiểm chứng khả năng của RT-DETR trong việc loại bỏ độ trễ biến thiên của NMS và thiết lập chuẩn mực mới cho bài toán thời gian thực.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Phương pháp tiếp cận

Đề tài áp dụng phương pháp nghiên cứu định lượng kết hợp thực nghiệm (Empirical & Quantitative Research). Quy trình nghiên cứu được thiết kế để kiểm chứng giả thuyết rằng kiến trúc Transformer lai (Hybrid Transformer) có thể thay thế hoàn toàn các kiến trúc CNN truyền thống (như YOLO) trong bài toán phát hiện đối tượng thời gian thực mà không cần phụ thuộc vào thuật toán hậu xử lý NMS.

2. Nội dung nghiên cứu chi tiết

Nghiên cứu được chia thành ba giai đoạn chính, tập trung vào việc phân tích kiến trúc, tối ưu hóa cơ chế khởi tạo và đánh giá thực nghiệm.

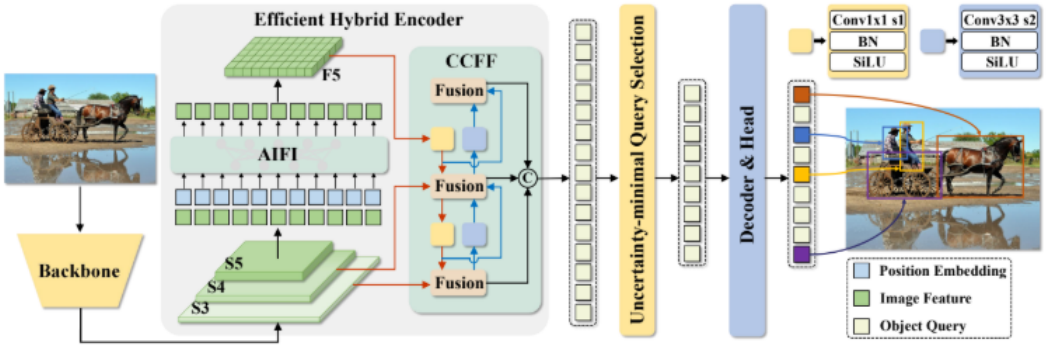
2.1. Phân tích và Triển khai Bộ mã hóa lai hiệu quả (Efficient Hybrid Encoder)

Trong các mô hình DETR truyền thống, việc xử lý các đặc trưng đa quy mô (multi-scale features) thường dẫn đến chi phí tính toán khổng lồ, tạo thành nút thắt cổ chai ngăn cản tốc độ thời gian thực.
Để giải quyết vấn đề này, nội dung đầu tiên của nghiên cứu sẽ tập trung phân tích và triển khai kiến trúc **Bộ mã hóa lai (Hybrid Encoder)** của RT-DETR.

Cụ thể, nghiên cứu sẽ đi sâu vào cơ chế tách biệt quá trình xử lý đặc trưng:

- **Tương tác nội bộ dựa trên Attention (AIFI)**: Chúng tôi sẽ phân tích hiệu quả của việc chỉ áp dụng cơ chế Self-attention lên tầng đặc trưng cao nhất S5 của Backbone. Đây là tầng chứa thông tin ngữ nghĩa phong phú nhất, giúp mô hình nắm bắt được mối quan hệ giữa các đối tượng mà không lãng phí tài nguyên tính toán vào các tầng thấp.
- **Hợp nhất đặc trưng chéo dựa trên CNN (CCFF):** Đối với các tầng đặc trưng

thấp hơn (S3, S4), nghiên cứu sẽ đánh giá việc thay thế Attention bằng các khối hợp nhất CNN (Fusion Blocks) để tăng tốc độ xử lý mà vẫn đảm bảo khả năng tổng hợp thông tin đa chiều .



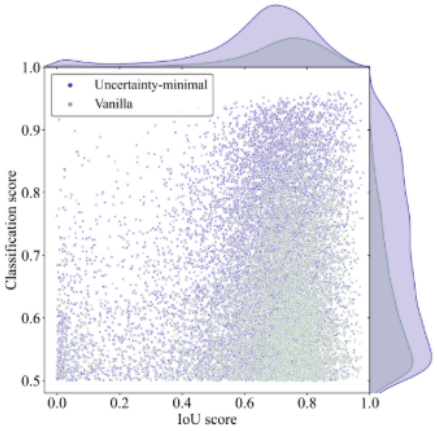
Hình 1. *Sơ đồ kiến trúc tổng quan của RT-DETR. Mô hình tiếp nhận đặc trưng từ 3 tầng cuối của Backbone, xử lý qua Bộ mã hóa lai (gồm module AIFI và CCFF) trước khi đưa vào Bộ giải mã Transformer để dự đoán trực tiếp*

2.2. Tối ưu hóa Cơ chế Chọn lọc Truy vấn (Uncertainty-minimal Query Selection)

Một hạn chế lớn của các mô hình DETR trước đây là việc lựa chọn truy vấn (Object Queries) chỉ dựa trên điểm số phân loại (Classification Score). Cách tiếp cận này thường dẫn đến việc chọn nhầm các đặc trưng có độ tin cậy phân loại cao nhưng độ chính xác định vị thấp, gây nhiễu cho Bộ giải mã.

Trong giai đoạn này, nghiên cứu sẽ tập trung đánh giá cải tiến cốt lõi thứ hai của RT-DETR: **Cơ chế chọn lọc giảm thiểu độ không chắc chắn.**

- Nghiên cứu sẽ tái hiện công thức tính độ không chắc chắn U dựa trên sự sai lệch giữa phân bố dự đoán vị trí và phân loại.
- Mục tiêu là chứng minh rằng việc đưa tham số này vào hàm mục tiêu (Loss function) sẽ ép mô hình phải chọn ra các đặc trưng "chất lượng kép" (tốt cả về vị trí lẫn nhận dạng) làm truy vấn khởi tạo.



Hình 2. *Biểu đồ phân tán so sánh chất lượng truy vấn. Các điểm màu tím (phương pháp đề xuất) tập trung ở vùng góc trên bên phải, minh chứng cho việc các đặc trưng được chọn có sự đồng nhất cao giữa điểm phân loại và điểm định vị (IoU), vượt trội hơn so với phương pháp cũ (điểm màu xanh)*

2.3. Thiết lập Thực nghiệm và Đánh giá Đối sánh (Benchmark)

Giai đoạn cuối cùng là thực hiện đánh giá định lượng để kiểm chứng hiệu năng thực tế.

- **Dữ liệu và Môi trường:** Thực nghiệm được tiến hành trên tập dữ liệu chuẩn MS COCO val2017. Môi trường phần cứng sử dụng GPU NVIDIA T4 và thư viện TensorRT FP16 để mô phỏng điều kiện triển khai thực tế.
- **Kịch bản so sánh:** Nghiên cứu sẽ thiết lập một bảng so sánh đối chứng (Benchmark) công bằng giữa các biến thể của RT-DETR (R50, R101) và các mô hình YOLO tiên tiến nhất hiện nay (YOLOv8-L, YOLOv8-X)11.
- **Chỉ số đánh giá:** Hai chỉ số trọng yếu sẽ được đo đạc là:
    1. **Độ chính xác trung bình (AP):** Để đánh giá khả năng phát hiện đúng vật thể.
    2. **Tốc độ suy luận đầu cuối (End-to-end Latency/FPS):** Đo lường tổng thời gian từ khi nhận ảnh đến khi ra kết quả cuối cùng. Đặc biệt, đối với YOLO, thời gian này sẽ bao gồm cả quá trình NMS để đảm bảo tính công bằng.

KẾT QUẢ MONG ĐỢI

Dựa trên cơ sở lý thuyết và phương pháp nghiên cứu đã đề xuất, đề tài đặt mục tiêu đạt được các kết quả cụ thể sau đây về cả mặt định lượng và định tính:

1. Chứng minh sự vượt trội về Hiệu năng (Quantitative Results) Kết quả quan trọng nhất của nghiên cứu là xác thực được giả thuyết: RT-DETR có khả năng thay thế các mô hình YOLO trong các tác vụ thời gian thực. Cụ thể, thông qua thực nghiệm đối sánh trên tập COCO val2017, chúng tôi kỳ vọng đạt được các chỉ số hiệu năng tương đương hoặc vượt trội so với công bố gốc của bài báo:

- **Về độ chính xác:** Mô hình RT-DETR-R50 dự kiến đạt độ chính xác trung bình (AP) khoảng **53.1%**, vượt qua đối thủ trực tiếp là YOLOv8-L (52.9%).
- **Về tốc độ:** Quan trọng hơn, nhờ loại bỏ hoàn toàn độ trễ của NMS, RT-DETR-R50 kỳ vọng đạt tốc độ suy luận khoảng **108 FPS** trên GPU T4, nhanh hơn đáng kể so với mức 71 FPS của YOLOv8-L. Điều này khẳng định tính ưu việt của kiến trúc Transformer lai trong việc xử lý luồng dữ liệu tốc độ cao.

Model	Backbone	#Epochs	#Params (M)	GFLOPs	$FPS_{b=1}$	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{S}	AP^{val}_{M}	AP^{val}_{L}
Real-time Object Detectors											
YOLOv5-L [11]	-	300	46	109	54	49.0	67.3	-	-	-	-
YOLOv5-X [11]	-	300	86	205	43	50.7	68.9	-	-	-	-
PPYOLOE-L [40]	-	300	52	110	94	51.4	68.9	55.6	31.4	55.3	66.1
PPYOLOE-X [40]	-	300	98	206	60	52.3	69.9	56.5	33.3	56.3	66.4
YOLOv6-L [16]	-	300	59	150	99	52.8	70.3	57.7	34.4	58.1	70.1
YOLOv7-L [38]	-	300	36	104	55	51.2	69.7	55.5	35.2	55.9	66.7
YOLOv7-X [38]	-	300	71	189	45	52.9	71.1	57.4	36.9	57.7	68.6
YOLOv8-L [12]	-	-	43	165	71	52.9	69.8	57.5	35.3	58.3	69.8
YOLOv8-X [12]	-	-	68	257	50	53.9	71.0	58.7	35.7	59.3	70.7
End-to-end Object Detectors											
DETR-DC5 [4]	R50	500	41	187	-	43.3	63.1	45.9	22.5	47.3	61.1
DETR-DC5 [4]	R101	500	60	253	-	44.9	64.7	47.7	23.7	49.5	62.3
Anchor-DETR-DC5 [39]	R50	50	39	172	-	44.2	64.7	47.5	24.7	48.2	60.6
Anchor-DETR-DC5 [39]	R101	50	-	-	-	45.1	65.7	48.8	25.8	49.4	61.6
Conditional-DETR-DC5 [27]	R50	108	44	195	-	45.1	65.4	48.5	25.3	49.0	62.2
Conditional-DETR-DC5 [27]	R101	108	63	262	-	45.9	66.8	49.5	27.2	50.3	63.3
Efficient-DETR [42]	R50	36	35	210	-	45.1	63.1	49.1	28.3	48.4	59.0
Efficient-DETR [42]	R101	36	54	289	-	45.7	64.1	49.5	28.2	49.1	60.2
SMCA-DETR [9]	R50	108	40	152	-	45.6	65.5	49.1	25.9	49.3	62.6
SMCA-DETR [9]	R101	108	58	218	-	46.3	66.6	50.2	27.2	50.5	63.2
Deformable-DETR [45]	R50	50	40	173	-	46.2	65.2	50.0	28.8	49.2	61.7
DAB-Deformable-DETR [23]	R50	50	48	195	-	46.9	66.0	50.8	30.1	50.4	62.5
DAB-Deformable-DETR++ [23]	R50	50	47	-	-	48.7	67.2	53.0	31.4	51.6	63.9
DN-Deformable-DETR [17]	R50	50	48	195	-	48.6	67.4	52.7	31.0	52.0	63.7
DN-Deformable-DETR++ [17]	R50	50	47	-	-	49.5	67.6	53.8	31.3	52.6	65.4
DINO-Deformable-DETR [44]	R50	36	47	279	5	50.9	69.0	55.3	34.6	54.1	64.6
Real-time End-to-end Object Detector (ours)											
RT-DETR	R50	72	42	136	**108**	**53.1**	**71.3**	**57.7**	34.8	58.0	70.0
RT-DETR	R101	72	76	259	**74**	**54.3**	**72.7**	58.6	36.0	58.8	**72.1**

Hình 3. *Bảng kết quả so sánh kỳ vọng giữa RT-DETR và các mô hình SOTA (YOLOv5, v6, v7, v8). Các chỉ số mục tiêu của nghiên cứu (dòng RT-DETR) cho thấy sự cải thiện đồng thời cả về độ chính xác (AP) và tốc độ (FPS) so với các mô hình cùng quy mô.*

2. Khả năng hoạt động bền vững trong môi trường phức tạp (Qualitative Results) Bên cạnh các con số, nghiên cứu dự kiến sẽ minh họa trực quan khả năng của RT-DETR trong các tình huống khó mà các mô hình CNN truyền thống thường gặp trở ngại (như vật thể bị che khuất, chuyển động nhòe hoặc mật độ vật thể dày đặc). Nhờ cơ chế Attention toàn cục và chọn lọc truy vấn thông minh, mô hình kỳ vọng sẽ giảm thiểu tỷ lệ bỏ sót (False Negatives) và nhận diện sai (False Positives) trong các điều kiện này.

Hình 4. *Minh họa kết quả phát hiện đối tượng dự kiến của RT-DETR trong các điều kiện thách thức như mật độ cao và bị che khuất, chứng minh khả năng tổng quát hóa tốt của mô hình.*

3. Sản phẩm đóng gói và Tài liệu khoa học

- **Mã nguồn tái hiện (Reproducible Code):** Một kho mã nguồn hoàn chỉnh trên GitHub, bao gồm các script huấn luyện, đánh giá và notebook demo, cho phép cộng đồng nghiên cứu kiểm chứng lại kết quả.
- **Báo cáo phân tích chuyên sâu:** Tài liệu tổng kết không chỉ dừng lại ở việc báo cáo số liệu mà còn phân tích sâu nguyên nhân gốc rễ (root-cause analysis) của sự chênh lệch hiệu năng giữa cơ chế Anchor-free của YOLO và cơ chế Set-prediction của RT-DETR.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, Yi Liu: **DETRs Beat YOLOs on Real-time Object Detection.** CVPR 2024: 16965-16974

[2] Glenn Jocher, Ayush Chaurasia, Jing Qiu: **Ultralytics YOLO.** CoRR abs/2304.00577 (2023)

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko: **End-to-End Object Detection with Transformers.** ECCV (1) 2020: 213-229

[4] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai: **Deformable DETR: Deformable Transformers for End-to-End Object Detection.** ICLR 2021