Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.240

HIỆU QUẢ CÁC NHÓM MÔ HÌNH HỌC SÂU TRONG BÀI TOÁN PHÁT HIỆN PHƯƠNG TIỆN GIAO THÔNG

Vũ Lê Quỳnh Phương[1*], Trần Nguyễn Minh Thư[2] và Phạm Nguyên Khang[2]
[1]*Trường Cao đẳng Sư Phạm Kiên Giang*
[2]*Trường Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
**Tác giả liên hệ (Corresponding author): vlqphuong@cdspkg.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 17/10/2023
Sửa bài (Revised): 01/11/2023
Duyệt đăng (Accepted): 03/11/2023

Title: *Efficiency of Deep Learning Model groups in vehicle detection*

Author(s): *Vu Le Quynh Phuong[1*], Tran Nguyen Minh Thu[2] and Pham Nguyen Khang[2]*

Affiliation(s): *[1]Kien Giang Teachers Training College, [2]Can Tho University*

TÓM TẮT

Các mô hình phát hiện đối tượng dựa trên mạng nơ-ron tích chập đang phát triển liên tục và được áp dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong hệ thống giao thông thông minh. Trong nghiên cứu này, các kỹ thuật học sâu đã được áp dụng, đặc biệt là các mô hình phát hiện phương tiện giao thông trong thời gian thực: dựa trên "anchor" (điển hình như mô hình You Only Look Once - YOLO), dựa trên "keypoint"(điển hình như mô hình CenterNet), và dựa trên "transformer"(điển hình như mô hình Detection Transformers - DETR). Các mô hình đã được tinh chỉnh và huấn luyện thông qua kỹ thuật học chuyển tiếp để cải thiện khả năng phát hiện phương tiện giao thông. Kết quả của các thử nghiệm đã chỉ ra rằng mô hình YOLO đạt được độ chính xác cao nhất (98,3%) với thời gian thực thi là 11,7 ms. Trong khi đó, mô hình DETR thực hiện thời gian thực thi nhanh nhất (2,3 ms), nhưng độ chính xác thấp nhất (62,4%). Mô hình CenterNet là lựa chọn tốt nhất (94,11% - 8 ms) vì cân đối được giữa độ chính xác và thời gian thực thi, có thể được sử dụng trong các ứng dụng thời gian thực.

Từ khoá: *Học sâu, phát hiện đối tượng, phát hiện phương tiện giao thông, "transformer"*

ABSTRACT

Object detection models based on convolutional neural networks are continuously evolving and widely applied in various domains, especially in intelligent transportation systems. In this study, the authors applied deep learning techniques, particularly real-time vehicle detection models: based on anchor-box (for example, You Only Look Once - YOLO), based on keypoint (for example CenterNet), and based on Transformers (for example, Detection Transformers - DETR) for detection vehicles. These models were fine-tuned and trained using transfer learning techniques to enhance vehicle detection capabilities. The results of the experiments indicated that the YOLO model achieved the highest accuracy (98.3%) with 11.7 ms time of detection. Meanwhile, the DETR model had the fastest execution time (2.3 ms) but the lowest accuracy (62.4%). The CenterNet model proved to be the best choice (94.11% - 8 ms) as it struck a balance between accuracy and execution time, making it suitable for real-time applications.

Keywords: *Deep learning, object detection, transformer, vehicle detection*

1. GIỚI THIỆU

Trong thời gian gần đây, lĩnh vực thị giác máy tính đã chứng kiến những tiến bộ đáng kể, đặc biệt là trong việc áp dụng các mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN) để phát hiện các đối tượng. Phát hiện đối tượng là quá trình kết hợp giữa việc phân loại và xác định vị trí của các đối tượng trong ảnh. Các thuật toán phát hiện đối tượng đã được áp dụng rộng rãi trong nhiều ứng dụng, đặc biệt là trong lĩnh vực hệ thống giao thông thông minh. Một trong những bài toán quan trọng của hệ thống giao thông thông minh là phát hiện và phân loại các phương tiện giao thông một cách chính xác và hiệu quả trong thời gian thực (Loce et al., 2023). Điều này đòi hỏi các giải thuật phát hiện đối tượng hoạt động ổn định và nhất quán trong các điều kiện ánh sáng và thời tiết khác nhau. Có nhiều nghiên cứu trên thế giới về việc phát hiện phương tiện giao thông dựa trên các mô hình CNN (Bautista et al., 2016; Hsu et al., 2018; Nguyen, 2019). Tập dữ liệu được gán nhãn trên các nghiên cứu trên thường là là xe ô tô, xe bus, xe tải,… Tuy nhiên, tại Việt Nam, xe máy là phương tiện phổ biến và thường là đối tượng gây ra tình trạng ùn tắc vào giờ cao điểm. Khi ùn tắc, các phương tiện xe máy chiếm đa số thì các hệ thống nhận diện khó áp dụng được vì sự chồng lắp và che khuất lẫn nhau.

Việc cải thiện phát hiện và đếm xe máy trong hình ảnh giao thông đang được đẩy mạnh thông qua sự phát triển của các mô hình sâu, đặc biệt là CNN. Mô hình CNN được áp dụng để phát hiện xe máy trong các góc ảnh đặt từ trên cao (1.000 ảnh) trong nghiên cứu của Huynh et al. (2016). Phương pháp đã được kiểm tra trên 200 hình ảnh thử nghiệm đạt được kết quả F1-score là 81%. Tuy nhiên, thời gian thực thi của mô hình là quá lớn (2 phút cho mỗi ảnh), không phù hợp để sử dụng phương pháp này trong các hệ thống có yêu cầu thời gian thực.

Mặc dù mô hình CNN có thể phát hiện đối tượng và đạt hiệu suất tốt nhưng tốc độ thực thi cao (Bautista et al., 2016). Mô hình Faster R-CNN (Ren et al., 2015) là một trong những mô hình đã cải thiện được vấn đề trên. Vì vậy, Phuong et al. (2021) áp dụng mô hình Faster-RCNN để phát hiện và đếm số lượng phương tiện giao thông. Hệ thống phát hiện các phương tiện, sau đó đếm số lượng xe hai bánh, bốn bánh và xe ưu tiên để xác định tình trạng giao thông. Độ chính xác của phát hiện đối tượng được ghi nhận 92% với mô hình Faster R-CNN và 92,53% với mô hình YOLOv4 trên tập dữ liệu thử nghiệm. Tuy nhiên, thời gian để phát hiện các đối tượng (xe hai bánh, bốn bánh và xe ưu tiên) vẫn cao (từ 1,5 giây đến 2,8 giây cho mỗi khung hình), khó triển khai trong việc phát hiện thời gian thực. Khi xe đông đúc, mô hình không thể nhận dạng chính xác (không thể phát hiện tất cả các xe trong khu vực). Trong lúc tắc đường, rất khó để nhận ra các phương tiện vì hình ảnh từ video là phương tiện chồng lên nhau. Vì vậy, việc đếm phương tiện trên địa bàn khi xảy ra tắc đường là không khả thi.

Để giải quyết vấn đề trên, mô hình CenterNet đã được đề xuất trong nghiên cứu của Zhou et al. (2019) nhằm phát hiện vật thể. CenterNet là một mạng nơ-ron tích chập với thiết kế đơn giản, nhưng có sự cân bằng tốt giữa tốc độ và độ chính xác. Tiếp cận mới của CenterNet là chuyển bài toán phát hiện đối tượng thành bài toán ước tính các điểm đặc trưng (keypoint estimation), từ đó dẫn đến việc tính toán kích thước và vị trí của hộp bao (bounding box) một cách hiệu quả cho bài toán phát hiện đối tượng. Nghiên cứu của Phuong et al. (2022) kết hợp mô hình CenterNet vào bài toán ước tính mật độ giao thông từ các camera giao thông để dự đoán tình trạng giao thông. Mô hình đầu tiên sử dụng mạng CNN (VGG-16, Inception-V3, ResNet-50) để phân loại tình trạng giao thông từ hình ảnh giao thông, xác định xem ảnh giao thông có đông đúc hay không. Nếu hình ảnh cho thấy giao thông đông đúc, mô hình trừ nền được áp dụng để đánh giá tốc độ của các phương tiện giao thông: chậm, bình thường, hoặc đang kẹt xe. Trong trường hợp không ùn tắc giao thông, mô hình CenterNet được sử dụng để phát hiện và đếm số lượng phương tiện giao thông. Các loại phương tiện được phân biệt gồm: xe hai bánh (bao gồm xe đạp và xe máy), xe bốn bánh (bao gồm xe tải, xe hơi và xe buýt) và xe ưu tiên. Hệ thống sẽ cung cấp thông tin về tình trạng giao thông cho người dùng, bao gồm tình trạng đường đông đúc (đi bình thường, đi chậm, hoặc kẹt xe) hoặc tình trạng đường không đông đúc (trung bình, thấp, hoặc thưa thớt). Hệ thống đạt được kết quả mAP là 93,13% với thời gian thực thi là 0,146 s cho mỗi ảnh.

Các mô hình phát hiện đối tượng dựa trên mạng CNN đã đạt được nhiều kết quả tích cực. Tuy nhiên, sự xuất hiện gần đây của các mô hình nhận dạng đối tượng dựa trên kiến trúc "transformer" đã thu hút sự quan tâm và tạo ra những đột phá đáng kể trong lĩnh vực thị giác máy tính. Kiến trúc "transformer", ban đầu được giới thiệu cho nhiệm vụ dịch máy và xử lý ngôn ngữ tự nhiên, nay đã được áp dụng thành công vào lĩnh vực thị giác máy tính. Tuy nhiên, vẫn chưa có nhiều nghiên cứu để phát hiện phương tiện giao thông với dữ liệu được thu thập trong thời gian thực dựa trên kiến trúc "transformer".

Để hiểu rõ hơn về hiệu quả của từng kiến trúc, các mô hình nhận dạng phương tiện giao thông dựa trên "anchor", "key-point", và "transformer" đã được áp dụng. Mục tiêu là xác định mô hình phát hiện phương tiện giao thông hoạt động tốt trong các bài toán thực tế.

2. NGHIÊN CỨU LIÊN QUAN

Phát hiện đối tượng là bài toán cơ bản nhưng đóng vai trò dùng để phân loại và xác định vị trí các đối tượng vật thể có trong ảnh hoặc video. Nhờ sự phát triển nhanh chóng về mặt dữ liệu cũng như sự xuất hiện thêm nhiều giải thuật mới, bài toán phát hiện đối tượng đã đạt được nhiều bước tiến đáng kể và được ứng dụng rất nhiều trong thực tế. Đặc biệt, sự xuất hiện của các bộ phát hiện đối tượng một giai đoạn đã thay đổi cách tiếp cận vấn đề phát hiện đối tượng. Các mô hình trước đó chia thành từng giai đoạn riêng biệt để tạo ra các đề xuất đối tượng, thì trong các mô hình một giai đoạn xem xét toàn bộ hình ảnh và thực hiện cả phân loại và hồi quy trực tiếp trong một mạng nơ-ron tổng thể. Các giải thuật này không phải tạo ra các đề xuất cố định trước, mà thay vào đó xem xét tất cả các vị trí tiềm năng trên ảnh và dự đoán xem chúng có chứa đối tượng hay không. Có hai loại giải thuật phát hiện một giai đoạn phổ biến là:

– Phát hiện dựa trên "anchor": Các giải thuật này sử dụng các hộp "anchor" có kích thước và tỷ lệ trước để tạo ra đề xuất. Sau đó, dự đoán xem mỗi hộp anchor có chứa đối tượng nào và cập nhật lại vị trí và kích thước của hộp dựa trên các dự đoán.

– Phát hiện dựa trên "key-point": Các giải thuật sử dụng các điểm "key-point" để xác định vị trí của các đối tượng. Thay vì sử dụng hộp giới hạn trước, giải thuật tập trung vào việc xác định các điểm quan trọng trên đối tượng và sau đó xây dựng các hộp từ các điểm này.

Bộ phát hiện đối tượng dựa trên kiến trúc mạng nơ-ron tích chập "anchor" và "key-point" đã đạt được sự chính xác đáng kể. Các bộ phát hiện hiện đại thường thực hiện hồi quy và phân loại trên một loạt đề xuất lớn. Do đó, hiệu suất của các giải thuật bị ảnh hưởng bởi các nhiệm vụ xử lý phức tạp như thuật toán chặn không cực đại (Non-Maximum Suppersion - NMS). Vì vậy, ứng dụng thị giác máy tính dựa vào "transformers" đã được giới thiệu như là một mô hình kiến trúc thay thế cho CNN.

2.1. Phát hiện đối tượng dựa trên "anchor"

Các bộ phát hiện dựa trên hộp "anchor" đã được xác định trước, với các tỷ lệ khung hình khác nhau, nhằm phục vụ cho việc phát hiện các đối tượng có hình dạng và kích thước đa dạng. Mô hình nổi bật về phương pháp này là YOLO (Redmon et al., 2016). YOLO hoạt động bằng cách biểu diễn hình ảnh đầu vào như một lưới các ô, trong đó mỗi ô có trách nhiệm dự đoán một hộp giới hạn nếu trung tâm của hộp nằm trong ô đó. Mỗi ô lưới dự đoán nhiều hộp giới hạn và đầu ra vị trí và nhãn lớp cùng với điểm tự tin (confidence score). YOLO được đánh giá tốt về tốc độ và đơn giản nhưng có nhược điểm là tỷ lệ thu hẹp tương đối thấp (Redmon & Farhadi, 2017). Để khắc phục những hạn chế này, YOLOv7 đã đưa ra nhiều cải tiến quan trọng so với các phiên bản trước đó, mang lại khả năng phát hiện đối tượng trong ảnh một cách hiệu quả hơn (Wang et al., 2023). Một trong những cải tiến quan trọng đó là việc sử dụng các hộp "anchor".

YOLOv7 sử dụng 9 hộp "anchor", cho phép thuật toán nắm bắt phạm vi hình dạng và kích thước của các đối tượng một cách rộng hơn so với các phiên bản trước đây (Wang et al., 2023). Điều này giúp giảm thiểu số lượng dự đoán sai. Một cải tiến quan trọng khác trong YOLOv7 là việc áp dụng hàm mất mát (loss function) mới gọi là "focal loss". Các phiên bản trước của YOLO đã sử dụng hàm mất mát "cross-entropy", nhưng nó không hiệu quả đối với việc phát hiện các đối tượng nhỏ. Vì vậy, mô hình YOLOv7 đã giải quyết vấn đề này bằng cách tập trung vào việc giảm trọng số cho các mẫu được phân loại đúng và thay vào đó, tập trung vào việc xử lý các mẫu khó (đối tượng khó phát hiện). YOLOv7 cũng cải thiện độ phân giải của hình ảnh đầu vào. Thay vì sử dụng độ phân giải 416 x 416 pixel như trong phiên bản YOLOv3 (Redmon & Farhadi, 2018), YOLOv7 xử lý hình ảnh ở độ phân giải cao hơn là 680 x 680 pixel. Điều này giúp YOLOv7 phát hiện và phân tích các đối tượng nhỏ hơn một cách hiệu quả hơn và nâng cao tỷ lệ chính xác trong quá trình phát hiện.

2.2. Phát hiện đối tượng dựa trên "key-point"

Các bộ phát hiện dựa trên "anchor" có nhược điểm là phải khắc phục các vấn đề với siêu tham số như số lượng, tỷ lệ cạnh và kích thước của các "anchor" và còn rất phụ thuộc vào tập dữ liệu. Điều này đã dẫn đến sự ra đời của một phương pháp hoàn toàn mới của bộ phát hiện không "anchor" (còn gọi là bộ phát hiện dựa trên "key-point"). Các phương pháp dựa trên "key-point" xem các đối tượng như các điểm thay vì các hộp giới hạn. Các điểm "key-point", như góc hoặc trung tâm của các đối tượng được ước tính, chiều rộng và chiều cao được hồi quy từ các điểm này thay vì các "anchor" được quy định

trước. Đã có nhiều mạng dựa trên điểm gốc được giới thiệu, gồm có CornerNet, CenterNet, FCOS, NanoDet và TTFNet (Law & Deng, 2018; Zhou et al., 2019; Tian et al., 2019; Liu et al., 2020). Trong số các phương pháp trên, mô hình CenterNet được lựa chọn và sử dụng, vì nó không chỉ đạt được độ chính xác cao hơn so với CornerNet, mà còn đơn giản hóa việc ước tính "keypoint".

CenterNet là một hệ thống phát hiện đối tượng một giai đoạn dựa trên "heatmap". Nguyên tắc của phương pháp này là dự đoán vị trí của trung tâm và kích thước của các đối tượng trong ảnh. Hình ảnh RGB đầu vào có chiều rộng w và chiều cao h, ký hiệu là $I \in R^{w \times h \times 3}$, mạng sẽ đầu ra một "heatmap" đã được giảm mẫu $\hat{Y} \in [0,1]^{\frac{w}{R} \times \frac{h}{R} \times C}$, trong đó R là bước đầu ra (stride) và C là số lớp. Ký hiệu $W = \frac{w}{R}$ và H= $\frac{h}{R}$ là kích thước không gian đầu ra. Dự đoán $\hat{Y}_{x,y,c} = 1$ tương ứng với trung tâm của một đối tượng thuộc lớp c tại (x, y), trong khi $\hat{Y}_{x,y,c} = 0$ tương ứng với nền. Để khắc phục trường hợp bị lỗi khi chuyển đổi giữa "ground truth heatmap" và ảnh đầu vào, đầu ra, bộ ước lượng độ lệch (offset predictor) được sử dụng $\hat{O} \in R^{W \times H \times 2}$ và một phép hồi quy $\hat{S} \in R^{W \times H \times 2}$ cho kích thước đối tượng.

2.3. Phát hiện đối tượng dựa trên "transformer"

Bộ phát hiện dựa trên "transformers" là một mô hình thiết kế mới trong thị giác máy tính, dựa vào cơ chế tự chú ý (self-attention) và đã được giới thiệu lần đầu trong bài toán nhận dạng đối tượng bởi thuật toán DETR (Carion et al., 2020). DETR là một thuật toán sử dụng kiến trúc mạng "transformer" để thực hiện phát hiện đối tượng trong ảnh. Điểm mới của DETR là nó coi việc phát hiện đối tượng như một bài toán dự đoán trực tiếp tập hợp các đối tượng có trong hình ảnh, thay vì tiếp cận theo các giai đoạn truyền thống. Thuật toán này bao gồm ba phần chính: trích xuất đặc trưng từ mạng CNN, bộ mã hóa-giải "transformer", và một mạng lan truyền xuôi (Feed Forward Network) cho việc dự đoán. Đầu tiên, DETR sử dụng một mạng CNN để trích xuất đặc trưng từ hình ảnh, tạo ra một bản đồ đặc trưng với độ phân giải thấp hơn nhưng chứa thông tin quan trọng về các đối tượng trong hình ảnh. Sau đó, bộ mã hóa-giải transformer thực hiện biến đổi tự chú ý trên bản đồ đặc trưng này và tạo ra các truy vấn đối tượng. Các truy vấn này sau đó được sử dụng để dự đoán lớp và vị trí của các đối tượng trong hình ảnh. DETR đã đạt được hiệu suất cao trong việc phát hiện đối tượng trong nhiều ứng dụng khác nhau và đánh dấu một bước tiến quan trọng trong lĩnh vực phát hiện đối tượng bằng cách kết hợp mạng transformer và thị giác máy tính để thực hiện nhiệm vụ này một cách hiệu quả và chính xác.

2.4. Các phương pháp đánh giá

Phương pháp phát hiện đối tượng phương tiện giao thông sử dụng độ đo mAP (Mean Average Precision). Độ đo mAP để đánh giá độ chính xác của việc nhận dạng đối tượng, là trung bình các giá trị AP (Average Precision) theo từng phân lớp đối tượng mà mô hình có khả năng nhận dạng được.

Các giá trị độ tin cậy (precision) và độ nhạy (recall) được tính toán bằng cách sử dụng các công thức đã được đề xuất trong nghiên cứu của Rocchio (Rocchio, 1971). Để tạo đường cong (PR curve) độ tin cậy theo độ nhạy cho mỗi lớp riêng biệt, lần lượt vẽ các điểm dữ liệu trên biểu đồ với tọa độ (P_n, R_n) (n = 1, 2,.., N), sau đó nối lại với nhau. Diện tích nằm dưới đường cong PR curve chính là giá trị AP. Một giá trị AP lớn đồng nghĩa với việc mô hình có chất lượng phát hiện tốt, khi độ tin cậy và độ nhạy đều cao. Độ chính xác trung bình (AP) được xác định theo công thức (1).

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \quad (1)$$

AP là độ chính xác trung bình cho lần lượt từng lớp nên mAP sẽ là trung bình cộng AP của tất cả các lớp theo công thức (2), với n là số lượng lớp và AP_k là giá trị AP ở lớp k.

$$mAP = \frac{1}{n} \sum_{k=0}^{n} AP_k \quad (2)$$

3. KẾT QUẢ VÀ THẢO LUẬN

Trong bài báo này, các mô hình được xử lý trên một máy tính đơn sử dụng nền tảng Google Colab với CPU Intel Xeon Processor with two cores 2.30 GHz đi kèm với 13GB ram cho phép huấn luyện với batch size lớn hơn, GPU có hiệu suất cao như GPU Nvidia Tesla T4 15102MiB giúp rút ngắn đáng kể thời gian tính toán. Các mô hình được huấn luyện với kích thước lô (batch size) được thiết lập là 8, thuật toán tối ưu ADAM với động lượng (momentum) 0,9 và tốc độ học ban đầu là 0,001. Quá trình đào tạo được thực hiện trong 25 vòng lặp. Tỷ lệ học tập sẽ giảm 10 lần nếu sau 10 vòng lặp mà độ chính xác trên tập dữ liệu xác thực không được cải thiện.

Dữ liệu được cung cấp bởi Công an phường Vĩnh Thanh Vân – Thành phố Rạch Giá – tỉnh Kiên

Giang. Dữ liệu gồm có 03 góc camera (CAM1, CAM3, CAM5). Vị trí đặt camera đường Lê Lợi (đối diện bệnh viện tỉnh Kiên Giang). Các video clip dữ liệu đều có thời lượng gần 1 tiếng. Độ phân giải ở góc CAM1 và CAM5 là 1920x1080, FPS lần lượt là: 15 khung hình/giây và 12 khung hình/giây. Riêng độ phân giải ở góc CAM3 là 1280x720, FPS: 10 khung hình/giây. Cảnh quay ở video clip góc CAM1 là vào ban ngày, trời nắng, thời gian từ 10 giờ 13 phút sáng ngày 01-11-2020 đến 11 giờ 13 phút sáng cùng ngày. Cảnh quay ở video clip góc CAM3 cũng vào ban ngày, trời nắng, thời gian từ 10 giờ tới 11 giờ sáng ngày 09-09-2019. Video clip góc CAM5 có thời gian từ 16 giờ 23 phút tới 17 giờ 23 phút chiều ngày 12-11-2020, cảnh quay ban đầu trời nắng nhưng tới khoảng 16 giờ 50 phút thì trời bắt đầu mưa trong vòng khoảng 15 phút.

Tập dữ liệu 1 bao gồm 2000 ảnh, được gán nhãn: "0" là xe ưu tiên, "1" là xe máy, xe đạp, "2" ứng với các loại xe ôtô 4 chỗ, 7 chỗ, xe bus,… Số lượng các mẫu đối tượng trên 3 nhãn (0, 1, 2) lần lượt là (275, 3910, 23810). Xét trên từng video gốc, tập dữ liệu đã gán nhãn được chia thành 3 tập con là "train", "test", "valid", tất cả các ảnh sẽ được sắp xếp theo thứ tự dòng thời gian như trong video gốc, 80% ảnh ở đầu video sẽ được sử dụng cho tập "train" và 20% đoạn sau được chia đều làm 2 phần cho tập "test" và "valid". Thực hiện như vậy trên từng video gốc sau đó tổng hợp lại ta thu được 3 tập dữ liệu "train", "valid", "test" tổng hợp.

Sau khi huấn luyện trên tập dữ liệu 1 bằng cả 3 mô hình: YOLO, CenterNet và DETR, thì độ chính xác của mô hình DETR cho kết quả không đạt mong đợi. Vì vậy, để kiểm tra giả định rằng việc tăng thêm dữ liệu có thể nâng cao độ chính xác của mô hình hay không, nhóm nghiên cứu đã bổ sung thêm hình ảnh có gán nhãn vào tập dữ liệu 1 và tiến hành huấn luyện lại.

Tập dữ liệu 2 bao gồm 2.946 ảnh (gồm 2.000 ảnh từ dữ liệu 1 và 946 ảnh được thêm vào từ các video thu thập), được gán nhãn và chia thành tập dữ liệu train, valid, test như tập dữ liệu 1. Số lượng các mẫu đối tượng trên 3 nhãn (0, 1, 2) lần lượt là (335, 4.295, 23.964).

3.1. Kết quả thực nghiệm

3.1.1. Mô hình YOLOv7

Huấn luyện tập dữ liệu 1 với mô hình YOLOv7 để thực hiện việc phát hiện phương tiện giao thông được đã được gán nhãn. Quá trình huấn luyện trong khoảng thời gian ngắn, chỉ mất 0,474 giờ (khoảng 28,4 phút) để hoàn thành quá trình học với mAP 98, 2%.

Mô hình YOLOv7 có tổng cộng 168 tầng, là các tầng của mạng nơ-ron sâu (Deep Neural Network) sử dụng trong thuật toán phát hiện đối tượng phương tiện giao thông. Số lượng tầng này có ảnh hưởng đến độ phức tạp của mô hình và khả năng phân tích đối tượng trong ảnh. Mô hình có tổng cộng 11.126.745 tham số. Tham số là các trọng số và thông số trong mô hình mà mạng nơ-ron sâu sử dụng để học và dự đoán. Số lượng tham số càng lớn thì mô hình càng phức tạp và có khả năng học tốt hơn từ dữ liệu, cũng đồng nghĩa với việc tốn thời gian và tài nguyên tính toán để huấn luyện và triển khai mô hình.

Bảng 1. Bảng kết quả đánh giá trên tập dữ liệu với mô hình YOLOv7

Lớp	Tập dữ liệu 1 mAP 50	Tập dữ liệu 2 mAP 50
Tất cả	98,2%	98,3%
Xe ưu tiên	99,4%	99,3%
Xe 2 bánh	97,2%	97,2%
Xe 4 bánh	98%	98,3%

Mô hình có khả năng phân tích mỗi khung hình trong một ảnh trong khoảng thời gian 11,7 ms. Điều này cho thấy mô hình đủ nhanh để sử dụng trong các ứng dụng thời gian thực như phát hiện đối tượng trong video trực tiếp. Một số ảnh minh hoạ về phát hiện phương tiện giao thông với mô hình YOLOv7 được hiển thị trong Hình 1. Mô hình YOLOv7 có thể phát hiện tốt 3 loại nhãn trong nhiều trường hợp.

Hình 1. Kết quả phát hiện phương tiện với mô hình YOLOv7

Tiếp tục huấn luyện mô hình YOLOv7 với tập dữ liệu 2, kết quả thu được là 98,3% sau 0,778 giờ huấn luyện. Bảng 1 so sánh mAP giữa 2 tập dữ liệu được huấn luyện. Kết quả thực nghiệm cho thấy rằng, khi tăng số lượng dữ liệu đối với mô hình YOLOv7, độ chính xác không đạt được sự cải thiện đáng kể (98,3%), nhưng thời gian huấn luyện lại tăng gấp đôi. Điều quan trọng là mô hình YOLOv7 vẫn đạt được độ chính xác cao trên cả 2 bộ dữ liệu, cho thấy rằng YOLOv7 vẫn là một giải pháp hiệu quả cho bộ dữ liệu nhỏ và có thể được triển khai trong các tình huống yêu cầu thời gian huấn luyện ngắn.

3.1.2. Mô hình CenterNet

Khi tiến hành huấn luyện tập dữ liệu 1 với mô hình CenterNet, mAP đạt 93,77%. Điều này cho thấy tính hiệu quả của mô hình trong việc phát hiện và xác định vị trí của các đối tượng trong ảnh. Ngoài ra, tốc độ dự đoán của mô hình khi xử lý một khung hình khoảng 8 ms, nhanh hơn so với mô hình YOLOv7. Điều này có nghĩa rằng CenterNet có khả năng xử lý hình ảnh một cách nhanh chóng và hiệu quả. Thời gian huấn luyện của mô hình CenterNet chỉ mất 1,04 giờ. Điều này cho thấy tính hiệu quả của thuật toán huấn luyện và khả năng học của mô hình.

Bảng 2. Bảng kết quả đánh giá trên tập dữ liệu với mô hình CenterNet

Lớp	Tập dữ liệu 1 mAP 50	Tập dữ liệu 2 mAP 50
Tất cả	93,7%	94,11%
Xe ưu tiên	94,1%	95,08%
Xe 2 bánh	93,8%	94,02%
Xe 4 bánh	93,3%	93,23%

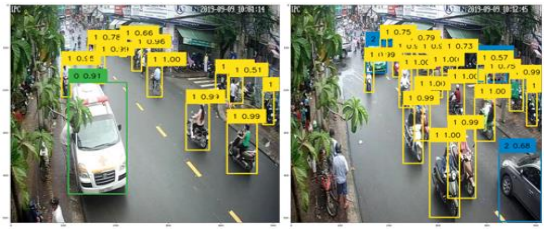
Tiếp tục huấn luyện mô hình trên tập dữ liệu thứ hai, kết quả thực nghiệm cho thấy có sự tăng nhẹ mAP lên 94,11%, nhưng thời gian huấn luyện đã tăng lên đáng kể, gần 2,5 giờ. Bảng 2 so sánh mAP trên 2 tập dữ liệu khi tiến hành huấn luyện với mô hình CenterNet. Vì sự phức tạp của tập dữ liệu thứ hai không khác biệt đáng kể so với tập dữ liệu ban đầu nên độ chính xác của mô hình đã đạt đến ngưỡng, khó tăng thêm nhiều. Tuy nhiên, thời gian huấn luyện tăng lên do mô hình phải xử lý một lượng dữ liệu lớn hơn, dẫn đến sự gia tăng trong quá trình học và cập nhật trọng số mô hình. Mặc dù không có sự cải thiện lớn về độ chính xác, mô hình CenterNet vẫn có thể được sử dụng hiệu quả trong việc phát hiện đối tượng trên bộ dữ liệu nhỏ, và thời gian huấn luyện có thể được chấp nhận nếu không có yêu cầu thời gian huấn luyện ngắn.

3.1.3. Mô hình DETR

Kết quả thực nghiệm cho thấy, mô hình DETR có thời gian huấn luyện mất 1,68 giờ. Bên cạnh đó, mô hình huấn luyện với 41,5 triệu tham số nên mô hình DETR được xem xét là một mô hình lớn và phức tạp. Số lượng tham số lớn này có thể làm gia tăng khả năng mô hình trong việc học và hiểu dữ liệu, nhưng cũng đặt ra thách thức về cách triển khai trên các nền tảng có tài nguyên hạn chế. Ngoài ra, mAP khá thấp (50,9%) thể hiện hiệu suất phân loại đối tượng của mô hình còn chưa cao, có thể do mô hình chưa hoàn thiện hoặc cần được điều chỉnh thêm để cải thiện khả năng phát hiện chính xác. Tuy nhiên, một điểm mạnh của mô hình DETR là thời gian nhận dạng nhanh, trung bình chỉ là 2,3 ms cho mỗi ảnh. Điều này cho thấy khả năng của nó trong việc xử lý ảnh nhanh chóng và phù hợp cho các ứng dụng đòi hỏi đáp ứng thời gian thực. Một số ảnh minh hoạ về phát hiện phương tiện giao thông với mô hình DETR được hiển thị trong Hình 2.

Bảng 3. Bảng kết quả đánh giá trên tập dữ liệu với mô hình DETR

Lớp	Tập dữ liệu 1 mAP 50	Tập dữ liệu 2 mAP 50
Tất cả	50,9%	62,4%
Xe ưu tiên	20,6%	22,7%
Xe 2 bánh	77,2%	87,8%
Xe 4 bánh	54,9%	76,7%



Hình 2. Kết quả phát hiện phương tiện với mô hình DETR

Khi huấn luyện trên tập dữ liệu 2, từ kết quả Bảng 3 cho thấy, độ chính xác của mô hình đã đạt 62,4%, tăng thêm hơn 11% so với khi chỉ sử dụng tập dữ liệu 1. Thời gian huấn luyện cũng tăng lên đến 2,69 giờ. Kết quả này do đã bổ sung gần 1.000 hình ảnh huấn luyện vào tập dữ liệu. Điều này chứng minh là mô hình DETR thích hợp cho các bài toán có dữ liệu lớn. Mô hình DETR sử dụng cơ chế tự chú ý để cải thiện khả năng hiểu và tương tác với các đối tượng trong hình ảnh. Khi bổ sung thêm dữ liệu, mô hình có thể học hỏi từ nhiều ví dụ khác nhau, từ đó nâng cao khả năng tổng quát hóa và độ chính xác của nó. Thời gian huấn luyện tăng do sự gia tăng trong số lượng dữ liệu và độ phức tạp của mô hình. Tuy nhiên, sự cải thiện độ chính xác đáng kể khi thêm dữ liệu là một kết quả quan trọng và cho thấy mô hình DETR phù hợp với dữ liệu lớn.

3.2. Đánh giá các mô hình

Bài báo cung cấp tổng thể về kết quả phát hiện phương tiện giao thông bằng các giải thuật học sâu dựa trên ba chỉ số: độ chính xác, tốc độ của suy luận và thời gian huấn luyện trên tập dữ liệu tự thu thập.

Trong số ba giải thuật phát hiện đối tượng được khảo sát, mô hình YOLO là giải thuật phát hiện dựa trên "anchor" có độ chính xác cao nhất (mAP 98,2%). Mô hình phát hiện dựa vào "key-point"

CenterNet đạt được độ chính xác cao thứ hai (mAP 93,7%). Giải thuật phát hiện này là dự đoán điểm "key-point" và không cần phải định nghĩa kích thước "anchor" dựa trên kích thước đối tượng. Các mô-đun tự chú ý được thêm vào trong mô hình DETR (mAP 50,9%) dẫn đến độ chính xác cải thiện, nhưng do mạng gốc vẫn là CNN, nên hiệu suất chưa cao do số lượng ảnh huấn luyện vẫn còn ít. Khi tăng số lượng dữ liệu huấn luyện, mAP của mô hình YOLO và CenterNet tăng không đáng kể. Trong khi đó, mô hình DETR có giá trị mAP tăng hơn 10% (62,4%). Kết quả này khẳng định mô hình DETR cần cung cấp nhiều dữ liệu học để đảm bảo độ chính xác cao. Trong nghiên cứu của Lin Matthieu và cộng sự, mô hình DETR được sử dụng để ước lượng mật độ và đếm đám đông (Lin et al., 2021). Kết quả thực nghiệm trên tập dữ liệu CrowdHunam với mô hình DETR có mAP thấp hơn (66,12%) so với mAP của mô hình Faster-RCNN (85%). Về nguyên lý, kiến trúc mạng DERT không có tính chất không biến dạng khi dịch chuyển như mạng CNN, nên mô hình huấn luyện cần được cung cấp nhiều dữ liệu để đáp ứng quá trình học các đặc trưng hình ảnh (Arkin et al., 2021).

Bảng 4. Bảng so sánh thời gian huấn luyện của mô hình YOLO, CenterNet và DETR trên 2 tập dữ liệu

Thời gian huấn luyện	YOLO	CenterNet	DETR
Tập dữ liệu 1	0,474 giờ	1,04 giờ	1,68 giờ
Tập dữ liệu 2	0,778 giờ	2,5 giờ	2,69 giờ

Tốc độ suy luận trong các giải cho thấy một xu hướng khác so với độ chính xác. Trong số ba giải thuật được áp dụng, mô hình DETR có tốc độ suy luận cao nhất, vì kiến trúc transformer không được tối ưu hóa phần cứng bằng CNN. Mô hình CenterNet là nhanh thứ hai vì không có hạn chế của NMS (CenterNet sử dụng "heatmap" dựa trên "peak" cho việc lọc bằng lớp "maxpooling" thay vì NMS dựa trên IoU), giúp tăng tốc độ suy luận. YOLO đạt được tốc độ thấp nhất trong 3 giải thuật.

Khi xem xét thời gian huấn luyện giữa các mô hình được minh học trong Bảng 4, mô hình YOLO nổi trội với tốc độ huấn luyện nhanh nhất trên cả hai tập dữ liệu. Điều này có thể cho thấy rằng kiến trúc của YOLO được tối ưu hóa cho tốc độ và có khả năng xử lý dữ liệu một cách hiệu quả hơn so với hai mô hình CenterNet và DETR. Trong khi đó, DETR chậm nhất trong việc huấn luyện ở cả hai tập dữ liệu. Điều này là do mô hình DETR sử dụng kiến trúc phức tạp hơn và cần thêm thời gian để hội tụ, thích hợp cho những tình huống yêu cầu độ chính xác cao hơn. CenterNet có thể được coi là sự cân bằng giữa tốc độ và độ chính xác, giữa YOLO và DETR. Một điểm quan trọng khác là thời gian huấn luyện tăng lên khi kích thước tập dữ liệu tăng vì khi có nhiều dữ liệu hơn, mô hình cần thêm thời gian để xử lý và học từ số lượng lớn thông tin đó.

4. KẾT LUẬN

Nghiên cứu này được tiến hành đã so sánh và đánh giá hiệu suất của các mô hình phát hiện đối tượng tiêu biểu, bao gồm các kiến trúc dựa trên "anchor" (YOLO), dựa trên "key-point" (CenterNet) và dựa trên "transformer" (DETR), trong việc phân loại các phương tiện giao thông thành ba nhóm: xe ưu tiên, xe 4 bánh và 2 xe bánh. Kết quả cho thấy rằng mô hình YOLOv7 đã đạt được độ chính xác cao nhất, lên đến 98,2%, vượt trội so với các mô hình khác trong việc phát hiện các loại phương tiện giao thông. Trong 3 mô hình huấn luyện, mô hình DETR có thời gian thực thi nhanh nhất (2,3 ms) nhưng độ chính xác thấp nhất (50,9%). Tuy nhiên, trong thực nghiệm cho thấy, độ chính xác của mô hình DETR có thể cải thiện nếu tăng dữ liệu huấn luyện. Trong khi đó, mô hình CenterNet đạt được sự cân bằng giữa độ chính xác và thời gian thực thi (93,7% - 8 ms). Sự so sánh này giúp rút ra các kết luận chính xác và đáng tin cậy về hiệu suất của từng mô hình trong bài toán giao thông thông minh.

Dựa trên những kết quả tích cực này, mô hình DETR được đề xuất mở rộng nghiên cứu để tăng cường độ chính xác và tính ổn định của trong việc phát hiện phương tiện giao thông. Điều này có thể được thực hiện thông qua việc gán nhãn thêm dữ liệu, tăng cường dữ liệu, thử nghiệm các siêu tham số khác nhau và điều chỉnh kiến trúc mô hình. Các vấn đề và kết quả trong nghiên cứu này đóng góp vào việc tạo ra nguồn cảm hứng và cung cấp thông tin hữu ích cho các nghiên cứu liên quan. Trong tương lai, nhiều nghiên cứu khác có thể được thực hiện để tiếp tục cải thiện hiệu suất của các mô hình trên các bộ dữ liệu về video giao thông.

TÀI LIỆU THAM KHẢO

Arkin, E., Yadikar, N., Muhtar, Y., & Ubul, K. (2021). A Survey of Object Detection Based on CNN and Transformer, *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)* (pp. 99-108). IEEE. https://doi.org/10.1109/PRML52754.2021.9520732

Bautista, C. M., Dy, C. A., Mañalac, M. I., Orbe, R. A., & Cordel, M. (2016). Convolutional neural

network for vehicle detection in low-resolution traffic videos. *In 2016 IEEE Region 10 Symposium (TENSYMP)* (pp. 277-281). IEEE. https://doi.org/10.1109/TENCONSpring.2016.7519418

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *In European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_13

Hsu, S. C., Huang, C. L., & Chuang, C. H. (2018). Vehicle detection using simplified fast R-CNN. *In 2018 International Workshop on Advanced Image Technology (IWAIT)* (pp. 1-3). IEEE. https://doi.org/10.48550/arXiv.2012.06785

Huynh, C. K., Le, T. S., & Hamamoto, K. (2016). Convolutional neural network for motorbike detection in dense traffic, *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)* (pp. 369-37). IEEE. https://doi.org/10.1109/CCE.2016.7562664

Law, H., & Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. *In Proceedings of the European conference on computer vision (ECCV)* (pp. 734-750). https://doi.org/10.1007/978-3-030-01264-9_45

Li, D., & Zhai, J. (2022). A real-time vehicle window positioning system based on nanodet. *In Chinese Intelligent Systems Conference* (pp. 697-705). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-6203-5_69

Lin, M., Li, C., Bu, X., Sun, M., Lin, C., Yan, J., Ouyang, W., & Deng, Z. (2021). Detr for crowd pedestrian detection. arXiv preprint arXiv:2012.06785.

Liu, Z., Zheng, T., Xu, G., Yang, Z., Liu, H., & Cai, D. (2020, April). Training-time-friendly network for real-time object detection. *In proceedings of the AAAI conference on artificial intelligence*, *34*(7), 11685-11692. https://doi.org/10.1609/aaai.v34i07.6838

Loce, R. P., Bernal, E. A., Wu, W., & Bala, R. (2013). Computer vision in roadway transportation systems: a survey. *Journal of Electronic Imaging*, *22*(4), 041121-041121. https://doi.org/10.1117/1.JEI.22.4.041121

Nguyen, H. (2019). Improving faster R-CNN framework for fast vehicle detection. *Mathematical Problems in Engineering 2019* (pp. 1-11). https://doi.org/10.1155/2019/3808064

Phuong, V. L. Q., Tai, B. N., Huy, N. K., Thu, T. N. M., & Khang, P. N. (2021). Estimating the traffic density from traffic cameras. In *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings 8*, 248-263. Springer Singapore. https://doi.org/10.1007/978-981-16-8062-5_17

Phuong, V. L. Q., Dong, N. V., Thu, T. N. M., & Khang, P. N. (2022). Combine Clasification Algorithm and Centernet Model to Predict Trafic Density. In *International Conference on Future Data and Security Engineering* (pp. 588-600). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8069-5_40

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271). IEEE

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint: arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137-1149. IEEE. https://doi.org/10.1109/TPAMI.2016.2577031

Rocchio, J. J. (1971). Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636). https://doi.org/10.1109/ICCV.2019.00972

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7464-7475). https://doi.org/10.1109/CVPR52729.2023.00721

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as Points. arXiv preprint: arXiv:1904.07850. https://doi.org/10.48550/arXiv.1904.07850