KHẢO SÁT MỘT SỐ MÔ HÌNH PHÁT HIỆN ĐỐI TƯỢNG DỰA VÀO HỌC SÂU

Nguyễn Dũng[1]*, Hoàng Văn Dũng[2], Lê Văn Tường Lân[1]

[1]Khoa Công nghệ thông tin, trường Đại học Khoa học, Đại học Huế

[2]Trường Đại học Sư phạm Kỹ thuật TP. HCM

*Email: nguyendung@hueuni.edu.vn

Ngày nhận bài: 18/01/2024; ngày hoàn thành phản biện: 19/02/2024; ngày duyệt đăng: 5/3/2024

TÓM TẮT

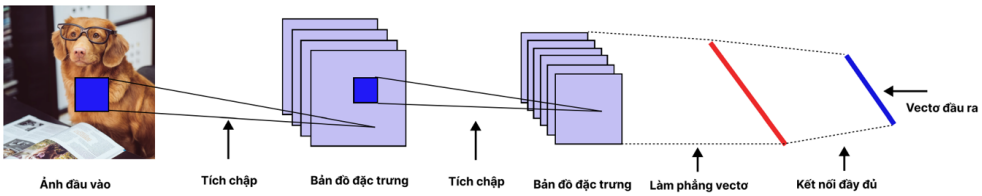
Phát hiện đối tượng là tác vụ phân loại và định vị các đối tượng trong hình ảnh hoặc video. Nó đã trở nên nổi bật trong những năm gần đây nhờ những ứng dụng rộng rãi của nó trong các bài toán trong thực tế. Bài viết này khảo sát những mô hình gần đây về tác vụ phát hiện đối tượng đặc biệt là các mô hình dựa trên học sâu (thời gian khảo sát từ những năm 2000 đến 2023). Đầu tiên bài báo giới thiệu tổng quan một cách ngắn gọn về tác vụ phát hiện đối tượng trong lĩnh vực thị giác máy tính, một số bộ dữ liệu chuẩn và độ đo phổ biến được sử dụng trong tác vụ này. Tiếp theo bài báo giới thiệu một số mô hình hay kiến trúc xương sống được sử dụng và cuối cùng, cũng là điểm chính của bài báo này là giới thiệu các phương pháp nổi trội sử dụng học sâu để phát hiện đối tượng.

Từ khóa: Thị giác máy tính, phát hiện đối tượng, phân loại, mạng tích chập, học sâu.

1. GIỚI THIỆU

Trong những năm gần đây, ngành thị giác máy tính đã chứng kiến sự tăng trưởng theo cấp số nhân với sự phát triển nhanh chóng của các công cụ, mô hình và kỹ thuật mới, trong đó có tác vụ phát hiện đối tượng. Phát hiện đối tượng là một trong những khía cạnh cơ bản và thách thức nhất của ứng dụng hiểu hình ảnh và thị giác máy tính. Nó đòi hỏi phải xác định loại đối tượng và vị trí của các đối tượng đó trong khung nhìn/ảnh/video. Những tiến bộ đáng kể trong việc phát hiện đối tượng đã đạt được thông qua việc cải thiện biểu diễn đối tượng và sử dụng các mô hình học sâu. Các nghiên cứu đột phá trong lĩnh vực này chia thành 2 giai đoạn, trước và sau năm 2014. Trước năm 2014, ngành này được đánh dấu bởi các thuật toán, mô hình truyền thống chủ yếu dựa vào các phương pháp trích chọn đặc trưng một cách thủ công và sau năm 2014 là

các thuật toán, mô hình dựa vào học sâu. Chính nhờ sự phát triển vượt bậc của mạng nơ-ron sâu (DNN-Deep Neural Network), đặc biệt là mạng tích chập sâu (Convolutional Neural Network) [1], như mô tả trong Hình 1, đã làm nên một cuộc cách mạng lớn trong ngành thị giác máy tính, đặc biệt là trong các tác vụ như phân loại, phân đoạn và phát hiện đối tượng. Các mô hình phát hiện đối tượng dựa vào học sâu, bằng cách sử dụng mạng nơ-ron tích chập và mạng transformer [2], hiện đang đóng một vai trò then chốt trong sự phát triển của lĩnh vực này. Ví dụ hệ thống hỗ trợ xe tự hành [3-6], phát hiện hành vi bất thường [7, 8], phát hiện khuôn mặt [9, 10], phân tích hành vi con người [11, 12] và hình ảnh y tế như phát hiện ung thư [12, 13],…

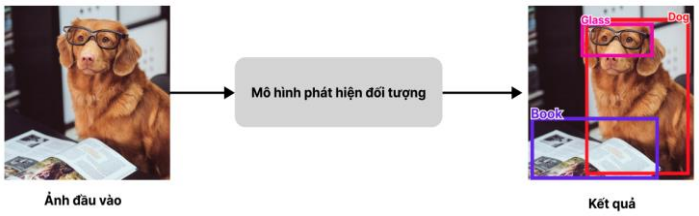


Hình 1. Mô hình mạng tích chập

Bài báo này khảo sát các phương pháp phát hiện đối tượng đã phát triển như thế nào trong kỷ nguyên học sâu những năm qua. Chúng tôi trình bày tổng quan tài liệu về các thuật toán phát hiện đối tượng tiên tiến khác nhau và các khái niệm cơ bản đằng sau các phương pháp này. Những đóng góp chính của bài viết này như sau: (1) Bài báo cung cấp một số vấn đề liên quan đến các mô hình, chẳng hạn như bộ dữ liệu chuẩn, độ đo thường được sử dụng trong tác vụ này. (2) Bài báo này cung cấp một cách tổng quan, khái quát về phương pháp sử dụng, ưu và nhược điểm của các mô hình xương sống của tác vụ và đặc biệt là mô hình phát hiện đối tượng.

2. PHÁT HIỆN ĐỐI TƯỢNG VÀ THÁCH THỨC CỦA NÓ

Phát hiện đối tượng là phần mở rộng tự nhiên của phân loại đối tượng, là tác vụ chỉ nhằm mục đích nhận dạng đối tượng trong ảnh. Mục tiêu của việc phát hiện đối tượng là phát hiện tất cả đối tượng được xác định trước (theo bộ dữ liệu huấn luyện) và vị trí của các đối tượng này theo hình hộp hay còn gọi là hộp giới hạn (bounding-box, bao gồm toạ độ tâm của hình và chiều rộng, chiều cao của hình hộp). Hình 2 dưới đây mô tả quy trình chung của tác vụ phát hiện đối tượng. Các mô hình phát hiện đối tượng được huấn luyện trên bộ dữ liệu đã được gắn nhãn và được đánh giá dựa trên độ đo khác nhau.



Hình 2. Tác vụ phát hiện đối tượng

Mặc dù lĩnh vực này đã có nhiều bước phát triển vượt bậc, tuy nhiên nó vẫn gặp một số thách thức mà lĩnh vực phải đối mặt trong các ứng dụng thực tế là: (1) Nhiều biến thể của các đối tượng trong cùng một lớp: Biến thể này có thể do nhiều lý do khác nhau như tầm nhìn bị che khuất, ảnh hưởng của ánh sáng, tư thế, góc nhìn, bị xoay, thu nhỏ hoặc mờ,... (2) Số lượng lớp đối tượng lớn: Việc này dẫn đến việc tài nguyên dành cho huấn luyện và suy luận trở nên cao hơn các tác vụ khác. (3) Hiệu quả: Các mô hình phát hiện đối tượng cần nguồn lực tính toán cao để tạo ra kết quả phát hiện chính xác. Vì vậy việc áp dụng các mô hình này lên các thiết bị di động, thiết bị có tài nguyên tính toán thấp sẽ trở nên khó khăn để cân bằng giữa độ chính xác và hiệu năng tính toán.

3. BỘ DỮ LIỆU VÀ PHƯƠNG PHÁP ĐÁNH GIÁ ĐỘ CHÍNH XÁC

3.1. BỘ DỮ LIỆU

Phần này trình bày tổng quan về các bộ dữ liệu chuẩn và được sử dụng phổ biến nhất cho tác vụ phát hiện đối tượng. Đồng thời cuối phần này, chúng tôi cũng trình bày bảng so sánh giữa các bộ dữ liệu thường dùng trong Bảng 2.

3.1.1. Bộ dữ liệu PASCAL VOC 2012

Thử thách Pascal Visual Object Class (VOC) [14] là một thử thách kéo dài trong nhiều năm nhằm đẩy nhanh sự phát triển trong lĩnh vực thị giác máy tính. Mục tiêu chính của thử thách này là nhận dạng các đối tượng từ một số lớp đối tượng trong các cảnh thực tế. Về cơ bản, đây là một vấn đề học tập có giám sát trong đó một tập huấn luyện các hình ảnh được dán nhãn. Nó bắt đầu vào năm 2005 với các nhiệm vụ phân loại và phát hiện trên bốn lớp đối tượng. Trong khi thử thách VOC07 có hơn 5000 hình ảnh huấn luyện và hơn 12.000 đối tượng được gắn nhãn thì thử thách VOC12 đã tăng chúng lên hơn 11.000 hình ảnh huấn luyện và hơn 27.000 đối tượng được gắn nhãn. Các lớp đối tượng đã được mở rộng thành 20 danh mục và các nhiệm vụ như phân đoạn và phát hiện hành động cũng được đưa vào, bao gồm các lớp đối tượng: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor. Pascal VOC đã giới thiệu độ chính xác trung bình (mAP) ở mức 0,5 IoU (Intersection over Union) để đánh giá hiệu suất của các mô hình.

3.1.2. Bộ dữ liệu ILSVRC/ImageNet

Thử thách nhận dạng hình ảnh quy mô lớn ImageNet, The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [15], là thử thách thường niên diễn ra từ năm 2010 đến năm 2017 và trở thành chuẩn mực để đánh giá hiệu suất của các mô hình. Kích thước tập dữ liệu được tăng lên tới hơn một triệu hình ảnh bao gồm 1000 lớp phân loại đối tượng. 200 lớp trong số này đã được chọn lọc kỹ lưỡng cho tác vụ phát hiện đối tượng, được tạo thành từ hơn 500 nghìn hình ảnh. Nhiều nguồn khác nhau bao gồm ImageNet và Flick, đã được sử dụng để xây dựng tập dữ liệu phát hiện. ILSVRC cũng cập nhật số liệu đánh giá bằng cách nới lỏng ngưỡng IoU để giúp phát hiện đối tượng nhỏ hơn.

3.1.3. Bộ dữ liệu MS-COCO

Microsoft Common Objects in Context (MS-COCO) [16] là một trong những bộ dữ liệu thách thức nhất hiện có. Nó có 91 loại đồ vật phổ biến được tìm thấy trong bối cảnh tự nhiên. Nó được ra mắt vào năm 2015 và mức độ phổ biến của nó ngày càng tăng. Nó cũng giới thiệu một phương pháp nghiêm ngặt hơn để đo hiệu suất của các mô hình phát hiện đối tượng. Không giống như Pascal VOC và ILSVCR, nó tính toán IoU từ 0.5 đến 0.95 theo các bước 0.05, sau đó sử dụng kết hợp 10 giá trị này làm số liệu cuối cùng, được gọi là độ chính xác Trung bình (AP). Ngoài ra, nó còn sử dụng AP riêng biệt cho các đối tượng nhỏ, vừa và lớn để so sánh hiệu suất ở các quy mô khác nhau.

3.1.4. OpenImage

Tập dữ liệu hình ảnh mở [17] của Google bao gồm hàng triệu hình ảnh, được sử dụng cho nhiều tác vụ khác nhau trong thị giác máy tính . Nó được ra mắt vào năm 2017 và đã nhận được bảy bản cập nhật (phiên bản OpenImageV7, tính đến tháng 11/2023). Để phát hiện đối tượng, Open Images có 16 triệu hộp giới hạn cho 600 danh mục đối tượng trên 1.9 triệu hình ảnh, khiến nó trở thành tập dữ liệu lớn nhất về localization đối tượng. Những người tạo ra nó đã hết sức cẩn thận khi chọn những hình ảnh thú vị, phức tạp và đa dạng.

3.2. PHƯƠNG PHÁP ĐÁNH GIÁ ĐỘ CHÍNH XÁC

Các mô hình phát hiện đối tượng sử dụng nhiều tiêu chí để đo lường hiệu suất của trình phát hiện, chẳng hạn như precision và recall. Tuy nhiên, độ chính xác trung bình (mAP) là thước đo đánh giá phổ biến nhất. Độ chính xác này có nguồn gốc từ độ đọ IoU (Intersection over Union), là tỷ lệ giữa diện tích giao (Area of Intersection) và diện tích hợp (Area of Union) giữa hộp giới hạn ground-truth và hộp giới hạn dự đoán.

$$IoU = \frac{Area\ of\ Union}{Area\ of\ Intersection}$$

Một ngưỡng được đặt ra để xác định xem mô hình phát hiện đối tượng có chính xác hay không. Nếu IoU lớn hơn ngưỡng, nó được phân loại là True Positive trong khi IoU thấp hơn được phân loại là False Positive. Nếu mô hình không phát hiện được đối tượng có trong ground-truth thì nó được gọi là False Negative. Precison đo lường tỷ lệ phần trăm dự đoán đúng trong khi Recall đo lường các dự đoán chính xác liên quan đến ground-truth.

Từ các độ đo trên, ta có thể tính được độ chính xác trung bình (AP) cho từng lớp theo công thức như sau:

$$AP_i = \frac{1}{N} * \ Precision_i$$

Tuy nhiên để có đánh giá cuối cùng, để so sánh độ hiệu quả các mô hình thì người ta sử dụng giá trị trung bình của các độ chính xác trung bình của tất cả các lớp, được gọi là mAP.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}$$

Trong đó: (N) là số lớp cần phân loại, (AP_i) độ chính xác trung bình của lớp thứ i

4. CÁC MÔ HÌNH XƯƠNG SỐNG CỦA TÁC VỤ PHÁT HIỆN ĐỐI TƯỢNG

Mô hình xương sống hay còn gọi là mô hình backbone, là một trong những thành phần trọng nhất của tác vụ phát hiện đối tượng. Các mô hình này có nhiệm vụ là trích xuất đặc trưng từ hình ảnh đầu vào được mô hình sử dụng. Ở đây, bài báo thảo luận về một số mô hình backbone quan trọng được sử dụng trong các mô hình phát hiện đối tượng hiện đại dựa vào học sâu, bao gồm: Mạng AlextNet được sử dụng mô hình R-CNN, mạng VGG hay các biến thể như Darknet được sử dụng trong mô hình Fast R-CNN, Faster R-CNN, các phiên bản Yolo và SSD, mạng GoogLeNet trong mô hình YoloV1, mạng RestNext trong mô hình Mask R-CNN, mạng ResNet trong mô hình DETR,…

4.1. Mạng AlexNet

Krizhevsky Alex và cộng sự đã đề xuất AlexNet [18], một kiến trúc dựa trên mạng nơ-ron tích chập để phân loại hình ảnh và đã giành chiến thắng trong thử thách nhận dạng hình ảnh quy mô lớn của ImageNet (ILSVRC) năm 2012. Nó đạt được độ chính xác cao hơn đáng kể (hơn 26%) so với các mô hình SOTA (State of the art) tại thời điểm đó. AlexNet bao gồm tám lớp có thể học được (năm lớp tích chập và ba lớp được kết nối đầy đủ), lớp cuối cùng của lớp được kết nối đầy đủ được kết nối với bộ phân loại softmax với vectơ N chiều (N: số lớp cần phân loại. Ở đây tác giả dùng ImageNet nên N

= 1000); và cuối cùng sử dụng hàm mất mát Cross-Entropy để xác định sai khác giữa nhãn dự đoán và nhãn thực tế.

4.2. Mạng VGG

VGGNet (còn được gọi là VGG) [19] là một trong những mô hình nổi tiếng trong lĩnh vực thị giác máy tính và nhận dạng hình ảnh. Mô hình này được phát triển bởi nhóm nghiên cứu Visual Geometry Group (VGG) tại Đại học Oxford và đã đạt được thành công lớn trong cuộc thi ImageNet Large Scale Visual Recognition Challenge (ILSVRC) năm 2014. VGGNet có kiến trúc đơn giản và dễ hiểu, với các tầng tích chập có kích thước nhỏ (3x3) liên tiếp nhau. Các tầng này được xếp chồng lên nhau để tạo thành các khối, và các khối này được kết hợp để tạo ra kiến trúc của mô hình. Số tầng của mô hình từ 16-19 tầng tùy phiên bản. VGGNet có số lượng tham số lớn do sử dụng các tầng tích chập với kích thước nhỏ và số lượng tầng lớn, do đó yêu cầu tài nguyên tính toán lớn. Tuy nhiên, điều này cũng giúp mô hình học được các đặc trưng chi tiết và phức tạp từ dữ liệu hình ảnh. Mô hình này nhanh chóng trở thành một trong những mạng xương sống được sử dụng nhiều nhất cho các mô hình phát hiện và phân loại đối tượng.

4.3. Mạng ResNet

ResNet (Residual Network) [20] là một kiến trúc được thiết kế và xuất bản vào năm 2016. Nó được biết đến với khả năng huấn luyện các mạng sâu mà không gặp vấn đề gradient vanishing (biến mất đạo hàm), một vấn đề phổ biến trong các mạng rất sâu. Bài viết gốc trên ResNet đề xuất năm kích cỡ khác nhau của mô hình: 18, 34, 50, 101 và 152 lớp. Kể từ đó, nhiều biến thể khác của ResNet đã được phát triển, chẳng hạn như ResNeXt . Kiến trúc ResNet sử dụng một khối được gọi là "khối Residual", chứa nhiều lớp tích chập và chuẩn hóa [21]. Lớp cuối cùng được kết nối với lớp được kết nối đầy đủ để phân loại hình ảnh.

5. CÁC MÔ HÌNH PHÁT HIỆN ĐỐI TƯỢNG

Bài báo này chúng tôi chia các mô hình phát hiện đối tượng thành mô hình phát hiện 2 giai đoạn, mô hình phát hiện 1 giai đoạn. Mô hình hai giai đoạn là các mô hình cố gắng tìm số lượng đối tượng tùy ý được đề xuất trong một hình ảnh trong giai đoạn đầu tiên, sau đó phân loại và vị trí hóa chúng trong giai đoạn thứ hai. Vì các mô hình này có hai bước riêng biệt nên chúng thường mất nhiều thời gian hơn để tạo đề xuất vùng chứa đối tượng, có kiến trúc phức tạp. Mô hình phát hiện một giai đoạn phân loại và vị trí hóa các đối tượng trong một lần bằng cách sử dụng phương pháp lấy mẫu dày đặc.

5.1. Họ mô hình R-CNN

Mạng R-CNN (Region-Based Convolutional Neural Networks) là bài báo đầu tiên trong họ R-CNN và đã chứng minh cách CNN có thể được sử dụng để cải thiện

đáng kể hiệu suất phát hiện đối tượng. Girshick và cộng sự đã công bố bài báo [22] vào năm 2012. R-CNN sử dụng phương pháp tìm kiếm chọn lọc để đề xuất khu vực chứa đối tượng RoI (region of interest) với mạng CNN để chuyển đổi việc phát hiện đối tượng thành bài toán phân loại và localization. Đầu tiên hình ảnh đầu vào được chuyển qua mô-đun đề xuất vùng, mô-đun này tạo ra khoảng 2000 đối tượng vùng khác nhau. Mô-đun này tìm các phần của hình ảnh có xác suất tìm thấy đối tượng cao hơn bằng cách sử dụng thuật toán Selective Search [23]. Sau đó, các RoI này được biến đổi và truyền qua mạng CNN, mạng này trích xuất ra một vectơ đặc trưng 4096 chiều cho mỗi RoI. Girshick và cộng sự đã sử dụng AlexNet [22] làm kiến trúc xương sống của mô hình. Sau đó, các vectơ đặc trưng được chuyển đến máy vectơ hỗ trợ (SVM) dành riêng cho từng lớp đã được đào tạo để tiến hành phân loại, xem thử nó có phải là đối tượng của lớp tương ứng không.

Một trong những vấn đề chính với R-CNN là sử dụng mạng CNN riêng biệt cho từng RoI được đề xuất, do đó mô hình chạy rất chậm. Fast R-CNN [24] được Girshick và cộng sự công bố năm 2015, một năm sau khi mô hình R-CNN ra đời, đã giải quyết vấn đề này bằng cách tạo ra một hệ thống có thể huấn luyện từ đầu đến cuối chỉ trong một lần duy nhất. Đầu tiên mạng lấy hình ảnh đầu vào và cho ra các đề xuất đối tượng vùng RoI. Đồng thời hình ảnh đầu vài cũng được truyền qua một mạng CNN để tìm ra được bản đồ đặc trưng. Sau đó các đề xuất vùng RoI được ánh xạ tới bản đồ đặc trưng thu được bằng lớp gộp RoI pooling. Kết quả của lớp gộp này được kết nối với 2 tầng Full Connected và sau đó phân thành 2 nhánh. Một nhánh sử dụng lớp SoftMax để phân loại đối tượng và lớp hồi quy hộp giới hạn để tìm hộp giới hạn của đối tượng. Fast R-CNN được giới thiệu như một sự cải thiện về tốc độ (gấp 146 lần trên R-CNN). Nó đơn giản hóa quy trình huấn luyện và đạt được tốc độ cao hơn R-CNN.

Mặc dù Fast R-CNN tiến gần hơn đến khả năng phát hiện đối tượng theo thời gian thực, nhưng việc tạo đề xuất vùng RoI của nó vẫn chậm hơn rất nhiều. Faster R-CNN ra đời và đã đánh dấu sự tiến bộ quan trọng trong lĩnh vực thị giác máy tính và đã trở thành một trong những kiến trúc phổ biến cho nhận dạng đối tượng trong hình ảnh và video. Tiếp nối thành công của mô hình Faster R-CNN, mô hình Mask R-CNN ra đời với mục đích để phân đoạn ảnh, nghĩa là ngoài việc phân loại đối tượng, tìm bounding-box thì mô hình còn cho phép tìm ra mặt nạ của đối tượng, tức là đối tượng ở cấp độ pixel. Mask R-CNN [25] được giới thiệu bởi Kaiming He và cộng sự vào năm 2017. Mô hình này thực chất là sự mở rộng của Faster R-CNN, trong đó tại bước tìm được vecto RoI đặc trưng, mô hình mở thêm một nhánh để tìm được mặt nạ của đối tượng. Mask R-CNN là một trong những mô hình mạng nơ-ron sâu đa nhiệm phức tạp, có khả năng xác định và tạo ra mặt nạ cho đối tượng trong hình ảnh. Điều này làm cho nó rất mạnh mẽ trong các ứng dụng như phân loại và định vị đối tượng cũng như phát hiện và tạo ra mặt nạ cho chúng, chẳng hạn trong việc xử lý ảnh y tế, xử lý hình ảnh video và nhiều ứng dụng khác.

5.2. Họ mô hình YOLO

YOLO (You Only Look Once) [26] là một trong những mô hình phát hiện đối tượng đầu tiên sử dụng mạng nơ-ron sâu để thực hiện phát hiện đối tượng và xác định đối tượng trong hình ảnh và video. Phiên bản đầu tiên của YOLO được gọi là YOLOv1, được giới thiệu bởi Joseph Redmon và Santosh Divvala vào năm 2016 trong bài báo. YOLOv1 đã đánh dấu sự tiến bộ đáng kể trong việc phát hiện đối tượng trong thị giác máy tính. Đầu tiên YOLO chia hình ảnh đầu vào thành lưới S × S. Nếu tâm của một đối tượng rơi vào một ô lưới thì ô lưới đó có nhiệm vụ phát hiện đối tượng đó. Mỗi ô lưới dự đoán các hộp giới hạn B và điểm tin cậy cho các hộp đó. Các điểm tin cậy này phản ánh mức độ tin cậy của mô hình rằng hộp chứa một đối tượng và mức độ chính xác mà mô hình cho rằng hộp được dự đoán. YOLO dự đoán nhiều hộp giới hạn trên mỗi ô lưới. Tại thời điểm huấn luyện, ta chỉ muốn một bộ dự đoán hộp giới hạn thể hiện cho từng đối tượng. YOLO chỉ định bộ dự đoán dựa trên chỉ số IOU hiện tại cao nhất với ground-truth. Một kỹ thuật quan trọng được sử dụng trong các mô hình YOLO là NMS (non-maximum suppression). NMS là một bước hậu xử lý được sử dụng để cải thiện độ chính xác và hiệu quả của việc phát hiện đối tượng. Trong phát hiện đối tượng, thông thường có nhiều hộp giới hạn được tạo cho một đối tượng trong một hình ảnh. Các hộp giới hạn này có thể chồng lên nhau hoặc nằm ở các vị trí khác nhau, nhưng tất cả chúng đều đại diện cho cùng một đối tượng. NMS được sử dụng để xác định và loại bỏ các hộp giới hạn dư thừa hoặc không chính xác và đề xuất một hộp giới hạn duy nhất cho từng đối tượng trong ảnh. Tuy nhiên nhược điểm của Yolov1 đó là không thể phát hiện được các đối tượng nhỏ, chồng lấp và chỉ dự đoán được tối đa là S x S đối tượng trong hình. Các phiên bản cải tiến của YOLO như: YOLOv2, YOLOv3,…, YOLOv7, YOLOv8.

5.3. Họ mô hình DETR

Trong những năm gần đây, mô hình Transformer đã ảnh hưởng sâu sắc đến toàn bộ lĩnh vực học sâu, đặc biệt là lĩnh vực thị giác máy tính. Mô hình Transformer [2] loại bỏ toán tử tích chập truyền thống thay vào đó chỉ tính toán dựa trên cơ chế tự chú ý nhằm khắc phục các hạn chế của CNN. Vào năm 2020, N. Carion và cộng sự đã đề xuất DETR [27], trong đó đề xuất một mạng phát hiện end-to-end với Transformers. Mô hình này dùng mạng CNN để trích xuất đặc trưng của ảnh, ở đây tác giả dùng mạng ResNet, kết quả ta thu được một bản đồ đặc trưng của ảnh đầu vào. Bản đồ đặc trưng này được cộng với vector mã hoá vị trí, nhằm xác định thứ tự các đặc trưng. Kết quả của phép cộng này ta chuyển vào mạng Encoder của Transformer để mã hoá. Kết quả của mạng Encoder (sau khi đã thực hiện 6 lần) được đưa vào mạng Decoder để giải mã. Kết quả đầu ta của mạng Decoder ta thu được một feature map, ta dùng nó để tiến hành phân loại và xác định bounding-box thông qua các lớp Full connected, Softmax và hồi quy. Cho đến nay, việc phát hiện đối tượng đã bước vào một kỷ nguyên mới trong đó các đối tượng có thể được phát hiện mà không cần sử dụng anchor box hoặc anchor point. Sau

đó, X. Zhu và cộng sự đã đề xuất [28] để giải quyết thời gian hội tụ dài của DETR và hiệu suất hạn chế trong việc phát hiện các vật thể nhỏ. Nó đạt được hiệu suất cao trên tập dữ liệu MSCOCO (COCO mAP@0,5=71,9%).

5.4. SO SÁNH KẾT QUẢ CÁC MÔ HÌNH

Trong tiến trình phát triển của các mô hình, có nhiều tác giả đã tiến hành cải tiến, chỉnh sửa, bổ sung để mô hình đạt được hiệu suất và kết quả cao hơn. Tuy nhiên, trong Bảng 4 dưới đây, chúng tôi trực tiếp lấy kết quả từ bài báo gốc của chính tác giả của mô hình để tiến hành so sánh độ chính xác của các mô hình dựa vào tiêu chí AP với mức IoU là 0,5.

Bảng 4. Bảng so sánh các mô hình trên bộ dữ liệu chuẩn COCO và PascalVOC-2012

Mô hình	Năm	Bộ dữ liệu	Mạng xương sống	Kích thước ảnh đầu vào	AP[0,5]
R-CNN	2014	Pascal VOC 2012	**AlexNet**	224	58,5%
Fast R-CNN	2015	Pascal VOC 2012	**VGG-16**	Nhiều size	65,7%
Faster R-CNN	2016	Pascal VOC 2012	**VGG-16**	600	67,0%
Mask R-CNN	2018	MS COCO	**ResNeXt**	800	62,3%
Yolo	2015	Pascal VOC 2012	**GoogLeNet**	448	57,9%
YoloV2	2016	MS COCO	**DarkNet-19**	352	44,0%
YoloV3	2018	MS COCO	**DarkNet-53**	320	51,5%
YoloV7	2022	MS COCO		640, 1280	69,7%
SSD	2016	MS COCO	**VGG-16**	300	41,2%
DETR	2020	MS COCO	**ResNet-50**	Nhiều size	42,0%
Deformale DETR	2020	MS COCO	**ResNet-50**	Nhiều size	44,5%

Với kết quả, chúng tôi thấy rằng, với chỉ số AP@0,5 nhóm các mô hình hai giai đoạn đạt được độ chính xác cao hơn hẳn (cao nhất là mô hình Faster CNN với 67,0%) các mô hình một giai đoạn (cao nhất là Yolov7 với 69,70%). Nhóm mô hình sử dụng Transformer mà chúng tôi khảo sát ở đây hiện đang ở mức khiêm tốn hơn, trong đó Deformale DETR chỉ đạt 44,50%.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Mặc dù tác vụ phát hiện đối tượng đã đi được một chặng đường dài trong thập kỷ qua, nhưng các mô hình phát hiện tốt nhất vẫn chưa bão hòa về hiệu suất hay độ

chính xác. Khi các ứng dụng của nó tăng lên trong thế giới thực, nhu cầu về các mô hình nhẹ có thể được triển khai trên các hệ thống di động và nhúng sẽ tăng theo cấp số nhân. Đã có sự quan tâm ngày càng tăng đối với lĩnh vực này, nhưng nó vẫn là một thách thức mở. Trong bài báo này, chúng tôi đã chỉ ra các mô hình hai giai đoạn và một giai đoạn phát triển như thế nào so với các mô hình tiền nhiệm của chúng.

Mặc dù mô hình hai giai đoạn nhìn chung chính xác hơn nhưng chúng chậm và không thể sử dụng cho các ứng dụng thời gian thực như ô tô tự lái hoặc giám sát an ninh. Tuy nhiên, điều này đã thay đổi trong vài năm qua khi mô hình một giai đoạn có độ chính xác tương đương và nhanh hơn nhiều so với mô hình trước đây. Với xu hướng tích cực hiện nay về độ chính xác của các mô hình, chúng tôi đặt nhiều hy vọng vào các mô hình trong tương lai sẽ chính xác hơn và nhanh hơn.

Một số hướng phát triển trong tương lai: (1) Đánh đổi giữa tốc độ và độ chính xác: Việc tăng độ chính xác của thuật toán phát hiện đối tượng đòi hỏi nhiều tài nguyên tính toán hơn và thời gian xử lý lâu hơn. Giảm độ chính xác có thể dẫn đến thời gian xử lý nhanh hơn nhưng hiệu suất phát hiện thấp hơn. (2) Phát hiện đối tượng nhỏ: Phát hiện đối tượng nhỏ là một trường hợp cụ thể của phát hiện đối tượng tập trung vào việc phát hiện và định vị các đối tượng rất nhỏ trong hình ảnh hoặc video. Nó vẫn còn nhiều thách thức vì việc trích xuất thông tin từ các vật thể nhỏ chỉ có vài pixel là rất khó. (3) Phát hiện đối tượng 3D: Phát hiện đối tượng 3D liên quan đến việc ước tính vị trí, hướng và kích thước của đối tượng trong không gian ba chiều. Phát hiện đối tượng 3D có thể hữu ích trong các ứng dụng như robot và lái xe tự động, trong đó cần có kiến thức chính xác về môi trường 3D để điều hướng và tương tác với thế giới vật lý.

TÀI LIỆU THAM KHẢO

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," (in eng), *Nature,* vol. 521, no. 7553, pp. 436-44, May 28 2015, doi: 10.1038/nature14539.

[2] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722-2730.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907-1915.

[5] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, "Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017: IEEE, pp. 1025-1032.

[6] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-14, 2022.

[7] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145-153.

[8] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 105-114.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730-3738.

[10] W. Liu, I. Hasan, and S. Liao, "Center and scale prediction: Anchor-free approach for pedestrian and face detection," *arXiv preprint arXiv:1904.02948*, 2019.

[11] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2*, 2011: Springer, pp. 29-39.

[12] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[13] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 127, p. 102276, 2022.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010/06/01 2010, doi: 10.1007/s11263-009-0275-4.

[15] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211-252, 2015.

[16] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014: Springer, pp. 740-755.

[17] A. Kuznetsova *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956-1981, 2020.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[21] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492-1500.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[23] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision,* vol. 104, pp. 154-171, 2013.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020: Springer, pp. 213-229.

[28] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159,* 2020.

A SURVEY OF OBJECT DETECTION MODELS BASED ON DEEP LEARNING

Nguyen Dung[1]*, Hoang Van Dung[2], Le Van Tuong Lan[1]

[1]Faculty of Information Technology, Hue University of Sciences, Hue University

[2]HCMC University of Technology and Education

*Email: nguyendung@hueuni.edu.vn

ABSTRACT

Object detection is the task of classifying and locating objects in images or videos. It has gained prominence in recent years due to its wide applications in real-world problems. This paper surveys recent models on the object detection task especially deep learning based models (survey period from 1990s to 2023). First, the article briefly introduces the object detection task in the field of computer vision, some benchmark datasets and popular evaluation metrics used in this task. Next, the article introduces some models or backbone architectures used and finally, the main point of this article is to introduce outstanding methods using deep learning to detect objects.

Keyword: Computer Vision, Object Detection, Classification, Convolutional Neural Networks, Deep Learning.

Nguyễn Dũng sinh ngày 13/06/1988 tại Thừa Thiên Huế. Ông là giảng viên Khoa Công nghệ thông tin, trường Đại học Khoa học, Đại học Huế. Ông tốt nghiệp cử nhân Tin học tại trường Đại học Khoa học, Đại học Huế năm 2010. Năm 2013, ông tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính tại trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Công nghệ phần mềm, trí tuệ nhân tạo, học máy, học sâu, cơ sở dữ liệu

Hoàng Văn Dũng nhận bằng tiến sĩ ngành Kỹ thuật điện tử và Hệ thống thông tin tại Trường Đại học Ulsan, Hàn Quốc, năm 2015. Ông công tác tại Khoa Công nghệ thông tin, Trường ĐH Sư phạm Kỹ thuật TP. HCM. Ông là thành viên tích cực của các hiệp hội khoa học kỹ thuật như IEEE, IEEE Computer, ICROS. Ông nghiên cứu và xuất bản các công trình liên quan đến Trí tuệ nhân tạo, thị giác máy tính, xử lý ảnh y tế, xe tự hành, các hệ thống hiểu biết thông minh và giám sát thông minh.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, Thị giác máy tính.

Lê Văn Tường Lân sinh ngày 10/11/1974 tại Thừa Thiên Huế. Năm 1996, ông tốt nghiệp Đại học ngành Toán - Tin tại Trường Đại học Khoa học, Đại học Huế. Ông nhận bằng thạc sỹ Công nghệ thông tin tại Trường Đại học Bách Khoa Hà Nội năm 2002 và nhận học vị Tiến sĩ ngành Khoa học máy tính tại Trường Đại học Khoa học, Đại học Huế năm 2018. Hiện công tác tại Ban Đào tạo và Công tác sinh viên, Đại học Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu, Công nghệ phần mềm, Khai phá dữ liệu.