# Khảo sát về đối tượng dựa trên Deep Learning hiện đại Mô hình phát hiện

Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam,
Nadia Kanwal, Mamoona Asghar và Brian Lee

Tóm tắt — Phát hiện đối tượng là nhiệm 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 do các ứng dụng rộng rãi của nó.
Bài viết này khảo sát những phát triển gần đây trong công cụ phát hiện đối tượng dựa trên học sâu. Tổng quan ngắn gọn về bộ dữ liệu điểm chuẩn và số liệu đánh giá được sử dụng trong phát hiện cũng được cung cấp cùng với một số kiến trúc xương sống nổi bật được sử dụng trong các nhiệm vụ nhận dạng. Nó cũng bao gồm các mod phân loại nhẹ đương đại được sử dụng trên các thiết bị cạnh. Cuối cùng, chúng tôi so sánh hiệu suất của các kiến trúc này trên nhiều chỉ số.

Điều khoản chỉ mục — Phát hiện và nhận dạng đối tượng, tích hợp mạng thần kinh (CNN), mạng nhẹ, học sâu

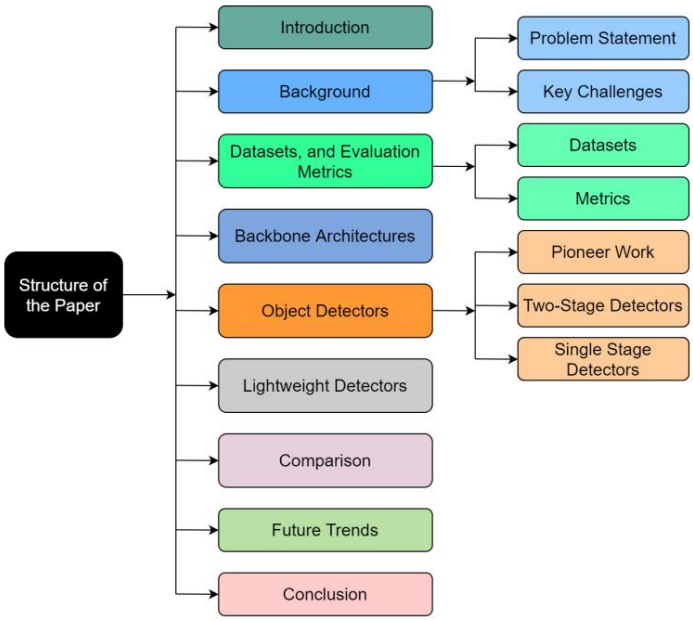
## I. GIỚI THIỆU

Phát hiện đối tượng là một nhiệm vụ tầm thường đối với con người. Một đứa trẻ vài tháng tuổi có thể bắt đầu nhận biết các đồ vật thông thường, tuy nhiên việc dạy nó trên máy tính là một nhiệm vụ khó khăn cho đến đầu thập kỷ trước. Nó đòi hỏi phải xác định và khoanh vùng tất cả các trường hợp của một đối tượng (như ô tô, con người, biển báo đường phố, v.v.) trong trường nhìn. Tương tự, các tác vụ khác như phân loại, phân đoạn, ước tính chuyển động, hiểu cảnh, v.v., là những vấn đề cơ bản trong thị giác máy tính.

Các mô hình phát hiện đối tượng ban đầu được xây dựng như một tập hợp các công cụ trích xuất tính năng thủ công như máy dò Viola-Jones [1], Biểu đồ phân định hướng (HOG) [2], v.v. Các mô hình này chậm, không chính xác và hoạt động kém trên các bộ dữ liệu không quen thuộc . Sự ra đời trở lại của mạng nơ-ron phức hợp (CNN) và học sâu để phân loại hình ảnh đã thay đổi toàn cảnh của nhận thức thị giác. Việc sử dụng nó trong thử thách nhận dạng hình ảnh quy mô lớn Im ageNet (ILSVRC) 2012 của AlexNet [3] đã truyền cảm hứng cho những nghiên cứu sâu hơn về ứng dụng của nó trong thị giác máy tính. Ngày nay, tính năng phát hiện đối tượng được ứng dụng từ xe tự lái và phát hiện danh tính cho đến các mục đích sử dụng an ninh và y tế. Trong những năm gần đây, nó đã chứng kiến sự phát triển theo cấp số nhân với sự phát triển nhanh chóng của các công cụ và kỹ thuật mới.

Cuộc khảo sát này cung cấp đánh giá toàn diện về các công cụ phát hiện đối tượng dựa trên học sâu và kiến trúc phân loại nhẹ. Trong khi các đánh giá hiện có khá kỹ lưỡng [4] - [7], hầu hết chúng đều thiếu những phát triển mới trong lĩnh vực này. Những đóng góp chính của bài báo này như sau:

SSA Zaidi, N. Kanwal, M Asghar và B. Lee đến từ Học viện Công nghệ Athlone, Ireland. MS Ansari thuộc Đại học Hồi giáo Aligarh, Ấn Độ. A. Aslam thuộc Trung tâm Phân tích Dữ liệu Insight, Đại học Quốc gia Ireland, Galway. (Email: sahilzaidi78@gmail.com, samar.ansari@zhect.ac.in, asra.aslam@insight-centre.org, nkanwal@ait.ie, masghar@ait.ie, blee@ait.ie)




Hình 1: Cấu trúc của bài báo.

1) Bài báo này cung cấp phân tích chuyên sâu về các máy dò đối tượng chính trong cả hai loại - máy dò một giai đoạn và hai giai đoạn. Hơn nữa, chúng tôi có cái nhìn lịch sử về sự phát triển của các phương pháp này.
2) Chúng tôi trình bày một đánh giá chi tiết về các kiến trúc xương sống và các mô hình nhẹ mang tính bước ngoặt. Chúng tôi không thể tìm thấy bất kỳ bài báo nào cung cấp một cái nhìn tổng thể về cả hai chủ đề này.

Trong bài báo này, chúng tôi đã xem xét một cách có hệ thống các kiến trúc phát hiện đối tượng khác nhau và các công nghệ liên quan của nó, như được minh họa trong hình 1. Phần còn lại của bài báo này được sắp xếp như sau. Trong phần II, vấn đề phát hiện đối tượng và các thách thức liên quan của nó sẽ được thảo luận. Các bộ dữ liệu điểm chuẩn và số liệu đánh giá khác nhau được liệt kê trong Phần III. Trong Phần IV, một số kiến trúc xương sống quan trọng được sử dụng trong các máy dò vật thể hiện đại được kiểm tra. Phần V được chia thành ba phần phụ chính, mỗi phần nghiên cứu một loại khác nhau của máy dò đối tượng. Tiếp theo là phân tích phân loại đặc biệt của các bộ phát hiện đối tượng, được gọi là mạng nhẹ trong phần VI và phân tích so sánh trong phần VII. Các xu hướng tương lai được đề cập trong Phần VIII trong khi bài báo được kết luận trong Phần IX.

(a) PASCAL VOC 12 (b) MS-COCO (c) ILSVRC (d) OpenImage

Hình 2: Hình ảnh mẫu từ các bộ dữ liệu khác nhau.

## II. LAI LỊCH

### A. Tuyên bố vấn đề

Phát hiện đối tượng là phần mở rộng tự nhiên của phân loại đối tượng, chỉ nhằm mục đích nhận dạng đối tượng trong ảnh. Mục tiêu của việc phát hiện đối tượng là phát hiện tất cả các phiên bản của các lớp được xác định trước và cung cấp bản địa hóa thô của nó trong hình ảnh bằng các hộp căn chỉnh theo trục. Bộ dò sẽ có thể xác định tất cả các trường hợp của các lớp đối tượng và vẽ hộp giới hạn xung quanh nó. Nó thường được coi là một vấn đề học tập có giám sát. Các mô hình phát hiện đối tượng hiện đại có quyền truy cập vào tập hợp lớn các hình ảnh được gắn nhãn để đào tạo và được đánh giá trên các tiêu chuẩn chuẩn khác nhau.

### B. Những thách thức chính trong phát hiện đối tượng

Thị giác máy tính đã trải qua một chặng đường dài trong thập kỷ qua, tuy nhiên nó vẫn còn một số thách thức lớn cần vượt qua. Một số thách thức chính mà mạng phải đối mặt trong các ứng dụng đời thực là:

- Biến đổi nội bộ lớp: Sự biến đổi nội bộ lớp giữa các thể hiện của cùng một đối tượng là tương đối phổ biến trong tự nhiên. Sự thay đổi này có thể do nhiều lý do khác nhau như khớp, độ chiếu sáng, tư thế, điểm nhìn, v.v ... Những tác động bên ngoài không bị hạn chế này có thể có tác động mạnh mẽ đến hình dáng đối tượng [5]. Dự kiến rằng các đối tượng có thể có biến dạng không cứng hoặc bị xoay, thu nhỏ hoặc mờ. Một số đối tượng có thể có xung quanh không dễ thấy, gây khó khăn cho việc khai thác.
- Số loại: Số lượng tuyệt đối các lớp đối tượng có sẵn để phân loại khiến nó trở thành một vấn đề khó giải quyết. Nó cũng yêu cầu nhiều dữ liệu được chú thích chất lượng cao hơn, điều này khó có được. Sử dụng ít ví dụ hơn để đào tạo một máy dò là một câu hỏi nghiên cứu mở. • Hiệu quả: Các mô hình ngày nay cần tài nguyên tính toán cao để tạo ra kết quả phát hiện chính xác. Với việc các thiết bị cạnh và góc cạnh trở nên phổ biến, các thiết bị phát hiện đối tượng hiệu quả là rất quan trọng để phát triển hơn nữa trong lĩnh vực thị giác máy tính.

## III. SỐ LIỆU VÀ PHƯƠNG PHÁP ĐÁNH GIÁ

### A. Tập dữ liệu

Phần này trình bày tổng quan về các tập dữ liệu có sẵn và được sử dụng phổ biến nhất cho các nhiệm vụ phát hiện đối tượng.

1) PASCAL VOC 07/12: Thử thách Pascal Visual Object Classes (VOC) là một nỗ lực trong nhiều năm nhằm đẩy nhanh sự phát triển trong lĩnh vực nhận thức thị giác. Nó bắt đầu vào năm 2005 với các nhiệm vụ phân loại và phát hiện trên bốn lớp đối tượng [8], nhưng hai phiên bản của thử thách này chủ yếu được sử dụng như một điểm chuẩn tiêu chuẩn. Trong khi thử thách VOC07 có 5k hình ảnh huấn luyện và hơn 12k đối tượng được gắn nhãn [9], thử thách VOC12 đã tăng chúng lên 11k hình ảnh huấn luyện và hơn 27k đối tượng được gắn nhãn [10]. Các lớp đối tượng đã được mở rộng thành 20 danh mục và các tác vụ như phân đoạn và phát hiện hành động cũng được bao gồm. Pascal VOC đã giới thiệu Độ chính xác trung bình trung bình (mAP) ở 0,5 IoU (Giao điểm qua Liên minh) để đánh giá hiệu suất của các mô hình. Hình 3 mô tả sự phân bố số lượng hình ảnh wrt cho các lớp khác nhau trong tập dữ liệu Pascal VOC.

2) ILSVRC: Thử thách nhận dạng hình ảnh quy mô lớn ImageNet (ILSVRC) [11] là một thử thách hàng năm kéo dài từ năm 2010 đến năm 2017 và đã trở thành tiêu chuẩn để đánh giá hiệu suất thuật toán. Kích thước tập dữ liệu đã được mở rộng lên đến hơn một triệu hình ảnh bao gồm 1000 lớp phân loại đối tượng. 200 trong số các lớp này đã được chọn lọc thủ công cho nhiệm vụ phát hiện đối tượng, tạo thành hơn 500 nghìn hình ảnh. Nhiều nguồn khác nhau bao gồm ImageNet [12] và Flikr, đã được sử dụng để xây dựng tập dữ liệu phát hiện. ILSVRC cũng cập nhật chỉ số đánh giá bằng cách nới lỏng ngưỡng IoU để giúp bao gồm phát hiện đối tượng nhỏ hơn. Hình 4 mô tả sự phân bố số lượng ảnh wrt cho các lớp khác nhau trong tập dữ liệu ImageNet.

3) MS-COCO: Microsoft Common Objects in Context (MS-COCO) [13] là một trong những bộ dữ liệu thách thức nhất hiện có. Nó có 91 vật thể phổ biến được tìm thấy trong bối cảnh tự nhiên của chúng mà một đứa trẻ 4 tuổi có thể dễ dàng nhận ra. Nó được ra mắt vào năm 2015 và mức độ phổ biến của nó chỉ tăng lên kể từ đó. Nó có hơn hai triệu phiên bản và một

trung bình 3,5 danh mục cho mỗi hình ảnh. Hơn nữa, nó chứa 7,7 phiên bản cho mỗi hình ảnh, thoải mái hơn so với các bộ dữ liệu phổ biến khác. MS COCO cũng bao gồm các hình ảnh từ các góc nhìn khác nhau. Nó cũng giới thiệu một phương pháp nghiêm ngặt hơn để đo hiệu suất của máy dò. Không giống như Pascal VOC và ILSVCR, nó tính toán IoU từ 0,5 đến 0,95 theo các bước 0,5, 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 cũng sử dụng AP cho các đối tượng nhỏ, vừa và lớn một cách riêng biệt để so sánh hiệu suất ở các quy mô khác nhau. Hình 5 mô tả sự phân bố số lượng ảnh wrt cho các lớp khác nhau trong tập dữ liệu MS-COCO.

4) Hình ảnh mở: Tập dữ liệu Hình ảnh mở [14] của Google bao gồm 9,2 triệu hình ảnh, được chú thích bằng nhãn cấp hình ảnh, hộp giới hạn đối tượng và mặt nạ phân đoạn, trong số những hình ảnh khác. Nó đã được ra mắt vào năm 2017 và đã nhận được sáu bản cập nhật. Để phát hiện đối tượng, Open Images có 16 triệu hộp giới hạn cho 600 danh mục trên 1,9 triệu hình ảnh, khiến nó trở thành tập dữ liệu lớn nhất về bản địa hóa đối tượng. Những người tạo ra nó đã rất cẩn thận để chọn những hình ảnh thú vị, phức tạp và đa dạng, có 8,3 danh mục đối tượng cho mỗi hình ảnh. Một số thay đổi đã được thực hiện đối với AP được giới thiệu trong Pascal VOC như bỏ qua lớp không được chú thích, yêu cầu phát hiện đối với lớp và lớp con của nó, v.v. Hình 6 mô tả sự phân bố số lượng hình ảnh wrt cho các lớp khác nhau trong tập dữ liệu Hình ảnh Mở.

5) Các vấn đề về Sai lệch / Sai lệch Dữ liệu: Trong khi quan sát Hình 3 đến Hình 6, người đọc cảnh báo chắc chắn sẽ nhận thấy rằng số lượng hình ảnh cho các lớp khác nhau thay đổi đáng kể trong tất cả các bộ dữ liệu [15]. Ba (Pascal VOC, MS-COCO và Open Images Dataset) trong số bốn tập dữ liệu được thảo luận ở trên có sự sụt giảm rất đáng kể về số lượng hình ảnh ngoài 5 lớp thường xuyên nhất. Có thể dễ dàng quan sát được trong Hình 3, có 13775 hình ảnh chứa 'người' và sau đó là 2829 hình ảnh chứa 'xe hơi'. Số lượng hình ảnh cho 18 lớp còn lại trong tập dữ liệu này gần như giảm tuyến tính với 55 hình ảnh 'cừu'. Tương tự, đối với tập dữ liệu MS-COCO, lớp 'người' có 262465 hình ảnh và lớp 'xe hơi' thường xuyên nhất tiếp theo có 43867 hình ảnh. Xu hướng giảm tiếp tục cho đến khi chỉ có 198 hình ảnh cho lớp 'máy sấy tóc'. Một hiện tượng tương tự cũng được quan sát thấy trong Tập dữ liệu hình ảnh mở, trong đó lớp 'Người đàn ông' là thường xuyên nhất với 378077 hình ảnh và lớp 'Máy cắt giấy' chỉ có 3 hình ảnh. Điều này rõ ràng thể hiện sự sai lệch trong bộ dữ liệu và bị ràng buộc tạo ra sai lệch trong quá trình đào tạo của bất kỳ mô hình phát hiện đối tượng nào. Do đó, một mô hình phát hiện đối tượng được đào tạo trên các tập dữ liệu sai lệch này sẽ cho thấy hiệu suất phát hiện tốt hơn đối với các lớp có nhiều hình ảnh hơn trong dữ liệu đào tạo. Mặc dù vẫn còn tồn tại, vấn đề này ít rõ ràng hơn một chút trong tập dữ liệu ImageNet, như có thể được quan sát từ Hình 4 từ đó có thể thấy rằng lớp thường xuyên nhất tức là 'koala' có 2469 ảnh và lớp ít thường xuyên nhất tức là 'giỏ hàng 'có 624 hình ảnh. Tuy nhiên, điều này dẫn đến một điểm đáng quan tâm khác trong tập dữ liệu ImageNet: lớp thường xuyên nhất là dành cho 'koala' và lớp xuất hiện nhiều nhất tiếp theo là 'bàn phím máy tính', rõ ràng không phải là đối tượng được săn lùng nhiều nhất trong thế giới thực kịch bản phát hiện đối tượng (nơi người, ô tô, biển báo giao thông, v.v. được quan tâm cao hơn).

B. Số liệu

Máy dò vật thể sử dụng nhiều tiêu chí để đo công suất của máy dò viz., Khung hình trên giây (FPS), độ chính xác và thu hồi. Tuy nhiên, Độ chính xác trung bình trung bình (mAP) là số liệu đánh giá phổ biến nhất. Độ chính xác có nguồn gốc từ Giao điểm trên Liên hợp (IoU), là tỷ lệ giữa diện tích chồng chéo và diện tích liên kết giữa chân lý cơ bản và hộp giới hạn dự đoán. Một ngưỡng được đặt để xác định xem phát hiện có chính xác hay không. Nếu IoU lớn hơn ngưỡng, nó được phân loại là Tích cực thực trong khi IoU dưới ngưỡng đó được phân loại là Dương tính giả. Nếu mô hình không phát hiện được một đối tượng hiện diện trong chân lý cơ bản, nó được gọi là Sai phủ định. Độ chính xác đo tỷ lệ phần trăm dự đoán đúng trong khi thu hồi đo lường các dự đoán chính xác đối với sự thật cơ bản.

$$\text{Phân tích P} = \frac{\text{T rue P ositive}}{\text{T rue P ositive + F alse P ositive}} = \frac{\text{T rue P ositive}}{\text{Tất cả các quan sát}} \quad (1)$$

$$\text{Nhớ lại} = \frac{\text{T rue P ositive}}{\text{T rue P ositive + F alse Âm tính}} = \frac{\text{T rue P ositive}}{\text{All Ground T ruth}} \quad (2)$$

Dựa trên phương trình trên, độ chính xác trung bình được tính riêng cho từng lớp. Để so sánh hiệu suất giữa các bộ phát hiện, giá trị trung bình của độ chính xác trung bình của tất cả các lớp, được gọi là độ chính xác trung bình trung bình (mAP) được sử dụng, hoạt động như một số liệu duy nhất để đánh giá cuối cùng.

## IV. KIẾN TRÚC NỀN TẢNG

Kiến trúc xương sống là một trong những thành phần quan trọng nhất của bộ phát hiện đối tượng. Các mạng này trích xuất tính năng từ hình ảnh đầu vào được sử dụng bởi mô hình. Ở đây, chúng tôi đã thảo luận về một số kiến trúc xương sống quan trọng được sử dụng trong các máy dò hiện đại:

A. AlexNet

Krizhevsky và cộng sự. đề xuất AlexNet [3], một kiến trúc dựa trên mạng nơ-ron phức hợp để phân loại hình ảnh và đã giành chiến thắng trong thử thách Nhận dạng Hình ảnh Quy mô lớn ImageNet (ILSVRC) 2012. Nó đạt được độ chính xác cao hơn đáng kể (hơn 26%) so với các mẫu hiện đại.
AlexNet bao gồm tám lớp có thể học được - năm lớp màu phức tạp 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 N-way (N: số lớp). Nó sử dụng nhiều hạt nhân phức hợp trong toàn mạng để có được các tính năng từ hình ảnh. Nó cũng sử dụng droppout và ReLU để chính quy hóa và hội tụ đào tạo nhanh hơn tương ứng. Các mạng nơ-ron tích tụ đã mang lại một sức sống mới khi nó được giới thiệu lại trong AlexNet và nó sớm trở thành kỹ thuật tiên tiến trong việc xử lý dữ liệu hình ảnh.