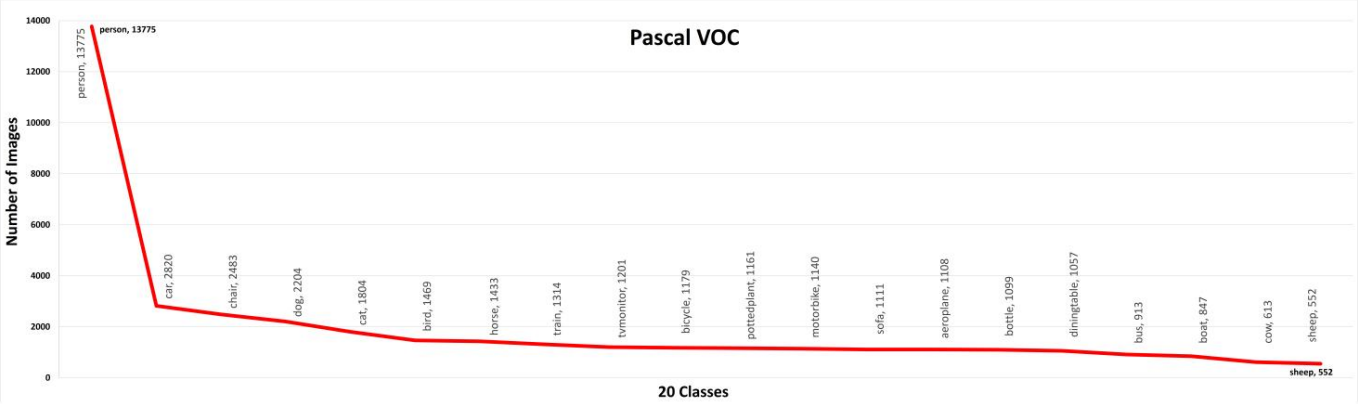
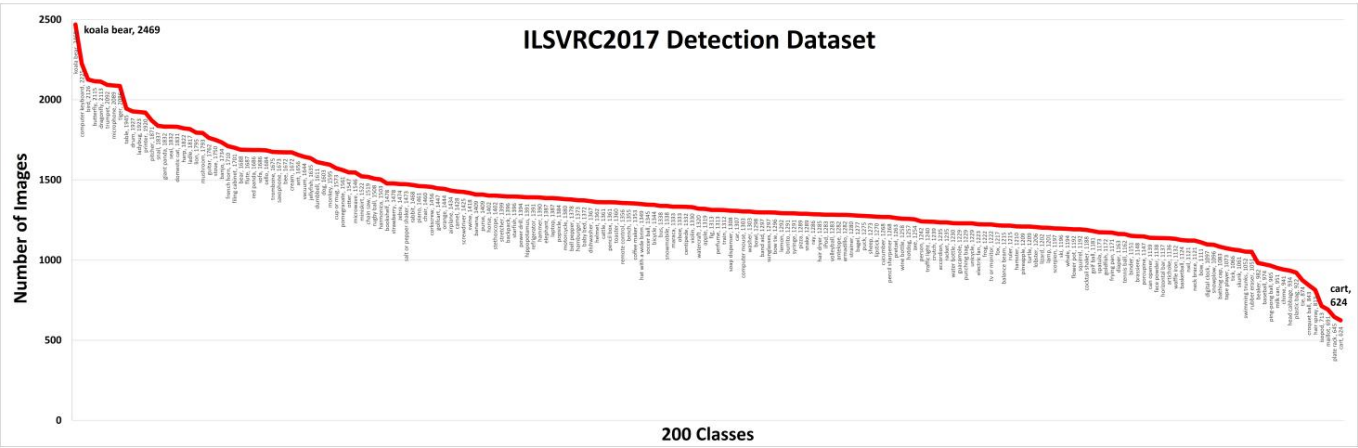
BẢNG I: So sánh các bộ dữ liệu phát hiện đối tượng khác nhau.

Dataset	Các lớp học	Xe lửa			Thẩm định			Bài kiểm tra
		Hình ảnh	Đối tượng	Đối tượng / Hình ảnh	Hình ảnh	Đối tượng	Đối tượng / Hình ảnh	
PASCAL VOC 12	20	5.717	13.609	2,38	5.823	13.841	2,37	10.991
MS-COCO	80	118.287	860.001	7,27	36.781	20.121	7,35	40.670
ILSVRC	200	456.567	1.743.042	1,05	204.621	415.526	2,76	40.152
Mở hình ảnh	600	1.478.807	14.610.229	8,38			4,92	125.436



Hình 3: (Hình ảnh này được xem tốt nhất ở dạng PDF với độ phóng đại) Số lượng hình ảnh cho các lớp khác nhau được chú thích trong Tập dữ liệu PascalVOC [15]



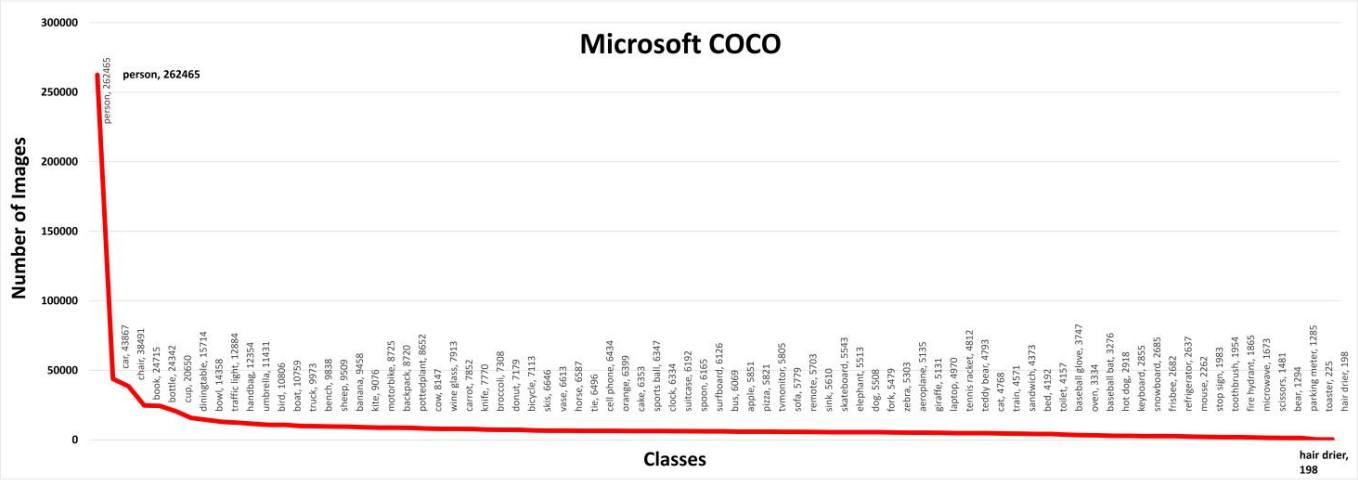
Hình 4: (Hình ảnh này được xem tốt nhất ở dạng PDF với độ phóng đại) Số lượng hình ảnh cho các lớp khác nhau được chú thích trong Tập dữ liệu ImageNet [15]

B. VGG

Trong khi AlexNet [3] và những người kế nhiệm của nó như [16] tập trung vào trên kích thước cửa sổ tiếp nhận nhỏ hơn để cải thiện độ chính xác, Si monyan và Zisserman đã nghiên cứu ảnh hưởng của mạng độ sâu trên đó. Họ đề xuất VGG [17], sử dụng bộ lọc chập để xây dựng các mạng có độ sâu khác nhau. Trong khi một trường tiếp nhận lớn hơn có thể được thu thập bởi một tập hợp bộ lọc chập nhỏ hơn, nó làm giảm đáng kể mạng tham số và hội tụ sớm hơn. Bài báo đã chứng minh kiến trúc mạng sâu (16-19 lớp) có thể được sử dụng như thế nào để thực hiện phân loại và bản địa hóa với độ chính xác vượt trội. VGG được tạo bằng cách thêm một chồng các lớp phức hợp với ba lớp được kết nối đầy đủ, tiếp theo là một softmax lớp. Số lượng các lớp phức tạp, theo tác giả, có thể thay đổi từ 8 đến 16. VGG được đào tạo trong nhiều các lần lặp lại; đầu tiên, kiến trúc 11 lớp nhỏ nhất được đào tạo với khởi tạo ngẫu nhiên có trọng số sau đó được sử dụng để huấn luyện các mạng lớn hơn để ngăn chặn sự mất ổn định của gradient. VGG vượt trội hơn người chiến thắng ILSVRC 2014 GoogLeNet [18] trong danh mục hiệu suất mạng đơn lẻ. Nó sớm trở thành một trong những xương sống mạng được sử dụng nhiều nhất để phân loại đối tượng và các mô hình phát hiện.

C. GoogLeNet / Inception

Mặc dù các mạng lưới phân loại đang xâm nhập hướng tới các mạng nhanh hơn và chính xác hơn, triển khai chúng trong các ứng dụng thế giới thực vẫn còn một chặng đường dài vì chúng sử dụng nhiều tài nguyên. Khi mạng được mở rộng quy mô để tốt hơn hiệu suất, chi phí tính toán tăng theo cấp số nhân. Szegedy và cộng sự. trong [18] mặc nhiên công nhận sự lãng phí tính toán trong mạng như một lý do chính cho nó. Các mô hình lớn hơn cũng có có một số lượng lớn các tham số và có xu hướng trang bị quá mức



Hình 5: (Hình ảnh này được xem tốt nhất ở dạng PDF với độ phóng đại) Số lượng hình ảnh cho các lớp khác nhau được chú thích trong Bộ dữ liệu MS-COCO [15]

dữ liệu. Họ đề xuất sử dụng ture architec được kết nối thưa thớt cục bộ thay vì một tệp được kết nối đầy đủ để giải quyết những vấn đề này. Do đó, GoogLeNet là một mạng sâu 22 lớp, được tạo thành bởi xếp chồng nhiều mô-đun Khởi động lên nhau.
Mô-đun khởi động là mạng có nhiều bộ lọc có kích thước Ở cùng cấp. Bản đồ tính năng đầu vào chuyển qua các bộ lọc này và được nối và chuyển tiếp đến lớp tiếp theo. Các mạng cũng có các bộ phân loại phụ trợ trong các lớp trung gian để giúp điều chỉnh và lan truyền gradient. GoogLeNet đã cho thấy cách sử dụng hiệu quả các khối tính toán có thể thực hiện ngang bằng với các mạng nhiều thông số khác. Nó đạt được độ chính xác trong top 5 là 93,3% trên tập dữ liệu ImageNet [11] mà không cần dữ liệu bên ngoài, trong khi nhanh hơn các mô hình cùng thời khác. Đã cập nhật các phiên bản của Inception như [19], [20] cũng được xuất bản trong những năm tiếp theo đã cải thiện hơn nữa hiệu suất của nó và đã cung cấp thêm bằng chứng về các ứng dụng của các kiến trúc kết nối.

D. ResNets

Khi mạng nơ-ron tích tụ trở nên sâu hơn và sâu hơn, Kaiming He et al. trong [21] cho thấy mức độ chính xác của chúng đầu tiên bão hòa và sau đó phân hủy nhanh chóng. Họ đề xuất sử dụng học tập còn sót lại vào các lớp xếp chồng lên nhau để giảm thiểu hiệu suất giảm dần. Nó được thực hiện bằng cách thêm vào một bỏ qua kết nối giữa các lớp. Kết nối này là một yếu tố bổ sung khôn ngoan giữa đầu vào và đầu ra của khối và không thêm tham số bổ sung hoặc tính toán phức tạp vào mạng. Một ResNet 34 lớp điển hình [21] về cơ bản là một bộ lọc tích chập lớn (7x7) theo sau là 16 nút cổ chai mô-đun (cặp bộ lọc 3x3 nhỏ với phím tắt nhận dạng trên chúng) và cuối cùng là một lớp được kết nối đầy đủ. Nút thắt cổ chai kiến trúc có thể được điều chỉnh cho các mạng sâu hơn bằng cách xếp chồng 3 lớp chập (1x1,3x3,1x3) thay vì 2. Kaiming Ông và cộng sự. cũng đã chứng minh cách mạng VGG 16 lớp có độ phức tạp cao hơn so với 101 và 152 sâu hơn đáng kể của chúng các kiến trúc ResNet lớp trong khi có độ chính xác thấp hơn. Trong bài báo tiếp theo, các tác giả đề xuất Resnetv2 [22] mà đã sử dụng chuẩn hóa hàng loạt và lớp ReLU trong các khối. Nó

BẢNG II: So sánh các kiến trúc Backbone.

Người mẫu	Tham số lớp năm	(Triệu)	Top 1 acc%	FLOPs (Tỷ)
AlexNet 2012		62,4	63,3	1,5
VGG-16 2014	7	138,4	73	15,5
GoogLeNet 2014	16	6,7	-	1,6
ResNet-50 2015	22	25,6	76	3.8
ResNeXt-50 2016	50	25	77,8	4.2
CSPResNeXt-50 2019	50	20,5	78,2	7.9
EfficientNet-B4 2019	59 160	19	83	4.2

là khái quát hơn và dễ đào tạo hơn. ResNets rộng rãi được sử dụng trong phân loại và phát hiện xương sống, và cốt lõi của nó các nguyên tắc đã truyền cảm hứng cho nhiều mạng ([20], [23], [24]).

E. ResNeXt

Các phương pháp thông thường hiện có để cải thiện độ chính xác của một mô hình là bằng cách tăng độ sâu hoặc chiều rộng của mô hình. Tuy nhiên, tăng bất kỳ dẫn đến độ phức tạp của mô hình và số lượng tham số cao hơn trong khi tỷ suất lợi nhuận giảm nhanh chóng. Xie và cộng sự. giới thiệu Kiến trúc ResNeXt [24] đơn giản hơn và hiệu quả hơn hơn các mô hình hiện có khác. ResNeXt được lấy cảm hứng từ xếp chồng các khối tương tự trong VGG / ResNet [3], [21] và hành vi "phân tách-hợp nhất" của mô-đun Inception [18]. Nó là về cơ bản là một ResNet trong đó mỗi khối ResNet được thay thế bằng một mô-đun ResNeXt giống như ban đầu. Phức tạp, phù hợp mô-đun chuyển đổi từ Khởi đầu được thay thế bằng các mô-đun giống nhau về mặt cấu trúc liên kết trong các khối ResNeXt, làm cho mạng dễ mở rộng và tổng quát hơn. Xie và cộng sự. cũng em phasize rằng cardinality (các đường dẫn topo trong ResNeXt khối) có thể được coi là chiều thứ ba, cùng với chiều sâu và chiều rộng, để cải thiện độ chính xác của mô hình. ResNeXt là thanh lịch và ngắn gọn hơn. Nó đạt được độ chính xác cao hơn trong khi có ít hyperparameters hơn đáng kể so với một tương tự chiều sâu kiến trúc ResNet. Nó cũng là người về nhì đầu tiên thách thức ILSVRC 2016.

Hình 6: (Hình ảnh này được xem tốt nhất ở dạng PDF với độ phóng đại) Số lượng hình ảnh cho các lớp khác nhau được chú thích trong tập dữ liệu Hình ảnh Mở [15]

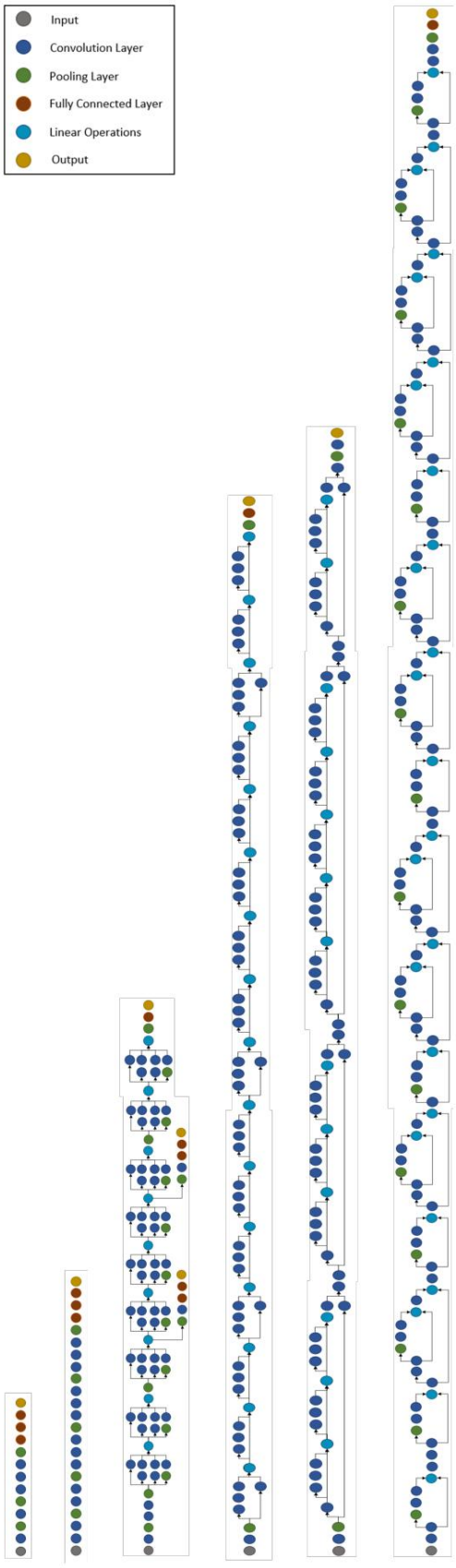
F. CSPNet

Các mạng nơ-ron hiện tại đã cho thấy những kết quả đáng kinh ngạc trong việc đạt được độ chính xác cao trong các tác vụ thị giác máy tính; tuy nhiên, họ dựa vào tài nguyên tính toán quá mức. Wang và cộng sự. tin rằng các phép tính suy luận nặng nề có thể được giảm bớt bằng cách cắt giảm thông tin gradient trùng lặp trong mạng. Họ đã đề xuất CSPNet [25] tạo ra các đường dẫn ent khác nhau cho luồng gradient trong mạng. CSPNet tách bản đồ đối tượng ở lớp cơ sở thành hai phần. Một phần được chuyển qua khối mạng tích chập từng phần (ví dụ, khối Dense và Transition trong DenseNet [23] hoặc khối Res (X) trong ResNeXt [24]) trong khi phần còn lại được kết hợp với các đầu ra của nó ở giai đoạn sau. Điều này làm giảm số lượng tham số, tăng việc sử dụng các đơn vị tính toán và giảm dung lượng bộ nhớ. Nó dễ thực hiện và đủ tổng quát để có thể áp dụng trên các kiến trúc khác như ResNet [21], ResNeXt [24], DenseNet [23], Scaled-YOLOv4 [26], v.v.

Việc áp dụng CSPNet trên các mạng này đã giảm các phép tính từ 10% xuống 20%, trong khi độ chính xác vẫn không đổi hoặc được cải thiện. Chi phí bộ nhớ và tắc nghẽn tính toán cũng được giảm đáng kể với phương pháp này. Nó được tận dụng trong nhiều mô hình máy dò hiện đại, đồng thời cũng được sử dụng cho các thiết bị di động và thiết bị cạnh.

G. EfficientNet

Tan và cộng sự. được nghiên cứu một cách có hệ thống về quy mô mạng và ảnh hưởng của nó đối với hoạt động của mô hình. Họ đã tóm tắt cách thay đổi các thông số mạng như độ sâu, chiều rộng và độ phân giải ảnh hưởng đến độ chính xác của nó. Chia tỷ lệ bất kỳ tham số riêng lẻ đi kèm với một chi phí liên quan. Việc tăng độ sâu của mạng có thể giúp nắm bắt các tính năng phong phú và phức tạp hơn, nhưng chúng rất khó đào tạo do vấn đề về độ dốc biến mất. Tương tự, việc mở rộng độ rộng mạng sẽ giúp dễ dàng nắm bắt các tính năng chi tiết nhưng gặp khó khăn trong việc thu thập các tính năng cấp cao. Thu được từ việc tăng độ phân giải hình ảnh, như chiều sâu và chiều rộng, bão hòa theo tỷ lệ mô hình. Trong bài báo [27], Tan et al. đã đề xuất việc sử dụng một hệ số phức hợp có thể chia tỷ lệ đồng nhất cả ba chiều. Mỗi tham số mô hình có một hằng số liên quan, được tìm thấy bằng cách cố định hệ số là 1 và thực hiện tìm kiếm lưới trên mạng cơ sở. Kiến trúc đường cơ sở, lấy cảm hứng từ công trình trước đây của họ [28], được phát triển bằng cách tìm kiếm kiến trúc thần kinh trên mục tiêu tìm kiếm đồng thời tối ưu hóa độ chính xác và tính toán. EfficientNet là một kiến trúc đơn giản và hiệu quả. Nó vượt trội hơn các mô hình hiện có về độ chính xác và tốc độ trong khi nhỏ hơn đáng kể. Bằng cách cung cấp sự gia tăng đáng kể về hiệu quả, nó có thể mở ra một kỷ nguyên mới trong lĩnh vực



Hình 7: Trực quan hóa các kiến trúc CNN1 Từ trái sang phải: AlexNet, VGG-16, GoogLeNet, ResNet-50, CSPResNeXt-50, EfficientNet-B4.

mạng hiệu quả.

V. NGƯỜI PHÁ T HIỆN ĐỐI TƯỢNG

Chúng tôi đã phân chia đánh giá này dựa trên hai loại máy dò - máy dò hai giai đoạn và một giai đoạn. Tuy nhiên, chúng tôi cũng đã thảo luận về công việc tiên phong, trong đó chúng tôi kiểm tra ngắn gọn một vài thiết bị dò tìm vật thể truyền thống. Một mạng có một mô-đun riêng biệt để tạo các đề xuất khu vực được gọi là một máy dò hai giai đoạn. Các mô hình này cố gắng tìm một số đề xuất đối tượng tùy ý trong một hình ảnh trong giai đoạn đầu tiên và sau đó phân loại và khoanh vùng chúng trong giai đoạn thứ hai. Vì các hệ thống này có hai bước riêng biệt nên chúng thường mất nhiều thời gian hơn để tạo các đề xuất, có kiến trúc phức tạp và thiếu bối cảnh toàn cầu. Máy dò một giai đoạn phân loại và khoanh vùng các đối tượng ngữ nghĩa trong một lần chụp bằng cách sử dụng lấy mẫu dày đặc. Họ sử dụng các hộp / điểm khóa được xác định trước với nhiều tỷ lệ và tỷ lệ khung hình khác nhau để bản địa hóa các đối tượng. Nó cạnh tranh với máy dò hai giai đoạn trong hiệu suất thời gian thực và thiết kế đơn giản hơn.

A. Công việc tiên phong

1) Viola-Jones: Được thiết kế chủ yếu để phát hiện khuôn mặt, máy dò vật thể Viola-Jones [1], được đề xuất vào năm 2001, là một máy dò ac curate và mạnh mẽ. Nó kết hợp nhiều kỹ thuật như các tính năng giống Haar, hình ảnh tích hợp, Adaboost và bộ phân loại cascad ing. Bước đầu tiên là tìm kiếm các tính năng giống Haar bằng cách trượt một cửa sổ trên hình ảnh đầu vào và sử dụng hình ảnh tích phân để tính toán. Sau đó, nó sử dụng một Adaboost được đào tạo để tìm bộ phân loại của từng tính năng haar và xếp tầng chúng. Thuật toán Viola Jones vẫn được sử dụng trong các thiết bị nhỏ vì nó rất hiệu quả và nhanh chóng.

2) HOG Detector: Năm 2005, Dalal và Triggs đề xuất bộ mô tả tính năng Histogram of Oriented Gradients (HOG) [2] được sử dụng để trích xuất các tính năng để phát hiện đối tượng. Đó là một chứng minh tốt hơn các máy dò khác như [29] - [32]. HOG trích xuất gradient và hướng của các cạnh để tạo một bảng tính năng. Hình ảnh được chia thành các lưới và bảng tính năng sau đó được sử dụng để tạo biểu đồ cho mỗi ô trong lưới. Các tính năng HOG được tạo cho vùng quan tâm và được đưa vào bộ phân loại SVM tuyến tính để phát hiện. Máy dò được đề xuất để phát hiện người đi bộ; tuy nhiên, nó có thể được đào tạo để phát hiện các lớp khác nhau.

3) DPM: Mô hình bộ phận có thể biến dạng (DPM) [33] được đưa ra bởi Felzenszwalb và cộng sự. và là người chiến thắng trong thử thách Pascal VOC vào năm 2009. Nó sử dụng "từng phần" riêng lẻ của đối tượng để phát hiện và đạt được độ chính xác cao hơn HOG. Nó tuân theo triết lý chia để trị; các bộ phận của đối tượng được phát hiện riêng lẻ trong thời gian suy luận và một sự sắp xếp có thể xảy ra của chúng được đánh dấu là phát hiện. Ví dụ, cơ thể con người có thể được coi là một tập hợp các bộ phận như đầu, tay, chân và thân. Một mô hình sẽ được chỉ định để chụp một trong các phần trong toàn bộ hình ảnh và quá trình này được lặp lại cho tất cả các phần đó. Sau đó, một mô hình sẽ loại bỏ các cấu hình không thể xảy ra của sự kết hợp của các bộ phận này để tạo ra khả năng phát hiện. Các mô hình dựa trên DPM [34], [35] là một trong những thuật toán thành công nhất trước kỷ nguyên học sâu.

1Tool được sử dụng: https://netron.app/

B. Máy dò hai giai đoạn

1) R-CNN: Công trình Mạng lưới thần kinh hợp pháp dựa trên khu vực (R-CNN) [36] là bài báo đầu tiên trong gia đình R-CNN, và đã chứng minh cách CNN có thể được sử dụng để chứng minh hiệu suất phát hiện vô cùng lớn. R-CNN sử dụng mô-đun đề xuất vùng bất khả tri lớp với CNN để chuyển đổi phát hiện thành vấn đề phân loại và bản địa hóa. Hình ảnh đầu vào đã trừ trung bình lần đầu tiên được chuyển qua mô-đun đề xuất vùng, mô-đun này tạo ra 2000 ứng viên đối tượng. Mô-đun này tìm các phần của hình ảnh có xác suất tìm thấy đối tượng cao hơn bằng cách sử dụng Tìm kiếm có chọn lọc [37]. Những ứng cử viên này sau đó được làm cong và truyền bá thông qua mạng CNN, mạng này trích xuất một vectơ đặc trưng 4096 chiều cho mỗi đề xuất.
Girshick và cộng sự. đã sử dụng AlexNet [3] làm kiến trúc xương sống của máy dò. Các vectơ đặc trưng sau đó được chuyển đến Máy vectơ hỗ trợ (SVM) được đào tạo, dành riêng cho từng lớp cụ thể để có được điểm tin cậy. Cấm không tối đa (NMS) sau đó được áp dụng cho các vùng được chấm điểm, dựa trên IoU và lớp của nó. Khi lớp đã được xác định, thuật toán dự đoán hộp giới hạn của nó bằng cách sử dụng một bộ hồi quy hộp giới hạn được đào tạo, dự đoán bốn tham số, tức là tọa độ tâm của hộp cùng với chiều rộng và chiều cao của nó.

R-CNN có một quá trình đào tạo nhiều tầng phức tạp. Giai đoạn đầu tiên là đào tạo trước CNN với một tập dữ liệu phân loại lớn. Sau đó, nó được tinh chỉnh để phát hiện bằng cách sử dụng các hình ảnh cụ thể của miền (các đề xuất bị cong vênh, đã trừ trung bình) bằng cách đặt lại lớp phân loại với bộ phân loại N + 1 chiều được khởi tạo ngẫu nhiên, N là số lớp, sử dụng gradient descent ngẫu nhiên (SGD) [38]. Một SVM lót và bộ hồi quy hộp giới hạn được đào tạo cho mỗi lớp.

R-CNN đã mở ra một làn sóng mới trong lĩnh vực phát hiện vật thể, nhưng nó chậm (47 giây cho mỗi hình ảnh) và tốn kém về thời gian và không gian [39]. Nó có quá trình đào tạo phức tạp và mất nhiều ngày để đào tạo trên các tập dữ liệu nhỏ ngay cả khi một số tính toán được chia sẻ.

2) SPP-Net: Ông và cộng sự. đề xuất sử dụng lớp Spatial Pyramid Pooling (SPP) [40] để xử lý hình ảnh có kích thước hoặc tỷ lệ co tùy ý. Họ nhận ra rằng chỉ phần được kết nối đầy đủ của CNN mới yêu cầu đầu vào cố định. SPP-net [41] chỉ đơn thuần chuyển các lớp tích chập của CNN trước mô-đun đề xuất vùng và thêm một lớp gộp, do đó làm cho mạng không phụ thuộc vào kích thước / tỷ lệ khung hình và giảm các tính toán. Thuật toán tìm kiếm chọn lọc [37] được sử dụng để tạo ra các cửa sổ ứng viên. Bản đồ đặc trưng thu được bằng cách chuyển hình ảnh đầu vào qua các lớp tích chập của mạng ZF-5 [16]. Các cửa sổ ứng viên sau đó được ánh xạ vào các bản đồ đối tượng, sau đó được chuyển đổi thành các biểu diễn có độ dài cố định bằng các thùng không gian của một lớp gộp hình chóp. Vectơ này được chuyển đến lớp được kết nối đầy đủ và cuối cùng, tới bộ phân loại SVM để dự đoán lớp và điểm. Tương tự như R-CNN [36], SPP-net có lớp xử lý hậu kỳ để cải thiện bản địa hóa bằng cách hồi quy hộp giới hạn. Nó cũng sử dụng cùng một quy trình đào tạo đa tầng, ngoại trừ việc tinh chỉnh chỉ được thực hiện trên các lớp được kết nối đầy đủ.

SPP-Net nhanh hơn đáng kể so với mô hình R-CNN với độ chính xác tương đương. Nó có thể xử lý hình ảnh của bất kỳ hình dạng / tỷ lệ khung hình nào và do đó, tránh biến dạng đối tượng do cong vênh đầu vào. Tuy nhiên, vì kiến trúc của nó tương tự như R-CNN, nó cũng chia sẻ những nhược điểm của R-CNN như đào tạo nhiều tầng, tốn kém về mặt tính toán và thời gian đào tạo.

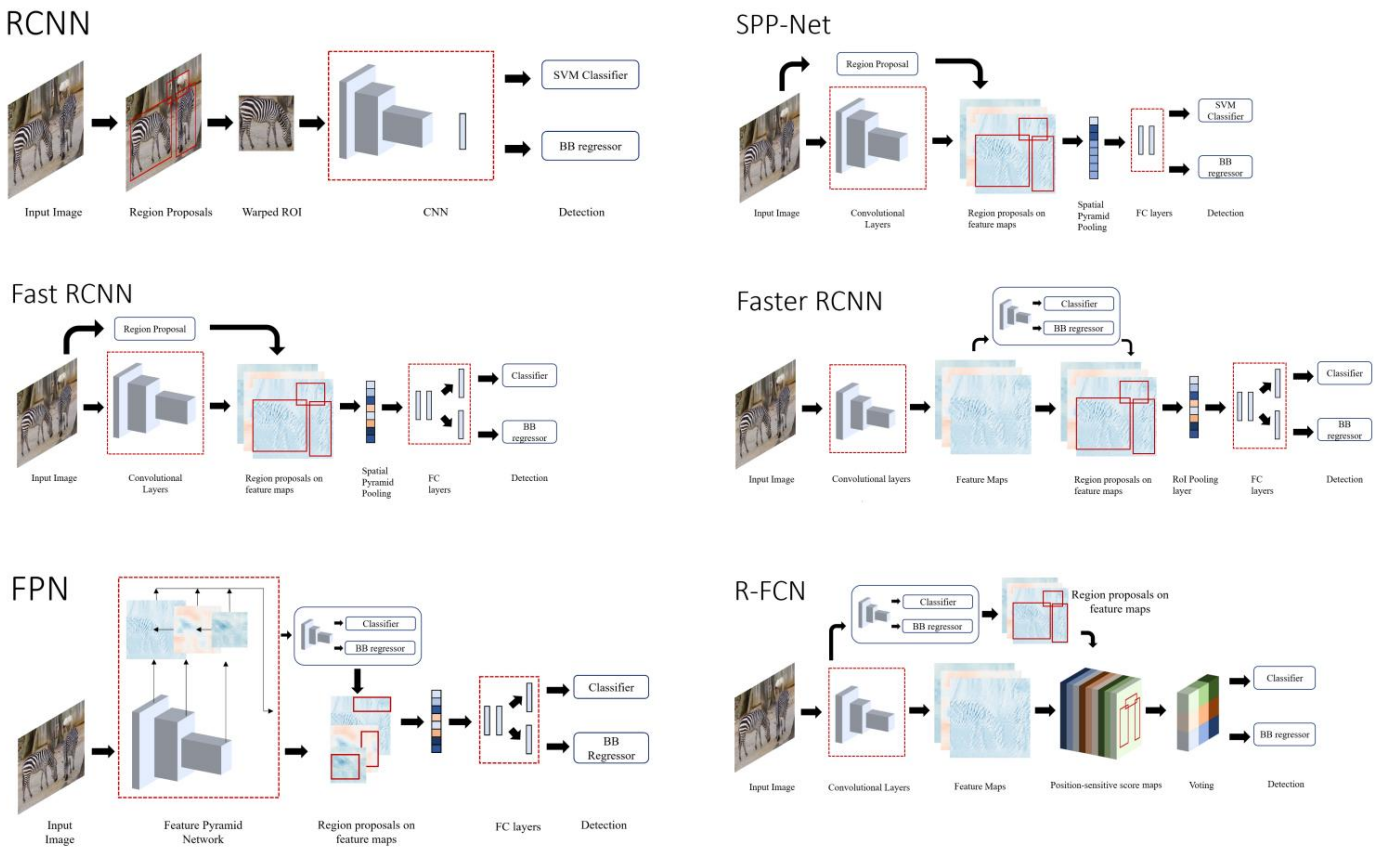
3) Fast R-CNN: Một trong những vấn đề lớn với R-CNN / SPP Net là nhu cầu đào tạo nhiều hệ thống riêng biệt. Fast R-CNN [39] đã giải quyết vấn đề này bằng cách tạo ra một hệ thống đầu cuối có thể đào tạo. Mạng lấy đầu vào một hình ảnh và các đề xuất đối tượng của nó. Hình ảnh được chuyển qua một tập hợp các lớp tích chập và các đề xuất đối tượng được ánh xạ tới các bản đồ đặc trưng thu được. Girshick đã thay thế cấu trúc hình chóp của các lớp gộp từ SPP-net [41] bằng một thùng không gian duy nhất, được gọi là lớp gộp RoI. Lớp này được kết nối với 2 lớp được kết nối đầy đủ và sau đó phân nhánh thành lớp N + 1 lớp SoftMax và lớp hồi quy hộp giới hạn, lớp này cũng có một lớp được kết nối đầy đủ. Mô hình cũng thay đổi chức năng mất mát của bộ hồi quy hộp giới hạn từ L2 thành L1 trơn để có hiệu suất tốt hơn, đồng thời giới thiệu tổn thất đa tác vụ để huấn luyện mạng.

Các tác giả đã sử dụng phiên bản sửa đổi của các mô hình được đào tạo trước hiện đại như [3], [17] và [42] làm xương sống. Mạng được đào tạo trong một bước duy nhất bằng cách giảm độ dốc ngẫu nhiên (SGD) và một loạt nhỏ gồm 2 hình ảnh. Điều này đã giúp mạng hội tụ nhanh hơn khi tính toán chia sẻ lan truyền ngược giữa các RoI từ hai hình ảnh.

Fast R-CNN được giới thiệu là một sự cải thiện về tốc độ (146 lần trên R-CNN) trong khi việc tăng độ chính xác là cần thiết. Nó đơn giản hóa quy trình đào tạo, loại bỏ tích lũy hình chóp và giới thiệu một chức năng mất mát mới. Đối tượng detec tor, không có mạng đề xuất vùng, đã báo cáo tốc độ gần thời gian thực với độ chính xác đáng kể.

4) R-CNN nhanh hơn: Mặc dù Fast R-CNN tiến gần hơn đến phát hiện đối tượng thời gian thực, việc tạo đề xuất vùng của nó vẫn chậm hơn theo thứ tự cường độ (2 giây mỗi hình ảnh so với 0,2 giây mỗi hình ảnh). Ren và cộng sự. đã đề xuất một mạng phức hợp đầy đủ [43] dưới dạng mạng đề xuất vùng (RPN) trong [44] lấy một hình ảnh đầu vào tùy ý và xuất ra một tập hợp các cửa sổ ứng viên. Mỗi cửa sổ như vậy có một
điểm đối tượng xác định khả năng có một đối tượng.
Không giống như những người tiền nhiệm của nó như [21], [34], [39] sử dụng kim tự tháp hình ảnh để giải quyết phương sai kích thước của các đối tượng, RPN giới thiệu hộp Anchor. Nó sử dụng nhiều hộp giới hạn có tỷ lệ khung hình khác nhau và hồi quy chúng để bản địa hóa đối tượng. Hình ảnh đầu vào đầu tiên được chuyển qua CNN để có được một tập hợp các bản đồ đặc trưng. Chúng được chuyển tiếp đến RPN, nơi tạo ra các hộp giới hạn và phân loại của chúng. Các đề xuất đã chọn sau đó được ánh xạ trở lại các bản đồ tính năng thu được từ lớp CNN trước đó trong lớp tổng hợp RoI và cuối cùng được đưa đến lớp được kết nối đầy đủ, lớp này được gửi tới trình phân loại và trình hồi quy hộp giới hạn. Faster R-CNN về cơ bản là Fast R-CNN với RPN là mô-đun đề xuất khu vực.

Việc đào tạo Faster R-CNN phức tạp hơn, do sự hiện diện của các lớp được chia sẻ giữa hai mô hình thực hiện các nhiệm vụ rất khác nhau. Thứ nhất, RPN được đào tạo trước về tập dữ liệu ImageNet [12] và được tinh chỉnh trên tập dữ liệu PASCAL VOC [8]. Một R-CNN nhanh được đào tạo từ các đề xuất khu vực của RPN từ bước đầu tiên. Cho đến thời điểm này, các mạng không có lớp tích chập được chia sẻ. Bây giờ, chúng tôi sửa các lớp chập của bộ dò và tinh chỉnh các lớp duy nhất trong RPN. Và cuối cùng,



Hình 8: Minh họa kiến trúc bên trong của hai bộ dò đối tượng giai đoạn khác nhau2

Fast R-CNN được tinh chỉnh từ RPN cập nhật.

Faster R-CNN đã cải thiện độ chính xác phát hiện so với loại hiện đại trước [39] hơn 3% và giảm thời gian suy luận theo một thứ tự độ lớn. Nó đã khắc phục sự cố tắc nghẽn của đề xuất vùng chậm và chạy trong thời gian gần thực với tốc độ 5 khung hình / giây. Một lợi thế khác của việc có một đề xuất khu vực của CNN là nó có thể học cách đưa ra các đề xuất tốt hơn và do đó tăng độ chính xác.

5) FPN: Sử dụng kim tự tháp hình ảnh để thu được kim tự tháp đặc trưng (hoặc kim tự tháp hình ảnh có lông vũ) ở nhiều cấp độ là một phương pháp phổ biến để tăng khả năng phát hiện các vật thể nhỏ. Mặc dù nó làm tăng Độ chính xác trung bình của máy dò, sự gia tăng thời gian suy luận là đáng kể. Lin và cộng sự. đã đề xuất Mạng Kim tự tháp Tính năng (FPN) [45], có cấu trúc lưu trữ từ trên xuống với các kết nối bên để xây dựng các tính năng ngữ nghĩa cấp cao ở các quy mô khác nhau. FPN có hai con đường, một con đường từ dưới lên là hệ thống phân cấp tính năng điện toán ConvNet ở một số tỷ lệ và một con đường từ trên xuống để nâng cao các bản đồ tính năng thô từ cấp cao hơn thành các tính năng có độ phân giải cao. Các đường dẫn này được kết nối với nhau bằng kết nối bên bằng phép toán tích chập 1x1 để nâng cao thông tin ngữ nghĩa trong các đối tượng địa lý. FPN được sử dụng làm mạng đề xuất vùng (RPN) của Faster R-CNN dựa trên ResNet-101 [21] tại đây.

FPN có thể cung cấp ngữ nghĩa cấp cao ở mọi quy mô, giúp giảm tỷ lệ lỗi khi phát hiện. Nó đã trở thành một khối xây dựng tiêu chuẩn trong các mô hình phát hiện trong tương lai và cải thiện độ chính xác độ chính xác của chúng trên bảng. Nó cũng dẫn đến sự phát triển của các mạng cải tiến khác như PANet [46], NAS-FPN [47] và EfficientNet [27], đây là bộ dò hiện đại.

6) R-FCN: Dai et al. được đề xuất Mạng hoàn toàn Convo dựa trên khu vực (R-FCN) [48] chia sẻ hầu hết tất cả các phép tính tổng hợp trong mạng, không giống như hai bộ dò giai đoạn trước áp dụng các kỹ thuật sử dụng nhiều tài nguyên trên mỗi đề xuất.
Họ lập luận chống lại việc sử dụng các lớp được kết nối đầy đủ và thay vào đó sử dụng các lớp phức hợp. Tuy nhiên, các lớp sâu hơn trong mạng phức hợp là bất biến dịch, làm cho chúng không hiệu quả cho các nhiệm vụ bản địa hóa. Các tác giả đề xuất việc sử dụng bản đồ điểm nhạy cảm về vị trí để khắc phục điều đó. Những bản đồ điểm nhạy cảm này mã hóa thông tin không gian tương đối của đối tượng và sau đó được tổng hợp lại để xác định bản địa hóa chính xác.
R-FCN thực hiện điều đó bằng cách chia vùng quan tâm thành lưới kxk và cho điểm mức độ thích hợp của từng ô với bản đồ tính năng lớp phát hiện. Những điểm số này sau đó được tính trung bình và được sử dụng để dự đoán lớp đối tượng. Bộ dò R-FCN là sự kết hợp của bốn mạng chập. Đầu tiên, hình ảnh đầu vào được chuyển qua ResNet-101 [21] để lấy bản đồ tính năng.
Một đầu ra trung gian (lớp Conv4) được chuyển đến Mạng đề xuất khu vực (RPN) để xác định các đề xuất RoI trong khi đầu ra cuối cùng được xử lý thêm thông qua một lớp phức hợp và được đầu vào cho bộ phân loại và bộ hồi quy. Lớp phân loại kết hợp bản đồ nhạy cảm với vị trí đã tạo với các đề xuất RoI để tạo ra các dự đoán trong khi mạng hồi quy xuất ra các chi tiết hộp giới hạn. R-FCN được đào tạo theo kiểu 4 bước tương tự như Faster-RCNN [44] trong khi sử dụng kết hợp entropy chéo và mất hồi quy hộp. Nó cũng thông qua

khai thác ví dụ trực tuyến (OHEM) [49] trong quá trình đào tạo.

Dai et al. đưa ra một phương pháp mới để giải quyết vấn đề bất biến dịch trong mạng nơ-ron tích chập. R FCN kết hợp Faster R-CNN và FCN để đạt được máy dò nhanh hơn, chính xác hơn. Mặc dù nó không cải thiện độ chính xác nhiều, nhưng nó nhanh hơn 2,5-20 lần so với đối tác của nó.

7) Mặt nạ R-CNN: Mặt nạ R-CNN [50] mở rộng trên Faster R-CNN bằng cách thêm một nhánh khác song song để phân đoạn cá thể đối tượng cấp pixel. Nhánh là một mạng được kết nối đầy đủ được áp dụng trên RoI để phân loại từng pixel thành các phân đoạn với chi phí tính toán tổng thể thấp. Nó sử dụng kiến trúc Faster R-CNN cơ bản tương tự cho đề xuất đối tượng, nhưng thêm một đầu mặt nạ song song với phân loại và đầu hồi quy hộp giới hạn. Một điểm khác biệt chính là việc sử dụng lớp RoIAlign, thay vì lớp RoIPool, để tránh sự sai lệch mức pixel do lượng tử hóa không gian. Các tác giả đã chọn ResNeXt-101 [24] làm xương sống của nó cùng với tính năng Mạng Kim tự tháp (FPN) để có độ chính xác và tốc độ tốt hơn. Chức năng mất mát của Faster R-CNN được cập nhật với mất mặt nạ và như trong FPN, nó sử dụng 5 hộp neo với 3 tỷ lệ khung hình. Nhìn chung quá trình đào tạo Mask R-CNN tương tự như R-CNN nhanh hơn.

Mask R-CNN hoạt động tốt hơn so với các kiến trúc mô hình đơn hiện đại hiện có, đã thêm một chức năng bổ sung là phân đoạn phiên bản với ít tính toán chi phí. Nó đơn giản để đào tạo, linh hoạt và khái quát tốt trong các ứng dụng như phát hiện điểm chính, ước tính tư thế người, v.v. Tuy nhiên, nó vẫn thấp hơn hiệu suất thời gian thực (> 30 khung hình / giây).

8) DetectoRS: Nhiều máy dò hai giai đoạn hiện đại như [44], [51], [52] sử dụng cơ chế nhìn và suy nghĩ hai lần, tức là tính toán các đề xuất đối tượng trước và sử dụng chúng để trích xuất các đặc điểm để phát hiện đối tượng. DetectoRS [53] áp dụng cơ chế này ở cả cấp vĩ mô và vi mô của mạng. Ở cấp vĩ mô, họ đề xuất Kim tự tháp tính năng đệ quy (RFP), được hình thành bằng cách xếp chồng mạng kim tự tháp nhiều tính năng (FPN) với kết nối phản hồi bổ sung từ đường dẫn cấp trên xuống trong FPN đến lớp dưới lên. Đầu ra của FPN được xử lý bởi lớp Atrous Spatial Pyramid Pooling (ASPP) [54] trước khi chuyển nó sang lớp FPN tiếp theo. Một mô-đun Fusion được sử dụng để kết hợp các đầu ra FPN từ các mô-đun khác nhau bằng cách tạo bản đồ chú ý. Ở cấp độ vi mô, Qiao et al. đã trình bày Chuyển đổi tốc độ có thể chuyển đổi (SAC) để điều chỉnh tốc độ giãn nở của tích chập. Một lớp gộp trung bình với bộ lọc 5x5 và tích chập 1x1 được sử dụng làm hàm chuyển đổi để quyết định tốc độ tích chập bất thường [55], giúp xương sống phát hiện các đối tượng ở nhiều quy mô khác nhau một cách nhanh chóng. Họ cũng đóng gói SAC vào giữa hai mô-đun ngữ cảnh toàn cầu [56] vì nó giúp thực hiện chuyển mạch ổn định hơn. Sự kết hợp của hai kỹ thuật này, Kim tự tháp tính năng đệ quy và Chuyển đổi hỗn hợp có thể chuyển đổi cho kết quả DetectoRS. Các tác giả đã kết hợp các kỹ thuật trên với Hybrid Task Cascade (HTC) [51] làm mô hình đường cơ sở và một đường trục ResNext-101.

DetectoRS kết hợp nhiều hệ thống để cải thiện hiệu suất của máy dò và thiết lập tính năng hiện đại cho hai máy dò giai đoạn. Các mô-đun RFP và SAC của nó được khái quát hóa tốt và có thể được sử dụng trong các mô hình phát hiện khác. Tuy nhiên, nó không thích hợp để phát hiện thời gian thực vì nó chỉ có thể xử lý khoảng 4 khung hình mỗi giây.

C. Máy dò một giai đoạn

1) YOLO: Hai bộ phát hiện giai đoạn giải quyết việc phát hiện đối tượng như một vấn đề phân loại, một mô-đun trình bày một số ứng cử viên mà mạng phân loại là đối tượng hoặc nền.
Tuy nhiên, YOLO hoặc You Only Look Once [57] đã sắp xếp lại nó như một bài toán hồi quy, dự đoán trực tiếp các pixel hình ảnh dưới dạng các đối tượng và các thuộc tính hộp giới hạn của nó. Trong YOLO, hình ảnh đầu vào được chia thành lưới S x S và ô nơi tâm của đối tượng rơi xuống có trách nhiệm phát hiện ra nó. Một ô lưới dự đoán nhiều ô giới hạn và mỗi mảng dự đoán bao gồm 5 phần tử: tâm ô giới hạn - x và y, kích thước của ô - w và h, và điểm tin cậy.

YOLO được lấy cảm hứng từ mô hình GoogLeNet để phân loại hình ảnh [18], sử dụng mô-đun xếp tầng của các mạng tích chập nhỏ hơn [58]. Nó được đào tạo trước trên dữ liệu ImageNet [12] cho đến khi mô hình đạt được độ chính xác cao và sau đó được sửa đổi bằng cách thêm tích chập được khởi tạo ngẫu nhiên và các lớp được kết nối đầy đủ. Tại thời điểm huấn luyện, các ô lưới chỉ dự đoán một lớp vì nó hội tụ tốt hơn, nhưng nó được tăng lên trong thời gian suy luận. Suy hao đa nhiệm, tổn thất kết hợp của tất cả các thành phần được dự đoán, được sử dụng để tối ưu hóa mô hình. Không triệt tiêu tối đa (NMS) loại bỏ nhiều phát hiện theo lớp cụ thể.

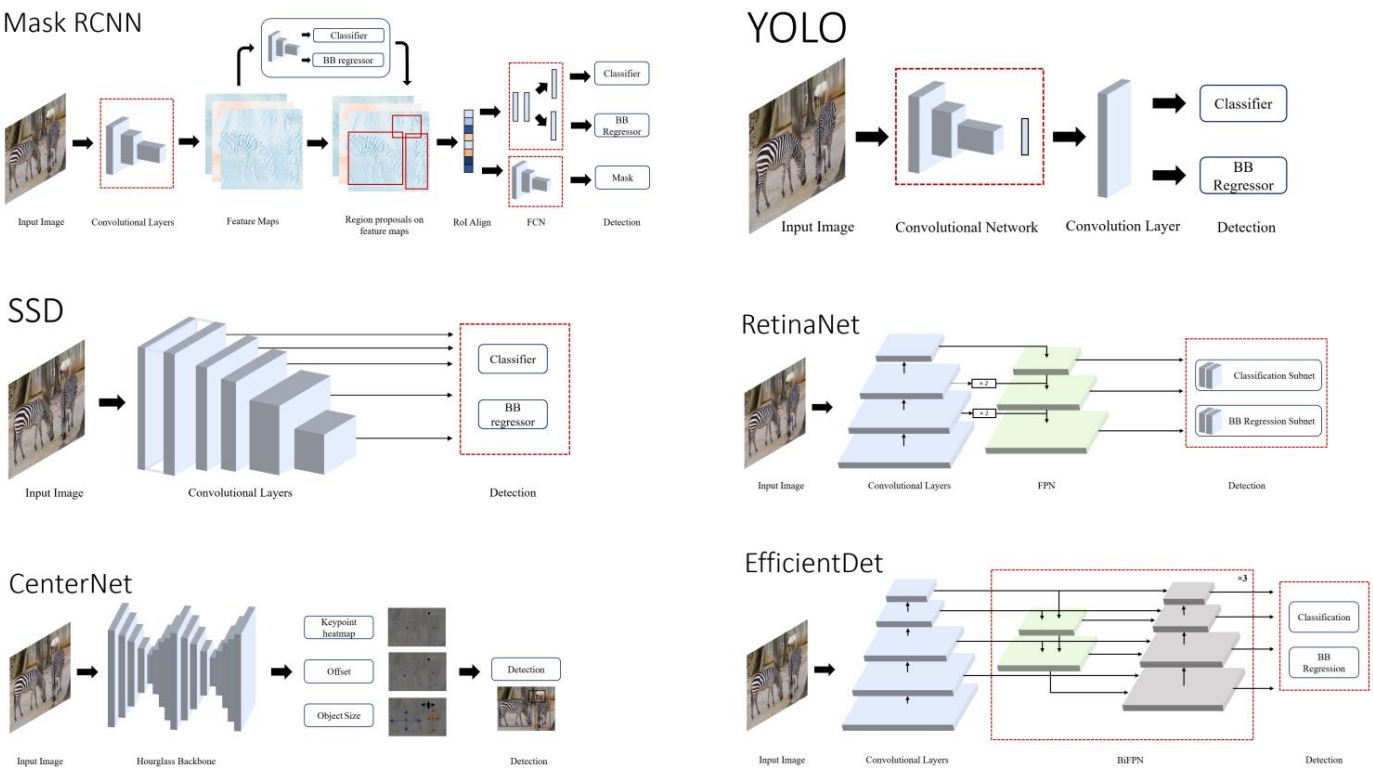
YOLO đã vượt qua các mô hình thời gian thực một giai đoạn hiện đại của mình bằng một lợi nhuận lớn cả về độ chính xác và tốc độ. Đã bao giờ, nó cũng có những thiếu sót đáng kể. Độ chính xác bản địa hóa cho các đối tượng nhỏ hoặc theo cụm và giới hạn về số lượng đối tượng trên mỗi ô là những nhược điểm lớn của nó. Các vấn đề này đã được khắc phục trong các phiên bản sau của YOLO [59] - [61].

2) SSD: Máy dò MultiBox Single Shot (SSD) [62] là máy dò một giai đoạn đầu tiên phù hợp với độ chính xác của các máy dò hai giai đoạn tạm thời như Faster R-CNN [44], trong khi vẫn duy trì tốc độ thời gian thực. SSD được xây dựng trên VGG-16 [17], với các cấu trúc phụ trợ bổ sung để cải thiện hiệu suất.
Các lớp tích chập bổ trợ này, được thêm vào cuối mô hình, giảm dần kích thước. SSD phát hiện các đối tượng nhỏ hơn trong mạng sớm hơn khi các đặc điểm hình ảnh không quá thô, trong khi các lớp sâu hơn chịu trách nhiệm bù đắp các hộp và tỷ lệ khung hình mặc định [63].

Trong quá trình đào tạo, SSD khớp từng hộp sự thật mặt đất với các hộp mặc định có chồng chéo jaccard tốt nhất và đào tạo mạng cho phù hợp, tương tự như Multibox [63]. Họ cũng sử dụng khai thác phủ định cứng và tăng dữ liệu nặng. Tương tự như DPM [33], nó sử dụng tổng trọng số của nội địa hóa và mất độ tin cậy để đào tạo mô hình. Đầu ra cuối cùng thu được bằng cách thực hiện triệt tiêu không tối đa.

Mặc dù SSD nhanh hơn và chính xác hơn đáng kể so với cả hai mạng hiện đại như YOLO và Faster R-CNN, nó vẫn gặp khó khăn trong việc phát hiện các vật thể nhỏ. Vấn đề này sau đó đã được giải quyết bằng cách sử dụng các kiến trúc xương sống tốt hơn như ResNet và các bản sửa lỗi nhỏ khác.

3) YOLOv2 và YOLO9000: YOLOv2 [59], một cải tiến trên YOLO [57], đưa ra sự cân bằng dễ dàng giữa tốc độ và độ chính xác trong khi mô hình YOLO9000 có thể đọc trước 9000 lớp đối tượng trong thời gian thực. Họ đã thay thế kiến trúc xương sống của GoogLeNet [18] bằng DarkNet-19 [64]. Nó kết hợp nhiều kỹ thuật ấn tượng như Batch Normalization [65] để cải thiện sự hội tụ, đào tạo chung các hệ thống phân loại và phát hiện để tăng khả năng phát hiện



Hình 9: Hình minh họa về kiến trúc bên trong của các bộ phát hiện đối tượng hai giai đoạn và một giai đoạn khác nhau2.

các lớp, loại bỏ các lớp được kết nối đầy đủ để tăng tốc độ và sử dụng các hộp neo đã học để cải thiện khả năng thu hồi và có mồi tốt hơn. Redmon và cộng sự. cũng đã kết hợp các bộ dữ liệu phân loại và phát hiện trong cấu trúc phân cấp bằng WordNet [66].
WordTree này có thể được sử dụng để dự đoán xác suất ẩn danh có điều kiện cao hơn, ngay cả khi từ ghép nghĩa không được phân loại chính xác, do đó làm tăng hiệu suất tổng thể của hệ thống.

YOLOv2 cung cấp sự linh hoạt tốt hơn để chọn mô hình về tốc độ và độ chính xác, đồng thời kiến trúc mới có ít tham số hơn. Như tiêu đề của bài báo cho thấy, nó "tốt hơn, nhanh hơn và mạnh hơn" [59].

4) RetinaNet: Đưa ra sự khác biệt giữa độ chính xác của máy dò một giai đoạn và hai giai đoạn, Lin et al. cho rằng lý do độ trễ của bộ phát hiện giai đoạn đơn là "sự mất cân bằng lớp nền ở tiền cảnh cực độ" [67]. Họ đề xuất một tổn thất entropy chéo được định hình lại, được gọi là tổn thất Tiêu điểm như một phương tiện để khắc phục sự mất cân bằng. Tham số tổn thất tiêu điểm làm giảm đóng góp tổn thất từ các ví dụ dễ hiểu. Các tác giả đã chứng minh tính hiệu quả của nó với sự trợ giúp của máy dò một giai đoạn đơn giản, được gọi là RetinaNet [67], dự đoán các đối tượng bằng cách lấy mẫu dày đặc của hình ảnh đầu vào theo vị trí, tỷ lệ và tỷ lệ khung hình. Nó sử dụng ResNet [21] được tăng cường bởi Feature Pyramid Network (FPN) [45] làm xương sống và hai mạng con tương tự - phân loại và bộ hồi quy hộp giới hạn. Mỗi lớp từ FPN được chuyển đến các mạng con, cho phép nó phát hiện các đối tượng dưới dạng các quy mô khác nhau. Mạng con phân loại dự đoán điểm đối tượng cho từng vị trí trong khi mạng con hồi quy hộp hồi quy phần bù cho mỗi điểm neo đối với sự thật mặt đất. Cả hai mạng con đều là FCN nhỏ và chia sẻ các thông số trên các mạng riêng lẻ. không giống như hầu hết các các tác phẩm trước đó, các tác giả sử dụng một công cụ hồi quy hộp giới hạn bất khả tri lớp và nhận thấy chúng có hiệu quả như nhau.

RetinaNet dễ đào tạo, hội tụ nhanh hơn và dễ thực hiện. Nó đạt được hiệu suất tốt hơn về độ chính xác và thời gian chạy so với hai máy dò giai đoạn. RetinaNet cũng thúc đẩy phong bì trong việc cải tiến các cách thức tối ưu hóa bộ phát hiện đối tượng bằng cách giới thiệu một chức năng mất mát mới.

5) YOLOv3: YOLOv3 đã có "những cải tiến gia tăng" so với các phiên bản YOLO trước đó [57], [59]. Redmon và cộng sự. đã thay thế mạng trích xuất tính năng bằng mạng Darknet-53 lớn hơn [64]. Họ cũng kết hợp các kỹ thuật khác nhau như tăng dữ liệu, đào tạo nhiều quy mô, chuẩn hóa hàng loạt, trong số những kỹ thuật khác. Softmax trong lớp phân loại đã được thay thế bằng một trình phân loại hậu cần.

Mặc dù YOLOv3 nhanh hơn YOLOv2 [59], nó không có bất kỳ thay đổi đột phá nào so với người tiền nhiệm. Nó thậm chí còn có độ chính xác thấp hơn máy dò hiện đại một năm tuổi [67].

6) CenterNet: Zhou và cộng sự. trong [68] có một cách tiếp cận rất khác về mô hình hóa các đối tượng dưới dạng điểm, thay vì biểu diễn hộp giới hạn thông thường. CenterNet dự đoán đối tượng là một điểm duy nhất ở tâm của hộp giới hạn.
Hình ảnh đầu vào được chuyển qua FCN để tạo ra bản đồ nhiệt, có các đỉnh tương ứng với tâm của đối tượng được phát hiện.
Nó sử dụng một Hourglass-101 xếp chồng lên nhau của ImageNet [69] làm mạng trích xuất tính năng và có 3 đầu - đầu bản đồ nhiệt để xác định tâm đối tượng, đầu kích thước để ước tính kích thước của đối tượng và đầu bù để hiệu chỉnh độ lệch của điểm đối tượng.
Đa nhiệm mất cả ba đầu được truyền lại thành tính năng

2Tính năng được tạo bằng cách sử dụng: https://poloclub.github.io/cnn-explainer/

vắt trong khi đào tạo. Trong quá trình suy luận, đầu ra từ đầu bù được sử dụng để xác định điểm đối tượng và cuối cùng một hộp được tạo ra. Vì các dự đoán, không phải kết quả, là các điểm và không phải các hộp giới hạn, không bắt buộc phải triệt tiêu tối đa (NMS) để xử lý hậu kỳ.

CenterNet mang đến một góc nhìn mới mẻ và dành nhiều năm tiến bộ trong lĩnh vực phát hiện đối tượng. Nó chính xác hơn và có ít thời gian suy luận hơn so với những người tiền nhiệm của nó. Nó có độ chính xác cao cho nhiều tác vụ như phát hiện đối tượng 3D, ước tính điểm chính, tư thế, phân đoạn phiên bản, phát hiện định hướng và các tác vụ khác. Tuy nhiên, nó yêu cầu các kiến trúc xương sống khác nhau vì các kiến trúc chung hoạt động tốt với các máy dò khác sẽ cho hiệu suất kém với nó và ngược lại.

7) EfficientDet: EfficientDet [70] được xây dựng theo hướng ý tưởng về bộ dò có thể mở rộng với độ chính xác và hiệu quả cao hơn. Nó giới thiệu các tính năng đa quy mô hiệu quả, BiFPN và mở rộng mô hình. BiFPN là mạng kim tự tháp tính năng hai chiều với các trọng số có thể học được để kết nối chéo các tính năng đầu vào ở các quy mô khác nhau. Nó cải thiện trên NAS-FPN [47], đòi hỏi đào tạo nặng và có mạng phức tạp, bằng cách loại bỏ một nút đầu vào và thêm một kết nối bên bổ sung. Điều này giúp loại bỏ các nút kém hiệu quả hơn và tăng cường kết hợp tính năng cấp cao. Không giống như các máy dò hiện có mở rộng quy mô với các lớp FPN lớn hơn, sâu hơn hoặc xếp chồng lên nhau, phần giới thiệu EfficientDet sử dụng một hệ số kép có thể được sử dụng để "cùng nhau mở rộng quy mô tất cả các kích thước của mạng đường trục, mạng BiFPN, mạng lớp / hộp và độ phân giải" [70] . EfficientDet sử dụng EfficientNet [27] làm mạng xương sống với nhiều tập hợp các lớp BiFPN xếp chồng lên nhau như một mạng khai thác tính năng.

Mỗi đầu ra từ lớp BiFPN cuối cùng được gửi đến mạng dự đoán lớp và hộp. Mô hình được đào tạo bằng cách sử dụng trình tối ưu hóa SGD cùng với chuẩn hóa hàng loạt đồng bộ và sử dụng kích hoạt swish [71], thay vì kích hoạt ReLU tiêu chuẩn, có thể phân biệt, hiệu quả hơn và có hiệu suất tốt hơn.

EfficientDet đạt được hiệu quả và độ chính xác tốt hơn so với các máy dò trước đó trong khi nhỏ hơn và rẻ hơn về mặt tính toán. Nó dễ dàng mở rộng quy mô, tổng quát hóa tốt cho các nhiệm vụ khác và là mô hình hiện đại nhất hiện nay để phát hiện đối tượng một giai đoạn.

8) YOLOv4: YOLOv4 [61] đã kết hợp rất nhiều ý tưởng thú vị để thiết kế một máy dò vật thể nhanh và dễ đào tạo có thể hoạt động trong các hệ thống sản xuất hiện có. Nó sử dụng "túi quà tặng", tức là, các phương pháp chỉ tăng thời gian đào tạo và không ảnh hưởng đến thời gian suy luận. YOLOv4 sử dụng các kỹ thuật tăng cường dữ liệu, phương pháp điều chỉnh, làm mịn nhãn lớp, mất CIoU [72], Chuẩn hóa hàng loạt mini chéo (CmBN), Huấn luyện tự đối đầu, Bộ lập lịch ủ Cosine [73] và các thủ thuật khác để cải thiện quá trình huấn luyện. Các phương pháp chỉ ảnh hưởng đến thời gian suy luận, được gọi là "Túi đặc biệt", cũng được thêm vào mạng, bao gồm kích hoạt Mish [74], Kết nối từng phần qua các giai đoạn (CSP) [25], SPP-Block [41], Đường dẫn PAN khối tổng hợp [46], Nhiều kết nối dư có trọng số đầu vào (Mi WRC), v.v. Nó cũng sử dụng thuật toán đặt trước để tìm kiếm tham số tối ưu và sử dụng CSPDarknet53 làm xương sống được đào tạo trước ImageNet với SPP và PAN block cổ và YOLOv3 làm đầu phát hiện.

Hầu hết các thuật toán phát hiện hiện tại yêu cầu nhiều GPU để đào tạo mô hình, nhưng YOLOv4 có thể dễ dàng được đào tạo trên một GPU. Nó nhanh gấp đôi so với EfficientDet với hiệu suất tương đương. Đây là công nghệ tiên tiến nhất dành cho máy dò một giai đoạn trong thời gian thực.

9) Máy biến áp Swin: Máy biến áp [75] đã có tác động chuyên nghiệp trong lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP) kể từ khi ra đời. Ứng dụng của nó trong các mô hình ngôn ngữ như BERT (Biểu diễn mã hóa hai chiều từ bộ chuyển đổi) [76], GPT (Biến áp được đào tạo trước tạo) [77], T5 (Biến áp chuyển văn bản thành văn bản) [78], v.v. đã thúc đẩy trạng thái của nghệ thuật trong lĩnh vực này. Transformers [75] sử dụng mô hình chú ý để thiết lập sự phụ thuộc giữa các phần tử của trình tự và có thể tham gia vào ngữ cảnh dài hơn so với các kiến trúc tuần tự khác. Sự thành công của máy biến áp trong NLP đã làm dấy lên sự quan tâm đến ứng dụng của nó trong thị giác máy tính. Trong khi CNN đã là trụ cột cho sự thăng tiến trong tầm nhìn, chúng có một số thiếu sót cố hữu như thiếu tầm quan trọng của bối cảnh toàn cầu, trọng số cố định sau đào tạo [79], v.v.

Swin Transformer [80] tìm cách cung cấp xương sống dựa trên máy biến áp cho các nhiệm vụ thị giác máy tính. Nó chia các hình ảnh đầu vào thành nhiều bản vá không chồng chéo và chuyển đổi chúng thành các bản nhúng. Sau đó, nhiều khối Swin Transformer được áp dụng cho các bản vá trong 4 giai đoạn, với mỗi giai đoạn kế tiếp sẽ giảm số lượng các bản vá để duy trì biểu diễn phân cấp. Khối Biến áp Swin bao gồm các mô-đun tự chú ý nhiều đầu cục bộ (MSA), dựa trên cửa sổ bản vá được dịch chuyển xen kẽ trong các khối liên tiếp. Độ phức tạp của việc đặt com trở nên tuyến tính với kích thước hình ảnh tự chú ý cục bộ trong khi cửa sổ được dịch chuyển cho phép kết nối nhiều cửa sổ. [80] cũng cho thấy cách các cửa sổ được dịch chuyển làm tăng độ chính xác của việc phát hiện với một ít chi phí.

Máy biến áp thể hiện một sự thay đổi mô hình so với mạng nơ-ron dựa trên CNN. Mặc dù ứng dụng của nó trong thị giác vẫn còn trong giai đoạn sơ khai, nhưng tiềm năng của nó để thay thế tích chập từ các nhiệm vụ này là rất thực tế. Swin Transformer đạt được trình độ tiên tiến nhất trên tập dữ liệu MS COCO, nhưng sử dụng các tham số tương đối cao hơn so với các mô hình tích tụ.

VI. MẠNG CHIẾU SÁ NG

Một nhánh nghiên cứu mới đã hình thành trong những năm gần đây, nhằm mục đích thiết kế các mạng nhỏ và hiệu quả cho các môi trường hạn chế tài nguyên như thường thấy trong việc triển khai Internet of Things (IoT) [81] - [84]. Xu hướng này cũng đã lan rộng đến việc thiết kế các công cụ phát hiện đối tượng mạnh mẽ. Có thể thấy rằng mặc dù một số lượng lớn máy dò đối tượng đạt được độ chính xác tuyệt vời và thực hiện suy luận trong thời gian thực, nhưng phần lớn các mô hình này đòi hỏi quá nhiều tài nguyên máy tính và do đó không thể triển khai trên các thiết bị biên.

Nhiều cách tiếp cận khác nhau đã cho thấy những kết quả thú vị trong quá khứ. Việc sử dụng các thành phần hiệu quả và các kỹ thuật nén như cắt tỉa ([85], [86]), lượng tử hóa ([87], [88]), băm [89], v.v. đã cải thiện hiệu quả của các mô hình học sâu. Việc sử dụng mạng lưới lớn được đào tạo để đào tạo các mô hình nhỏ hơn, được gọi là chưng cất [90], cũng đã cho thấy những kết quả thú vị. Tuy nhiên, trong phần này, chúng tôi khám phá một số ví dụ nổi bật về thiết kế mạng nơ-ron hiệu quả để đạt được hiệu suất cao trên các thiết bị cạnh.

A. SqueezeNet

Những tiến bộ gần đây trong lĩnh vực CNN chủ yếu tập trung vào việc cải thiện độ chính xác hiện đại của bộ dữ liệu điểm chuẩn, dẫn đến sự bùng nổ về kích thước mô hình và các thông số của chúng. Nhưng vào năm 2016, Iandola et al. đã đề xuất một mạng nhỏ hơn, thông minh hơn được gọi là SqueezeNet [91], giúp giảm các thông số trong khi vẫn duy trì hiệu suất. Họ đã đạt được điều đó bằng cách sử dụng ba chiến lược thiết kế chính. sử dụng các bộ lọc nhỏ hơn, giảm số lượng kênh đầu vào xuống bộ lọc 3x3 và đặt các lớp lấy mẫu xuống sau này trong mạng. Hai chiến lược đầu tiên làm giảm số lượng tham số trong khi cố gắng duy trì độ chính xác và chiến lược thứ ba làm tăng độ chính xác của mạng. Khối xây dựng của SqueezeNet được gọi là mô-đun lửa, bao gồm hai lớp: lớp ép và lớp mở rộng, mỗi lớp có kích hoạt ReLU. Lớp ép được tạo thành từ nhiều bộ lọc 1x1 trong khi lớp mở rộng là sự kết hợp của các bộ lọc 1x1 và 3x3, do đó hạn chế số lượng kênh đầu vào. Kiến trúc SqueezeNet bao gồm một chồng 8 mô-đun Fire nằm gọn giữa các lớp chập. Lấy cảm hứng từ ResNet [21], SqueezeNet với các kết nối dư cũng được đề xuất để tăng độ chính xác so với mô hình vani. Các tác giả cũng đã thử nghiệm với Deep Compression [87] và đạt được kích thước mô hình giảm 510 lần so với AlexNet, trong khi vẫn duy trì độ chính xác của đường cơ sở. SqueezeNet đã giới thiệu một ứng cử viên sáng giá để cải thiện hiệu quả phần cứng của kiến trúc mạng thần kinh.

B. MobileNets

MobileNet [92] đã rời xa các mô hình nhỏ thông thường như thu nhỏ, cắt tỉa, lượng tử hóa hoặc nén, và thay vào đó sử dụng kiến trúc mạng hiệu quả.
Mạng đã sử dụng tích chập phân tách theo chiều sâu, tính toán nhân tử của một tích chập tiêu chuẩn thành một tích chập theo chiều sâu và một tích chập theo chiều kim điểm 1x1. Phép chập tiêu chuẩn sử dụng các hạt nhân trên tất cả các kênh đầu vào và kết hợp chúng trong một bước trong khi phép chập theo chiều sâu sử dụng các hạt nhân khác nhau cho mỗi kênh đầu vào và sử dụng phép chập điểm để kết hợp các đầu vào. Việc tách lọc và kết hợp các tính năng này làm giảm chi phí tính toán và kích thước mô hình. MobileNet bao gồm 28 lớp tích hợp riêng biệt, mỗi lớp tiếp theo là chuẩn hóa hàng loạt và chức năng kích hoạt ReLU. Howard và cộng sự. cũng giới thiệu hai siêu tham số thu nhỏ mô hình: chiều rộng và hệ số phân giải, nhằm cải thiện hơn nữa tốc độ và giảm kích thước của mô hình. Hệ số độ rộng điều khiển độ rộng của mạng một cách đồng nhất bằng cách giảm các kênh đầu vào và đầu ra trong khi hệ số độ phân giải ảnh hưởng đến kích thước của hình ảnh đầu vào và các biểu diễn của nó trong toàn mạng. MobileNet đạt được khả năng quản lý ac tương đương với một số mô hình chính thức trong khi kích thước chỉ bằng một phần nhỏ của chúng. Howard và cộng sự. cũng cho thấy cách nó có thể tổng quát hóa qua các ứng dụng khác nhau như phân bổ khuôn mặt, định vị địa lý và phát hiện đối tượng. Tuy nhiên, nó quá đơn giản và tuyến tính như VGG và do đó có ít lối đi xen kẽ nhiều hơn. Những điều này đã được sửa chữa trong các lần lặp lại sau này của mô hình này [93], [94].

C. ShuffleNet

Năm 2017, Zhang et al. đã giới thiệu ShuffleNet [95], một kiến trúc mạng thần kinh vô cùng hiệu quả về mặt tính toán, được thiết kế đặc biệt cho các thiết bị di động. Họ nhận ra rằng nhiều mạng hiệu quả trở nên kém hiệu quả hơn khi chúng thu nhỏ quy mô và cho rằng nó là do các chập 1x1 đắt tiền gây ra. Cùng với việc xáo trộn kênh, họ đã đặt ra việc sử dụng tích chập nhóm để tránh nhược điểm của luồng thông tin hạn chế. ShuffleNet chủ yếu bao gồm một tích chập tiêu chuẩn theo sau là các ngăn xếp của các đơn vị ShuffleNet được nhóm lại trong ba giai đoạn. Đơn vị ShuffleNet tương tự như khối ResNet trong đó chúng sử dụng tích chập theo chiều sâu trong lớp 3x3 và thay thế lớp 1x1 bằng tích chập nhóm theo chiều kim loại. Lớp tích chập theo chiều sâu được đặt trước bởi thao tác trộn kênh. Chi phí tính toán của ShuffleNet có thể được quản lý bởi hai siêu tham số: số nhóm để kiểm soát độ thưa thớt của kết nối và hệ số tỷ lệ để thao tác kích thước mô hình. Khi số lượng nhóm trở nên lớn, tỷ lệ lỗi bão hòa khi các kênh đầu vào cho mỗi nhóm giảm và do đó có thể làm giảm khả năng biểu diễn. ShuffleNet làm tốt hơn các mô hình đương đại ([3], [18], [91], [92]) trong khi có kích thước nhỏ hơn đáng kể.

Vì tiến bộ duy nhất trong ShuffleNet là xáo trộn kênh, nên không có bất kỳ sự cải thiện nào về tốc độ suy luận của mô hình.

D. MobileNetv2

Cải tiến trên MobileNetv1 [92], Sandler et al. đề xuất MobileNetv2 [93] vào năm 2018. Nó giới thiệu phần dư ngược với nút cổ chai tuyến tính, một mô-đun lớp mới để giảm đặt com và cải thiện độ chính xác. Mô-đun mở rộng biểu diễn chiều thấp của đầu vào thành chiều cao, lọc với tích chập theo chiều sâu và sau đó chiếu nó trở lại kích thước thấp, không giống như khối dư thông thường thực hiện các hoạt động nén, tích chập và sau đó mở rộng. MobileNetv2 chứa một lớp tích chập theo sau là 19 mô-đun nút cổ chai còn lại và sau đó là hai lớp tích chập. Mô-đun nút cổ chai còn lại chỉ có kết nối phím tắt khi sải chân là 1. Đối với sải chân cao hơn, lối tắt không được sử dụng vì sự khác biệt về kích thước. Họ cũng sử dụng ReLU6 làm hàm phi tuyến tính, thay vì ReLU đơn giản, để hạn chế tính toán. Để phát hiện đối tượng, các tác giả đã sử dụng MobileNetv2 làm trình trích xuất tính năng của một biến thể hiệu quả về mặt tính toán của SSD [62]. Mô hình này, được gọi là SSDLite, được tuyên bố có ít thông số hơn 8 lần so với SSD ban đầu trong khi vẫn đạt được độ chính xác cạnh tranh. Nó khái quát tốt hơn trên các bộ dữ liệu khác, dễ thực hiện và do đó, được cộng đồng đón nhận.

E. PeleeNet

Các mô hình học sâu nhẹ hiện tại như [92], [93], [95] chủ yếu dựa vào phép tích chập phân tách theo chiều sâu, vốn thiếu việc triển khai hiệu quả. Wang và cộng sự. đã đề xuất một kiến trúc hiệu quả mới dựa trên tích chập thông thường, được đặt tên là PeleeNet [96], sử dụng một loạt các kỹ thuật bảo toàn tính toán. PeleeNet tập trung xung quanh DenseNet [23] nhưng đã xem xét nhiều hơn và tìm kiếm cảm hứng. Nó giới thiệu các lớp dày đặc hai chiều, khối gốc, động

số lượng kênh trong một nút cổ chai, lớp chuyển tiếp yêu cầu kích hoạt bài đăng và thông thường để giảm chi phí tính toán và tăng tốc độ. Lấy cảm hứng từ [18], lớp dày đặc hai chiều giúp nhận được các quy mô khác nhau của trường tiếp nhận, giúp dễ dàng xác định các đối tượng lớn hơn. Để giảm mất mát thông tin, một khối gốc đã được sử dụng theo cách tương tự đối với [20], [97]. Họ cũng chia tay với hệ số nén được sử dụng trong [23] vì nó làm tổn hại đến biểu thức tính năng và làm giảm độ chính xác. PeleeNet bao gồm một khối gốc, bốn giai đoạn của các lớp chuyển tiếp và dày đặc đã được sửa đổi, và cuối cùng là lớp phân loại. Các tác giả cũng đề xuất một hệ thống phát hiện đối tượng thời gian thực, được gọi là Pelee, dựa trên PeleeNet và một biến thể của SSD [62]. Hiệu suất của nó so với các thiết bị phát hiện vật thể hiện đại trên thiết bị di động và thiết bị cạnh đã tăng lên nhưng cho thấy các lựa chọn thiết kế đơn giản có thể tạo ra sự khác biệt lớn về hiệu suất tổng thể như thế nào.

F. ShuffleNetv2

Vào năm 2018, Ningning Ma et al. trình bày một bộ hướng dẫn toàn diện để thiết kế kiến trúc mạng hiệu quả trong ShuffleNetv2 [98]. Họ lập luận về việc sử dụng các số liệu trực tiếp như tốc độ hoặc độ trễ để đo độ phức tạp tính toán, thay vì các số liệu gián tiếp như FLOP. ShuffleNetv2 được xây dựng dựa trên bốn nguyên tắc hướng dẫn - 1) độ rộng bằng nhau cho các kênh đầu vào và đầu ra để giảm thiểu chi phí truy cập bộ nhớ, 2) lựa chọn cẩn thận tích chập nhóm dựa trên nền tảng và nhiệm vụ mục tiêu, 3) cấu trúc đa đường dẫn đạt được độ chính xác cao hơn với chi phí hiệu quả và 4) các hoạt động khôn ngoan của phần tử như add và ReLU về mặt tính toán là không đáng kể. Tuân theo các nguyên tắc trên, họ thiết kế một khối công trình mới. Nó chia đầu vào thành hai phần bằng một lớp tách kênh, tiếp theo là ba lớp chập, sau đó được nối với kết nối dư và được chuyển qua một lớp trộn kênh. Đối với mô hình lấy mẫu xuống, tách kênh bị loại bỏ và kết nối còn lại có các lớp tích chập phân tách theo chiều sâu. Một tập hợp các khối này nằm giữa một vài lớp phức hợp dẫn đến ShuffleNetv2. Các au thors cũng đã thử nghiệm với các mô hình lớn hơn (50/162 lớp) và thu được độ chính xác vượt trội với ít FLOP hơn đáng kể.

ShuffleNetv2 đã vượt qua trọng lượng của nó và vượt trội hơn các mô hình hiện đại khác ở độ phức tạp tương đương.

G. MnasNet

Với nhu cầu ngày càng tăng về các mô hình chính xác, nhanh chóng và độ trễ thấp cho các thiết bị biên khác nhau, việc thiết kế một mạng nơ-ron như vậy đang trở nên khó khăn hơn bao giờ hết. Vào năm 2018, Tan et al. đề xuất Mnasnet [28] được thiết kế từ cách tiếp cận tìm kiếm kiến trúc thần kinh tự động (NAS). Họ hình thành vấn đề tìm kiếm dưới dạng tối ưu hóa đa đối tượng nhằm đạt được độ chính xác cao và độ trễ thấp. Nó cũng phân chia dữ liệu của không gian tìm kiếm bằng cách phân chia CNN thành các khối duy nhất và sau đó tìm kiếm các hoạt động và kết nối trong các khối đó một cách riêng biệt, do đó giảm không gian tìm kiếm. Điều này cũng cho phép mỗi khối có một thiết kế đặc biệt, không giống như các mô hình trước đó [99] - [101] xếp chồng các khối giống nhau.

Các tác giả đã sử dụng tác nhân học tập củng cố dựa trên RNN làm bộ điều khiển cùng với một người huấn luyện để đo độ chính xác và tính di động thiết bị cho độ trễ. Mỗi mô hình được lấy mẫu được đào tạo về một nhiệm vụ để có được độ chính xác của nó và chạy trên các thiết bị thực để có độ trễ. Điều này được sử dụng để đạt được mục tiêu phần thưởng mềm và bộ điều khiển được cập nhật. Quá trình được lặp lại cho đến khi số lần lặp tối đa hoặc một ứng cử viên phù hợp được dẫn xuất. Nó bao gồm 16 khối đa dạng, một số có các kết nối còn lại. MnasNet nhanh hơn gần như gấp đôi MobileNetv2 trong khi có độ chính xác cao hơn. Tuy nhiên, giống như các mô hình tìm kiếm kiến trúc thần kinh dựa trên học tập củng cố khác, thời gian tìm kiếm của MnasNet yêu cầu các nguồn tài nguyên tính toán thiên văn.

H. MobileNetv3

Trung tâm của MobileNetv3 [94] là phương pháp tương tự được sử dụng để tạo MnasNet [28] với một số sửa đổi. Một tìm kiếm kiến trúc thần kinh tự động nhận biết nền tảng được thực hiện trong không gian tìm kiếm phân cấp theo thừa số và do đó được NetAdapt [102] tối ưu hóa, giúp loại bỏ các thành phần chưa được sử dụng hết của mạng trong nhiều lần lặp lại. Sau khi nhận được một đề xuất kiến trúc, nó sẽ cắt các kênh, chạy khởi tạo các trọng số và sau đó tinh chỉnh nó để tôi chứng minh các chỉ số mục tiêu. Mô hình đã được sửa đổi thêm để loại bỏ một số lớp đắt tiền trong kiến trúc và cải thiện độ trễ bổ sung. Howard và cộng sự. lập luận rằng các bộ lọc trong kiến trúc thường là hình ảnh phản chiếu của nhau và độ chính xác có thể được duy trì ngay cả sau khi giảm một nửa số bộ lọc này. Sử dụng kỹ thuật này làm giảm các tính toán. MobileNetv3 đã sử dụng sự pha trộn giữa ReLU và hard swish làm bộ lọc kích hoạt, bộ lọc sau chủ yếu được sử dụng ở phần cuối của mô hình. Hard swish không có sự khác biệt đáng chú ý so với hàm swish nhưng rẻ hơn về mặt tính toán trong khi vẫn giữ được độ chính xác. Đối với các trường hợp sử dụng tài nguyên khác nhau, [94] đã giới thiệu hai mô hình - MobileNetv3-Large và MobileNetv3-Small. MobileNetv3-Large bao gồm 15 khối nút cổ chai trong khi MobileNetv3-Small có 11. Nó cũng bao gồm lớp ép và kích thích [56] trên các khối xây dựng của nó. Tương tự như [93], mô hình này hoạt động như một bộ phát hiện tính năng trong SSDLite và nhanh hơn 35% so với các lần lặp trước đó [28], [93], đồng thời đạt được mAP cao hơn.

I. Một lần cho Tất cả (OFA)

Việc sử dụng tìm kiếm kiến trúc thần kinh (NAS) để thiết kế kiến trúc đã tạo ra các mô hình hiện đại trong vài năm qua, tuy nhiên, chúng rất tốn kém về tính toán do việc đào tạo mô hình được lấy mẫu. Cai và cộng sự. [103] đã đề xuất một phương pháp mới để tách giai đoạn đào tạo mô hình và giai đoạn tìm kiếm kiến trúc thần kinh. Mô hình chỉ được đào tạo một lần và các mạng con có thể được chắt lọc từ nó theo yêu cầu.

Mạng một lần cho tất cả (OFA) cung cấp tính linh hoạt cho các mạng con như vậy theo bốn chiều quan trọng của mạng nơ-ron tích tụ - chiều sâu, chiều rộng, kích thước hạt nhân và thứ nguyên. Khi chúng được lồng trong mạng OFA và gây trở ngại cho việc đào tạo, việc thu nhỏ dần dần đã được giới thiệu. Đầu tiên, mạng lớn nhất được huấn luyện với tất cả các tham số được đặt ở mức tối đa.
Sau đó, mạng được tinh chỉnh bằng cách giảm dần các kích thước tham số như kích thước hạt nhân, chiều sâu và chiều rộng. Đối với nhân đàn hồi, một tâm của nhân lớn được sử dụng làm nhân nhỏ. Khi trung tâm được chia sẻ, một sự chuyển đổi hạt nhân

BẢNG III: So sánh hiệu suất của các bộ phát hiện đối tượng khác nhau trên bộ dữ liệu MS COCO và PASCAL VOC 2012 ở mức tương tự
kích thước hình ảnh đầu vào.

Mô	Năm	Xương sống	Kích thước 224	AP [0,5: 0,95]	AP0,5	FPS
hình R-	2014	AlexNet		-	58,50%	0.02
CNN * SPP-	2015	ZF-5	Biến đổi	-	59,20%	0,23
Net * R-CNN	2015	VGG-16	Biến 600	-	65,70%	0.43
nhanh * R-CNN	2016	VGG-16	600	-	67,00%	5
nhanh hơn * R-	2016	ResNet-101	800	31,50%	53,20%	3
FCN FPN Mask R-	2017	ResNet-101	800	36,20%	59,10%	5
CNN DetectoRS	2018	ResNeXt-101-FPN	1333	39,80%	62,30%	5
YOLO * SSD	2020	ResNeXt-101	448	53,30%	71,60%	4
YOLOv2 RetinaNet	2015	(Đã sửa đổi) GoogLeNet	300	-	57,90%	45
YOLOv3 CenterNet	2016	VGG-16	352	23,20%	41,20%	46
EfficientDet-D2	2016	DarkNet-19	400	21,60%	44,00%	81
YOLOv4 Swin-L	2018	ResNet-101-FPN	320	31,90%	49,50%	12
aModels được	2018	DarkNet-53	512	28,20%	51,50%	45
đánh dấu bằng *	2019	Đồng hồ cát-104	768	42,10%	61,10%	7.8
được so sánh	2020	Hiệu quả-B2	512	43,00%	62,30%	41,7
trên PASCAL VOC	2020	CSPDarkNet-53		43,00%	64,90%	31
2012, trong khi	2021	HTC ++	-	57,70%	-	-

những cái khác trên MS COCO.Rows có màu xám là máy dò thời gian thực (> 30 FPS).

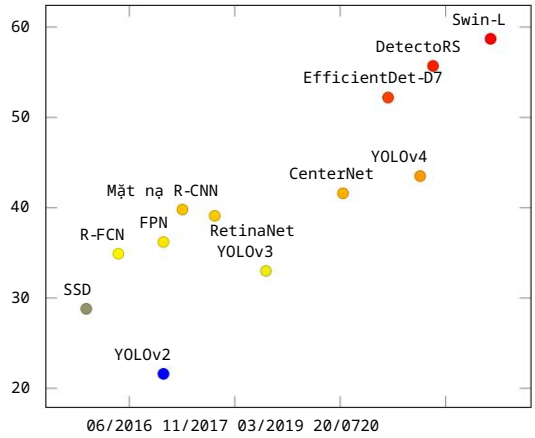
BẢNG IV: So sánh các mô hình hạng nhẹ.

Người mẫu	Top-1 của năm Acc%		Độ trễ (mili giây)	Thông số (Triệu)	FLOPs (Triệu)
SqueezeNet 2016	60,5	MobileNet	-	3,2	833
2017 70,6 Shuffle	MobileNetv2	2018	113	4,2	569
74,7 PeleeNet	ShuffleNetv2	2018	108	5,4	524
75,4 MnasNet	2018	76,7	143	6,9	300
MobileNetv3	2019	75,2 OFA	-	2,8	508
			178	7,4	597
			103	5,2	403
			58	5,4	219
			58	7,7	595

ma trận được sử dụng để duy trì hiệu suất. Để thay đổi độ sâu, đầu tiên một vài lớp được sử dụng và phần còn lại được bỏ qua từ lớp lớn mạng. Chiều rộng đàn hồi sử dụng thao tác sắp xếp kênh để tổ chức lại các kênh và sử dụng những kênh quan trọng nhất trong các mô hình nhỏ hơn. OFA đạt tỷ lệ hiện đại là 80% trong Tỷ lệ phần trăm chính xác top 1 của ImageNet và cũng giành được mức thấp thứ 4 Power Computer Vision Challenge (LPCVC) trong khi giảm nhiều thứ tự độ lớn của giờ đào tạo GPU. Nó cho thấy một mô hình mới về thiết kế các mô hình nhẹ cho nhiều loại yêu cầu phần cứng.

VII. KẾT QUẢ SO SÁ NH

Chúng tôi so sánh hiệu suất của cả giai đoạn đơn và giai đoạn hai máy dò trên PASCAL VOC 2012 [10] và Microsoft COCO [13] bộ dữ liệu. Hiệu suất của máy phát hiện đối tượng bị ảnh hưởng bởi một số yếu tố như kích thước và tỷ lệ hình ảnh đầu vào, tính năng trình giải nén, kiến trúc GPU, số lượng đề xuất, đào tạo phương pháp luận, hàm mất mát, v.v., điều này gây khó khăn để so sánh các mô hình khác nhau mà không có điểm chuẩn chung Môi trường. Ở đây trong bảng III, chúng tôi đánh giá hiệu suất của mô hình dựa trên kết quả từ các bài báo của họ. Mô hình là so với độ chính xác trung bình (AP) và khung hình đã xử lý mỗi giây (FPS) tại thời điểm suy luận. AP0,5 là mức trung bình độ chính xác của tất cả các lớp khi hộp giới hạn dự đoán có IoU> 0,5 với sự thật cơ bản. Bộ dữ liệu COCO được giới thiệu một chỉ số hiệu suất khác AP [0,5: 0,95] hoặc đơn giản là AP, là AP trung bình cho IoU từ 0,5 đến 0,95 ở kích thước bước của 0,5. Chúng tôi cố tình so sánh hiệu suất của các máy dò



Hình 10: Hiệu suất của các công cụ dò tìm đối tượng trên MS COCO tập dữ liệu.

trên hình ảnh đầu vào có kích thước tương tự, nếu có thể, để cung cấp tài khoản hợp lý, vì các tác giả thường giới thiệu một loạt các mô hình để cung cấp sự linh hoạt giữa độ chính xác và suy luận thời gian. Trong bộ lễ phục. 10, chúng tôi chỉ sử dụng mô hình hiện đại nhất từ mảng có thể có của họ mô hình máy dò đối tượng. Trọng lượng nhẹ các mô hình được so sánh trong bảng IV, nơi chúng tôi so sánh chúng về độ chính xác, độ trễ, số lượng phân loại Top-1 của ImageNet tham số và độ phức tạp trong MFLOPs. Các mô hình có MFLOPs ít hơn 600 dự kiến sẽ hoạt động đầy đủ trên thiết bị di động các thiết bị.

VIII. XU HƯỚNG TƯƠNG LAI

Tính năng phát hiện đối tượng đã có tiến bộ vượt bậc trong thời gian qua thập kỷ. Thuật toán gần như đã đạt đến cấp độ con người độ chính xác trong một số miền hẹp, tuy nhiên nó vẫn có nhiều những thách thức thú vị để giải quyết. Trong phần này, chúng tôi thảo luận về một số của vấn đề mở trong lĩnh vực phát hiện đối tượng.

AutoML: Việc sử dụng tìm kiếm kiến trúc thần kinh tự động (NAS) để xác định các đặc tính của máy dò đối tượng đã là một khu vực đang phát triển tích cực. Chúng tôi đã chỉ ra một số

Tuy nhiên, các công cụ dò tìm được thiết kế bởi NAS trong các phần trước, nó vẫn còn sơ khai. Tìm kiếm một thuật toán rất phức tạp và tốn nhiều tài nguyên.

Bộ phát hiện nhẹ: Trong khi các mạng nhẹ đã cho thấy nhiều hứa hẹn bằng cách khớp lỗi phân loại với các mô hình chính thức, chúng vẫn thiếu độ chính xác phát hiện hơn 50%. Khi ngày càng có nhiều ứng dụng học máy trên thiết bị được thêm vào thị trường, nhu cầu về các mô hình nhỏ, hiệu quả và không kém phần chính xác sẽ tăng lên.

Được giám sát yếu / phát hiện một vài cảnh quay: Hầu hết các mô hình phát hiện đối tượng hiện đại đều được đào tạo dựa trên hàng triệu dữ liệu được chú thích trong hộp giới hạn, không thể thay đổi tỷ lệ vì dữ liệu nhập chú thích yêu cầu thời gian và tài nguyên. Khả năng đào tạo trên dữ liệu được giám sát yếu, tức là dữ liệu được gắn nhãn mức hình ảnh, dẫn đến giảm đáng kể các chi phí này.

Chuyển miền: Chuyển miền đề cập đến việc sử dụng một mô hình được đào tạo dựa trên hình ảnh được gắn nhãn của một tác vụ nguồn cụ thể trên một tác vụ đích riêng biệt, nhưng có liên quan. Nó khuyến khích sử dụng lại mô hình đã được đào tạo và giảm sự phụ thuộc vào sự sẵn có của một tập dữ liệu lớn để đạt được độ chính xác cao.

Phát hiện đối tượng 3D: Phát hiện đối tượng 3D là một vấn đề đặc biệt quan trọng đối với lái xe tự hành. Mặc dù các mô hình đã đạt được độ chính xác cao, nhưng việc triển khai bất cứ thứ gì dưới hiệu suất của con người sẽ gây ra những lo ngại về an toàn.

Phát hiện đối tượng trong video: Công cụ phát hiện đối tượng được thiết kế để thực hiện trên từng hình ảnh thiếu sự tương quan giữa chúng. Sử dụng mối quan hệ không gian và thời gian giữa các khung để nhận dạng đối tượng là một vấn đề mở.

IX. PHẦN KẾT LUẬN

Mặc dù khả năng phát hiện vật thể đã trải qua một chặng đường dài trong thập kỷ qua, nhưng các thiết bị phát hiện tốt nhất vẫn chưa bão hòa về hiệu suất. 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 thiết bị di động và các hệ thống nhúng sẽ tăng lên theo cấp số nhân.

Ngày càng có nhiều mối quan tâm đến lĩnh vực này, nhưng nó vẫn còn là một thách thức còn bỏ ngỏ. Trong bài báo này, chúng tôi đã chỉ ra cách phát triển của máy dò hai giai đoạn và một giai đoạn so với các thiết bị tiền nhiệm của chúng. Mặc dù hai thiết bị phát hiện giai đoạn thường chính xác hơn, nhưng chúng chậm và không thể được sử dụng cho các ứng dụng thời gian thực như ô tô tự lái hoặc bảo mật. Tuy nhiên, điều này đã thay đổi trong vài năm qua khi máy dò một giai đoạn có độ chính xác không kém và nhanh hơn nhiều so với máy dò trước đây. Như rõ ràng trong Hình 10, Máy biến áp Swin là máy dò chính xác nhất cho đến nay. Với xu hướng tích cực hiện nay về độ chính xác của máy dò, chúng tôi có nhiều hy vọng về máy dò chính xác hơn và nhanh hơn.

NGƯỜI GIỚI THIỆU

[1] P. Viola và M. Jones, “Phát hiện đối tượng nhanh chóng bằng cách sử dụng một loạt các tính năng đơn giản được tăng cường,” trong Kỷ yếu của Hội nghị Hiệp hội Máy tính IEEE năm 2001 về Thị giác Máy tính và Nhận dạng Mẫu. CVPR 2001, tập. 1. Máy tính IEEE. Soc, 2001, trang I – 511 – I – 518. [Trực tuyến]. Có sẵn: http://ieeexplore.ieee.org/document/990517/

[2] N. Dalal và B. Triggs, “Biểu đồ của độ dốc định hướng để phát hiện con người,” trong Hội nghị của Hiệp hội Máy tính IEEE về Thị giác Máy tính và Nhận dạng Mẫu (CVPR'05), tập. 1, 2005-06, trang 886- 893 quyển. 1, ISSN: 1063-6919.

[3] A. Krizhevsky, I. Sutskever và GE Hinton, “Phân loại mạng hình ảnh với mạng nơ-ron phức hợp sâu”, trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, F. Pereira, CJC Burges, L. Bottou, và KQ Weinberger, Eds. Curran Associates, Inc.

[4] Z. Zou, Z. Shi, Y. Guo, và J. Ye, "Phát hiện vật thể trong 20 năm: Một cuộc khảo sát." [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1905.05055 [5] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu và M. Pietikainen, “Học sâu cho chung phát hiện đối tượng: Một cuộc khảo sát, ”phiên bản: 1. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1809.02165 [6] KS Chahal và K. Dey, “Một cuộc khảo sát về tài liệu phát hiện đối tượng hiện đại sử dụng học sâu”. [Trực tuyến]. Có sẵn: http://arxiv.org/ abs / 1808.07256

[7] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, và R. Qu, “Một cuộc khảo sát về phát hiện đối tượng dựa trên học sâu,” vol. 7, pp. 128 837–128 868. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1907.09408

[8] M. Everedham, L. Van Gool, CKI Williams, J. Winn và A. Zisserman, “Thử thách các lớp đối tượng hình ảnh pascal (VOC),” vol . 88, không. 2, trang 303–338. [Trực tuyến]. Có sẵn: http://link.springer.com/10.1007/ s11263-009-0275-4 [9] ——, “PASCAL Visual Object Classes Chal lenge 2007 (VOC2007)

Kết quả, ”http: //www.pascal network.org/challenges/VOC/voc2007/workshop/index.html.

[10] M. Everedham và J. Winn, “Bộ công cụ phát triển các lớp đối tượng trực quan PASCAL thách thức 2012 (VOC2012),” tr. 32.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, AC Berg, và L. Fei -Fei, “Thử thách nhận dạng hình ảnh quy mô lớn ImageNet,” vol. 115, không. 3, trang 211–252. [Trực tuyến]. Có sẵn: https://doi.org/10.1007/ s11263-015-0816-y [12] J. Deng, W. Dong, R. Socher, L. Li, Kai Li và Li Fei-Fei, “ImageNet: Một cơ sở dữ liệu hình ảnh phân cấp quy mô lớn, ”trong Hội nghị IEEE năm 2009 về Thị giác Máy tính và Nhận dạng Mẫu, trang 248–255, ISSN: 1063-6919.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar và CL Zitnick, “Microsoft coco: Các đối tượng phổ biến trong ngữ cảnh,” trong Computer Vision - ECCV 2014, D. Fleet , T. Pajdla, B. Schiele và T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, trang 740–755.

[14] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, và V Ferrari, “Tập dữ liệu hình ảnh mở v4,” vol. 128, không. 7, trang 1956– 1981. [Trực tuyến]. Có sẵn: https://doi.org/10.1007/s11263-020-01316-z [15] A. Aslam và E. Curry, “Một cuộc khảo sát về phát hiện đối tượng cho internet vạn vật đa phương tiện (iomt) sử dụng học sâu và sự kiện dựa trên phần mềm trung gian: Phương pháp tiếp cận, thách thức và hướng đi trong tương lai, ”Image and Vision Computing, vol. 106, tr. 104095, năm 2021.

[16] MD Zeiler và R. Fergus, “Hình dung và hiểu các mạng tích chập”, trong Computer Vision - ECCV 2014, ser. Ghi chú bài giảng trong Khoa học máy tính, D. Fleet, T. Pajdla, B. Schiele, và T. Tuytelaars, Eds. Nhà xuất bản Quốc tế Springer, trang 818–833.

[17] K. Simonyan và A. Zisserman, “Các mạng tích tụ rất sâu để nhận dạng hình ảnh quy mô lớn.” [Trực tuyến]. Có sẵn: http: // arxiv. org / abs / 1409.1556 [18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke và A. Rabinovich, “Tìm hiểu sâu hơn với chứng co giật . ” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1409.4842 [19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens và Z. Wojna, “Suy nghĩ lại kiến trúc khởi động cho tầm nhìn máy tính” trong IEEE 2016 Hội nghị về Thị giác Máy tính và Nhận dạng Mẫu (CVPR). IEEE, trang 2818–2826. [Trực tuyến]. Có sẵn: http: // ieeexplore. ieee.org/document/7780677/ [20] C. Szegedy, S. Ioffe, V. Vanhoucke và A. Alemi, “Inception-v4, inception-ResNet và tác động của các kết nối còn lại đối với việc học.”

[Trực tuyến]. Sẵn có: http://arxiv.org/abs/1602.07261

[21] K. He, X. Zhang, S. Ren và J. Sun, “Học sâu còn sót lại để nhận dạng hình ảnh,” trong Hội nghị IEEE 2016 về Thị giác Máy tính và Nhận dạng mẫu (CVPR), 2016, trang 770–778.

[22] K. He, X. Zhang, S. Ren, và J. Sun, “Á nh xạ danh tính trong các mạng dư sâu”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1603.05027 [23] G. Huang, Z. Liu, L. van der Maaten và KQ Weinberger, “Các mạng phức hợp được kết nối dày đặc”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1608.06993

[24] S. Xie, R. Girshick, P. Dollar, Z. Tu và K. He, “Các phép biến đổi phần dư tổng hợp cho các mạng nơron sâu”. [Trực tuyến]. Có sẵn: http://arxiv.org/ abs/1611.05431

[25] C.-Y. Wang, H.-YM Liao, I.-H. Đúng, Y.-H. Wu, P.-Y. Chen và J.-W. Hsieh, “CSPNet: Một xương sống mới có thể nâng cao khả năng học tập của CNN,” phiên bản: 1. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1911.11929

[26] C.-Y. Wang, A. Bochkovskiy và H.-YM Liao, “Scaled YOLOv4: Scaled cross-stage part network.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/2011.08036 [27] M. Tan và QV Le, “EfficientNet: Suy nghĩ lại về tỷ lệ mô hình cho các mạng nơ-ron phức hợp.” [Trực tuyến]. Có sẵn: http://arxiv.org/ abs / 1905.11946

[28] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard và QV Le, “MnasNet: Tìm kiếm kiến trúc thần kinh nhận biết nền tảng dành cho thiết bị di động.” [Trực tuyến]. Có sẵn: https://arxiv.org/abs/1807.11626v3 [29] DG Lowe, “Các tính năng hình ảnh khác biệt với các điểm chính bất biến theo tỷ lệ,” vol. 60, không. 2, trang 91–110. [Trực tuyến]. Có sẵn: https://doi.org/10.1023/B:VISI.0000029664.99615.94 [30] ——, “Nhận dạng đối tượng từ các tính năng bất biến ở quy mô cục bộ,” trong Pro ceedings của Hội nghị quốc tế IEEE lần thứ bảy về thị giác máy tính, vol . 2, trang 1150–1157 quyển 2.

[31] A. Mohan, C. Papageorgiou, và T. Poggio, “Phát hiện đối tượng dựa trên ví dụ trong ảnh theo thành phần,” vol. 23, không. 4, trang 349–361. [Trực tuyến]. Có sẵn: http://ieeexplore.ieee.org/document/917571/ [32] Yan Ke và R. Sukthankar, “PCA-SIFT: một bản gửi lại đặc biệt hơn cho các bộ mô tả hình ảnh cục bộ,” trong Kỷ yếu của Hiệp hội Máy tính IEEE năm 2004 Hội nghị về Thị giác Máy tính và Nhận dạng Mẫu, 2004. CVPR 2004., vol. 2, trang II – II, ISSN: 1063-6919.

[33] P. Felzenszwalb, D. McAllester và D. Ramanan, “Một mô hình bộ phận có thể biến dạng, được đào tạo phân biệt đối xử, đa cấp độ,” trong Hội nghị IEEE về Thị giác Máy tính và Nhận dạng Mẫu năm 2008, trang 1–8, ISSN: 1063-6919 .

[34] PF Felzenszwalb, RB Girshick, D. McAllester, và D. Ramanan, “Phát hiện đối tượng với các mô hình dựa trên một phần được đào tạo phân biệt,” vol. 32, không. 9, trang 1627–1645.

[35] PF Felzenszwalb, RB Girshick và D. McAllester, “Phát hiện đối tượng xếp tầng với các mô hình bộ phận có thể biến dạng,” trong Hội nghị IEEE Computer Society về Thị giác Máy tính và Nhận dạng Mẫu năm 2010. IEEE, trang 2241–2248. [Trực tuyến]. Có sẵn: http: //ieeexplore.ieee.org/document/5539906/

[36] R. Girshick, J. Donahue, T. Darrell và J. Malik, “Hệ thống phân cấp tính năng phong phú để phát hiện đối tượng chính xác và phân đoạn ngữ nghĩa,” trong Kỷ yếu của Hội nghị IEEE về Thị giác Máy tính và Nhận dạng Mẫu (CVPR), tháng 6 2014.

[37] JRR Uijlings, T. Gevers, và AWM Smeulders, “Tìm kiếm có chọn lọc để nhận dạng đối tượng,” tr. 18.

[38] Y. LeCun, B. Boser, JS Denker, D. Henderson, RE Howard, W. Hubbard, và LD Jackel, “Backpropagation áp dụng cho nhận dạng mã zip viết tay,” vol. 1, không. 4, trang 541–551, nhà xuất bản: MIT Press. [Trực tuyến]. Có sẵn: https://doi.org/10.1162/neco.1989.1.4.541 [39] R. Girshick, “Fast r-cnn,” trong Hội nghị Quốc tế IEEE về Thị giác Máy tính (ICCV) năm 2015, trang 1440–1448 .

[40] K. Grauman và T. Darrell, “Hạt nhân đối sánh kim tự tháp: phân loại phân biệt với các tập hợp các tính năng hình ảnh,” trong Hội nghị Quốc tế IEEE lần thứ mười về Thị giác Máy tính (ICCV'05) Tập 1, tập. 2, trang 1458– 1465 Vol. 2, ISSN: 2380-7504.

[41] K. He, X. Zhang, S. Ren, và J. Sun, “Kim tự tháp không gian tích hợp trong các mạng phức hợp sâu để nhận dạng trực quan,” vol. 37, không. 9, trang 1904–1916, Tên hội nghị: Giao dịch IEEE về Phân tích Mẫu và Trí tuệ Máy móc.

[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama và T. Darrell, “Caffe: Kiến trúc phù hợp để nhúng tính năng nhanh chóng,” trong Kỷ yếu của hội nghị quốc tế ACM lần thứ 22 về Đa phương tiện, ser. THÁ NG 14. Hiệp hội Máy tính, trang 675–678. [Trực tuyến]. Có sẵn: https: //doi.org/10.1145/2647868.2654889

[43] J. Long, E. Shelhamer và T. Darrell, “Các mạng phức hợp hoàn toàn để phân đoạn ngữ nghĩa,” tr. 10.

[44] S. Ren, K. He, R. Girshick và J. Sun, “Nhanh hơn r-cnn: Hướng tới phát hiện đối tượng trong thời gian thực với mạng đề xuất khu vực,” 2016.

[45] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan và S. Belongie, “Tính năng mạng kim tự tháp để phát hiện đối tượng,” năm 2017.

[46] S. Liu, L. Qi, H. Qin, J. Shi và J. Jia, “Mạng tổng hợp đường dẫn để phân đoạn ví dụ”. [Trực tuyến]. Có sẵn: http: //arxiv.org/abs/1803.01534 [47] G. Ghiasi, T.-Y. Lin và QV Le, “NAS-FPN: Học hỏi kiến trúc kim tự tháp tính năng có thể mở rộng để phát hiện đối tượng,” trong Hội nghị IEEE / CVF 2019 về Thị giác Máy tính và Nhận dạng Mẫu (CVPR). IEEE, trang 7029–7038. [Trực tuyến]. Có sẵn: https://ieeexplore.ieee.org/document/8954436/

[48] J. Dai, Y. Li, K. He và J. Sun, “R-fcn: Phát hiện đối tượng thông qua mạng tích hợp hoàn toàn dựa trên vùng,” 2016.

[49] A. Shrivastava, A. Gupta và R. Girshick, “Đào tạo người phát hiện đối tượng dựa trên khu vực với khai thác ví dụ trực tuyến,” 2016.

[50] K. He, G. Gkioxari, P. Dollar và R. Girshick, “Mask r-cnn,” 2018.

[51] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, CC Loy và D. Lin , “Dòng tác vụ kết hợp để phân đoạn ví dụ.” [Trực tuyến]. Có sẵn: https: //arxiv.org/abs/1901.07518v2 [52] Z. Cai và N. Vasconcelos, "Cascade r-CNN: Đi sâu vào phát hiện đối tượng chất lượng cao." [Trực tuyến]. Có sẵn: https://arxiv.org/abs/ 1712.00726v1 [53] S. Qiao, L.-C. Chen, và A. Yuille, “DetectoRS: Phát hiện các đối tượng với kim tự tháp tính năng đệ quy và tích chập bất thường có thể chuyển đổi.”

[Trực tuyến]. Có sẵn: http://arxiv.org/abs/2006.02334

[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy và AL Yuille, “DeepLab: Phân đoạn hình ảnh ngữ nghĩa với lưới chập sâu, chập chập chờn và CRF được kết nối đầy đủ.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1606.00915

[55] M. Holschneider, R. Kronland-Martinet, J. Morlet, và P. Tchamitchian, “Một thuật toán thời gian thực để phân tích tín hiệu với sự trợ giúp của phép biến đổi wavelet,” trong Wavelets, ser. các vấn đề nghịch đảo và hình ảnh lý thuyết, J.-M. Combes, A. Grossmann và P. Tchamitchian, Eds. Springer, trang 286–297.

[56] J. Hu, L. Shen, S. Albanie, G. Sun và E. Wu, “Các mạng ép và kích thích”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1709.01507 [57] J. Redmon, S. Divvala, R. Girshick và A. Farhadi, “Bạn chỉ nhìn một lần: Phát hiện đối tượng hợp nhất, thời gian thực” vào năm 2016 Hội nghị IEEE về Thị giác Máy tính và Nhận dạng Mẫu (CVPR), 2016, trang 779–788.

[58] M. Lin, Q. Chen, và S. Yan, “Mạng trong mạng”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1312.4400

[59] J. Redmon và A. Farhadi, "YOLO9000: Tốt hơn, nhanh hơn, mạnh hơn." [Trực tuyến]. Có sẵn: https://arxiv.org/abs/1612.08242v1

[60] ——, “Yolov3: Một cải tiến gia tăng,” 2018.

[61] A. Bochkovskiy, C.-Y. Wang và H.-YM Liao, “YOLOv4: Tốc độ phát hiện đối tượng và độ chính xác tối ưu.” [Trực tuyến]. Có tại: http://arxiv.org/abs/2004.10934 [62] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu và AC Berg, “SSD: Máy dò MultiBox bắn một lần” trong Computer Vision - ECCV 2016, B. Leibe, J. Matas, N. Sebe và M. Welling, Eds.

Nhà xuất bản Quốc tế Springer, trang 21–37.

[63] D. Erhan, C. Szegedy, A. Toshev và D. Anguelov, “Phát hiện đối tượng có thể mở rộng bằng cách sử dụng mạng nơ ron sâu”, 2013.

[64] J. Redmon, “Darknet: Mạng nơ-ron mã nguồn mở trong c,” http: // pjreddie.com/darknet/, 2013–2016.

[65] K. He, X. Zhang, S. Ren, và J. Sun, “Tìm hiểu sâu về bộ chỉnh lưu: Vượt qua hiệu suất cấp con người trong phân loại ImageNet,” trong Hội nghị Quốc tế IEEE về Thị giác Máy tính (ICCV) năm 2015. IEEE, trang 1026–1034. [Trực tuyến]. Có sẵn: http://ieeexplore.ieee.org/document / 7410480 /

[66] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, và K. Miller, “Giới thiệu về Mạng từ: Cơ sở dữ liệu từ vựng trực tuyến *,” vol. 3.

[67] T. Lin, P. Goyal, R. Girshick, K. He và P. Dollar, “Mất tiêu điểm để phát hiện đối tượng dày đặc,” Giao dịch IEEE về Phân tích mẫu và Trí thông minh máy, vol. 42, không. 2, trang 318–327, 2020. …

[68] X. Zhou, D. Wang, và P. Krahenb ¨ uhl, “Các đối tượng là điểm.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1904.07850 [69] A. Newell, K. Yang và J. Deng, “Các mạng đồng hồ cát xếp chồng để ước tính tư thế con người,” trong Computer Vision - ECCV 2016, ser. Ghi chú bài giảng trong Khoa học máy tính, B. Leibe, J. Matas, N. Sebe, và M. Welling, Eds. Nhà xuất bản Quốc tế Springer, trang 483–499.

[70] M. Tan, R. Pang, và QV Le, “EfficientDet: Phát hiện đối tượng có thể mở rộng và hiệu quả,” vào Hội nghị IEEE / CVF năm 2020 về Nhận dạng Mẫu và Thị giác Máy tính (CVPR). IEEE, trang 10 778–10 787. [Trực tuyến]. Có sẵn: https://ieeexplore.ieee.org/document/9156454/ [71] P. Ramachandran, B. Zoph và QV Le, “Tìm kiếm các chức năng kích hoạt”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1710.05941 [72] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye và D. Ren, “Mất khoảng cách-IoU: Nhanh hơn và tốt hơn học cho hồi quy hộp giới hạn. ” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1911.08287 [73] I. Loshchilov và F. Hutter, “SGDR: Giảm độ dốc ngẫu nhiên khi khởi động lại ấm.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1608.03983 [74] D. Misra, “Mish: Một chức năng kích hoạt không đơn điệu tự điều chỉnh.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1908.08681 [75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, L. Kaiser và I. Polosukhin, "Sự chú ý là tất cả những gì bạn cần", năm 2017.

[76] J. Devlin, M.-W. Chang, K. Lee và K. Toutanova, “Bert: Đào tạo trước về máy biến áp hai chiều sâu để hiểu ngôn ngữ,” 2019.

[77] A. Radford, K. Narasimhan, T. Salimans và I. Sutskever, “Nâng cao hiểu biết ngôn ngữ bằng cách đào tạo trước chung”, 2018.

[78] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, và PJ Liu, “Khám phá các giới hạn của việc học chuyển giao với một văn bản thống nhất -to-text biến áp, ”Tạp chí Nghiên cứu Máy học, tập. 21, không. 140, trang 1–67, 2020. [Trực tuyến].
Có sẵn: http://jmlr.org/papers/v21/20-074.html [79]
S. Khan, M. Naseer, M. Hayat, SW Zamir, FS Khan và M. Shah, “Người vận chuyển trong tầm nhìn: Một cuộc khảo sát, ”arXiv preprint arXiv: 2101.01169, 2021.

[80] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin và B. Guo, “Máy biến áp Swin: Máy biến áp thị giác phân cấp sử dụng các cửa sổ được dịch chuyển,” 2021.
[81] MN Abbas, MS Ansari, MN Asghar, N. Kanwal, T. O'Neill và B. Lee, “Mô hình học sâu nhẹ để phát hiện giả mạo hình ảnh sao chép di chuyển với các cuộc tấn công được xử lý sau”, IEEE năm 2021 Hội nghị chuyên đề thế giới lần thứ 19 về Tin học và Trí tuệ Máy ứng dụng (SAMI).
IEEE, 2021, trang 000 125–000 130.
[82] S. Karakanis và G. Leontidis, “Các mô hình học sâu nhẹ để phát hiện covid-19 từ hình ảnh X-quang ngực,” Máy tính trong Sinh học và Y học, tập. 130, tr. 104181, năm 2021.
[83] A. Jadon, A. Varshney và MS Ansari, “Mô hình học sâu hiệu suất cao có độ phức tạp thấp dành cho các hệ thống phát hiện cháy được nhúng chi phí thấp trong thời gian thực”, Khoa học Máy tính Thủ tục, vol. 171, trang 418–426, 2020.
[84] A. Jadon, M. Omama, A. Varshney, MS Ansari và R. Sharma, “Firenet: mô hình phát hiện khói & lửa hạng nhẹ chuyên dụng cho các ứng dụng iot thời gian thực”, arXiv preprint arXiv: 1905.11922, 2019.
[85] YL Cun, JS Denker, và SA Solla, Tổn thương não tối ưu.
San Francisco, CA, Hoa Kỳ: Morgan Kaufmann Publishers Inc., 1990, tr. 598–605.
[86] B. Hassibi, DG Stork, và GJ Wolff, “Bác sĩ phẫu thuật não tối ưu và cắt tỉa mạng nói chung,” trong Hội nghị quốc tế IEEE về mạng thần kinh, 1993, trang 293–299 quyển 1.
[87] S. Han, H. Mao, và WJ Dally, “Nén sâu: Nén mạng nơ-ron sâu bằng cách cắt tỉa, lượng tử hóa được đào tạo và mã hóa huffman.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1510.00149
[88] M. Courbariaux, Y. Bengio, và J.-P. David, “BinaryConnect: Đào tạo mạng nơ-ron sâu với trọng số nhị phân trong quá trình truyền bá.”
[Trực tuyến]. Có sẵn: http://arxiv.org/abs/1511.00363
[89] W. Chen, JT Wilson, S. Tyree, KQ Weinberger và Y. Chen, “Nén mạng nơ-ron bằng thủ thuật băm”. [Trực tuyến].
Có sẵn: http://arxiv.org/abs/1504.04788 [90]
G. Hinton, O. Vinyals và J. Dean, “Chắt lọc kiến thức trong mạng nơ-ron”. [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1503.02531

[91] FN Iandola, S. Han, MW Moskewicz, K. Ashraf, WJ
Dally và K. Keutzer, “SqueezeNet: Độ chính xác cấp AlexNet với tham số ít hơn 50 lần và kích thước mô hình <0,5mb.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1602.07360 [92] AG Howard, M. Zhu, B. Chen, D. Kalenichenko,
W. Wang, T. Weyand, M. Andreetto và H. Adam, “MobileNets : Mạng nơ-ron phức hợp hiệu quả cho các ứng dụng thị giác di động. ”

[Trực tuyến]. Có sẵn: http://arxiv.org/abs/1704.04861
[93] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov và L.-C. Chen, “MobileNetV2: Phần dư ngược và tắc nghẽn tuyến tính.” [Trực tuyến].
Có sẵn: http://arxiv.org/abs/1801.04381 [94]
A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, QV Le và H. Adam, “Tìm kiếm MobileNetV3.” [Trực tuyến].
Có sẵn: https://arxiv.org/abs/1905.02244v5 [95] X. Zhang, X. Zhou, M. Lin và J. Sun, “ShuffleNet: Một mạng nơ-ron phức hợp cực kỳ hiệu quả dành cho thiết bị di động” vào năm 2018 Hội nghị IEEE / CVF về Thị giác Máy tính và Nhận dạng Mẫu. IEEE, trang 6848–6856. [Trực tuyến]. Có sẵn: https: //ieeexplore.ieee.org/document/8578814/

[96] RJ Wang, X. Li, và CX Ling, “Pelee: Một hệ thống phát hiện đối tượng thời gian thực trên thiết bị di động,” tr. 10.
[97] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, và X. Xue, “DSOD: Tìm hiểu sâu về công cụ phát hiện đối tượng được giám sát từ đầu”. [Trực tuyến].
Có tại: http://arxiv.org/abs/1708.01241 [98]
N. Ma, X. Zhang, H.-T. Zheng và J. Sun, “ShuffleNet v2: Hướng dẫn thực tế để thiết kế kiến trúc CNN hiệu quả.” [Trực tuyến]. Có sẵn: http://arxiv.org/abs/1807.11164 [99] B. Zoph và QV Le, “Tìm kiếm kiến trúc thần kinh với học tăng cường”. [Trực tuyến]. Có tại: https://arxiv.org/abs/1611.01578v2 [100] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J.
Huang và K. Murphy, "Tìm kiếm kiến trúc thần kinh tiến bộ." [Trực tuyến]. Có sẵn: https://arxiv.org/abs/1712.00559v3 [101] E. Real, A. Aggarwal, Y. Huang và QV Le, “Sự phát triển quy định hóa cho tìm kiếm kiến trúc bộ phân loại hình ảnh,” vol. 33, không. 1, trang 4780–4789, số: 01. [Trực tuyến]. Có sẵn: https://ojs.aaai.org/index.php/AAAI/ article / view / 4405

[102] T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sandler, V. Sze và H. Adam, “NetAdapt: Thích ứng mạng nơ-ron nhận biết nền tảng cho các ứng dụng di động.” [Trực tuyến]. Có sẵn: http: //arxiv.org/abs/1804.03230 [103] H. Cai, C. Gan, T. Wang, Z. Zhang và S. Han, “Một lần cho tất cả: Đào tạo một mạng và chuyên môn hóa nó để triển khai hiệu quả, ”năm 2020.