

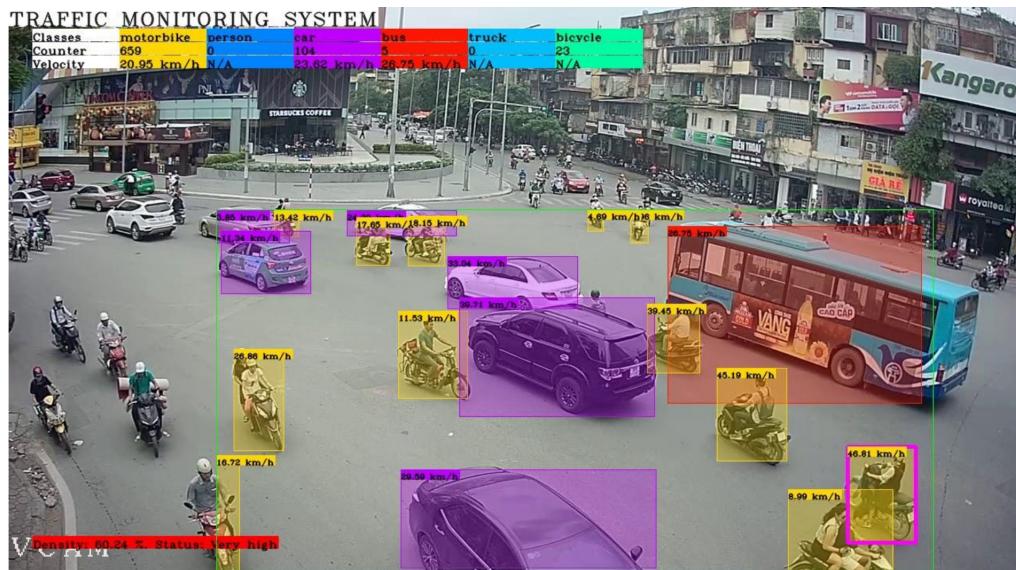
Phát hiện đối tượng - Object Detection

Người viết : Trịnh Anh Phúc

Người kiểm tra : x

1 Giới thiệu

Bài toán phát hiện đối tượng (Object Detection) trong ảnh là một trong những bài toán quan trọng trong lĩnh vực Thị Giác Máy Tính (Computer Vision) ngoài ra cũng là bài toán gần với các bài toán khác như phân đoạn ảnh (segmentation). Nó giúp cho người dùng phát hiện các đối tượng đã định nghĩa có hay không xuất hiện trong khung hình. Trong trường hợp này, có thể ảnh tĩnh hoặc ảnh động trong các đoạn CCTV giám sát giao thông hoặc tại các tụ điểm công cộng.



Hình 1: Ứng dụng bài toán phát hiện đối tượng giao thông trong khung hình của camera giám sát, ví trí ngã ba các đường Chùa Bộc - Phạm Ngọc Thạch - Tôn Thất Tùng. Các đối tượng phát hiện được đánh dấu bằng các hộp bao chử nhật (bounding box) cho phép xác định nhanh chóng vị trí của đối tượng trong khung hình, được gán nhãn (legend) tương ứng tên của đối tượng, được xác định cùng độ đo (score) trong trường hợp này là vận tốc (km/h) của đối tượng. Khung giám sát thức tế được giới hạn bởi đường kẻ màu xanh lá có kích thước nhỏ hơn khung hình theo ví dụ minh họa nằm phía dưới góc quay CCTV.

Ngoài ra còn có bài toán mang tính riêng biệt như phát hiện người, phát hiện mặt người đã có trong hầu hết ứng dụng điện thoại thông minh, phát hiện văn bản trong các hoá đơn, thư báo và còn nhiều ứng dụng khác nữa. Bài toán phát hiện đối tượng là bài toán kinh điển nên có nhiều nghiên cứu và hướng tiếp cận khác nhau để giải quyết chúng

Nội dung bài đọc sẽ được chia tuần tự thành ba phần

1. Giới thiệu một số mạng đề xuất vùng (two-stage object detectors)
 - (a) Hướng tiếp cận cửa sổ trượt (sliding window based frameworks)
 - (b) Hướng tiếp cận đề xuất vùng (region proposal based frameworks)
2. Giới thiệu một số mạng không đề xuất vùng (one-stage detectors)
 - YOLO- You Only Look Once

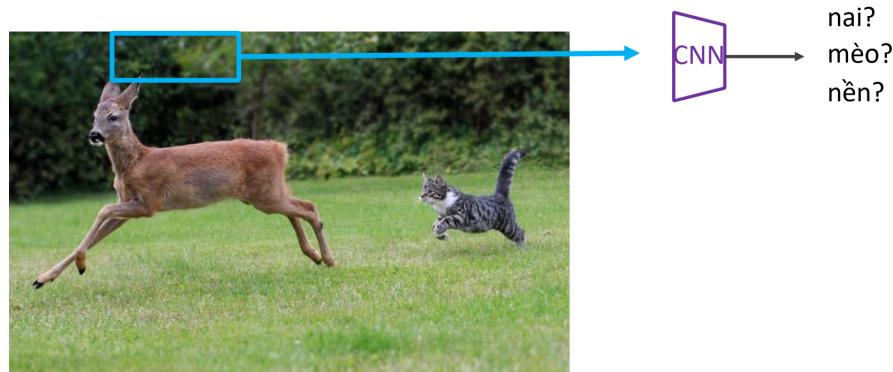
2 Giới thiệu một số mạng đề xuất vùng (two-stage object detectors)

Về cơ bản đối với bài toán phát hiện đối tượng, ta cần sớm xác định được hộp bao tại đó xuất hiện đối tượng sau đó mới phân loại để biết cụ thể tên của đối tượng. Thông thường gồm theo thứ tự liên quan đến hai bài toán con như sau

- Bài toán hồi quy để xác định hộp bao (bounding box regression) cần để xác định các biên của hình chưa nhật bao quanh đối tượng. Dĩ nhiên nó khá gần với bài toán phân đoạn ảnh, ta cần dùng phương pháp hồi quy để biên của nó áp sát đối tượng sau đó tinh chỉnh dần vị trí, kích thước của hộp bao.
- Bài toán phân loại ảnh xác định đối tượng trong hộp bao là gì ? các mô hình phân loại như Support Vector Machine (SVM), AdaBoost hoặc Deformable Part-based Model (DPM) thường là lựa chọn tốt.

2.1 Hướng tiếp cận cửa sổ trượt

Theo hướng tiếp cận truyền thống, do các đối tượng khác nhau có thể xuất hiện tại bất cứ vị trí nào trong ảnh cũng như có tỷ lệ và kích thước rất khác nhau dẫn đến hướng tiếp cận tự nhiên là phải quét toàn bộ ảnh với một cửa sổ trượt kích thước thay đổi (multi-scaled sliding window).



Hình 2: Quét cửa sổ từ trái sang phải, từ trên xuống dưới. Tại mỗi vị trí thực hiện bài toán phân loại vùng cửa sổ hiện tại thành nhiều lớp đối tượng cộng thêm lớp nền. Mạng tích chập CNN sẽ thực phân lì ngoài các đối tượng cần phát hiện cần phát hiện cả thêm nền (background) để xác định xem cửa sổ

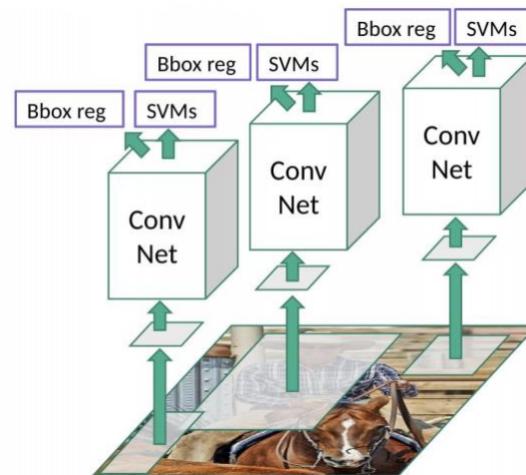
Rõ ràng đây là chiến lược tìm kiếm vét cạn dẫn đến có quá nhiều cửa sổ, chi phí tính toán đắt đỏ và cũng sinh ra rất nhiều cửa sổ dư thừa.

2.2 Hướng tiếp cận đề xuất vùng

Hướng tiếp cận đề xuất vùng cho phép giải quyết được vấn đề nêu trên, trước hết

1. Bước 1 : Ta cần tìm kiếm vùng đề xuất. Chẳng hạn dùng phương pháp tìm kiếm có lựa chọn - Selective Search¹ mà không cần thực hiện vét cạn.
2. Bước 2 : xử lý từng vùng để phân loại và hiệu chỉnh tọa độ hộp bao. Để phân loại ta có khá nhiều mô hình còn bài toán hiệu chỉnh chính là bài toán hồi quy để xác định các hộp bao

Mô hình R-CNN (Region-Based ConvNet)² là mô hình tiêu biểu cho mô hình hai giai đoạn để giải quyết bài toán phát hiện đối tượng. Đề xuất vùng tiềm năng bằng phương pháp thô như phương pháp tìm kiếm có lựa chọn (2K vùng đề xuất), sau đó dùng mạng nơ ron tích chập ConvNet để trích xuất đặc trưng của từng vùng rồi phân loại bằng SVM. Việc hiệu chỉnh vị trí biên của hộp bao sẽ dùng các phương pháp hồi quy, ở đây thường thống nhất tọa độ trên cùng, bên trái cùng kích thước rộng, dài của hình chữ nhật tương ứng. Mô hình cải tiến Faster R-CNN³ được dùng cho bài toán phát hiện đối tượng thời gian thực, theo đó tất cả 2000 vùng đề xuất sẽ được qua mạng CNN cùng một lúc. Cắt ảnh thông tin ở lớp đầu ra của CNN thay vì cắt vùng trên ảnh gốc như R-CNN. Cuối cùng, đẩy qua nhánh phân loại và nhánh hiệu chỉnh tọa độ hộp bao (bounding box regression).



Hình 3: Mô hình R-CNN dùng bài toán phát hiện đối tượng trong ảnh. Ban đầu, phương pháp thô tìm kiếm có lựa chọn sẽ đề xuất 2K vùng tiềm năng. Mạng tích chập Conv Net sẽ trích xuất các đặc trưng trên vùng đề xuất sau đó dùng SVMs để phân loại còn hiệu chỉnh là bài toán hồi quy hộp bao (Bbox reg).

¹K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers and A. W. M. Smeulders, "Segmentation as selective search for object recognition," 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 1879-1886, doi: 10.1109/ICCV.2011.6126456

²R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

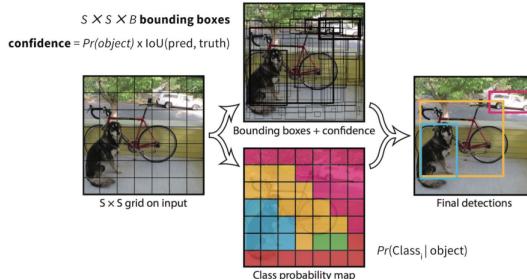
³S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards realtime object detection with region proposal networks," in NIPS, 2015, pp. 91-99.

3 Giới thiệu một số mạng không đề xuất vùng (one-stage detectors)

Đặc điểm của mạng một giai đoạn là thường có tốc độ phát hiện nhanh hơn so với mạng hai giai đoạn tuy nhiên độ chính xác là không tốt bằng. Về cơ bản thay vì đề xuất vùng hay dùng của sổ trượt, mạng dùng luôn bộ lọc tích chập hộp bao với cơ chế trượt đều của bộ lọc này làm tác tử phát hiện đối tượng trong ảnh. Lưới các hộp sẽ xuất hiện dày đặc sau đó sẽ phân loại và hiệu chỉnh nếu chúng chưa đối tượng cần phát hiện đều bằng mạng tích chập.

3.1 YOLO

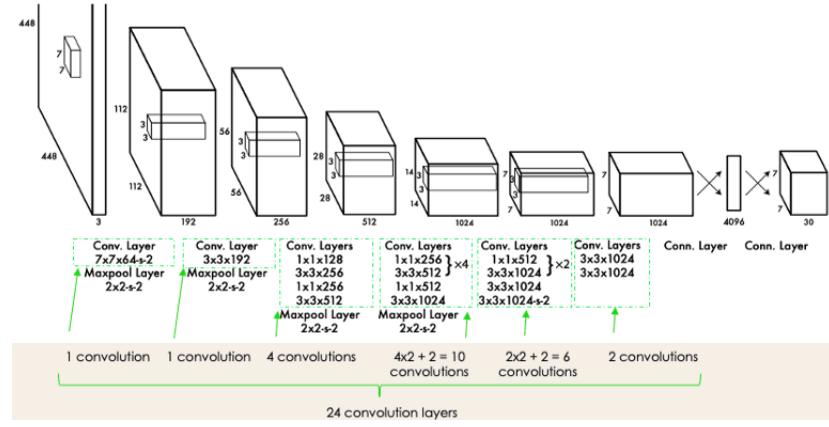
Mạng YOLO - You Only Look Once là mạng nơ ron tiên phong trong hướng tiếp cận này⁴, về cơ bản nó dùng một mạng nơ ron duy nhất để huấn luyện và dự đoán (phát hiện) nên có tốc độ nhanh hơn so với mô hình hai giai đoạn



Hình 4: Mô hình YOLO phát hiện đối tượng dưới dạng bài toán hồi quy. Nó chia hình ảnh thành lưới kích thước $S \times S$ và với mỗi ô sẽ dự đoán B hộp bao, độ tin tưởng mỗi hộp bao, và C lớp dự đoán. Các thành phần dự đoán sẽ là ma trận tích chập $S \times S \times (B * 5 + C)$. Trong bài báo $S = 7, B = 2, C = 20$ đối với bộ dữ liệu VOC PASCAL nên ma trận tích chập dự đoán kích thước $7 \times 7 \times 30$

Kiến trúc mạng nơ ron của YOLO được xác định là giống mạng GooleNet dùng làm "khung xương" backbone cho phép trích xuất thông tin cũng như thực hiện quá trình phát hiện, kích thước hình ảnh đầu vào lớn hơn 448×448 có tầng tích chập 7×7 đầu vào. Như trong chú thích minh họa mô hình YOLO, đầu ra ma trận tích chập có kích thước $7 \times 7 \times 30$

⁴J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91



Hình 5: Mô hình YOLO phát hiện đối tượng một giai đoạn dùng mạng nơ ron dự đoán duy nhất không dùng thêm giai đoạn đề xuất vùng làm giảm thời gian phát hiện xuống, tăng tốc 45 khung hình/giây và với mô hình YoLo sau đó là 155 khung hình/giây dùng để phát hiện thời gian thực.

Hàm mất mát được tính khá phức tạp bao gồm nhiều chỉ số cả phân loại lẫn hồi quy cộng thêm độ tin tưởng, xem minh họa Hình 6

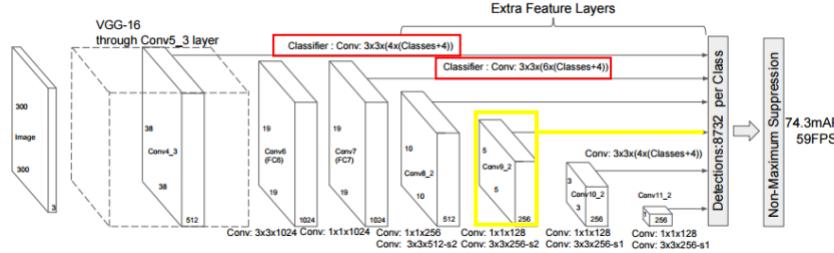
$$\begin{aligned}
 & \text{1 when there is object, 0 when there is no object} \\
 & \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \quad \text{Bounding Box Location (x, y) when there is object} \\
 & + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \quad \text{Bounding Box size (w, h) when there is object} \\
 & + \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 \quad \text{Confidence when there is object} \\
 & + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \quad \text{1 when there is no object, 0 when there is object} \\
 & + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{noobj}} \left(\sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \right) \quad \text{Confidence when there is no object} \\
 & + \sum_{i=0}^{S^2} \mathbb{1}_i^{\text{obj}} \left(\sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \right) \quad \text{Class probabilities when there is object}
 \end{aligned}$$

Hình 6: Trong công thức hàm lỗi, hai thành phần đầu là bài toán hồi quy (x_i, y_i) là toạ độ trái trên cùng của hộp bao (w_i, h_i) là chiều cao và rộng, đều dùng hàm bình phương lỗi. Các công thức sau liên quan đến bài toán phân loại để xác định ô tương ứng thuộc lớp dự đoán nào và giá trị độ tin tưởng cũng dùng bình phương lỗi. Hai siêu tham số $\lambda_{\text{coord}} = 5$ và $\lambda_{\text{noobj}} = 0.5$ để xác định trọng số cho hai khối hồi quy và phân loại.

Đây là mô hình phát hiện đối tượng được dùng phổ biến nhất YOLO đã đến Version 8 (cập nhật 9/2024), đã có bản thương mại hóa <https://yolov8.com/>

3.2 SSD : Single Shot Detector

Tương tự YOLO nhưng lưỡi hộp bao dày đặc hơn, có nhiều lưỡi với các kích thước hộp khác nhau. Kiến trúc mạng SSD là mạng VGG-16 dùng làm "khung xương" có dùng tăng cường dữ liệu cùng các mẫu âm. Khác với YOLO kiến trúc SSD có gắn thêm các tầng tích chập sau VGG-16. Cho phép phát hiện đối tượng ở nhiều mức khác nhau trong mạng.



Hình 7: Mô hình SSD phát hiện đối tượng gắn thêm các tầng tích chập phía sau VGG-16 cho phép phát hiện đối tượng nhiều độ dung sai khác nhau. Mô hình có tốc độ đo được 59 khung hình/giây tương đương YOLO v1

4 Tổng kết

Bảng sau dùng để tóm tắt đặc điểm của hai hướng tiếp cận

Mạng một giai đoạn	Mạng hai giai đoạn
Nhanh và đơn giản hơn	Độ chính xác cao hơn
YOLO versions 1-8	R-CNN
SSD DSSD	Faster R-CNN
Squizze RetinaNet	Feature Pyramid Network (FPN)
CornerNet CenterNet	Mask R-CNN
EfficientNet	

Bảng 1: Bảng tổng kết dung để so sánh hai hướng tiếp cận của ứng dụng phát hiện đối tượng.

Về cơ bản bài toán dựa trên các mạng nơ ron tích chập phân loại phổ biến cho phép ta trích xuất tự động các đặc trưng các đối tượng trong ảnh như GoogleNet hay VGG-16, sau đó thêm các lớp tích chập và hàm mục tiêu cũng như kết quả dự đoán tích chập đầu ra. Đối với hướng tiếp cận hai giai đoạn, bản chất giai đoạn đầu, là phương pháp tích xuất vùng để giảm thiểu chi phí, sau đó mới là mạng nơ ron tích chập để thực hiện bài toán hồi quy và phân loại. Trong các bài toán ứng dụng trong lĩnh vực thị giác máy tính, bài toán phát hiện đối tượng áp dụng tại nhiều nơi kể cả lĩnh vực quân sự, có tính ứng dụng rất cao.