Trần Minh Hoàn   
MSSV: HE170962

YOLO(Individual-Options)

**1. Introduction**

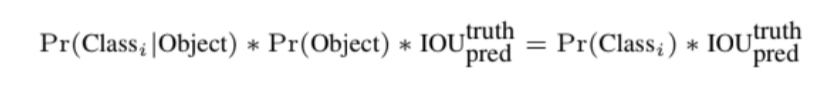
YOLO, một phương pháp mới trong phát hiện đối tượng. Khác với các hệ thống hiện tại, YOLO đưa ra khái niệm phát hiện đối tượng như một vấn đề hồi quy, giúp tạo ra các hộp giới hạn và xác suất lớp trực tiếp từ ảnh gốc mà không cần các bước trung gian. Mô hình YOLO làm việc nhanh chóng và chính xác, có thể được tối ưu hóa end-to-end trực tiếp trên hiệu suất phát hiện. So với các phương pháp khác, YOLO chạy rất nhanh và học được biểu diễn tổng quát về đối tượng. Mặc dù có nhược điểm về địa lý hóa chính xác, YOLO vẫn là một tiến bộ đáng kể trong lĩnh vực phát hiện đối tượng với tiềm năng ứng dụng rộng rãi trong lái xe tự động, thiết bị hỗ trợ, và hệ thống robot đa dụng.  
  
**2. Unified Detection**

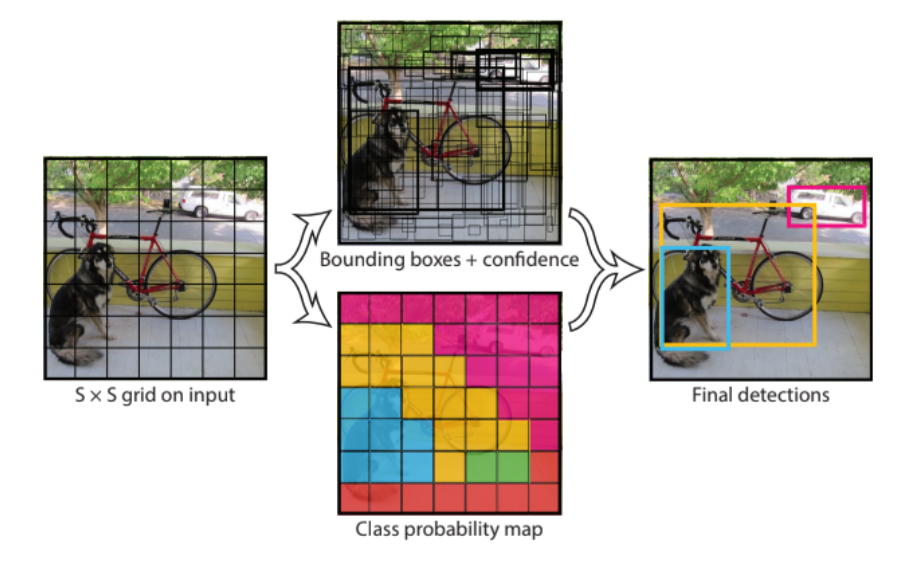
YOLO sử dụng single neural network, mô hình sẽ quét qua toàn bộ ảnh và dự đoán các bounding box cho tất cả các đối tượng xuất hiện trong ảnh.Thiết kế YOLO cho phép huấn luyện từ đầu đến cuối và tốc độ thời gian thực trong khi duy trì độ chính xác trung bình cao.

Hệ thống chia ảnh đầu vào thành lưới S × S.Nếu trung tâm của một đối tượng nằm trong ô lưới, ô lưới đó chịu trách nhiệm phát hiện đối tượng đó. Mỗi ô lưới dự đoán B hộp giới hạn và điểm tin cậy cho những hộp đó. Những điểm tin cậy này phản ánh mức độ tự tin của mô hình rằng hộp đó chứa một đối tượng và cũng mức độ chính xác mà nó nghĩ rằng hộp đó dự đoán. Điểm tin cậy (confidence scores) được định nghĩa : Pr(Object) ∗ IOU(truth/pred). Nếu không đối tượng tồn tại trong ô đó, điểm tin cậy sẽ là số không. Ngược lại , điểm tin cậy bằng giao điểm trên liên kết (IOU) giữa hộp dự đoán và hộp thực tế.

Mỗi hộp giới hạn bao gồm 5 dự đoán: x, y, w, h và điểm tin cậy. Các tọa độ (x, y) đại diện cho trung tâm của hộp liên quan đến biên của ô lưới. Chiều rộng và chiều cao được dự đoán liên quan đến toàn bộ ảnh.

Ở thời gian kiểm tra, xác suất lớp có điều kiện và các dự đoán điểm tin cậy của từng hộp được nhân tích hợp, điều này mang lại các điểm tin cậy cụ thể cho từng hộp theo từng lớp. Những điểm này mã hóa cả xác suất của lớp đó xuất hiện trong hộp và cách mà hộp dự đoán phù hợp với đối tượng.





Những dự đoán này được mã hóa dưới dạng tensor S × S × (B ∗ 5 + C).Đối với việc đánh giá YOLO trên PASCAL VOC, chúng tôi sử dụng S = 7, B = 2. PASCAL VOC có 20 lớp được gán nhãn nên C = 20. Dự đoán cuối cùng của chúng tôi là một tensor 7 × 7 × 30.

**2.1. Network Design**Mô hình YOLO triển khai dưới dạng mạng nơ-ron tích chập và kiểm tra nó trên bộ dữ liệu phát hiện PASCAL VOC. Mạng lấy cảm hứng từ GoogLeNet với 24 lớp tích chập và 2 lớp kết nối đầy đủ. Thay vì sử dụng mô-đun inception, lớp giảm kích thước 1 × 1 và lớp tích chập 3 × 3 3 được . 1 phiên bản nhanh của YOLO huấn luyện phiên bản nhanh của YOLO, gọi là Fast YOLO, với ít lớp tích chập hơn và ít bộ lọc. Cả hai phiên bản đều chia sẻ cùng một kiến trúc và tất cả các tham số huấn luyện và kiểm tra giống nhau. Đầu ra của mạng là tensor 7 × 7 × 30 dự đoán.

**2.2. Training**

YOLO tiền huấn luyện mạng trên bộ dữ liệu ImageNet với 1000 lớp, sau đó chuyển đổi mô hình để thực hiện phát hiện đối tượng. Mạng dựa trên GoogLeNet với 24 lớp tích chập và 2 lớp kết nối đầy đủ. Sau tiền huấn luyện, thêm các lớp tích chập và kết nối đầy đủ để cải thiện hiệu suất phát hiện. Độ phân giải đầu vào tăng từ 224 × 224 lên 448 × 448 để cung cấp thông tin chi tiết hơn.

Lớp cuối cùng của mô hình dự đoán xác suất lớp và tọa độ hộp giới hạn. Chúng tôi tối ưu hóa bằng tổng bình phương lỗi và điều chỉnh mất từ dự đoán tọa độ hộp và "confidence" cho các hộp không chứa đối tượng. Để giảm chệch lớn giữa các hộp lớn và nhỏ, chúng tôi dự đoán căn bậc hai của chiều rộng và chiều cao của hộp giới hạn.

Trong quá trình huấn luyện, chúng tôi tối ưu hóa các thành phần của hàm mất mát để chỉ áp dụng lỗi phân loại và tọa độ hộp cho các ô lưới chứa đối tượng. Chúng tôi cũng sử dụng các biện pháp tránh quá mức và tăng cường dữ liệu để tránh việc quá mức.

Chúng tôi huấn luyện mạng trên bộ dữ liệu PASCAL VOC 2007 và 2012 trong khoảng 135 epochs với batch size là 64, sử dụng hằng số học làm tăng từ 10^(-3) lên 10^(-2) trong các epochs đầu, sau đó giảm dần về 10^(-3) và 10^(-4). Để tránh quá mức, chúng tôi áp dụng dropout và tăng cường dữ liệu như thay đổi ngẫu nhiên kích thước và vị trí ảnh, điều chỉnh sáng tối và bão hòa màu sắc.

**2.4. Limitations of YOLO**

YOLO áp đặt ràng buộc không gian mạnh mẽ trên dự đoán hộp giới hạn vì mỗi ô lưới chỉ dự đoán hai hộp và một lớp. Ràng buộc này hạn chế khả năng dự đoán của mô hình đối với số lượng đối tượng gần nhau. Mô hình gặp khó khăn khi xử lý đối tượng nhỏ xuất hiện theo nhóm và khó tổng quát hóa đối với các đối tượng có tỷ lệ hoặc cấu hình mới. Sử dụng đặc trưng tương đối thô và xử lý lỗi giống nhau trong các hộp giới hạn nhỏ và lớn cũng là nguồn chính của sai sót của mô hình, chủ yếu là trong việc định vị không chính xác.

**3. Comparison to Other Detection Systems**

Nhận diện đối tượng là một thách thức cơ bản trong thị giác máy tính, thường bao gồm việc trích xuất đặc trưng và sau đó là phân loại hoặc địa chỉ hóa. YOLO, hoặc "You Only Look Once," nổi bật bằng cách sử dụng một mạng nơ-ron tích chập duy nhất để đồng thời thực hiện trích xuất đặc trưng, dự đoán hộp giới hạn, kiểm soát suy giảm cực đại và suy luận ngữ cảnh. Kiến trúc thống nhất này dẫn đến kết quả nhanh chóng và chính xác hơn so với các phương pháp truyền thống.

So sánh với các khung công việc hiện tại làm nổi bật những ưu điểm của YOLO. Mô hình Deformable Parts (DPM) và R-CNN sử dụng các đường ống phức tạp, trong khi mô hình YOLO với mạng nơ-ron duy nhất giản đơn quá trình, đạt được tốc độ và chính xác tốt hơn. YOLO có điểm chung với R-CNN nhưng áp dụng ràng buộc không gian cho các đề xuất ô lưới, giảm thiểu việc dự đoán nhiều và đề xuất ít hộp giới hạn hơn. Các mô hình nhanh khác, như Fast R-CNN, Faster R-CNN và OverFeat, tập trung vào việc tối ưu hóa các thành phần cụ thể nhưng không đạt được hiệu suất thời gian thực.

YOLO nổi bật với thiết kế chuyên cho tốc độ và sự đa dạng, có khả năng phát hiện đồng thời nhiều đối tượng khác nhau. Nó khác biệt với các phương pháp như MultiBox, Deep MultiBox và MultiGrasp, cung cấp một hệ thống phát hiện đầy đủ và hiệu quả. Trong khi lấy cảm hứng từ việc nhận diện cầm tay, nhiệm vụ của YOLO là dự đoán hộp giới hạn và xác suất lớp cho nhiều đối tượng thuộc nhiều lớp khác nhau trong một hình ảnh, là một nhiệm vụ phức tạp và linh hoạt hơn.

**4. Experiments**

Đầu tiên, so sánh YOLO với các hệ thống nhận diện thời gian thực khác trên PASCAL VOC 2007. Để hiểu sự khác biệt giữa YOLO và các biến thể R-CNN, chúng ta khám phá các lỗi trên VOC 2007 được YOLO và Fast R-CNN thực hiện, một trong những phiên bản R-CNN hiệu suất cao nhất. Dựa trên các hồ sơ lỗi khác nhau, ta chỉ ra rằng YOLO có thể được sử dụng để đánh điểm lại các phát hiện Fast R-CNN và giảm lỗi từ các dương tính giả mạo nền, mang lại một đòn bẩy hiệu suất đáng kể. Chúng ta cũng trình bày kết quả VOC 2012 và so sánh mAP với các phương pháp hiện đại nhất. Cuối cùng, YOLO được chỉ ra rằng tổng quát hóa tốt hơn vào các lĩnh vực mới hơn so với các bộ dò khác trên hai bộ dữ liệu nghệ thuật.

**5. Real-Time Detection In The Wild**

YOLO là một bộ phát hiện đối tượng nhanh chóng và chính xác, làm cho nó trở thành lựa chọn lý tưởng cho các ứng dụng thị giác máy tính. Chúng ta kết nối YOLO với một webcam và xác nhận rằng nó duy trì hiệu suất thời gian thực, bao gồm cả thời gian để lấy ảnh từ máy ảnh và hiển thị các phát hiện.

Hệ thống kết quả là tương tác và hấp dẫn. Trong khi YOLO xử lý ảnh cá nhân, khi kết nối với một webcam, nó hoạt động như một hệ thống theo dõi, phát hiện các đối tượng khi chúng di chuyển và thay đổi về hình dạng

**6. Conclusion**

Chúng tôi giới thiệu YOLO, một mô hình thống nhất cho việc phát hiện đối tượng. Mô hình của chúng tôi dễ dàng xây dựng và có thể được huấn luyện trực tiếp trên toàn bộ hình ảnh. Khác với các phương pháp dựa trên bộ phân loại, YOLO được huấn luyện trên một hàm mất mát tương ứng trực tiếp với hiệu suất phát hiện và toàn bộ mô hình được huấn luyện cùng một lúc.

Fast YOLO là bộ phát hiện đối tượng đa dụng nhanh chóng nhất trong văn học và YOLO đẩy ranh giới hiện đại trong việc phát hiện đối tượng thời gian thực. YOLO cũng tổng quát tốt cho các lĩnh vực mới, làm cho nó lý tưởng cho các ứng dụng dựa trên việc phát hiện đối tượng nhanh chóng và mạnh mẽ.