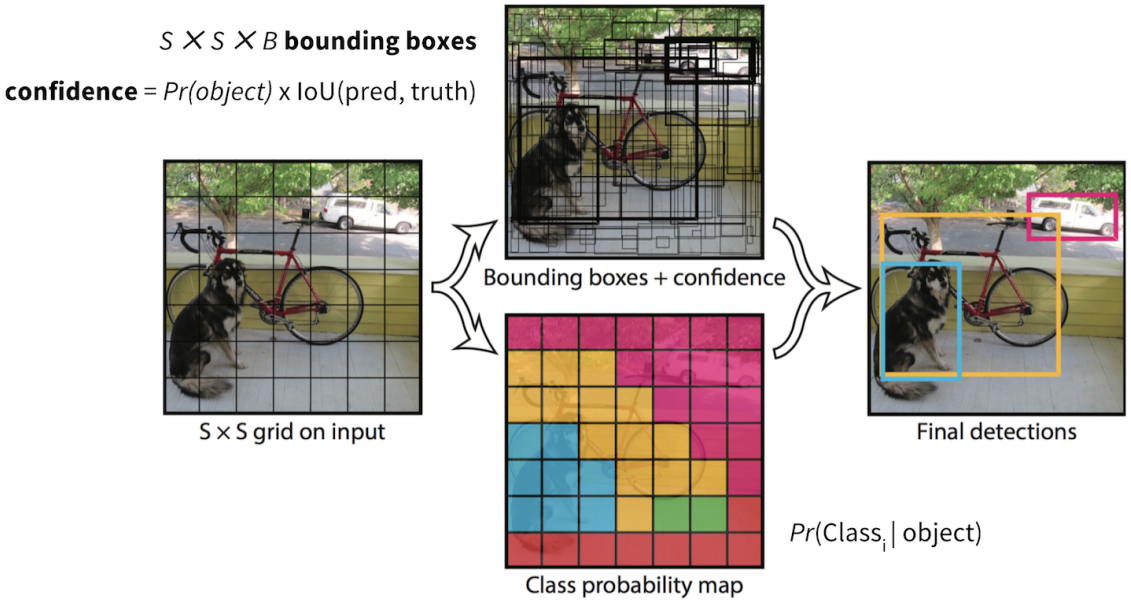
YOLOv1

Paper: https://arxiv.org/pdf/1506.02640.pdf

1. Thuật toán YOLO



   a. S x S ô vuông
      - Mỗi hình ảnh sẽ được chia thành (SxS) các ô vuông. Ở YOLOv1 tác giả cho S = 7. Mỗi ô vuông phát hiện 1 vật thể dựa vào chính điểm tâm của nó. Nếu tâm của object thuộc ô vuông nào thì nó sẽ chịu trách nhiệm phát hiện vật thể đó.
   b. Bounding box
      - Bounding box là khung hình bao quanh vật thể. Một bounding box sẽ được hình thành từ tọa độ tâm (center x, center y) và kích thước box (width, height).
      - (x,y) là tọa độ tâm của bounding box so với ô vuông và được chuẩn hóa để có giá trị từ 0 đến 1.
      - (w, h) là width và height của bounding box so với width và height của toàn bộ ảnh.
      - Đối với mỗi ô vuông sẽ dự đoán B bounding box (ở YOLOv1 thì B=2) và xác suất các classe với điều kiện tồn tại object trong ô vuông đó Pr(class i | obj). Ở đây tác giả sử dụng bộ dữ liệu PASCAL VOC có số class C=20.

2. Kiến trúc mạng YOLO

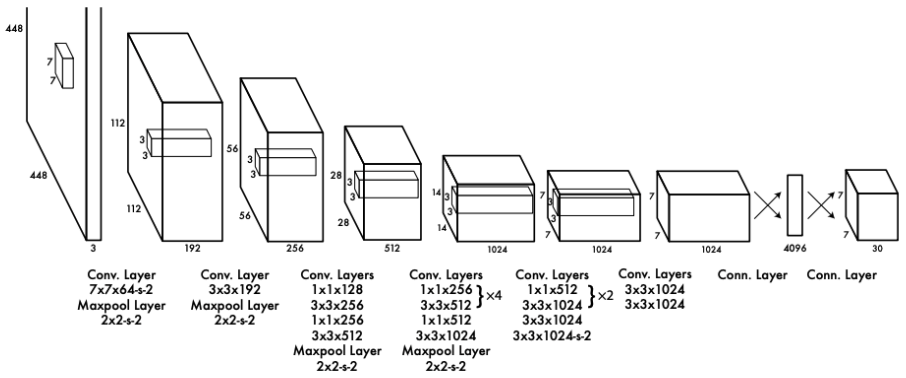


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

- Kiến trúc mạng bao gồm 24 lớp Convolutional và 2 lớp Fully Connected.
- Đầu vào của mạng là những vector 3 x 448 x 448 đại diện cho một ảnh RGB.
- Đầu ra của mạng là những vector 7 x 7 x 30, trong đó:
    - 7x7: tác giả chia ảnh thành 7x7 ô vuông, và cho phép 2 bounding box trên mỗi ô vuông.
    - 30:
        - [0:20]: xác suất dự đoán vật thể xuất hiện trong bounding box. (do có 20 class để phân loại)
        - [20]: độ confidence của box1
        - [21:25]: tọa độ của box1
        - [25]: độ confidence của box2
        - [26:30]: tọa độ của box2

3. Độ Confidence

Độ Confidence có thể hiểu là khả năng mà một vật thể nào đó có thể có trong một bounding box.

$$Pr(Object) * IOU^{truth}_{pred},$$

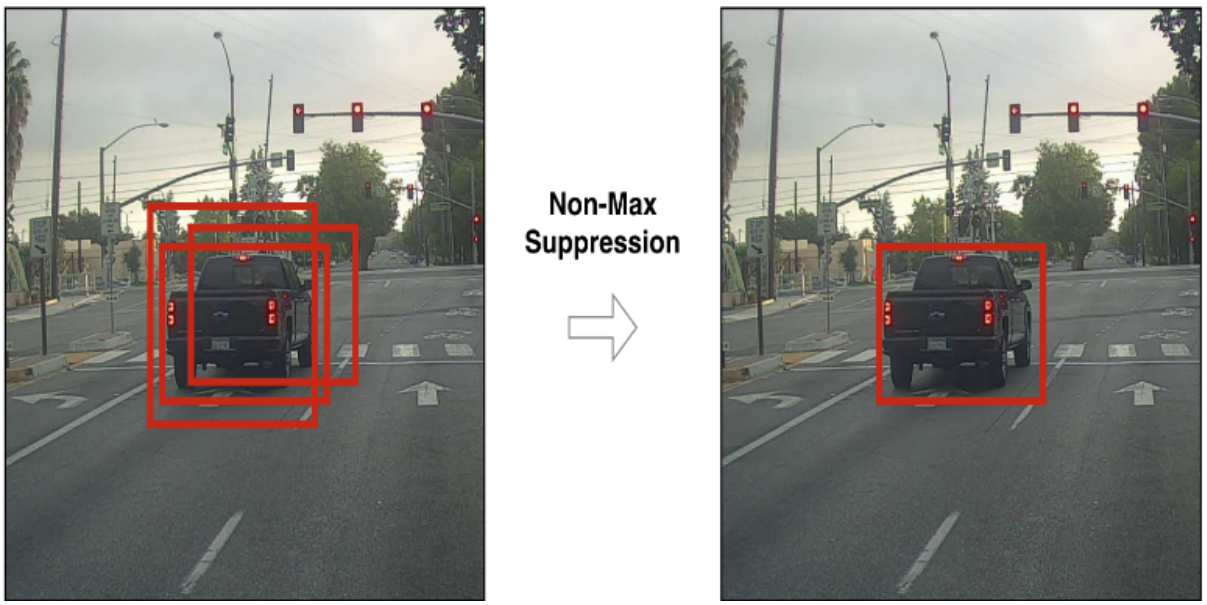
- Pr(Object) biểu thị xác suất có vật thể
- biểu thị IoU giữa box dự đoán và box chính xác.

4. Intersection Over Union (IOU)

IOU đo lượng sự chênh lệch giữa bounding box dự doán và box chính xác. IOU được tính bằng thương của vùng giao nhau và vùng liên hợp của 2 box này.

5. Non Max Suppression

Non-max suppression được dùng để giảm bớt các bounding box quá gần nhau, dễ bị chồng lên nhau.



- Bước 1: Giảm bớt các bounding box bằng cách lọc bỏ toàn bộ những bounding box có xác suất chứa vật thể nhỏ hơn một ngưỡng threshold (thường là 0.5).
- Bước 2: Đối với các bouding box giao nhau, non-max suppression sẽ lựa chọn ra một bounding box có xác suất chứa vật thể là lớn nhất. Sau đó tính toán chỉ số giao thoa IoU với các bounding box còn lại. Nếu chỉ số này lớn hơn ngưỡng threshold thì điều đó chứng tỏ 2 bounding boxes đang overlap nhau rất cao. Ta sẽ xóa các bounding có có xác suất thấp hơn và giữ lại bounding box có xác suất cao nhất.

6. Forward Pass Prediction
   Sau khi đi qua các lớp trong mạng, các vector ảnh trở thành output 7x7x30 có chứa thông tin của 2 box đã được dự đoán. Đầu tiên, tất cả các bounding box có độ confidence < ngưỡng sẽ bị loại bỏ.
   Sau đó, non-max suppression được thực hiện
   Cuối cùng, chọn ra box có độ confidence cao hơn.
7. Training: hàm Loss

$$\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \qquad (3)$$

a. Coordinate Loss

$$\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

Hàm loss tọa độ được tính khi so sánh tọa độ của bounding box dự đoán và bounding box chính xác. Tham số λ_{coord} làm cho độ loss không quá nhỏ, ở YOLOv1 tác giả để bằng 5.

1_{ij}^{obj} : Hàm indicator có giá trị 0,1 nhằm xác định xem ô vuông ii có chứa vật thể hay không. Bằng 1 nếu chứa vật thể và 0 nếu ngược lại.
Với các tọa độ tâm (x, y), ta tìm Mean Squared Error (MSE).
Với tọa độ kích thước (w, h), ta tìm Mean Squared Error (MSE) của căn bậc hai của (w, h).

b. Confidence Loss

$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{\text{obj}}\left(C_i-\hat{C}_i\right)^2$$

$$+\lambda_{\text{noobj}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbb{1}_{ij}^{\text{noobj}}\left(C_i-\hat{C}_i\right)^2$$

Confidence Loss tính khá giống với Coordinate Loss, nhưng có điểm khác sau:

- Nếu có vật thể trong ô vuông, Confidence Loss tương ứng với công thức bên trên. $C_i = 1$
- Nếu không có vật thể thì Confidence Loss tương ứng với công thức bên dưới, $C_i = 0$. Tham số λ_{noobj} được lấy bằng 0.5 để làm cho Confidence Loss bé hơn.

c. Class Loss

$$+\sum_{i=0}^{S^2}\mathbb{1}_{i}^{\text{obj}}\sum_{c\in\text{classes}}\left(p_i(c)-\hat{p}_i(c)\right)^2$$

Đối với mỗi lớp, tính MSE giữa xác suất lớp dự đoán và xác suất lớp chính xác (sử dụng one-hot vector).