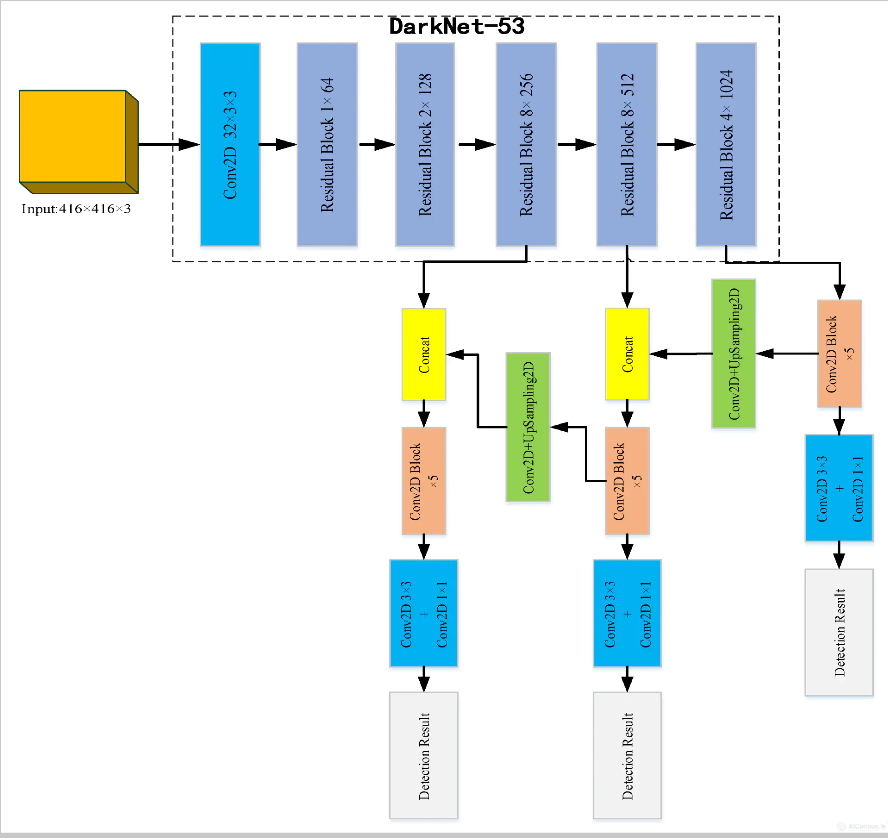
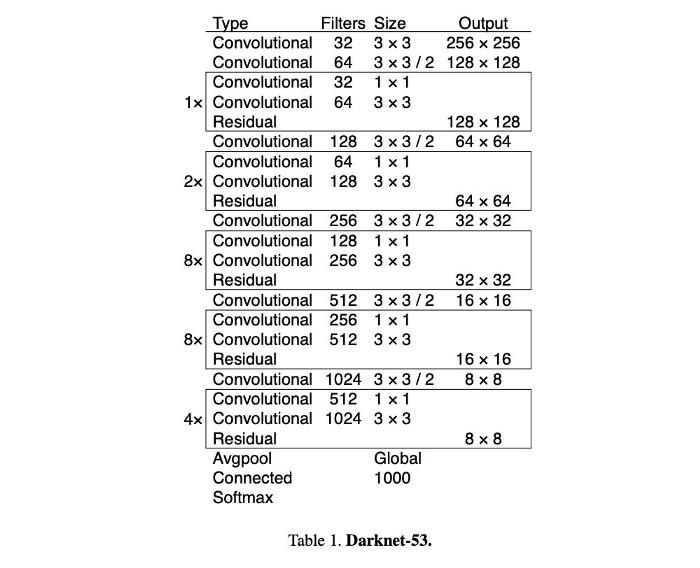
1.Kiến trúc

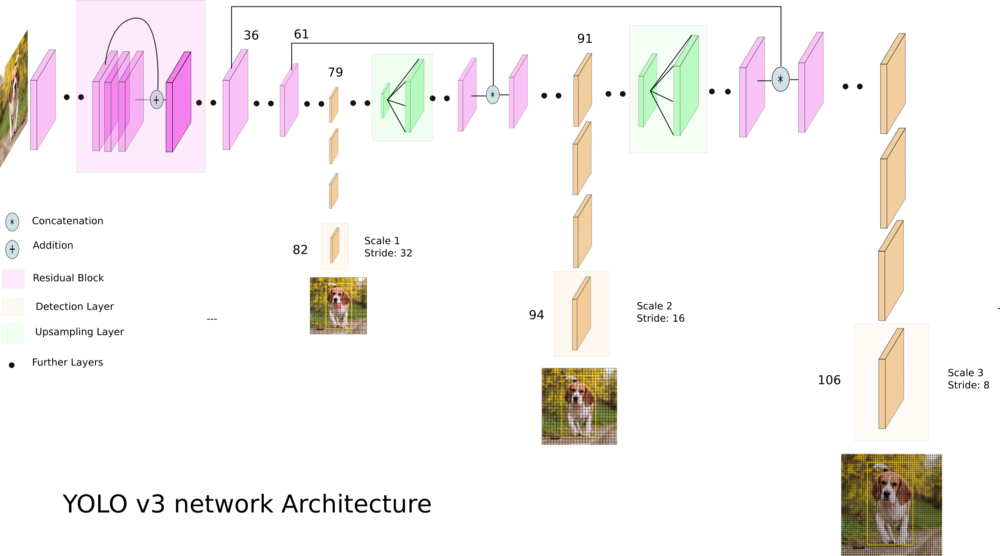


* **Logistic regression cho confidence score:** predict độ tự tin của bounding box (có chứa vật hay không) sử dụng logistic regression
* **Thay softmax bằng các logistic classifier rời rạc:** sử dụng các logistic classifier thay vì softmax cho việc classify đối tượng. Việc này cho hiệu quả tốt hơn nếu các label không "mutually exclusive", tức là có thể có đối tượng cùng thuộc 2 hay nhiều class khác nhau. Ví dụ với bài toán cần phát hiện 2 đối tượng là "chó" và "chó Phú Quốc". Rõ ràng nếu đối tượng là "chó Phú Quốc" thì nó cũng thuộc class "chó", và việc sử dụng softmax là không phù hợp trong trường hợp này.
* **Backbone mới - Darknet-53:** Backbone được thiết kế lại với việc thêm các residual blocks (kiến trúc sử dụng trong ResNet).
* **Multi-scale prediction:** sử dụng kiến trúc Feature Pyramid Networks (FPN) để đưa ra các dự đoán từ nhiều scale khác nhau của feature map. Việc này giúp tận dụng các feature map với độ thô - tinh khác nhau cho việc dự đoán.
* **Skip-layer concatenation:** cũng thêm các liên kết giữa các lớp dự đoán. Mô hình upsample các lớp dự đoán ở các tầng sau và sau đó concatenate với các lớp dự đoán ở các tầng trước đó. Phương pháp này giúp tăng độ chính xác khi predict các object nhỏ.

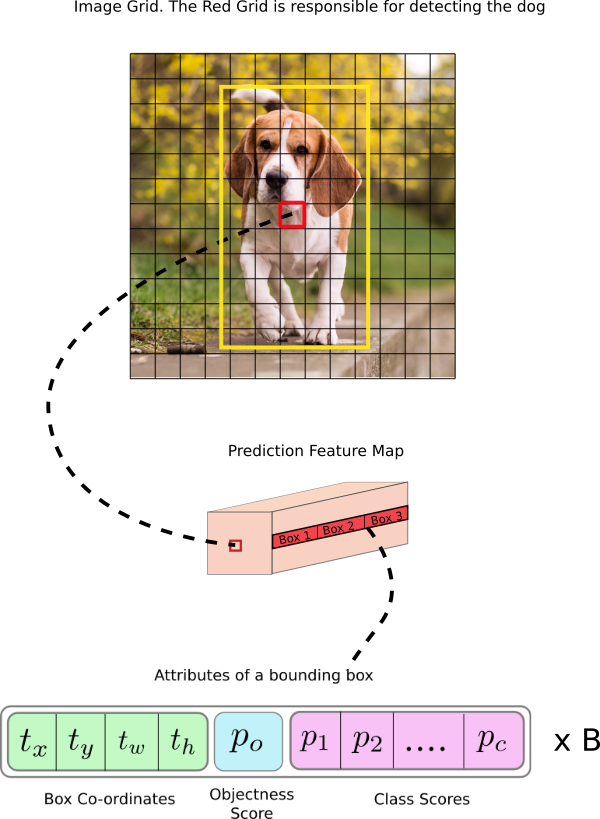


YOLO v3 sử dụng một biến thể của Darknet thêm các residual blocks (kiến trúc sử dụng trong ResNet), ban đầu có mạng 53 lớp được đào tạo trên Imagenet. Để có nhiệm vụ phát hiện, 53 lớp nữa được xếp chồng lên nhau, cho chúng ta một kiến trúc gồm 106 lớp convolutional cho YOLO v3. Đây là lý do đằng sau sự chậm chạp của YOLO v3 so với YOLO v2.

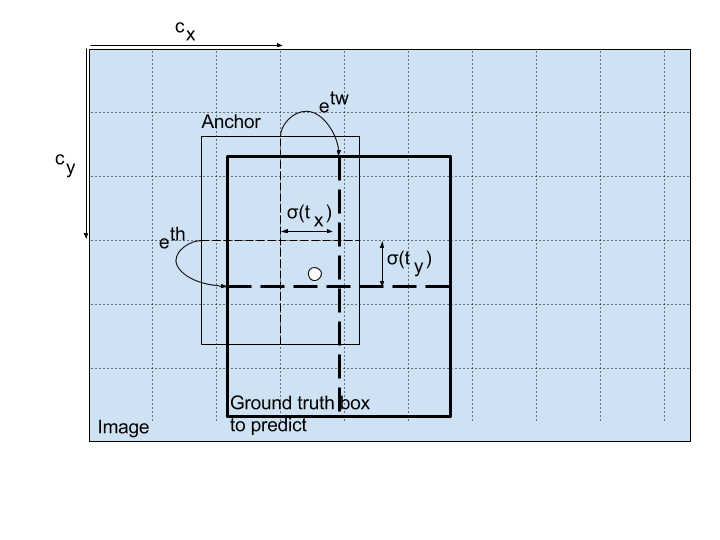
3.Quá trình phát hiện vật thể

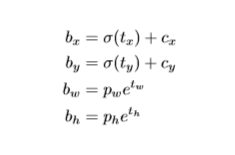


* Để phát hiện được các vật thể nhỏ trên Images thì Yolo v3 đã xử lý bằng cách nhận dạng 3 lần trên 1 Image
* Giả sử với kích thước ảnh đầu vào 416 x 416:
  + Tại lớp chập thứ 82: Bức ảnh được chia thành các ô lưới với kích thước 13×13 ( bức ảnh đã được chia với 32). Tại đây, các ô lưới sẽ có trách nhiệm tìm các vật thể có kích thước lớn trong bức hình.
  + Tại lớp chập thứ 94: Bức ảnh được chia thành các ô lưới với kích thước 26 x 26 ( bức ảnh được chia với 16). Và có trách nhiệm tìm các vật thể có kích thước trung bình.
  + Tương tự, tại lớp chập 106, bức ảnh được chia thành các ô lưới với kích thước 52×52( ảnh được chia với 8) và có trách nhiệm tìm các vật thể có kích thước bé.
  + Bức ảnh sau khi được chia thành các ô lưới, thì được gọi là Downsample images( ảnh giảm mẫu)



* Kích thước của đầu ra là :S x S x 255
  + S lần lượt được gắn với các giá trị:13,26,52. Và
  + Giá trị 255 = B x (5+C) với B = 3 là số lượng bounding box, và C =80 là số lớp vật thể.
  + Mỗi bounding box chứa 5 giá trị bao gồm: x , y , w , h và p( xác suất ô lưới có chứa vật thể)
  + Để tính ra kết quả cuối cùng thì mỗi bounding box sẽ nhân giá trị P lần lượt với P1,P2,…Pc ( là xác suất ô lưới đó chứa các vật thể cụ thể) và lựa chọn ra kết quả cao nhất.
  + Với mỗi bức ảnh được nhận dạng, thì YOLOv3 sẽ sinh ra (13 x 13 + 26 x 26 + 52 x 52) x3 = 10,647 bounding box. Từ đó sẽ tính xác suất và xuất ra kết quả là những vật thể có xác suất cao nhất
* Lựa chọn anchor box :





* + Anchor box thực ra là các bounding box nhưng được tạo sẵn ( chứ không phải kết quả của quá trình nhận dạng – prediction).
  + Với mỗi ô lưới sẽ 9 anchor box khác nhau với kích thước:
    - Ô lưới 13×13: (116×90) , (156 × 198) , (373 × 32)
    - Ô lưới 26×26: (30×61) , (62×45) , (59 × 119).
    - Ô lưới 52×52: (10×13) , (16×30) , (33 x 23).
  + Trong quá trình đào tạo sẽ kết hợp với thuật toán Kmean cluster( thuật toán phân cụm). Cùng với ground truth để tính ra sai sót giữa ground truth và anchor box bằng cách điều chỉnh các giá trị x,y,w,h, từ đó học được các đặc điểm của vật thể.