YOLO

YOLO(you only look once)란?

이미지 전체를 한번에 분석해 객체의 위치와 종류를 분석하는 알고리즘.

논문 링크 : https://arxiv.org/pdf/1506.02640

YOLO의 특징

- end-to-end 학습
  객체 입력부터 검출까지의 단계가 하나의 신경망으로 구성됨
  모델의 구조가 단순하고 빠름
- 주변 정보까지 학습하며 이미지 전체를 처리하기 때문에 background error 가 Fast R-CNN에 비해 적음
  background error : 배경에 노이즈나 반점이 있는 경우, 이를 물체로 인식하는 것
- 검출 정확도는 높으나, SOTA 객체 검출 모델에 비해 정확도(mAP)는 떨어짐 (작은 물체일수록 더욱더)
  SOTA(state-of-the-art) : 현재 수준에서 정확도가 가장 높은 모델

작동 원리

입력 이미지(input images)를 S x S 그리드(S x S grid)로 나눔.

각각 그리드 셀은 B개의 바운딩 박스, 바운딩 박스에 대한 confidence score를 예측한다.

- 그리드 셀 내 아무 객체가 없으면 Pr(Obejct)=0, confidence score도 0

- 그리드 셀에 어떤 객체가 확실히 있다고 예측하면 Pr(Object)=1.
  이 때가 가장 이상적이며, confidence score가 IOU와 같다면 가장 이상적인 score

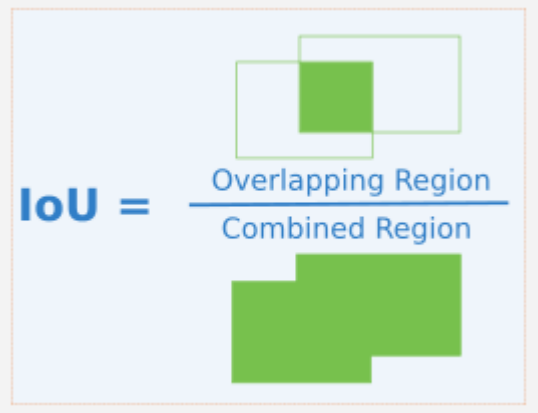
confidence score : bounding box가 객체를 포함한다는 것이 얼마나 믿을만한지, 예측한 bounding box가 얼마나 정확한지를 나타냄.

$$\Pr(Object) * IOU_{pred}^{truth}$$

confidence score 정의

IOU(intersection over union) = 교집합 영역 넓이 / 합집합 영역 넓이

IOU는 객체 인식 모델의 성능 평가를 하는 과정에서 사용되는 도구로 정답 영역과 예측 영역이 얼마나 겹쳐있는지 평가하는 지표.



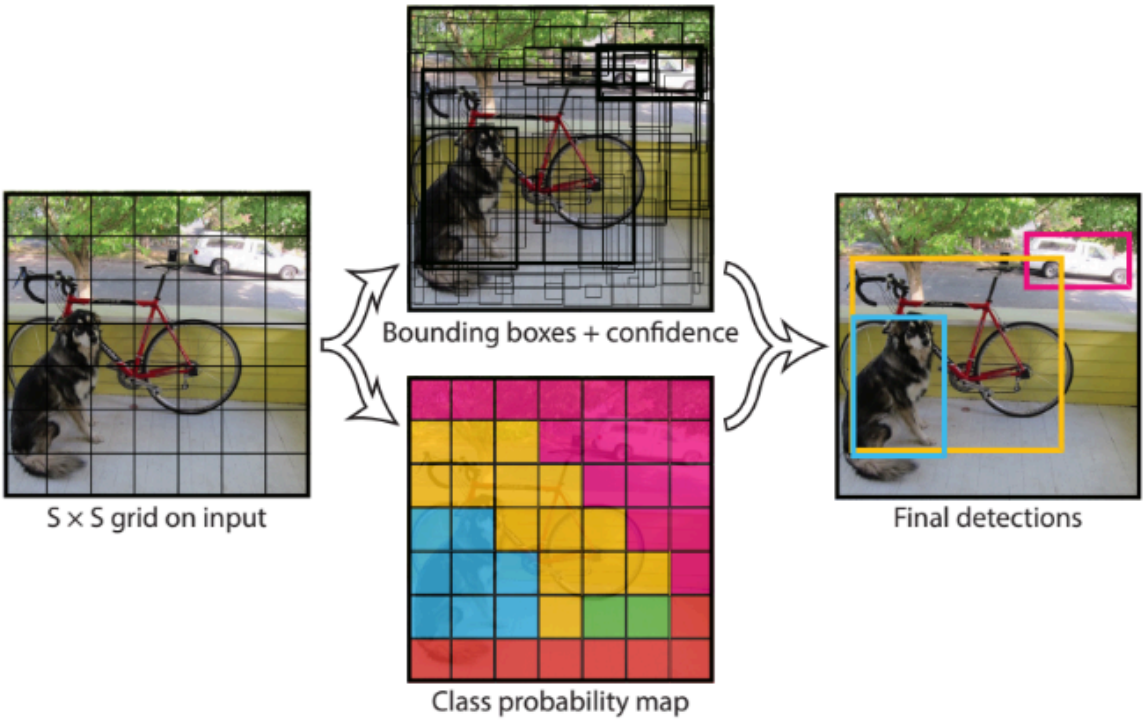
- IoU = 1.0 → 완벽히 일치
- IoU = 0.5 → 절반 정도 겹침
- IoU = 0 → 전혀 겹치지 않음

중복 박스 제거(NMS) 에서도 사용됨.

그리드 셀의 예측치
x, y, w, h, confidence로 구성.
x, y : 그리드 셀 내 상대 위치
w, h : 그리드 셀 내 상대 넓이와 상대 높이

confidence : confidence score와 동일
클래스 확률 : 개체가 어떤 클래스일 확률

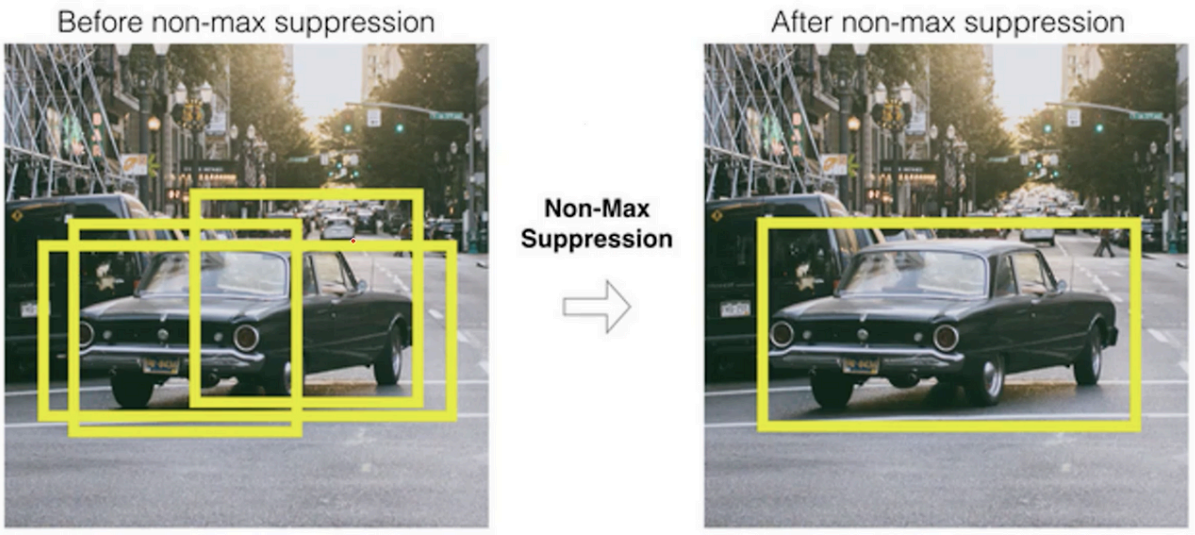
x, y, w, h는 전체 넓이 높이, 가로, 세로를 1이라고 했을 때 셀의 넓이, 높이, 가로, 세로를 0~1 사이의 값으로 나타냄. (정규화)



그리드 셀은 클래스 확률(conditional class probabilities(C))를 예측.
그리드 셀 내 객체가 있다는 조건 하에 객체가 어떤 Class인지 조건부 확률을 의미하며, 바운딩 박스 개수와 무관하게 클래스 하나만 예측한다.

Class probability map은 그리드 셀별로 어떤 클레스의 객체가 있는지 색으로 표시한 이미지이다.

NMS(Non-Maximu Suppression, 중복 박스 제거)



YOLO는 같은 객체에 대해서 여러 바운딩 박스를 예측할 수 있음.

NMS는 모델에서 중복된 바운딩 박스를 제거하고 가장 신뢰도가 높은 박스만 남기는 것이다.

confidence score가 높은 방식으로 박스를 정렬 후 가장 높은 Score 박스 선택하고, 나머지 박스 중 선택된 박스와 비교해 IoU가 임계값 이상인 박스를 제거한다.

NMS를 통해 남은 바운딩 박스와 Class probability map을 합치면 어느 위치에 어떤 객체가 있는지 표시 및 결과 이미지를 얻을 수 있다.

네트워크 구조

1개의 CNN 구조로 디자인 되었음.

- 앞단 conv 계층(특징 추출) - 전결합 계층(클래스 확률, 바운딩 박스 좌표 예측)
- 신경망 구조는 GoogleNet에서 따왔다. 인셉션 구조 대신 1 x 1 축소 계층과 3 x 3 conv 계층의 결합을 사용함.
- 모델의 최종 아웃풋은 7x7×30의 예측 텐서.
  - 이미지를 7x7의 그리드로 나눔. (총 49개의 그리드)
  - 각 그리드 셀이 출력하는 예측 값이 30, 여러 정보를 합친 결과.
  바운딩 박스 5개의 값(총 2개의 박스에 대한 정보) + 클래스 확률 20 (PASCAL VOC 데이터 셋이 20개의 클래스를 가짐)

학습

객체 검출을 위해서는 이미지 정보와 해상도가 높아야 함. 입력 이미지의 해상도를 244x244 > 448×448로 증가시킴.