

YOLO v2

국민대 소프트웨어 학부
UROP(학부생 연구 참여기회)
김대희
(2019.4.4)

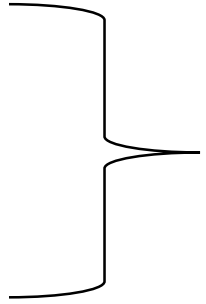
INDEX

1. YOLO v1의 한계점
2. YOLO v2 : Better
3. YOLO v2 : Faster
4. YOLO 9000 : stronger
5. Conclusion

1. YOLO v1의 한계점

- 분류 능력 자체는 다른 모델(Faster R-CNN)보다 떨어짐
- 빠르고 배경정보에 강하지만,
작은 물체를 Detection하는 데 애를 먹음
- 군집 되어 있는 작은 물체들을 잘 잡아내지 못함

2. YOLO 9000: Better, Faster, Stronger

-Better  **YOLO v2**

-Faster

-Stronger - **YOLO 9000**

2. Better : Batch Normalization

- 배치정규화 : 공변량 시프트 현상을 누그러뜨리기위해 정규화를 모든 층에 적용

$$z = \mathbf{w}^T \tilde{\mathbf{x}} + b$$
$$y = \tau(z) \quad \Rightarrow z \text{에 적용}$$

- 2% improvement in mAP
- 사용함으로써 규제역할을 해줘서 Yolo v1의 dropout을 제거함

+ 공변량 시프트(Covariate shift)

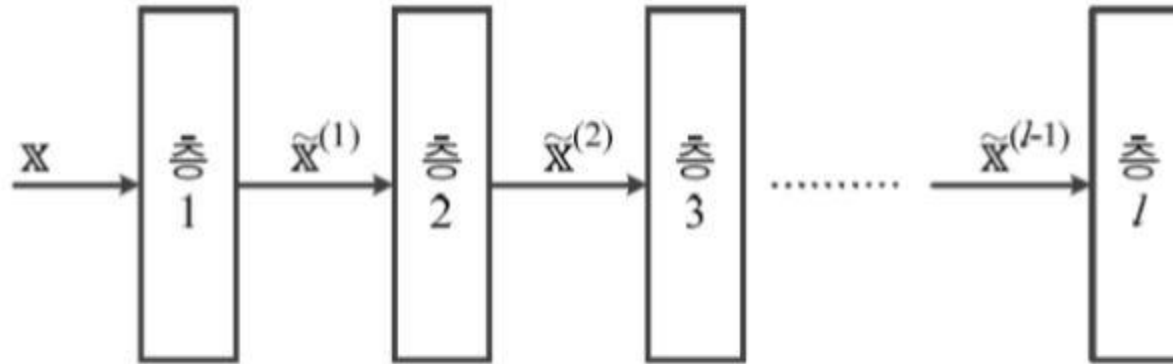


그림 5-17 공변량 시프트 현상

- 학습이 진행되면서 층1의 매개변수가 바뀔에 따라 $\hat{x}^{(1)}$ 이 바뀔
- 층이 깊어짐에 따라 심각해지고, 학습을 방해하는 요인으로 작용

2. Better : High Resolution Classifier

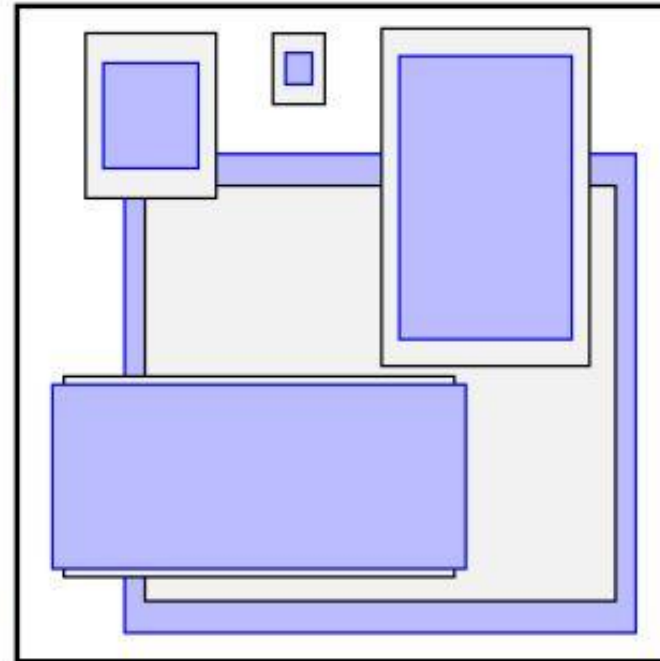
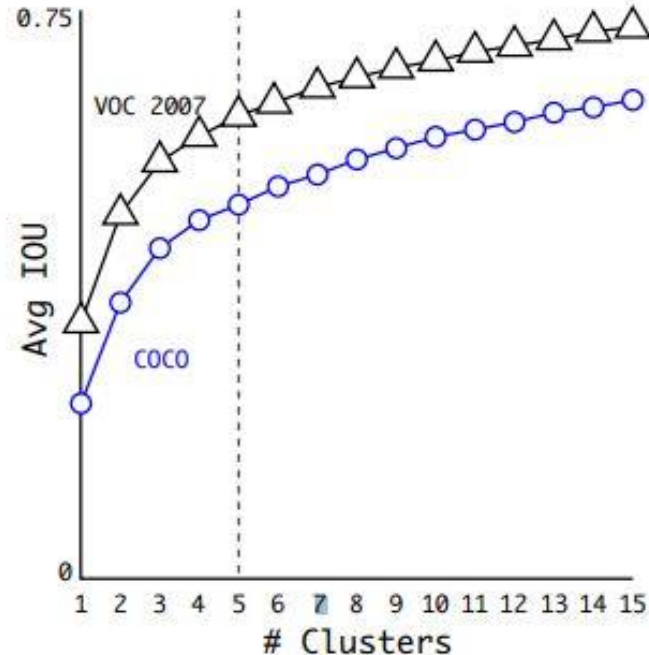
- YOLO는 224×224 로 학습 후 448로 디텍션
- V2는 448×448 사이즈를 10epoch동안 fine tuning
- 약 4% improvement in mAP

2. Better : Convolutional With Anchor Boxes

- YOLO는 앵커박스를 사용하지 않았음
- Faster R-CNN 계열 처럼 앵커박스 추가
- 인풋 416x416 으로 축소 → 마지막 featuremap 사이즈를 홀수로 함
큰 이미지는 가운데에 있을 가능성이 높기때문에 홀수 그리드셀이 유리
- 69.5 mAP → 69.2 mAP로 하락 / recall 82% → 88%로 상승

2. Better : Dimension Clusters

- 앵커박스를 수작업 할 것인가? -> No, k-means clustering 사용
- Pascal voc, coco의 ground truth box를 클러스터링
$$d(\text{box}, \text{centroid}) = 1 - \text{IOU}(\text{box}, \text{centroid})$$



2. Better : Direct location prediction

RCNN

$$x = (t_x * w_a) - x_a$$

$$y = (t_y * h_a) - y_a$$

yolo v2

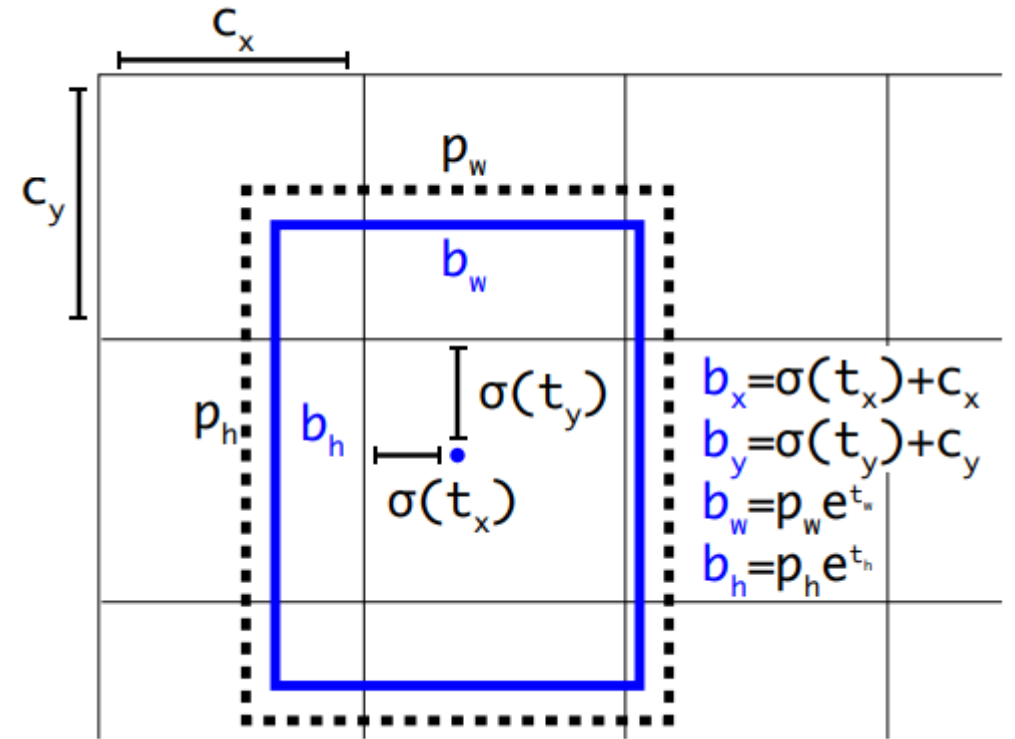
$$b_x = \sigma(t_x) + c_x$$

$$b_y = \sigma(t_y) + c_y$$

$$b_w = p_w e^{t_w}$$

$$b_h = p_h e^{t_h}$$

- t_x, t_y 에 시그모이드를 취한것은
범위를 그리드 셀 안으로 제한하기 위함

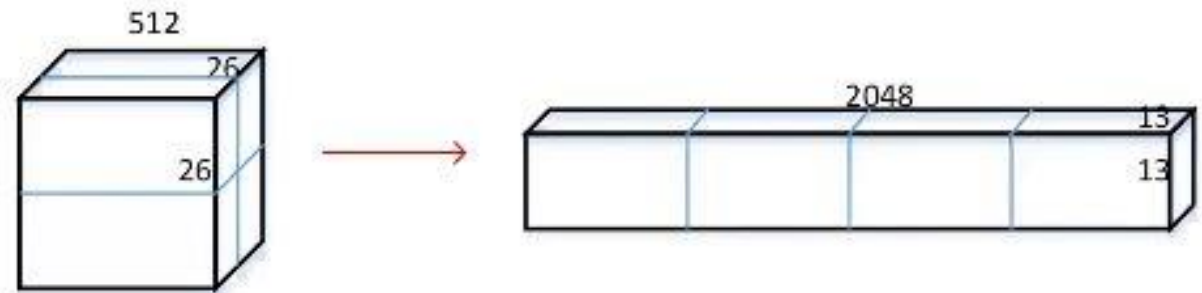


- 학습 초기에 위치가 엉뚱한 곳으로 옮겨갈 가능성이 있다고 함

2. Better : Fine-Grained Features

- 13 x 13 feature map에서 예측하도록 수정
- 작은 물체를 더 잘 찾기 위해 마지막 단 이전의 feature map (26 x 26 x 512) 을 (13 x 13 x 2048)로 나누어 마지막 feature map에 concat함
- 이는 ResNet의 identity mapping과 유사함
- SSD처럼 다양한 사이즈의 feature map을 활용하기 위함(작은 물체)

- 1% performance increase



2. Better : Multi-Scale Training

- fully connected layer 제거,
따라서 다양한 input 사이즈 가능

- 매 10 batch 마다 랜덤한
image dimension size 선택
{320, 352, ... , 608}

- 해상도가 커지면 성능 상승, 느려짐
작아지면 성능 하락, 빨라짐

Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19
YOLOv2 288 × 288	2007+2012	69.0	91
YOLOv2 352 × 352	2007+2012	73.7	81
YOLOv2 416 × 416	2007+2012	76.8	67
YOLOv2 480 × 480	2007+2012	77.8	59
YOLOv2 544 × 544	2007+2012	78.6	40

->tradeoff

3. Faster : Darknet-19

- 모델을 새로 디자인해서 사용
(VGG와 비슷) 3x3 필터 주로 사용
- 19 conv layers, 5 maxpooling layers
- 전역 평균 풀링 사용, 매개변수 줄임

Type	Filters	Size/Stride	Output
Convolutional	32	3×3	224×224
Maxpool		$2 \times 2/2$	112×112
Convolutional	64	3×3	112×112
Maxpool		$2 \times 2/2$	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		$2 \times 2/2$	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		$2 \times 2/2$	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		$2 \times 2/2$	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

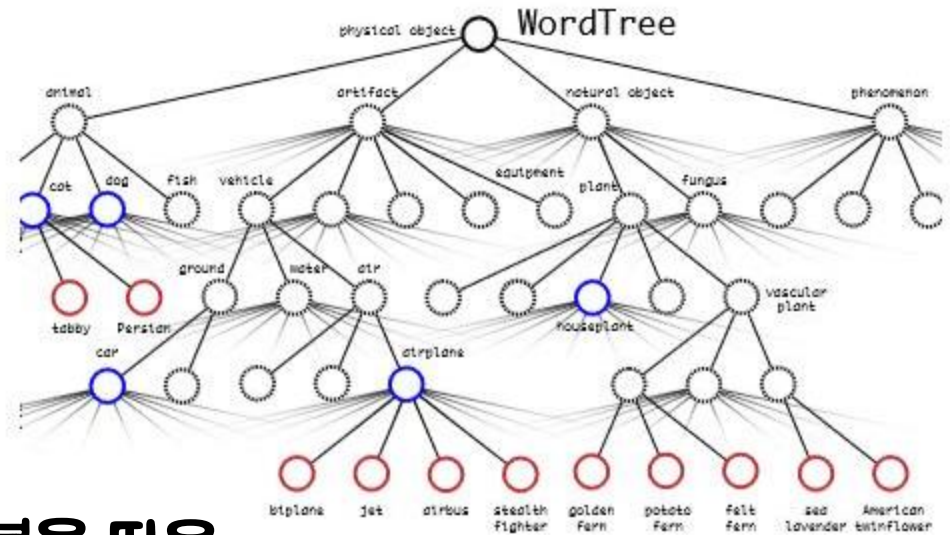
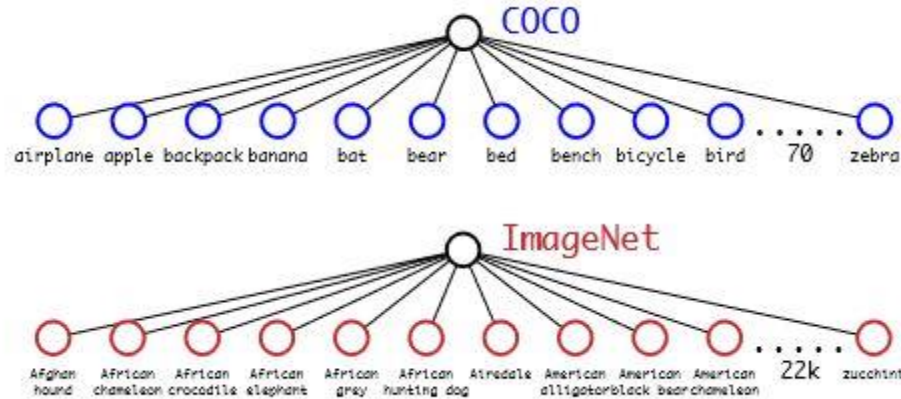
3. Faster : Training for classification

- imagenet 1000 class classification dataset for 160epochs**
- Random crops, rotations, hue, saturation, and exposure shifts
(Standard data augmentation tricks)**
- 224 x 224 -> 448x448 fine-tuning for 10epochs**

3. Faster : Training for detection

- 3 x 3 conv layers with 1024 filters each followed by a final 1x1 conv layer
- Pascal voc 기준,
- (좌표 4개 + confidence 1 + 클래스 20개) * 5개 바운딩 박스 = 125채널

4. Stronger



- ImageNet은 WordNet 라는 곳에서 레이블을 따옴

ex) "Norfolk terrier" and "Yorkshire terrier" are both hyponyms of "terrier" which is a type of "hunting dog", which is a type of "dog", which is a "canine", etc.

- WordNet은 트리 구조가 아니므로, 중복을 제거하면서

COCO와 ImageNet의 레이블을 트리로 합치는 시도

5. Conclusion

- YOLO v2는 이미지 사이즈에 따라 속도와 정확도 사이의 tradeoff 선택가능
- YOLO 9000은 9000개의 object detection과 classification 가능
- 계층적 분류를 이용해 다양한 데이터셋을 조합시키는 방법은 분류나 세그멘테이션 문제에서도 유용할 것이다.

끝. 감사합니다.

- 참고 : <https://www.youtube.com/watch?v=6fdclSGgeio>