SSD: Single Shot MultiBox Detector

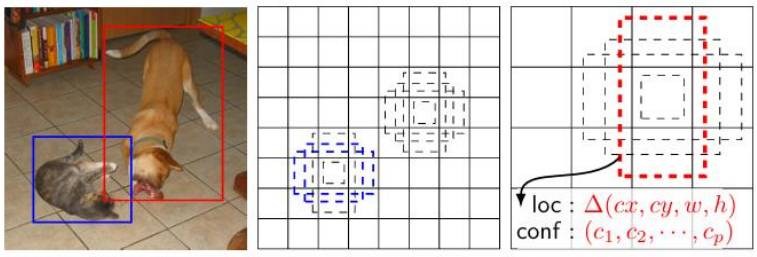
Abstract

- 아웃풋을 만드는 공간(multi feature map)을 분할한 뒤, 각 피쳐맵(output map)에서 각각의 비율과 스케일로 default box를 생성하고, 생성된 모델을 통해 계산된 좌표와 클래스 값에 default box를 활용해 최종 bounding box를 생성하는 모델

Intruduction

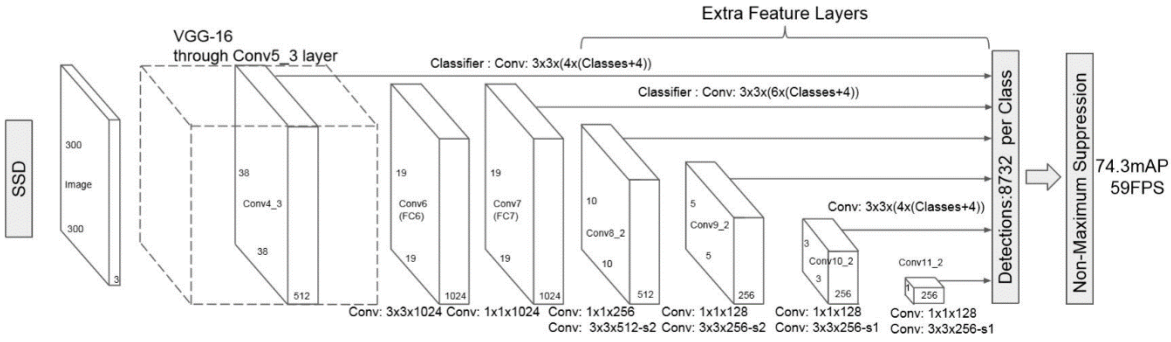
- 기본적으로 물체를 인식하는 과정에서, 물체의 후보 박스를 탐지 (hypothesize bounding boxes)한 후 해당 영역에서 픽셀과 특징을 정렬한 후, 더 좋은 분류기에 앞의 값들을 적용시키는 방식이 주로 사용되어왔음
- 하지만 이러한 방식은 복잡한 계산과정을 거쳐야 하기에 실사용에 문제가 더러 존재함
- 이러한 문제를 해결하기 위해 논문의 저자는 SSD network 중간에 존재하는 conv layer의 feature map들을 추출하여 detection 시 사용하는 방법을 제안 -> The Single Shot Detector(SSD)

- Model



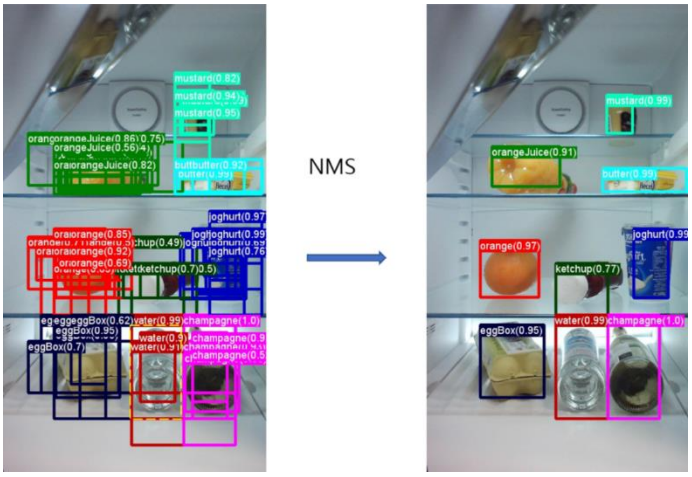
(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

*SSD 프레임워크

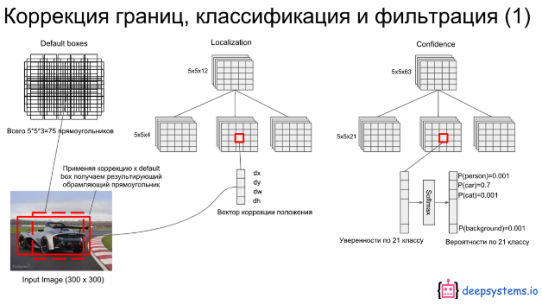


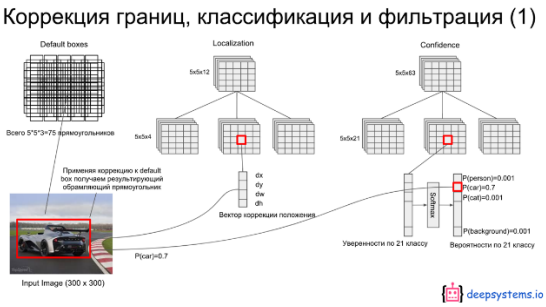
- 모델 구조: SSD에서는 VGG-16모델을 가져와 conv4_3까지 적용하는 것을 base network로 두는데, 이를 통하여 저해상도 파일을 분석하는 데 이점을 가짐(300x300x3이 38x38x512)

- conv filter size는 3 x 3 x (#바운딩박스 개수 x (class score + offset))이다. 이 6개의 피쳐맵 각각에서 예측된 바운딩박스의 총 합은 8732개
- 다른 해상도의 여러 feature map에 적용한다는 점에서, 특성 맵에서 다른 default box를 허용하게 될 시 output box 모양의 공간을 효율적으로 구분할 수 있음



- 여러 개의 값 중, 하나만 차용하는 방식





모델 작동 방식 요약

1. 모델에 입력된 이미지를 Convolution과 Pooling을 하면서 점점 작아지는 필터들을 만들어낸다.

2. 각각의 필터들로 어느 위치에서 어떤 물체를 인식하였는지에 대한 데이터들을 받아온다.

3. 출력된 데이터 중 어떤 데이터가 실제 물체를 인식했는지 확인하기 위해 IOU값을 사용한다. (NMS알고리즘 사용) 주변의 데이터들과 비교했을 때 최대값인 데이터만 추출한다.

Experimental Results

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast [6]	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
Fast [6]	07+12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster [2]	07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
Faster [2]	07+12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Faster [2]	07+12+COCO	78.8	84.3	82.0	77.7	68.9	65.7	88.1	88.4	88.9	63.6	86.3	70.8	85.9	87.6	80.1	82.3	53.6	80.4	75.8	86.6	78.9
SSD300	07	68.0	73.4	77.5	64.1	59.0	38.9	75.2	80.8	78.5	46.0	67.8	69.2	76.6	82.1	77.0	72.5	41.2	64.2	69.1	78.0	68.5
SSD300	07+12	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD300	07+12+COCO	79.6	80.9	86.3	79.0	**76.2**	57.6	87.3	88.2	88.6	60.5	85.4	**76.7**	**87.5**	**89.2**	84.5	81.4	55.0	81.9	**81.5**	85.9	78.9
SSD512	07	71.6	75.1	81.4	69.8	60.8	46.3	82.6	84.7	84.1	48.5	75.0	67.4	82.3	83.9	79.4	76.6	44.9	69.9	69.1	78.1	71.8
SSD512	07+12	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
SSD512	07+12+COCO	**81.6**	**86.6**	**88.3**	**82.4**	76.0	**66.3**	**88.6**	**88.9**	**89.1**	**65.1**	**88.4**	73.6	86.5	88.9	**85.3**	**84.6**	**59.1**	**85.0**	80.4	**87.4**	**81.2**

Table 1: **PASCAL VOC2007 `test` detection results.** Both Fast and Faster R-CNN use input images whose minimum dimension is 600. The two SSD models have exactly the same settings except that they have different input sizes (300×300 vs. 512×512). It is obvious that larger input size leads to better results, and more data always helps. Data: "07": VOC2007 `trainval`, "07+12": union of VOC2007 and VOC2012 `trainval`. "07+12+COCO": first train on COCO `trainval35k` then fine-tune on 07+12.

Method	data	Avg. Precision, IoU:			Avg. Precision, Area:			Avg. Recall, #Dets:			Avg. Recall, Area:		
		0.5:0.95	0.5	0.75	S	M	L	1	10	100	S	M	L
Fast [6]	train	19.7	35.9	-	-	-	-	-	-	-	-	-	-
Fast [24]	train	20.5	39.9	19.4	4.1	20.0	35.8	21.3	29.5	30.1	7.3	32.1	52.0
Faster [2]	trainval	21.9	42.7	-	-	-	-	-	-	-	-	-	-
ION [24]	train	23.6	43.2	23.6	6.4	24.1	38.3	23.2	32.7	33.5	10.1	37.7	53.6
Faster [25]	trainval	24.2	45.3	23.5	7.7	26.4	37.1	23.8	34.0	34.6	12.0	38.5	54.4
SSD300	trainval35k	23.2	41.2	23.4	5.3	23.2	39.6	22.5	33.2	35.3	9.6	37.6	56.5
SSD512	trainval35k	**26.8**	**46.5**	**27.8**	**9.0**	**28.9**	**41.9**	**24.8**	**37.5**	**39.8**	**14.0**	**43.5**	**59.0**

Table 5: **COCO `test-dev2015` detection results.**

- PASCAL VOC, COCO Dataset에서 속도, 정확도 성능이 가장 좋게 나타남