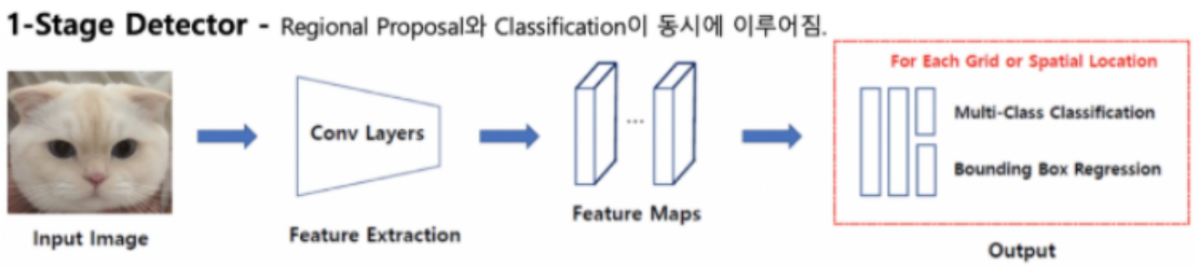
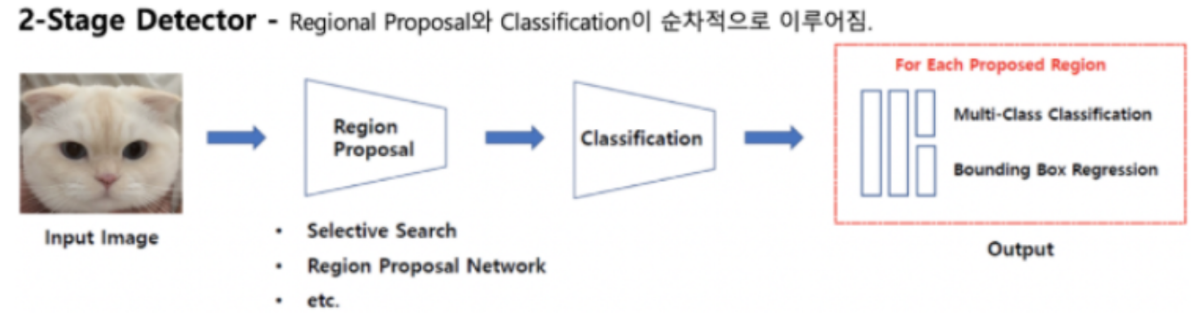
R-CNN, SPPNet

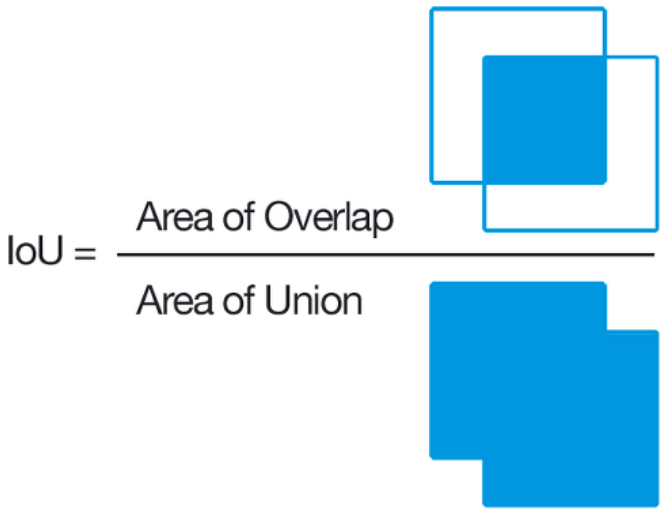
Created At	@2021년 7월 23일 오전 10:05
Created By	조승제
Topics	Deep Learning, Object Detection
Type	Lesson
발표일	@2021년 7월 23일
발표자	조승제
참고 링크 1	
참고 링크 2	
참석자	유정민, 엄현식, 조건우
첨부 자료 1	
첨부 자료 2	
첨부 자료 3	

Background





- IoU (Intersection Over Union)
  - Bounding Box를 얼마나 잘 예측하였는지를 판단하는 지표
  - 교집합 / 합집합



- mAP

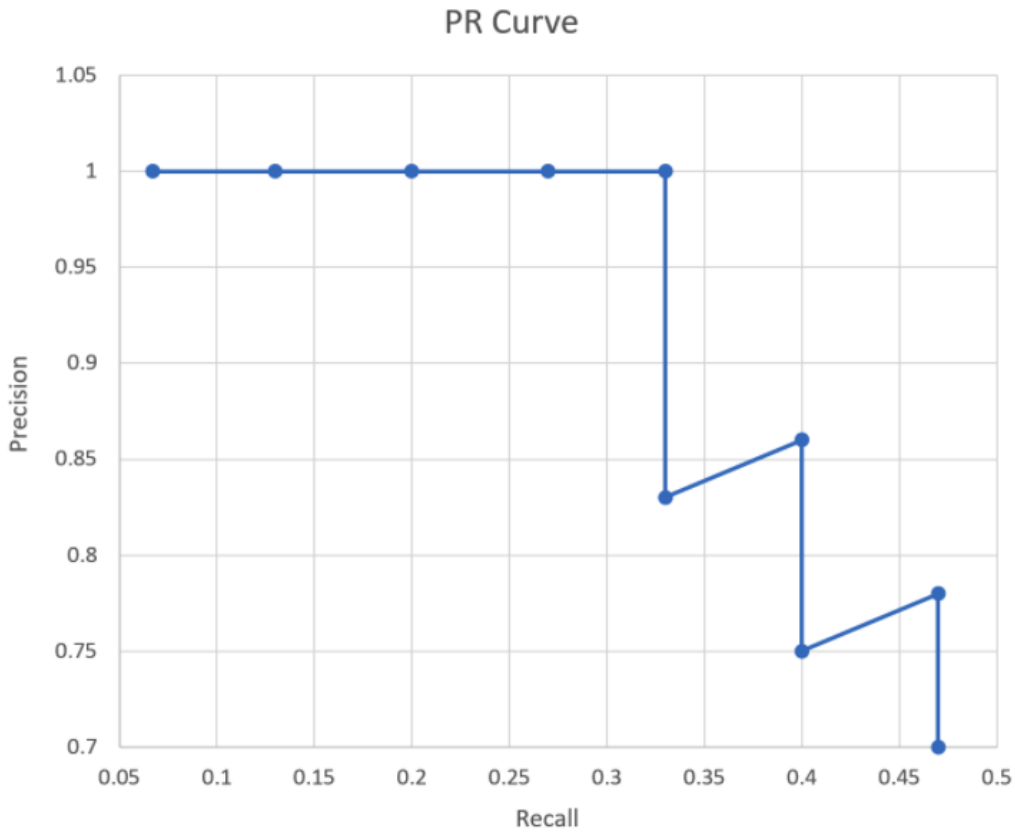
  : AP 값의 평균

  ▼ AP

실제 상황 (ground truth)	예측 결과 (predict result)	
	Positive	Negative
Positive	TP(true Positive) 옳은 검출	FN(false negative) 검출되어야 할 것이 검출되지 않음
Negative	FP(false positive) 틀린 검출	TN(true negative) 검출되지 말아야할 것이 검출되지 않음

- precision = TP / (TP + FP)
- recall = TP / (TP + FN)
- precision, recall 값은 confidence 값의 threshold 값에 따라 달라짐

- threshold 값을 0 ~ 1.0까지 0.1 단위로 증가시키면서 precision, recall 값을 계산



- AP = 그래프 선의 아래쪽 면적으로 계산

- Bounding Box


Open
Open Dir
Change Save Dir
Next Image
Prev Image
Verify Image
Save
yolo
YOLO
Create RectBox
Duplicate RectBox
Delete RectBox
Zoom In

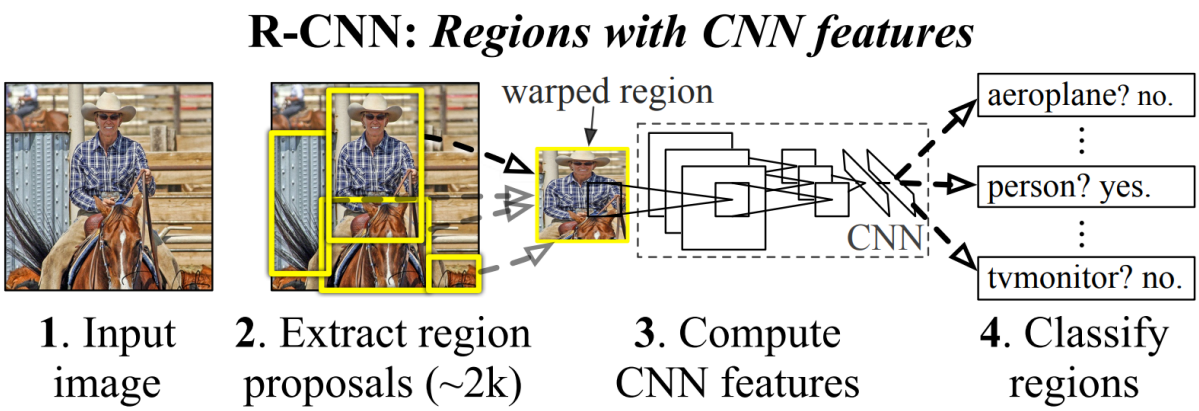
- NMS


mustard(0.82)
mustard(0.94)
mustard(0.95)
orangorangeJuice(0.86)0.75)
orangeJuice(0.56)4)
oraiorangeJuice(0.82)
buttbutter(0.92)
joghurt(0.97
oraiorange(0.85)
chup(0.49)
joghjoghurt(0.99
joghurt(0.76
oraioraiorange(0.92)
ororange(0.69)
ketchup(0.7)0.5)
eggeggeggBox(0.62)
vater(0.99)
champagne(1.0)
eggBox(0.95)
water(0.9)
champagne(0.9
eggBox(0.7)
champagne(0.5)

R-CNN (2014)

1. Introduction

- The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT and HOG
- It is generally acknowledged that progress has been slow during 2010-2012
- 기존 방법들은 비효율적 → CNN 도입 → 기존 연구들보다 월등한 성능 개선



2. Object detection with R-CNN

1) Generates category-independent region proposals.

2) A large convolutional neural network that extracts a fixed-length feature vector from each region.

3) Classification을 위한 linear SVMs

2.1) Region Proposals.

- 기존에 category-independent region proposals 연구가 활발했음
  - ex) objectness, selective search, category-independent object proposals, CPMC 등등
- 본 논문에선 Selective Search 기법을 사용
- 2000개의 독립적인 region proposal 생성

💡 Selective Search

: Segmentation 분야에 많이 쓰이는 알고리즘이며, 객체와 주변 간의 색감, 질감 차이 등을 파악해서 물체 위치를 파악

1. 이미지의 초기 segment를 지정하고, region 영역 생성
2. greedy 알고리즘을 이용해서 각 region을 기준으로 주변의 유사한 영역을 결합
3. 결합된 region을 최종 region proposal로 제안

2.2) Feature Extraction

- Pre-trained CNN Model (AlexNet) 사용
- 서로 다른 크기를 가진 region proposal들을 warp → (227, 227) RGB pixel

💡 warp
(x, y) 좌표의 픽셀을 (x', y') 좌표로 대응시키는 작업

참고

- 2000개의 region proposal들을 CNN을 통과시켜 4096차원의 feature vector를 추출
- CNN → 5개의 conv layer → 2개의 fully connected layer → feature vector 추출

2.3) Test-time detection

- Fully connected layer를 통과한 feature들은 SVM을 통해 각 class로 분류됨
- SVM을 통과한 region proposal은 NMS를 적용하여 하나의 bounding box만 남김
- NMS를 적용하여 IoU가 가장 높은 bounding box를 선택
- 이후 bounding box regression 적용 → ground-truth box와 비슷하게 조정

$$\hat{G}_x = P_w d_x(P) + P_x \quad (1)$$
$$\hat{G}_y = P_h d_y(P) + P_y \quad (2)$$
$$\hat{G}_w = P_w \exp(d_w(P)) \quad (3)$$
$$\hat{G}_h = P_h \exp(d_h(P)). \quad (4)$$

$$t_x = (G_x - P_x)/P_w \quad (6)$$
$$t_y = (G_y - P_y)/P_h \quad (7)$$
$$t_w = \log(G_w/P_w) \quad (8)$$
$$t_h = \log(G_h/P_h). \quad (9)$$

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t_\star^i - \hat{\mathbf{w}}_\star^{\mathrm{T}} \boldsymbol{\phi}_5(P^i))^2 + \lambda \left\| \hat{\mathbf{w}}_\star \right\|^2$$

$$d_\star(P) = \mathbf{w}_\star^{\mathrm{T}} \boldsymbol{\phi}_5(P)$$

3. Training

Supervised pre-training

- CNN 모델은 ILSVRC2012 classification 데이터셋으로 훈련

Domain-specific fine-tuning

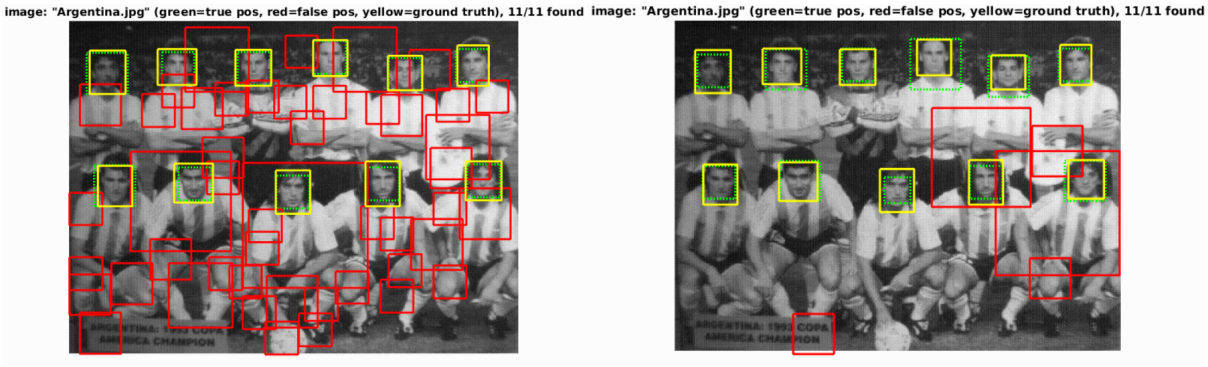
- CNN을 detection task와 warped proposal windows에 적응시키기 위해, warped region proposals을 사용하여 SGD로 50000번 훈련시킴

Object category classifiers

- Ground truth box → positive sample, IoU 값이 0.3 보다 작은 것은 negative sample 로 지정 (IoU 값이 0.3보다 큰 경우 무시)
- positive sample 32개, negative sample 96개, 총 128개의 mini-batch를 구성
- mini-batch → CNN → 4096 차원 feature vector 추출
- 추출된 벡터로 linear SVMs training (hard negative mining 기법 적용)

Hard negative mining

사람을 탐지하면 positive sample, 배경을 탐지하면 negative sample



4. Experiments

VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

Table 1: Detection average precision (%) on VOC 2010 test. R-CNN is most directly comparable to UVA and Regionlets since all methods use selective search region proposals. Bounding-box regression (BB) is described in Section C. At publication time, SegDPM was the top-performer on the PASCAL VOC leaderboard. †DPM and SegDPM use context rescoring not used by the other methods.

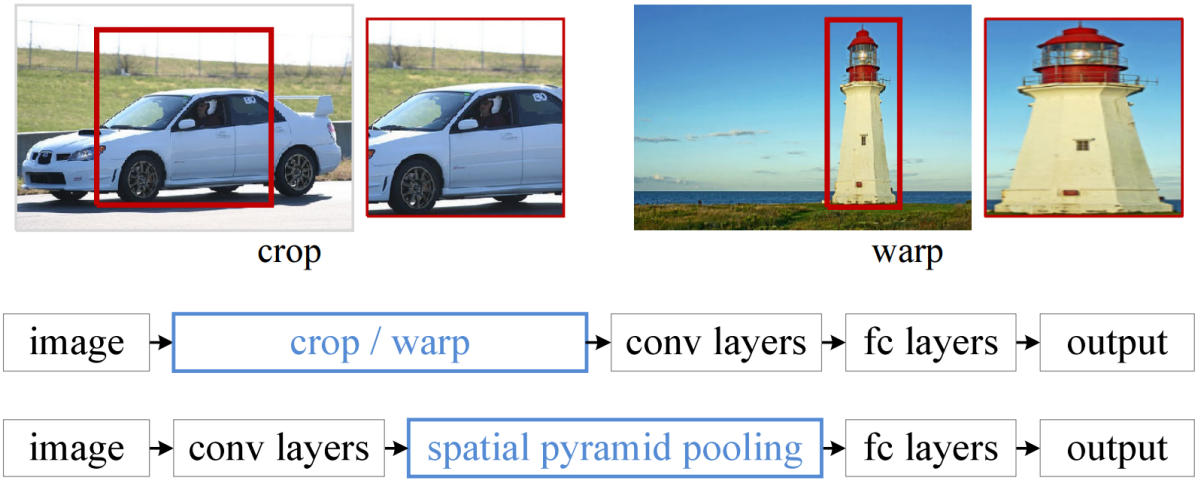
SPPNet (2015)

1. Introduction

- R-CNN의 문제점
  - 1개의 이미지에 대해 2000번의 CNN을 수행 → 시간적 비용 손해
  - Selective Search 이후 wrap 과정에서 이미지 왜곡 발생 → 성능 저하 가능성
  - CNN에 227x227의 고정된 input이 필요한 이유에 대한 의문

- In fact, convolutional layers do not require a fixed image size and can generate feature maps of any sizes.
- On the other hand, the fully-connected layers need to have fixedsize/length input by their definition.

- spatial pyramid pooling (SPP) layer to remove the fixed-size constraint of the network
- Specifically, we add an SPP layer on top of the last convolutional layer.



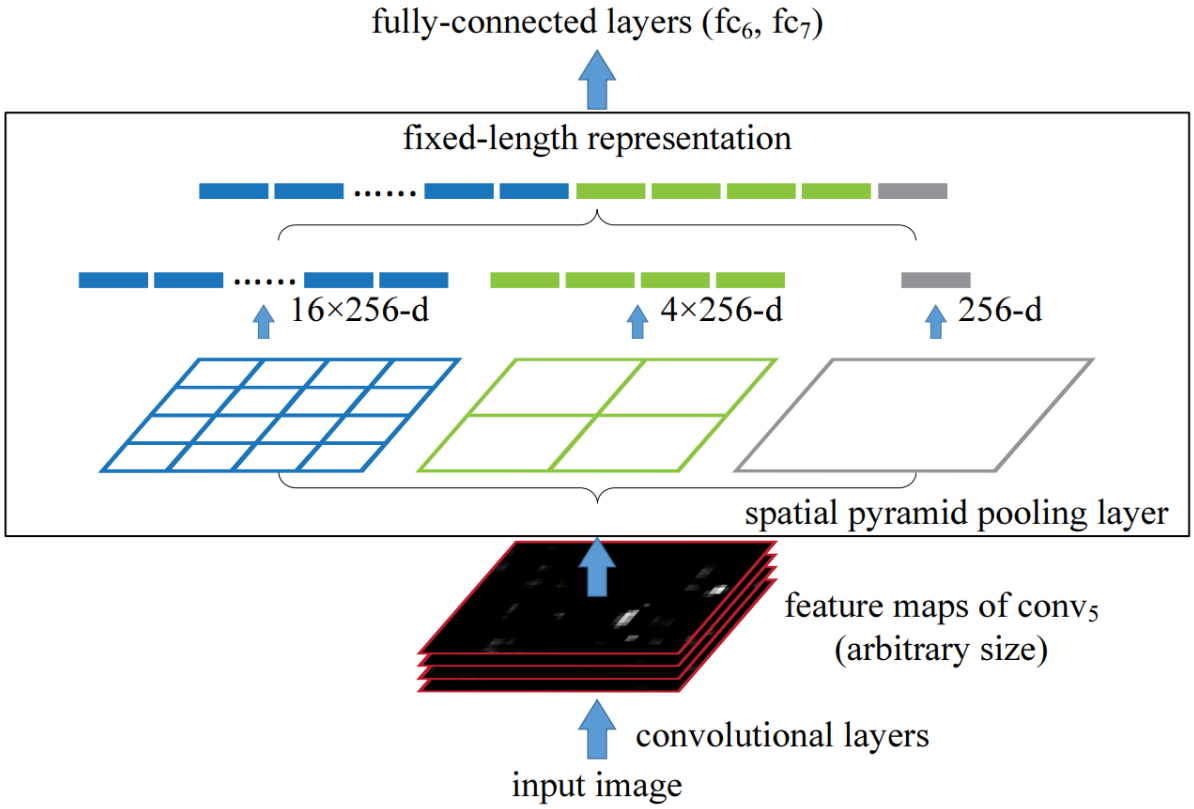
- The SPP layer pools the features and generates fixedlength outputs, which are then fed into the fullyconnected layers (or other classifiers). → 처음부터 이미지를 crop, warp 하지 않아도 되게 됨

2. The Spatial Pyramid Pooling Layer

Figure 2: Visualization of the feature maps. (a) Two images in Pascal VOC 2007. (b) The feature maps of some $conv_5$ filters. The arrows indicate the strongest responses and their corresponding positions in the images. (c) The ImageNet images that have the strongest responses of the corresponding filters. The green rectangles mark the receptive fields of the strongest responses.

코드 참고

- Bag-of-Words(BoW)에서 파생



- (?, ?, 256)의 feature map이 들어오면 filter size, stride 값을 조절해서 Max Pooling
- 위 사진에서는 {4x4, 2x2, 1x1}(totally 21 bins) 적용

- SPP layer 덕분에 input image의 크기에 제한 없이 특징을 잘 반영할 수 있게 되었음
- 또한 pooling은 resolution을 감소시키는데, 여러 pooling을 수행하면서 다양한 resolution을 가지게 됨

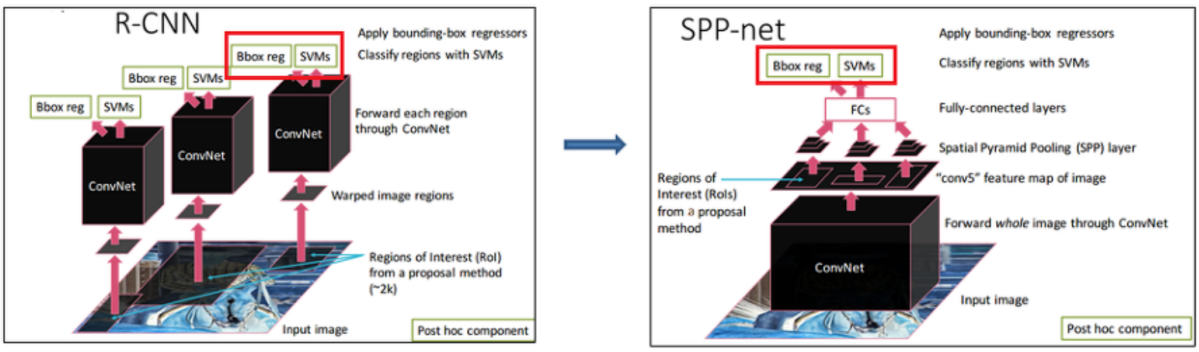
3. Experiments

method	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
DPM [23]	33.7	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5
SS [20]	33.8	43.5	46.5	10.4	12.0	9.3	49.4	53.7	39.4	12.5	36.9	42.2	26.4	47.0	52.4	23.5	12.1	29.9	36.3	42.2	48.8
Regionlet [39]	41.7	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	49.1	26.6	57.0	54.5	43.4	16.4	36.6	37.7	59.4	52.3
DetNet [40]	30.5	29.2	35.2	19.4	16.7	3.7	53.2	50.2	27.2	10.2	34.8	30.2	28.2	46.6	41.7	26.2	10.3	32.8	26.8	39.8	47.0
RCNN ftfc$_7$ (A5)	54.2	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7
RCNN ftfc$_7$ (ZF5)	55.1	64.8	68.4	47.0	39.5	30.9	59.8	70.5	65.3	33.5	62.5	50.3	59.5	61.6	67.9	54.1	33.4	57.3	52.9	60.2	62.9
SPP ftfc$_7$ (ZF5)	55.2	65.5	65.9	51.7	38.4	32.7	62.6	68.6	69.7	33.1	66.6	53.1	58.2	63.6	68.8	50.4	27.4	53.7	48.2	61.7	64.7
RCNN bb (A5)	58.5	68.1	72.8	56.8	**43.0**	36.8	**66.3**	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8
RCNN bb (ZF5)	**59.2**	68.4	**74.0**	54.0	40.9	35.2	64.1	**74.4**	69.8	**35.5**	66.9	53.8	**64.2**	69.9	69.6	**58.9**	**36.8**	**63.4**	**56.0**	62.8	64.9
SPP bb (ZF5)	**59.2**	**68.6**	69.7	**57.1**	41.2	**40.5**	**66.3**	71.3	**72.5**	34.4	**67.3**	**61.7**	63.1	**71.0**	**69.8**	57.6	29.7	59.0	50.2	**65.2**	**68.0**

Table 11: Comparisons of detection results on Pascal VOC 2007.

4. Conclusion

- SPP를 통해 RCNN에서의 warping 작업을 없애서 이미지 왜곡을 없앰
- RCNN은 CNN 연산을 2000번 했지만, SPP에서는 1번만 하면서 train, test 시간 단축



의문

☑ ~~CNN fine tuning 과정?~~

- For VOC, N = 20 and for ILSVRC2013, N = 200. We treat all region proposals with ≥ 0.5 IoU overlap with a ground-truth box as positives for that box's class and the rest as negatives.
- SVM과 비슷하게 진행됨, IoU값이 0.5 이상이면 Positive sample, 이하이면 negative sample

☑ ~~SVM이 왜 좋은 성능을 내는가?~~

- Network의 Overfitting을 피하기 위해 Positive 데이터가 많아야 하는데, 시가상으로 데이터가 많지 않아 softmax classifier을 적용했을 때 성능이 좋지 않았음

☑ ~~BoW랑 무슨 관계?~~

- BoW → Bag of Visual Words → Spatial Pyramid Matching