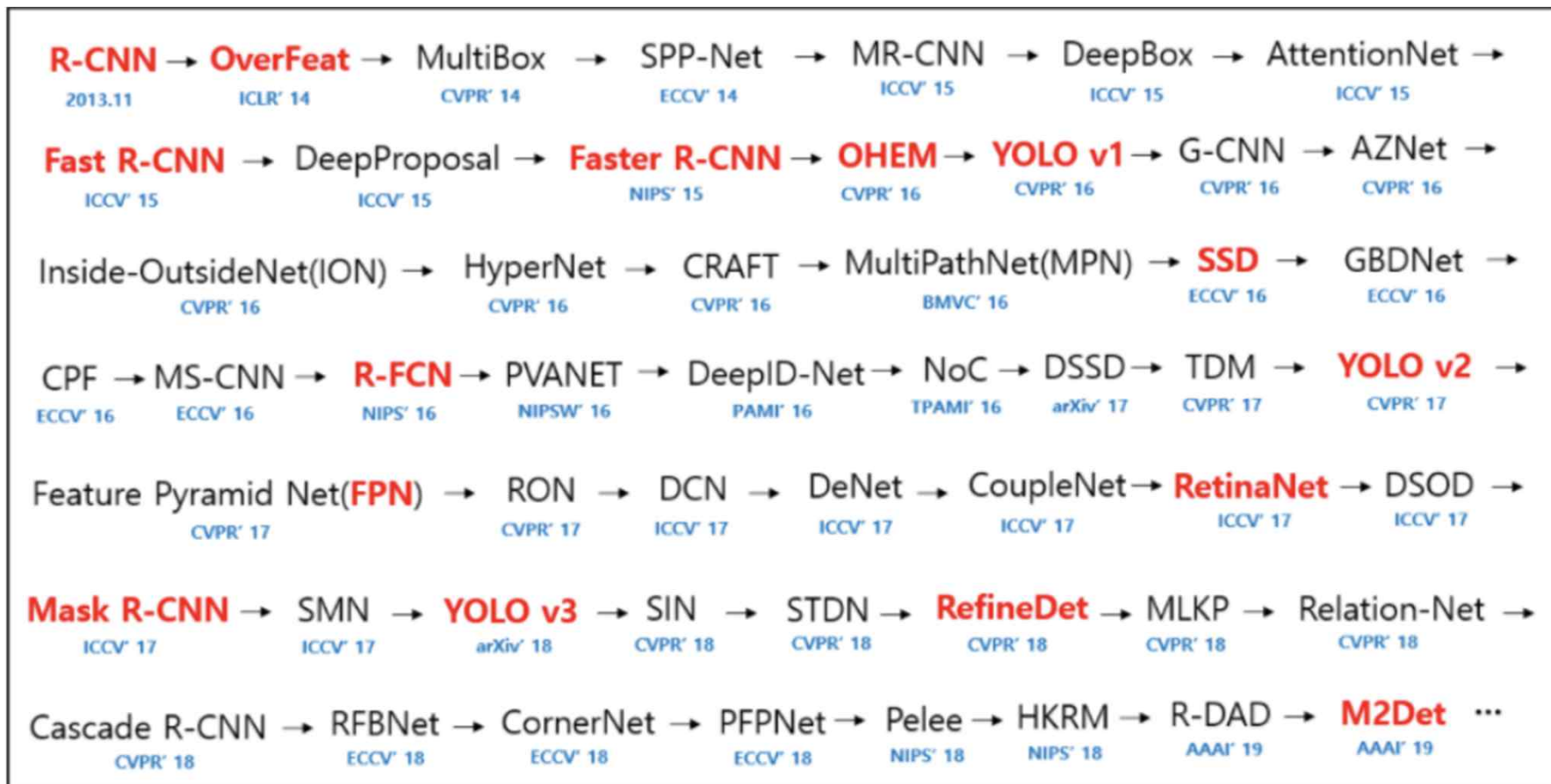




R-CNN : Rich feature hierarchies for accurate object detection and semantic segmentation Tech report



<https://arxiv.org/abs/1311.2524>



목차

A table of Contents

#1. Introduction

#2. Object detection with R-CNN

#3. Visualization, ablation, and modes of error

#4. The ILSVRC2013 detection dataset

#5. Semantic segmentation

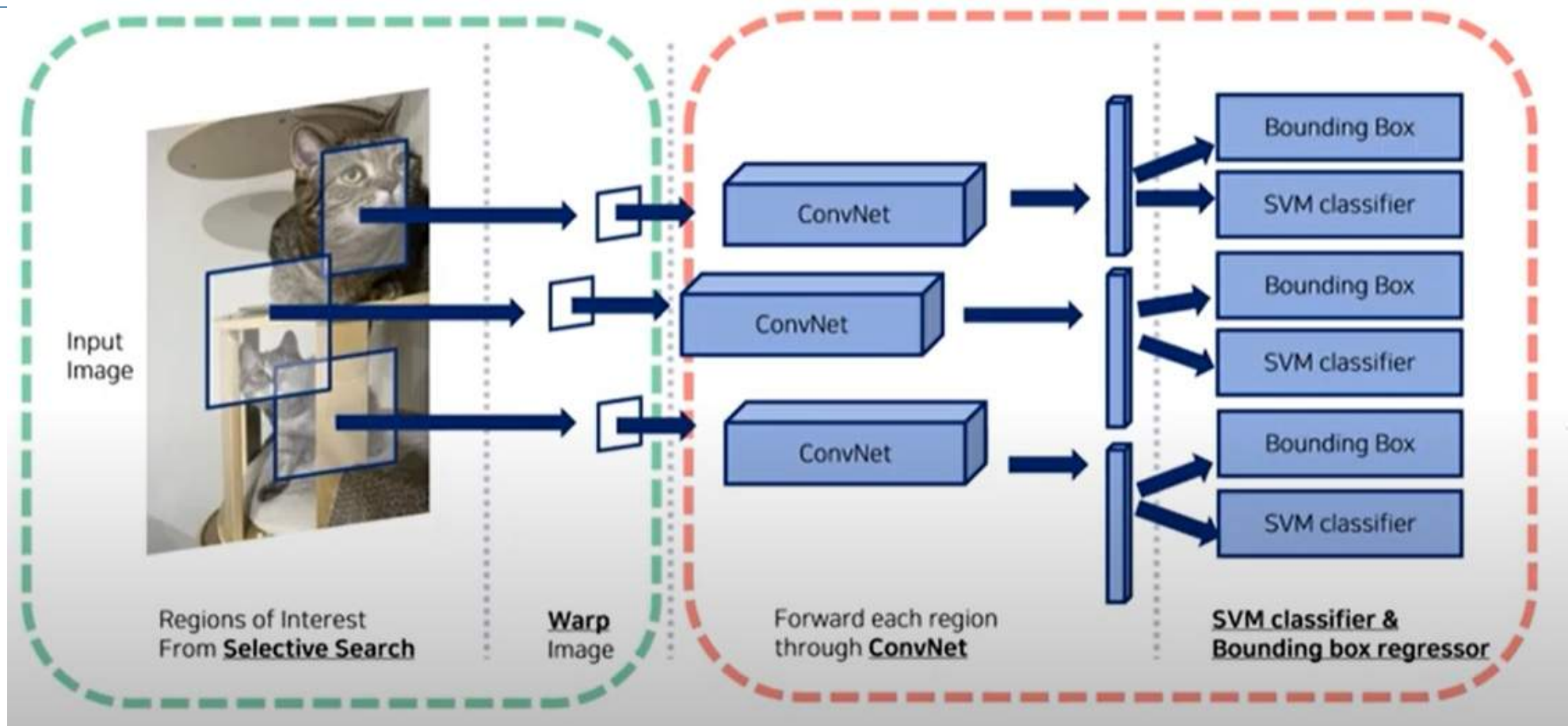
#6. Conclusion

Part 1, Introduction



Introduction

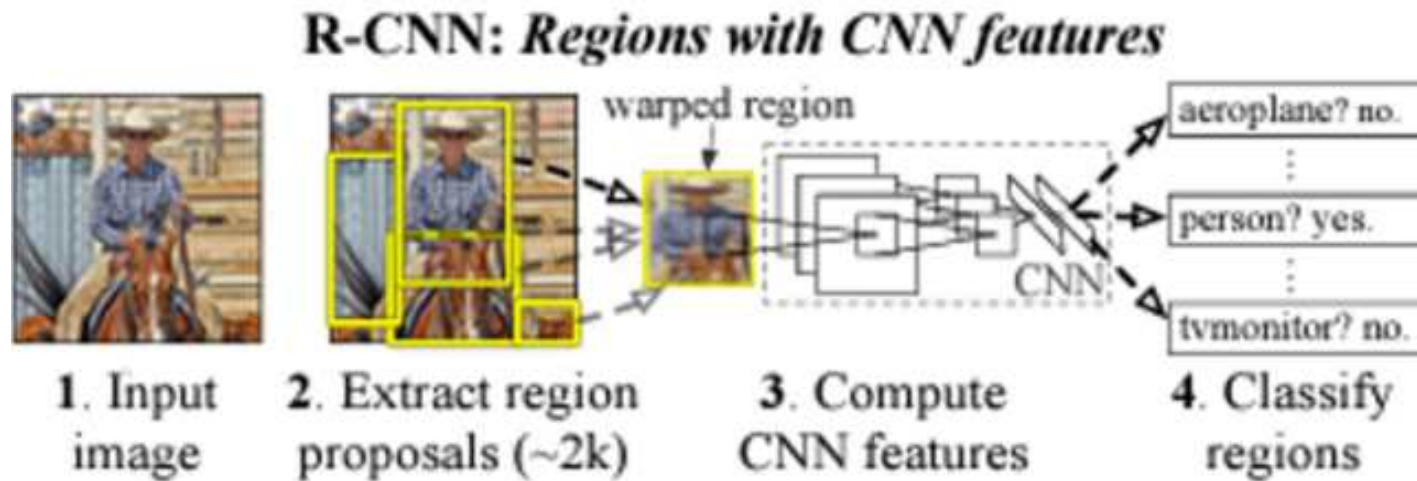
- Object detection 분야에서 최초로 convolution based Network 을 적용 시켰고, CNN을 이용한 검출 방식이 Classification 뿐만 아니라 object detection 분야에도 높은 수준의 성능을 이끌어 낼 수 있다는 것을 보여준 모델
- ILSVRC의 Classification result를 PASCAL VOC Challenge의 Object detection task에 확장하고자 연구를 진행하였습니다.
- 본 논문에서는 CNN을 이용하여 기존의 HOG 기반의 시스템에 비해 PASCAL VOC에서의 object detection의 성능을 높게 이끌 수 있다는 것을 보여줍니다.
- deep network를 통한 localization / 적은 양의 labeled 된 data로 모델을 학습시키는 것





Part 2, Object detection with R-CNN

Object detection with R-CNN

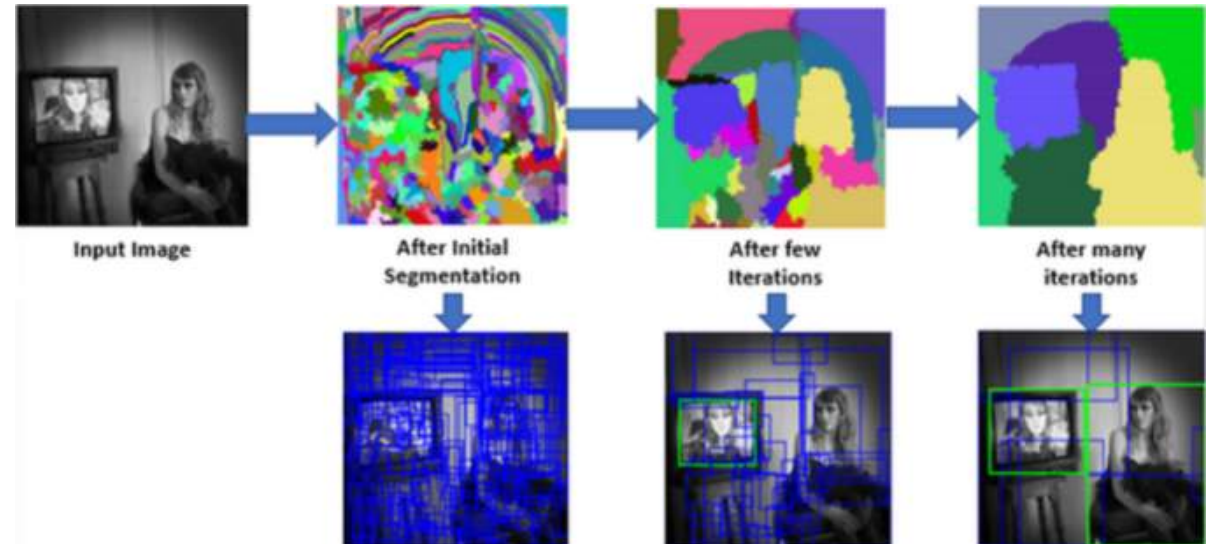


1. 독립적인 region proposals을 생성
→ proposals은 detector에서 사용할 수 있는 candidate detections 세트를 정의
2. 각 region에서 고정 길이의 feature vector를 추출하는 대형 convolutional neural network
3. 클래스 별 linear SVM 세트

2.1 Module design

Region proposals.

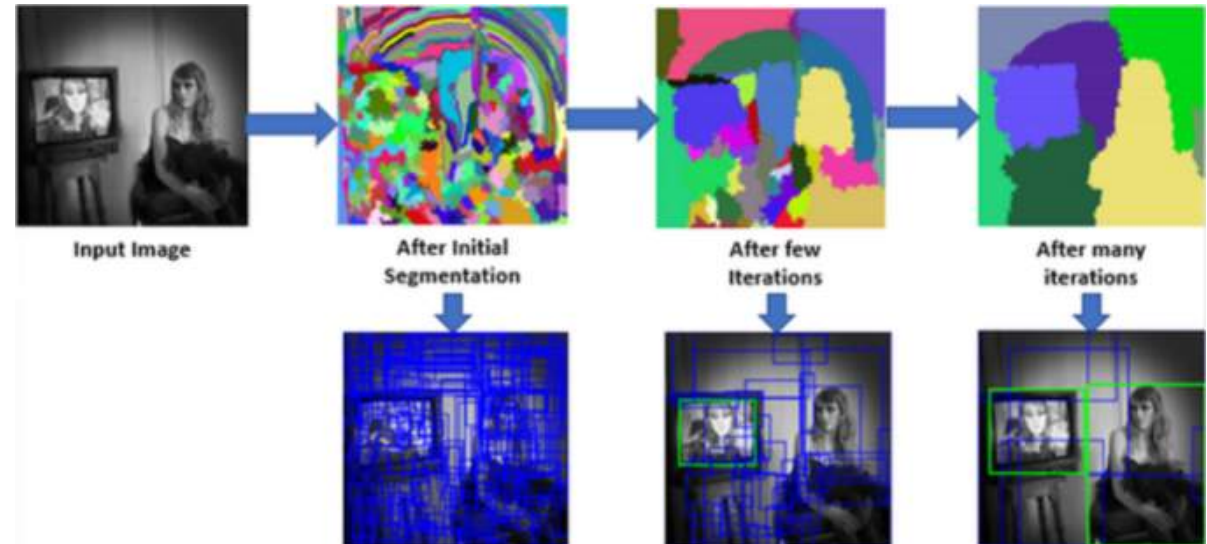
- 이미지 안에서 객체가 있을 만한 후보 영역을 먼저 찾아주는 방법
- 해당 논문에서는 Selective Search 라는 region Proposal을 사용
- Selective Search : object가 있을 법한 영역만 찾는 방법



2.1 Module design

Region proposals.

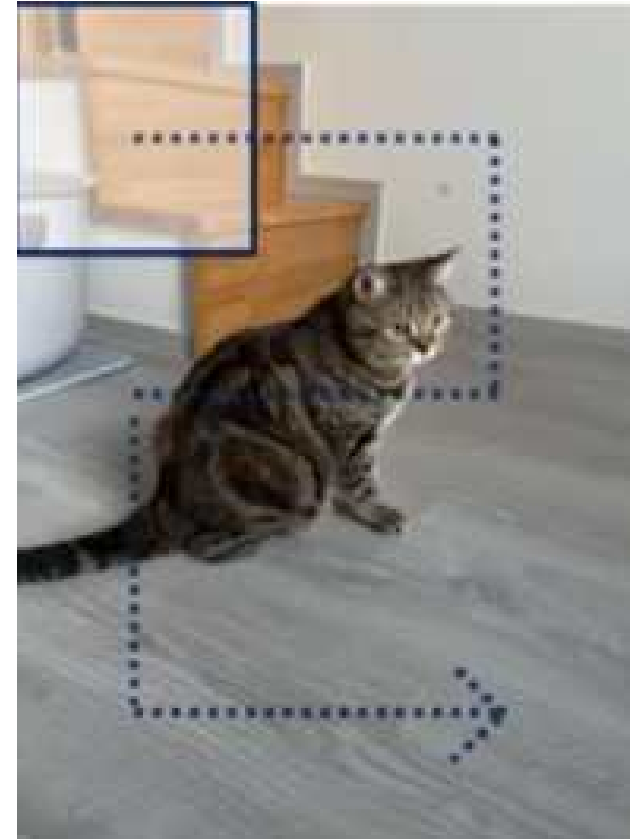
- 1) input에 대해 object가 있을 법한 부분 추정
- 2) 객체가 있을 법한 영역 탐지
- 3) 유사한 영역, 색, 질감을 가지는 픽셀끼리 묶어가며 순차적으로 넓혀감
- 4) 객체가 있을 법한 위치를 Bounding Box 형태로 추출함



2.1 Module design

sliding window

- 다양한 크기의 window를 이동시키며 객체가 있을 법한 위치를 찾는 것
 - but. 물체가 존재할 수 있는 모든 영역에 대해 탐색하기에 탐색 영역이 많음
- Background 처럼 필요 없는 부분도 진행해서 시간, 연산 낭비가 존재



2.1 Module design



Feature extraction

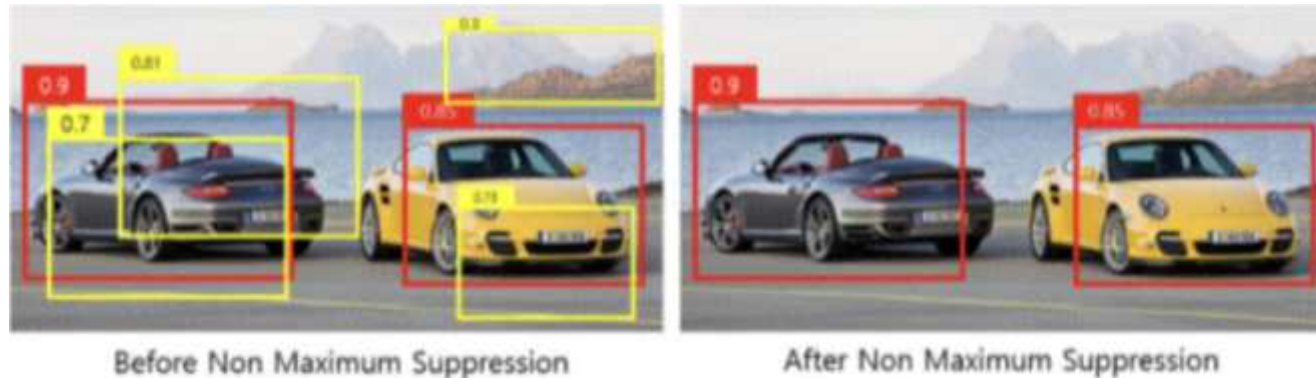
- AlexNet을 활용해 제안된 각 후보 영역별로 feature 4,096개를 추출
- 추출한 feature를 기반으로 SVM이 최종 분류 작업을 수행함
- 제안된 후보 영역(region proposal)에서 feature를 추출하려면 image 데이터를 CNN 구조에 삽입 가능한 일정한 크기로 변경하여야 함

2.1 Module design

Warp

- 해당 논문에서는 227×227 픽셀로 영역의 크기를 고정
- warping 하기 전에 16 픽셀만큼 주변 배경을 살려서 warping
- -feature은 227×227 RGB image를 5개의 convolution layer와 2개의 fc layer를 forward propagating하여 계산
- 2000개의 region 후보를 모두 고정된 길이의 고정된 size로 변경함
- 원래 이미지의 크기와 가로, 세로 비율은 고려하지 않음
- warp하여 고정된 사이즈로 만든 image들을 pre-train 된 CNN 구조를 통해 고정된 길이의 feature vector을 추출함

2.2 Test-time detection



- 2000개의 region proposal를 추출하기 위해 테스트 이미지에서 selective search 진행
 - feature를 계산하기 위해 각 proposal을 warp 하고 CNN을 통해 전달
 - 각 클래스에 대해 해당 클래스에 대해 훈련된 SVM을 사용하여 추출된 각 feature vector의 점수를 매김
 - 이미지에 대한 모든 scored regions가 주어지면, NMS를 이용하여 임계값 이상인 영역을 제거
- Non Maximum suppression (NMS) : bounding box 가운데 가장 확실한 bounding box만 남기고 나머지 bounding box는 제거하는 기법
- 점수가 높은 후보 영역 bounding box를 기준으로 IoU가 특정 임계값을 넘는 다른 영역 박스는 모두 제거

2.3. Training

Supervised pre-training

- 대규모 보조 데이터셋(ILSVRC2012 분류 데이터셋)으로 R-CNN을 사전 훈련
- ILSVRC2012 데이터셋에는 경계 박스 정보가 없으므로, 오직 이미지 레이블만 사용해 CNN을 훈련

2.3. Training

Domain-specific fine-tuning

- 사전 훈련한 CNN 모델을 새로운 객체 탐지 영역에 적용하기 위해, 해당 도메인에 특화되게끔 fine-tuning
- warping한 후보 영역 이미지만 활용해서 CNN 파라미터를 갱신
- ILSVRC2012의 ImageNet 데이터셋은 클래스가 총 1,000개 → CNN 모델의 최종 출력층은 1,000개 값을 내놓음

객체 탐지에 맞게 구조를 바꾸려면 최종 출력층이 $(N + 1)$ 개 값을 내놓도록 변경

- CNN을 detection 그리고 새로운 도메인에 적용하기 위해서 warped region proposals만을 이용해서 SGD를 사용해서 CNN 파라미터들을 훈련

- object detection을 하기 위해 classification layer를 (Object class 개수(N) + background(1))으로 바꾸고 초기

2.3. Training

Domain-specific fine-tuning

- Positive Sample : ground-truth box와 IoU 값이 0.5이상인 region (class와 무관하게)
- Negative Sample : Positive Sample 이외의 나머지 region
- 128개의 Mini-Batch를 구성하기 위해 SGD iteration 마다, 모든 class의 positive sample 32개, 그리고 96개의 background(negative sample)를 사용한다.

2.3. Training

Object category classifiers

- Positive sample : 각 클래스별 object의 ground-truth bounding boxes. (실제 object의 박스)
- Negative sample : 각 클래스별 object의 ground-truth와 IoU가 0.3 미만인 region

positive인지 negative인지 구분하기 애매 → IoU 임계값을 활용해 문제를 해결

부록 B : 임계값을 서로 같게 설정하니 오히려 성능이 떨어짐.

성능 향상을 위해 IoU 임계값을 각각 0.5, 0.3으로 설정

2.3. Training

Object category classifiers

- feature 추출을 마친 뒤에는 선형 SVM으로 클래스를 분류
- 훈련 데이터가 너무 커서 메모리 용량이 꽉 차는 문제가 생겨, hard negative mining 기법을 적용

hard negative mining

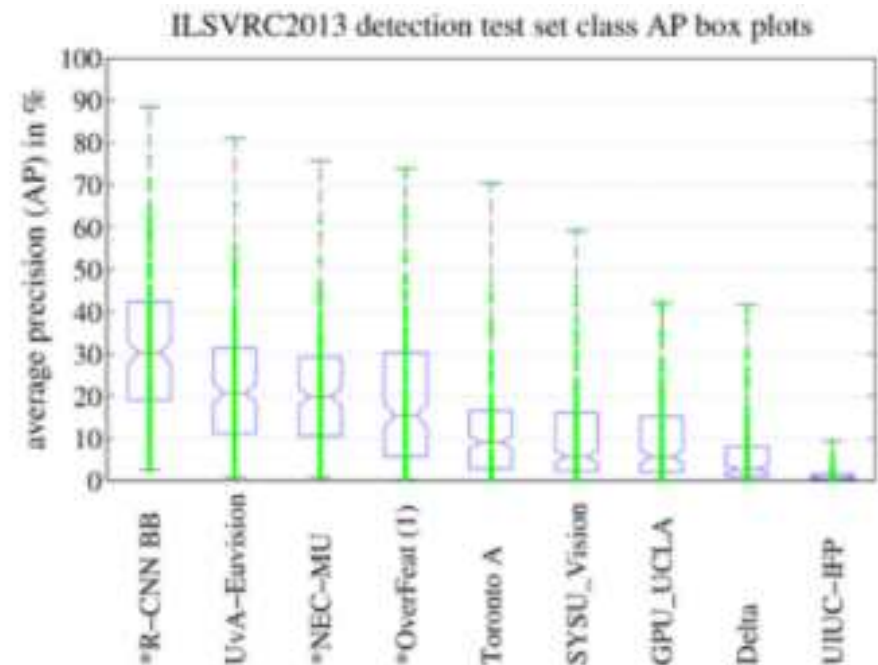
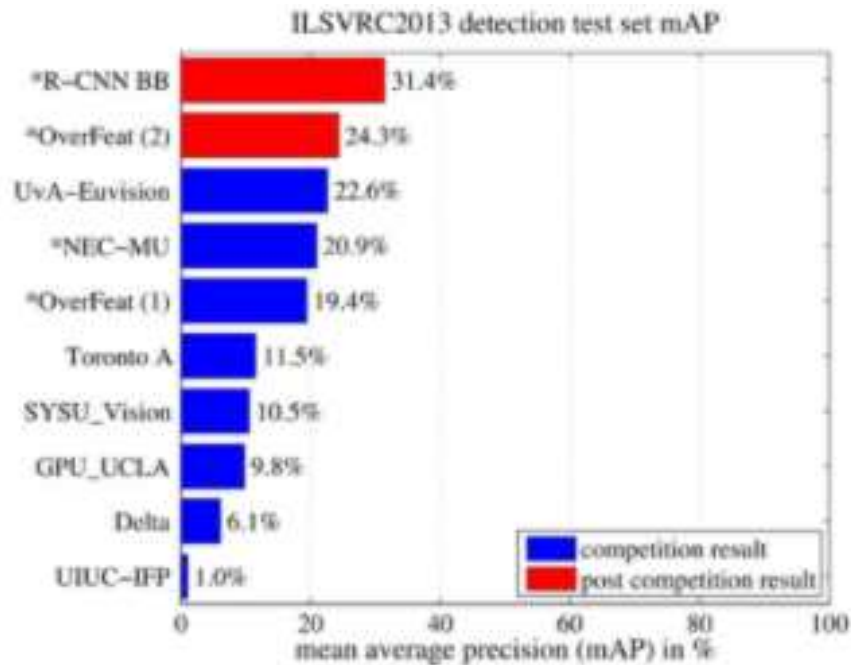
- positive 샘플과 negative 샘플의 개수를 균일하게 만드는 방법
- 신뢰도 점수(confidence score)를 활용해 negative 샘플을 선정

2.4. Results on PASCAL VOC 2010-12

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

- PASCAL VOC 2010 테스트 데이터셋으로 성능을 평가해본 결과, 다른 모델과 비교해서 성능이 꽤 높음
- 기존에 가장 성능이 좋던 SegDPM은 PASCAL VOC 2010에서 mAP 40.4%를 기록
- Bounding-box regression을 사용하지 않은 R-CNN은 50.2%, Bounding-box regression을 사용한 R-CNN은 53.7%의 mAP를 달성
- PASCAL VOC 2011-12에서도 53.3%로 높은 성능

Part 2 2.5. Results on ILSVRC2013 detection



- ASCAL VOC에서 사용한 파라미터를 그대로 사용하여 200개의 클래스를 갖는 ILSVRC2013 데이터셋에서도 R-CNN을 테스트
- R-CNN BB가 mAP 31.4%로 가장 높은 성능을 보임. 두 번째로 성능이 좋은 OverFeat의 mAP(24.3%)를 크게 앞섬
- R-CNN BB는 Bounding-box regression를 사용한 R-CNN을 뜻합니다.

**Part 3,
Visualization, ablation
and modes of error**



Experimental Results

3.1. Visualizing learned features

- first-layer filters, first features는 직관적으로 visualized 될 수 있고 이해하기 쉬움
(oriented edge, opponent colors)
- layer가 깊어지면서 차후의 layer에 대해서 이해하기 더 어려움
→ 이 논문에서는 non-parametric method를 이용
- non-parametric method로 non-maximum suppression을 사용
→ 가장 스코어가 높은 bounding box를 제외하고 다 삭제하는 것

Experimental Results



Figure 4: Top regions for six pool₅ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).

object를 분류할 때, 모양, 텍스처, 색상 및 재료의 특성에 영향을 받음

3.2. Ablation studies

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN pool ₅	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc ₆	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc ₇	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT pool ₅	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc ₆	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc ₇	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	54.2
R-CNN FT fc ₇ BB	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	58.5
DPM v5 [20]	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	33.7
DPM ST [28]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [31]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

- pool5 는 9 X 9 X 256으로 9216 차원이고, fc6은 4096차원, fc7은 4096차원

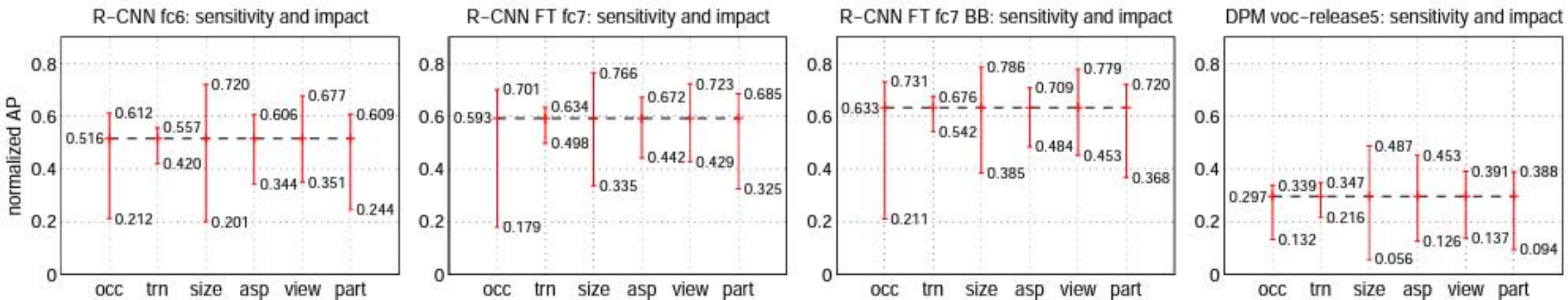
3.3. Network architectures

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN T-Net	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	54.2
R-CNN T-Net BB	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	58.5
R-CNN O-Net	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	35.7	62.1	64.0	66.5	71.2	62.2
R-CNN O-Net BB	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1	66.0

Table 3: Detection average precision (%) on VOC 2007 test for two different CNN architectures. The first two rows are results from Table 2 using Krizhevsky et al.'s architecture (T-Net). Rows three and four use the recently proposed 16-layer architecture from Simonyan and Zisserman (O-Net) [43].

- 어떤 architecture로 구성하느냐가 R-CNN에 큰 영향
- O-Net을 이용한 R-CNN이 T-Net을 이용한 R-CNN의 성능을 증가

3.4. Detection error analysis



- occlusion (occ), truncation (trn), bounding-box area (size), aspect ratio (asp), viewpoint (view), part visibility (part) 등의 문제가 있을 때의 성능
- DPM보다는 R-CNN이 더 나옴
- R-CNN 중에서도 fine-tuning, bounding-box regression을 활용했을 때 더 좋은 성능을 보임

3.5. Bounding-box regression

predicted box가 ground truth box와 유사하도록 학습하는 것.

($P \rightarrow G$ transform)

Predicted box

$$P^i = (P_x^i, P_y^i, P_w^i, P_h^i)$$

Predict bounding box

Ground truth box

$$G = (G_x, G_y, G_w, G_h)$$

Ground truth bounding box

(x 좌표, y 좌표, width, height)

3.5. Bounding-box regression

$d_x(P)$, $d_y(P)$, $d_w(P)$, and $d_h(P)$

bounding box transformation

$$\hat{G}_x = P_w d_x(P) + P_x \quad (1) \quad \boxed{\text{최0}}$$

$$\hat{G}_y = P_h d_y(P) + P_y \quad (2)$$

$$\hat{G}_w = P_w \exp(d_w(P)) \quad (3)$$

$$\hat{G}_h = P_h \exp(d_h(P)). \quad (4)$$

최0

bounding box transformation

최소연, 2023-01-06T10:12:34.855

3.5. Bounding-box regression

$$d_{\star}(P) = \mathbf{w}_{\star}^T \phi_5(P)$$

transformation 함수

$$\mathbf{w}_{\star} = \operatorname{argmin}_{\hat{\mathbf{w}}_{\star}} \sum_i^N (t_{\star}^i - \hat{\mathbf{w}}_{\star}^T \phi_5(P^i))^2 + \lambda \|\hat{\mathbf{w}}_{\star}\|^2.$$

최0

bounding box transformation

최소연, 2023-01-06T10:12:34.855

3.5. Bounding-box regression

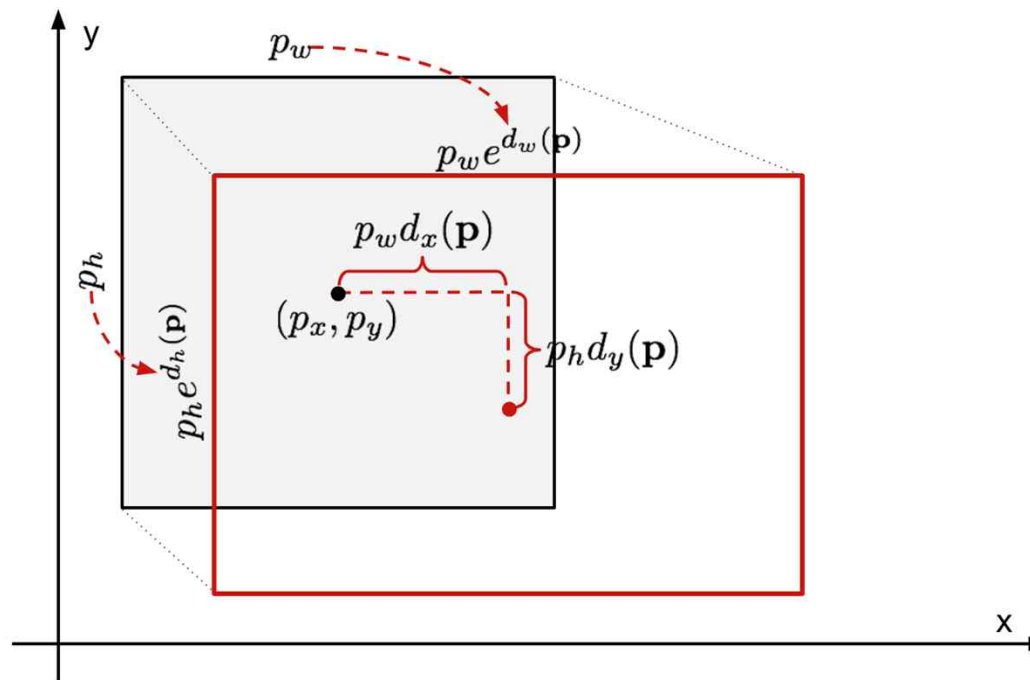
$$t_x = (G_x - P_x)/P_w \quad (6)$$

$$t_y = (G_y - P_y)/P_h \quad (7)$$

$$t_w = \log(G_w/P_w) \quad (8)$$

$$t_h = \log(G_h/P_h). \quad (9)$$

transformation 함수



최0

bounding box transformation

최소연, 2023-01-06T10:12:34.855

A modern interior space with a blue wall on the left and white horizontal-slatted walls on the right. Two large, black, dome-shaped pendant lights hang from the ceiling. In the foreground, a long wooden table is partially visible. In the background, there are large windows and glass doors with black frames. One of the glass doors has a blue and green curved graphic on it. The floor is made of light-colored wood.

Part 4, The ILSVRC2013 detection dataset

Part 4 **The ILSVRC2013 detection dataset**

4.1. Dataset overview

ILSVRC2013 데이터셋

- 훈련 데이터 395,918개, 검증 데이터 20,121개, 테스트 데이터 40,152

4.2. Region proposals

selective search 결과 이미지당 2403개의 region proposals가 생성.

Part 4 **The ILSVRC2013 detection dataset**

4.3. Training data

Training data는 R-CNN에서 3가지 부분에 요구됨

- (1) CNN fine-tuning
- (2) detector SVM training
- (3) bounding-box regressor training

4.4. Validation and evaluation

val1 + train 1k로 fine-tuned 된 CNN을 사용하여 fine-tuning과 feature computation이 재실행되는 것을 피함

Part 4 The ILSVRC2013 detection dataset

test set	val ₂	val ₂	val ₂	val ₂	val ₂	val ₂	test	test
SVM training set	val ₁	val ₁ +train _{5k}	val ₁ +train _{1k}	val ₁ +train _{1k}	val ₁ +train _{1k}	val ₁ +train _{1k}	val+train _{1k}	val+train _{1k}
CNN fine-tuning set	n/a	n/a	n/a	val ₁	val ₁ +train _{1k}	val ₁ +train _{1k}	val ₁ +train _{1k}	val ₁ +train _{1k}
bbox reg set	n/a	n/a	n/a	n/a	n/a	val ₁	n/a	val
CNN feature layer	fc ₆	fc ₆	fc ₆	fc ₇	fc ₇	fc ₇	fc ₇	fc ₇
mAP	20.9	24.1	24.1	26.5	29.7	31.0	30.2	31.4
median AP	17.7	21.0	21.4	24.8	29.2	29.6	29.0	30.3

Table 4: ILSVRC2013 ablation study of data usage choices, fine-tuning, and bounding-box regression.

4.5. Ablation study

training data, fine-tuning, and bounding-box regression의 양이 다름에 따른 결과

**Part 5,
Semantic segmentation**



Part 5 CNN features for segmentation.

	<i>full</i> R-CNN		<i>fg</i> R-CNN		<i>full+fg</i> R-CNN	
O_2P [4]	fc_6	fc_7	fc_6	fc_7	fc_6	fc_7
46.4	43.0	42.5	43.7	42.1	47.9	45.8

Table 5: Segmentation mean accuracy (%) on VOC 2011 validation. Column 1 presents O_2P ; 2-7 use our CNN pre-trained on ILSVRC 2012.


- full : region의 shape을 무시하고 CNN features를 warped window에 바로 연산
- fg : region 의 가장 앞쪽 mask에만 CNN features를 연산한다. 이 때, 배경을 평균으로 변경
- Full + fg : full과 fg를 합친 것

Results on VOC 2011

VOC 2011 test	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
R&P [3]	83.4	46.8	18.9	36.6	31.2	42.7	57.3	47.4	44.1	8.1	39.4	36.1	36.3	49.5	48.3	50.7	26.3	47.2	22.1	42.0	43.2	40.8
O ₂ P [4]	85.4	69.7	22.3	45.2	44.4	46.9	66.7	57.8	56.2	13.5	46.1	32.3	41.2	59.1	55.3	51.0	36.2	50.4	27.8	46.9	44.6	47.6
ours (full+fg R-CNN f ₆)	84.2	66.9	23.7	58.3	37.4	55.4	73.3	58.7	56.5	9.7	45.5	29.5	49.3	40.1	57.8	53.9	33.8	60.7	22.7	47.1	41.3	47.9

Table 6: Segmentation accuracy (%) on VOC 2011 test. We compare against two strong baselines: the “Regions and Parts” (R&P) method of [3] and the second-order pooling (O₂P) method of [4]. Without any fine-tuning, our CNN achieves top segmentation performance, outperforming R&P and roughly matching O₂P.

- full : region의 shape을 무시하고 CNN features를 warped window에 바로 연산
- 이전의 방법론 보다는 full+fg R-CNN fc6이 성능이 더 좋다.

A modern interior space featuring a bright blue wall on the left and white horizontal wooden slats on the right. Two large, black, dome-shaped pendant lights hang from the ceiling. In the foreground, a long wooden table is partially visible. In the background, there are large windows and glass doors with black frames. One of the glass doors has a blue and green curved graphic. The floor is made of light-colored wood. A semi-transparent blue rectangle is overlaid on the left side of the image, containing the text "Part 6, Conclusions".

Part 6, Conclusions

Conclusions

- object를 localize하고, 분할하기 위하여 bottom-up Region proposal을 CNN에 적용
- label 된 training data 가 부족한 경우 대규모 CNN을 훈련하기 위한 paradigm
- 파스칼 VOC 2012 에서 이전의 최상의 결과에 비해 30% 향상된 결과를 보여줌
- region proposal 에 대한 CNN 학습, SVM Classification, Bounding Box regression을 통하여 이전의 object detection 방법론들보다도 큰 성능을 보임
- 연산 속도는 overfeat에 비해 느림, 그러나 detection task에 있어서는 2배 이상 성능 향상
- 이후 R-CNN을 수정, 보완하여 성능과 속도를 향상시킨 수많은 모델들이 탄생하는 등 다양한 모델의 기초가 됨

Conclusions

- Fine-tuning 된 CNN으로 image를 넣기 위하여 고정된 크기의 이미지로 warp 과정을 통해 image가 손상된다는 단점이 있습니다.
- 학습이 Conv Fine-tuning, SVM classification, Bounding box regression 총 3단계로 이루어지는데, 이 과정에서 긴 학습 시간과 대용량의 저장 공간이 요구되고, 학습 시간이 너무 길다는 단점이 있습니다.
- 2000개의 region에 대해 각각 CNN을 수행하기에 실제 학습시간이 너무 길다는 단점이 있습니다.