INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

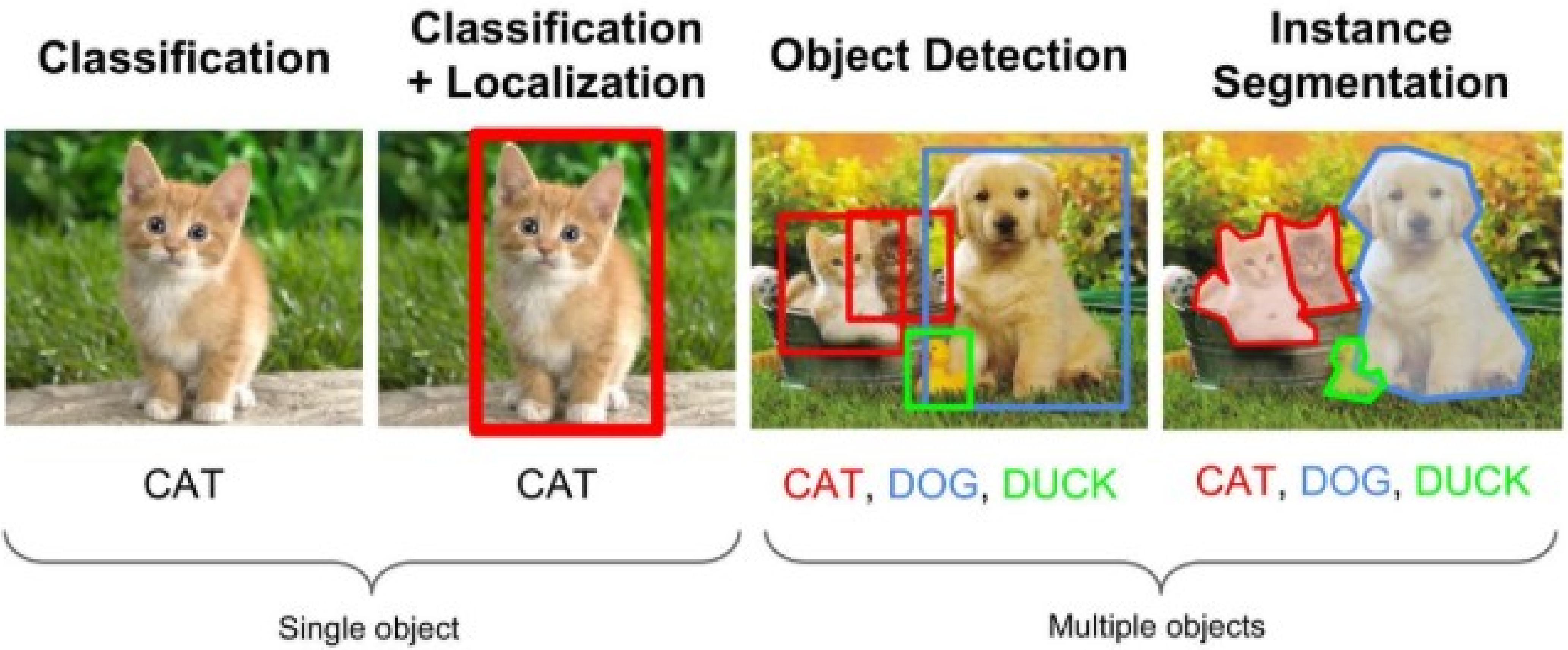
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Undergraduate Researcher at CVLab
Lee Dohyeong

2025.1.17

Contents

Introduction – Task

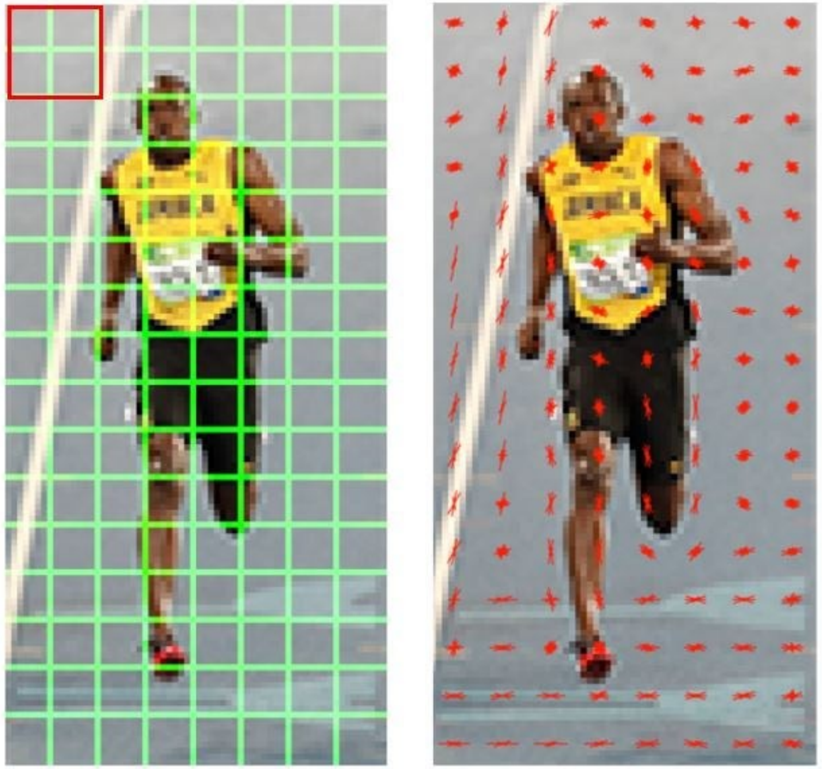


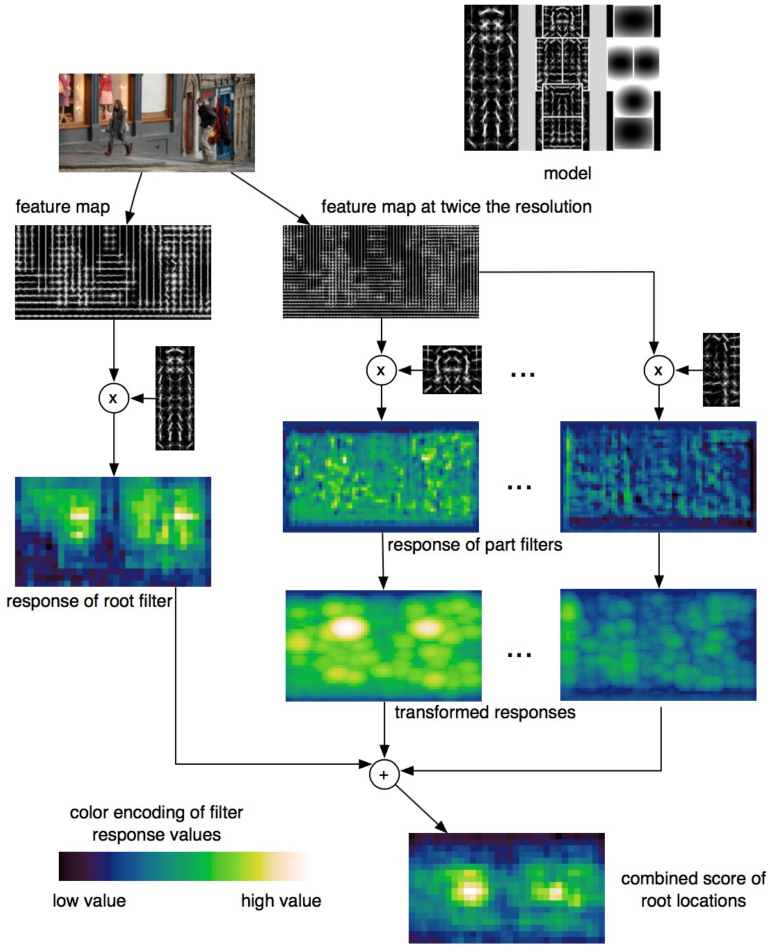
Introduction – Sliding Window

- Sliding Window
  - Fixed-size window is moved across an image at regular intervals
  - Disadvantages : High Cost, Inefficiency

Introduction

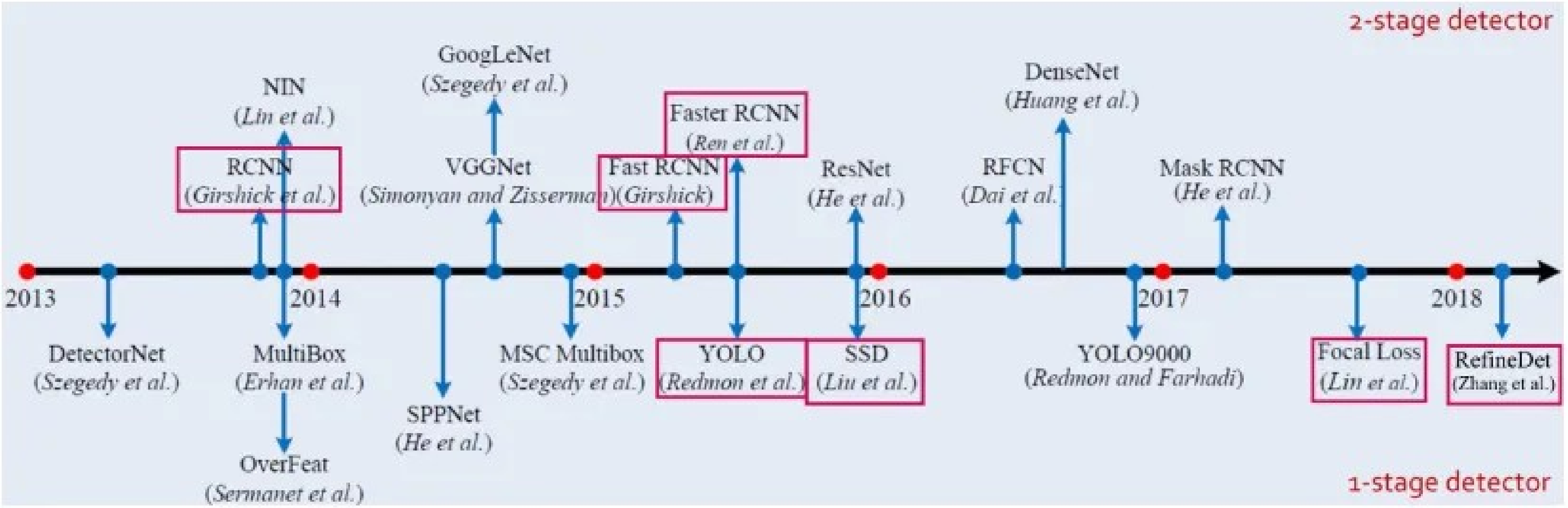
- HOG, Deformable Part Model
- HOG : local gradient orientation information in an image
- DPM : Divides an object into parts



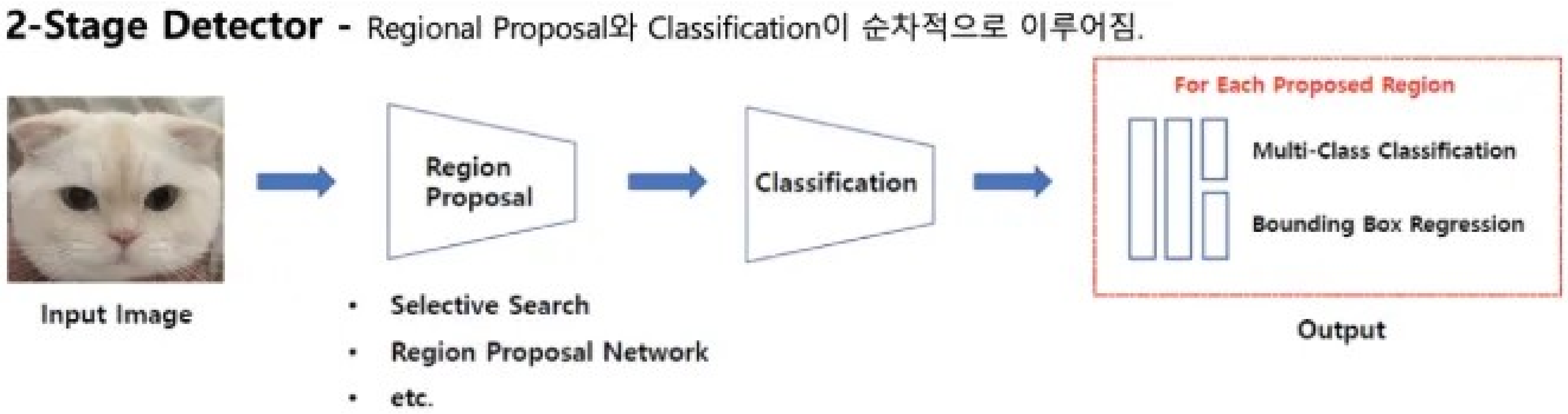


Introduction

- Object Detection
  - 1 – stage Detector
  - 2 – stage Detector



Introduction – 2 Stage Detector



Introduction – Selective Search

● Selective Search：

1. Divide the image into small segments.
2. Merge segments with similar characteristics (color, texture, size, etc.)
   - Using a Greedy Algorithm.
3. Generate candidate regions of various sizes and shapes during the merging process.
4. Select regions with a high likelihood of containing objects.

Advantages:

- Efficiency：Specific regions
- Diversity：Considering diverse features

Disadvantage:

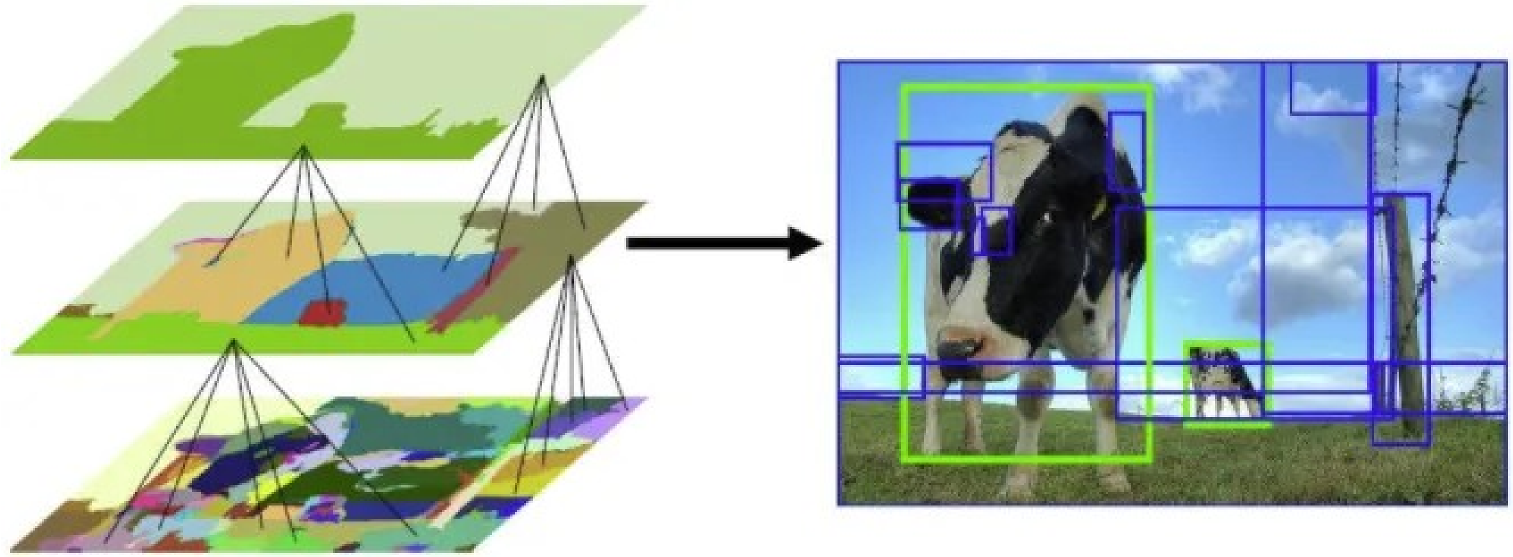
- Not end-to-end , Difficult Real time

(b)

Introduction

- ROI(Region of Interest)

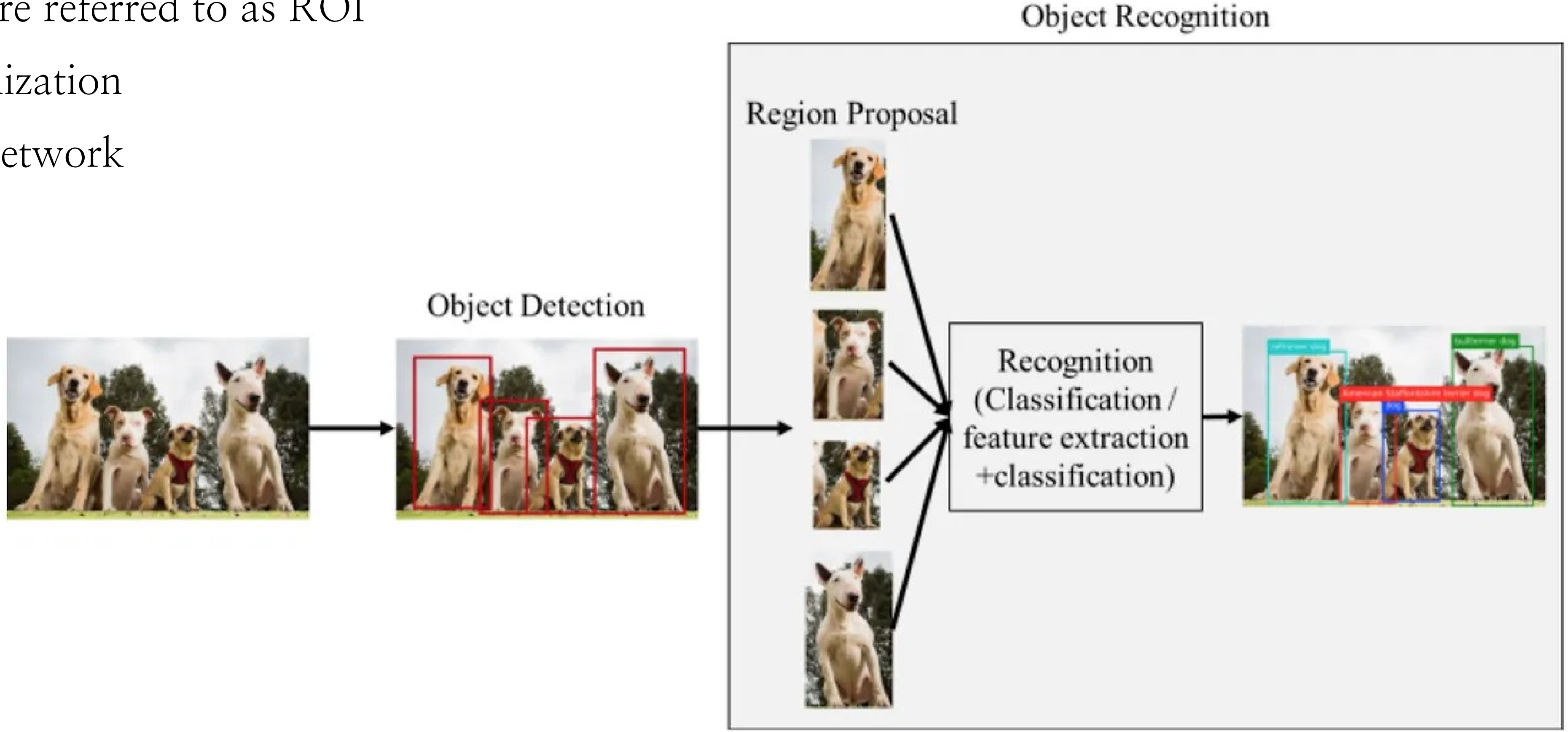
– Using the Selective Search algorithm, ROI candidate regions are generated.



Introduction

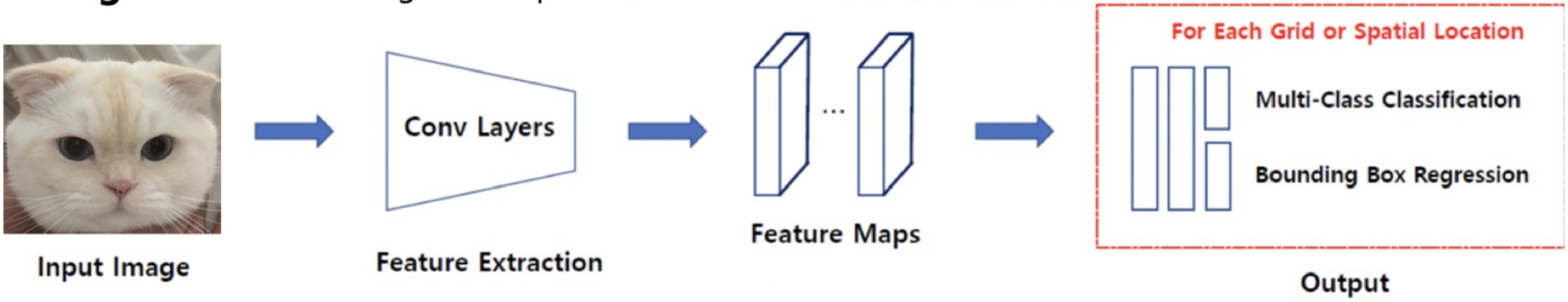
- 2 Stage Detector Process

1. Region Proposal : Selective Search, Region Proposal Network
   - Extracted regions are referred to as ROI
2. Classification & Localization
   - Use Convolution Network



Introduction – 1 Stage Detector

1-Stage Detector - Regional Proposal와 Classification이 동시에 이루어짐.



Ex) **YOLO 계열** (YOLO v1, v2, v3) , **SSD 계열** (SSD, DSSD, DSOD, RetinaNet, RefineDet ...)

Introduction

- 1, 2 Stage Detector

1 Stage Detector:

- Relatively fast, low accuracy(difficult to capture fine details)

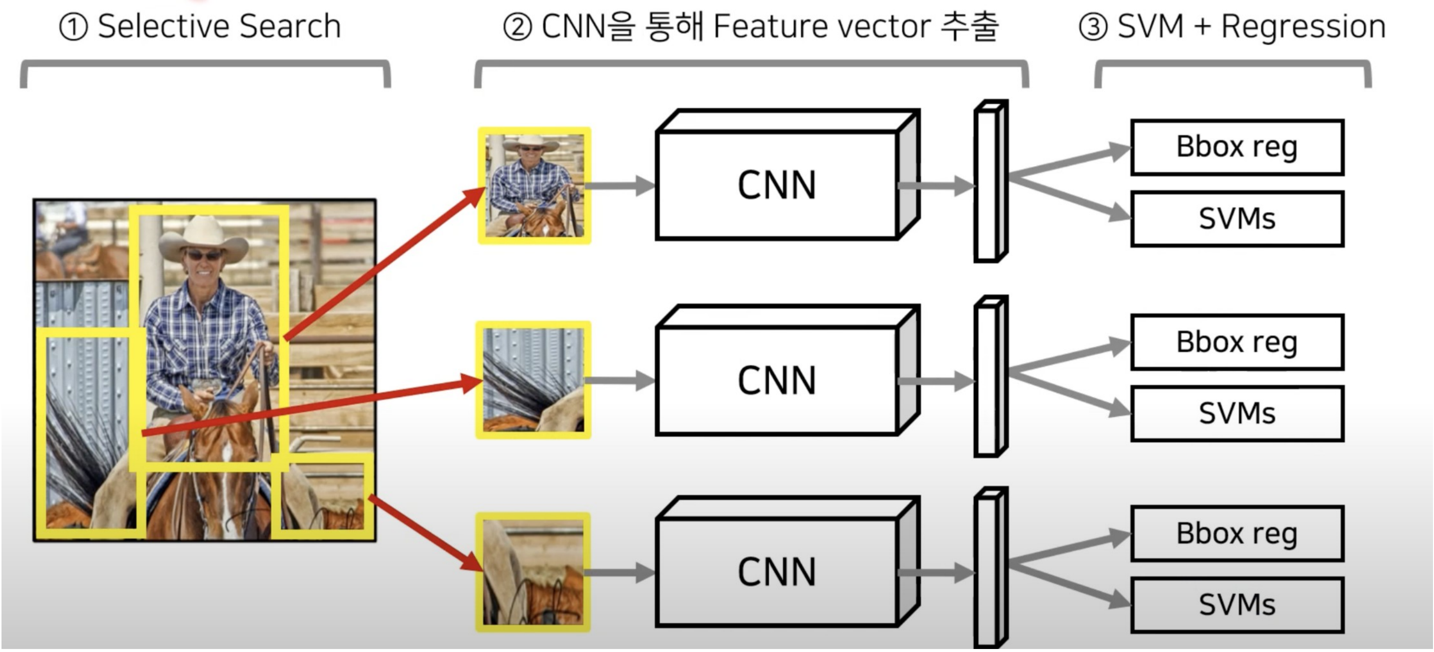
- imbalanced(Background > Object)

2 Stage Detector:

- Slower(complex structure), high accuracy

Related Work

- R- CNN
  - Region proposal : Selective Search Algorithm(2k)
  - Warping: Region proposals into fixed-size inputs
  - CNN : 2,000 images are fed into the CNN for processing.
    - Classification : SVM
    - Regression : B box Reg

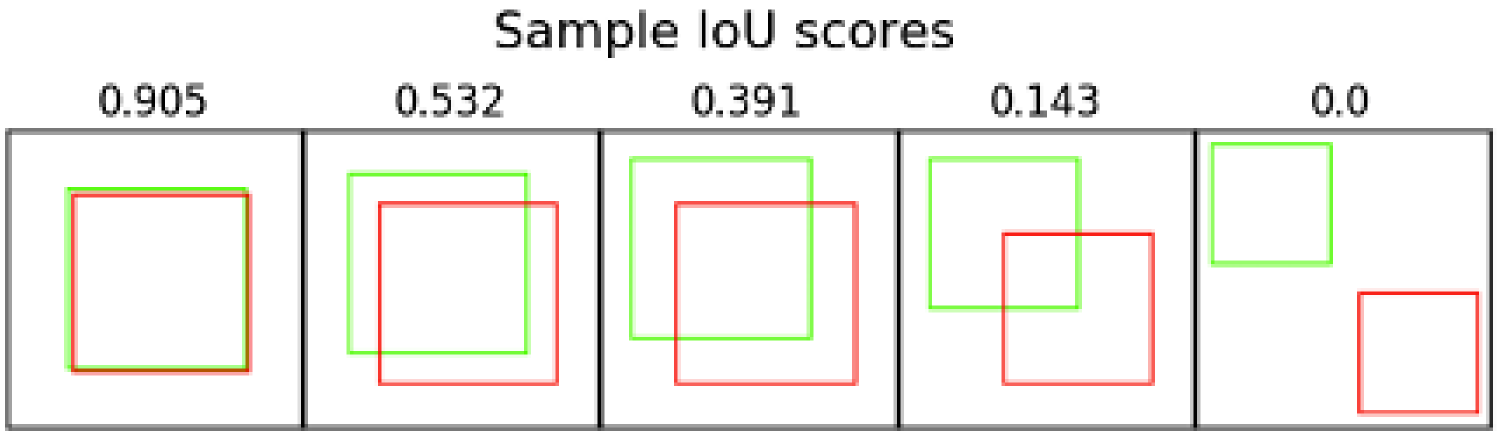


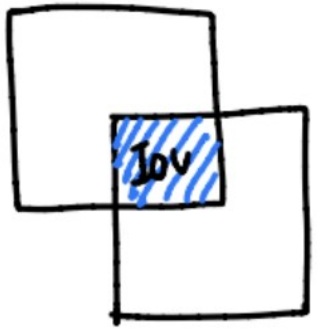
Related Work

- R‑ CNN
  - IoU : Intersection over Union
  - Selective search 2000k

IoU filitering

- IoU >= 0.5 : positive
- IoU < 0.5 : negative

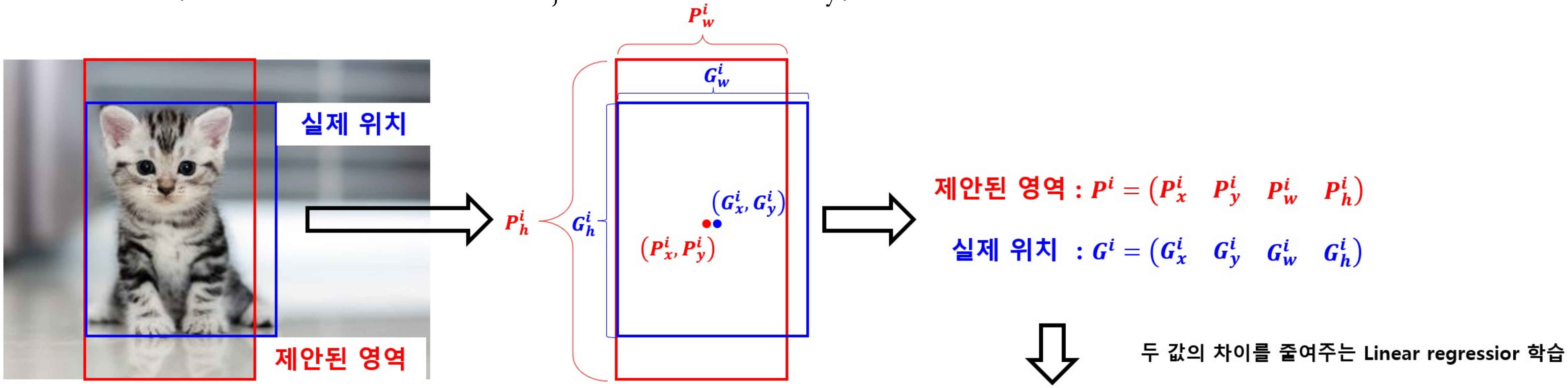




$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Related Work

- R – CNN
  - Bounding Box Regression(Bbox Regression)
    1. Regions suggested by Selective Search don't perfectly match
    2. Adjusts these regions to better align with the objects.
    3. Make the boxes fit the objects more accurately.

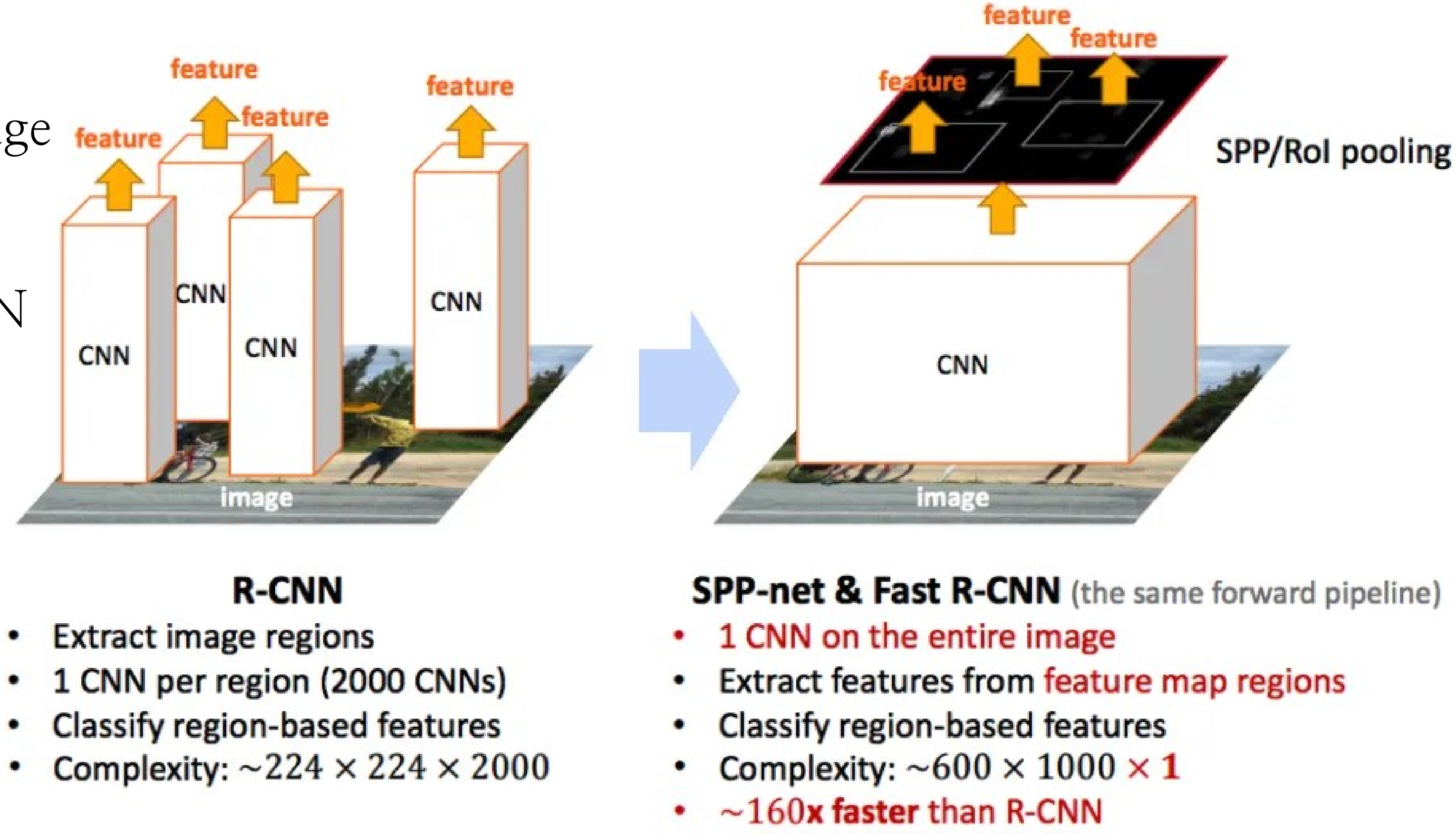


Related Work

- R – CNN
  - Problem: Time cost
    - Selective Search: -> CPU bottleneck
    - Independent training: Region Proposal, SVM, Regression -> not end-to-end
    - Each 2k proposal passing CNN >> Time, Computational Cost

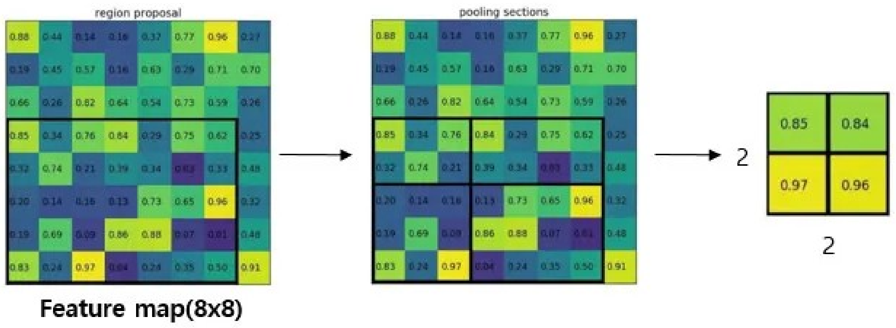
Related Work

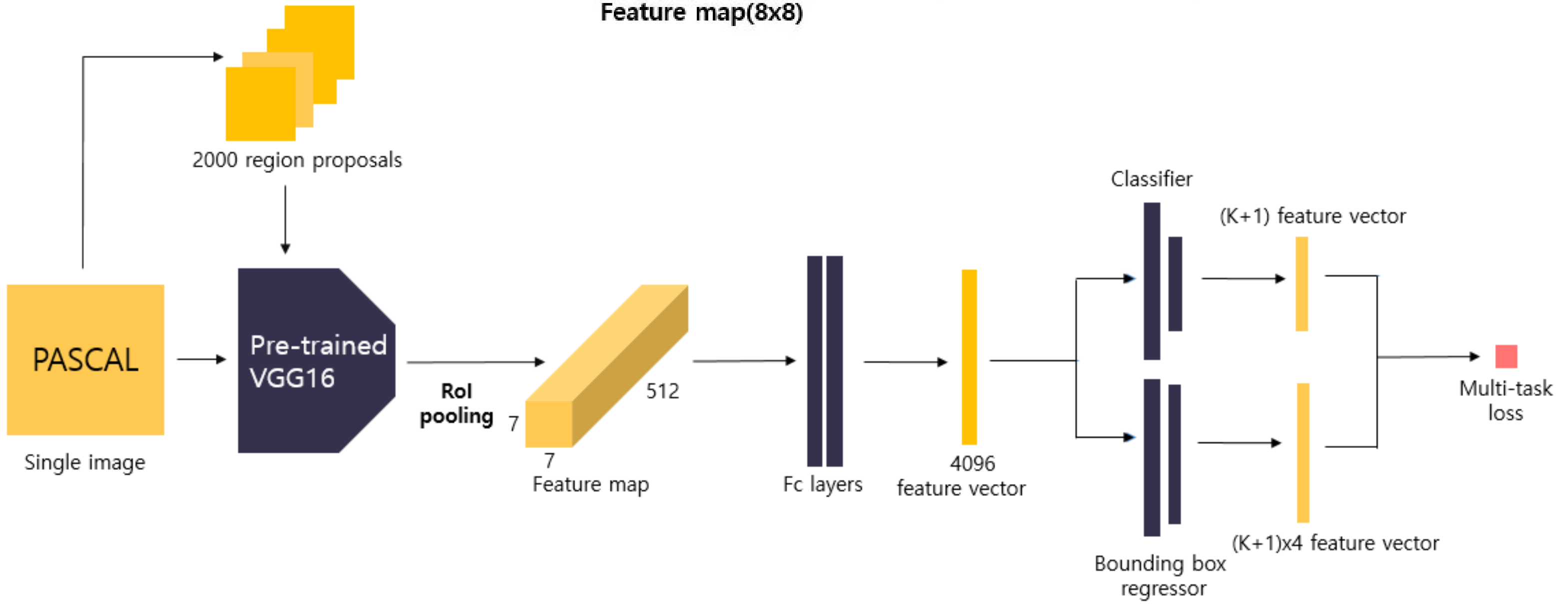
- Fast R – CNN
  - Difference Between R-CNN
    - 1 CNN on the entire Image
    - ROI Pooling
    - 160 x faster than R-CNN



Related Work

- Fast R – CNN
  - Region proposal : Selective Search -> Pre-trained model
  - ROI pooling : Replace the max-pooling layer
  - Multi-task loss
    - Classifier : SVM
    - Regressor : B Box Regressor





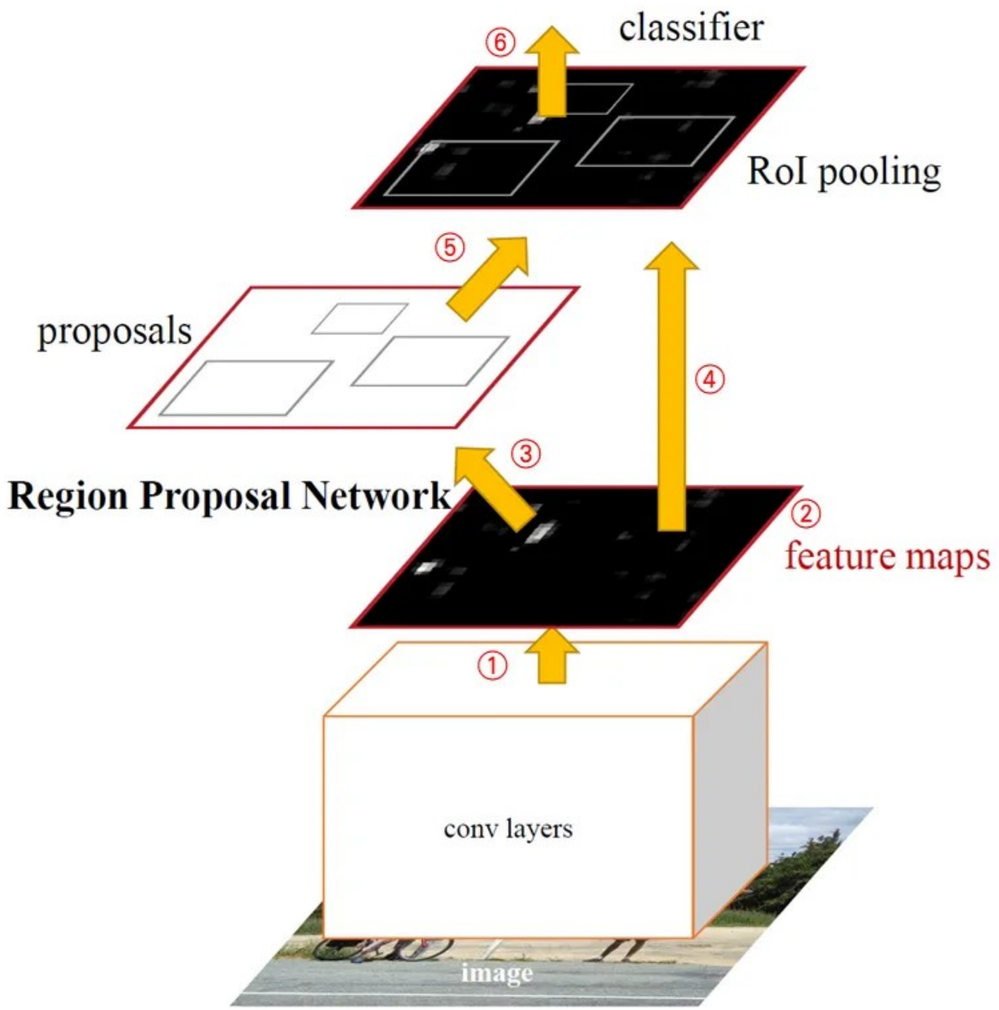
Related Work

- Fast R – CNN
  - Resolution：
    - ROI Pooling 〉〉 Computational Efficiency
    - End to end training >> Group RoI pooling, Region Classification, B Box Re-gression
  - Problems：
    - Selective Search：Using cpu 〉〉 bottleneck

Method

- Faster R – CNN

  1. CNN
  2. Feature map -> RPN
  3. RPN -> Region proposal(Anchor)
  4. Feature map -> RoI Pooling
  5. Region Proposal -> RoI pooling
  6. Classifier(Classification, Regression)

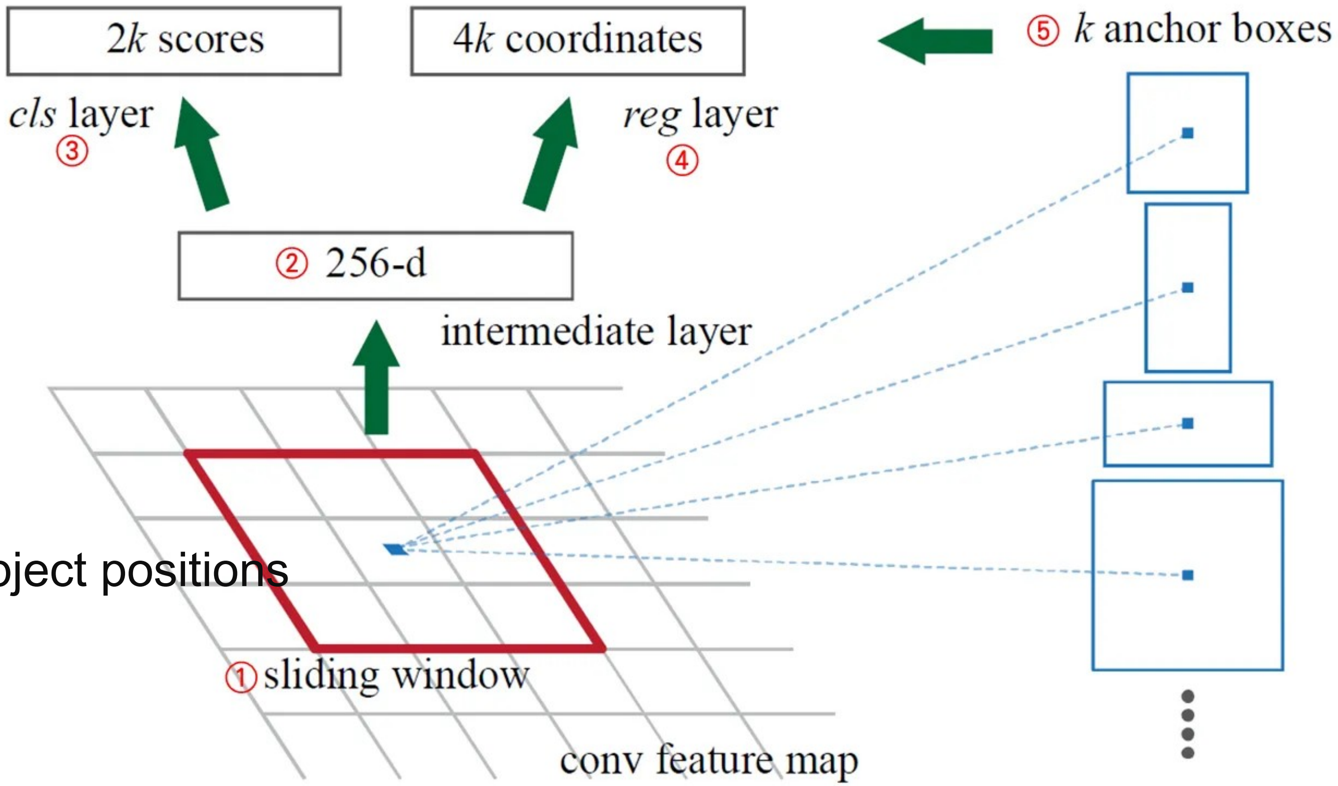


Method

- Faster R － CNN
  - RPN(Region Proposal Network)
    - K(=9) Anchor Boxes
    - Sliding window on feature map
    - Cls layer：Object or Background

      2k scores(positive or negative)
    - Reg layer：Bbox ⟩better match the object positions

      4k coordinates(dx, dy, dw, dh)



Method

- Faster R – CNN
  - Anchor Box
    - Objects of Different Sizes and Ratios
    - Facilitating Bounding Box Refinement

Method

- Faster R － CNN
  - Loss Func
    - Cls
      - Positive label : Ground Truth Box + IoU>=0.7 OR Ground Truth Box + Most high IoU Anchor
      - Negative label : IoU < 0.3
    - Reg

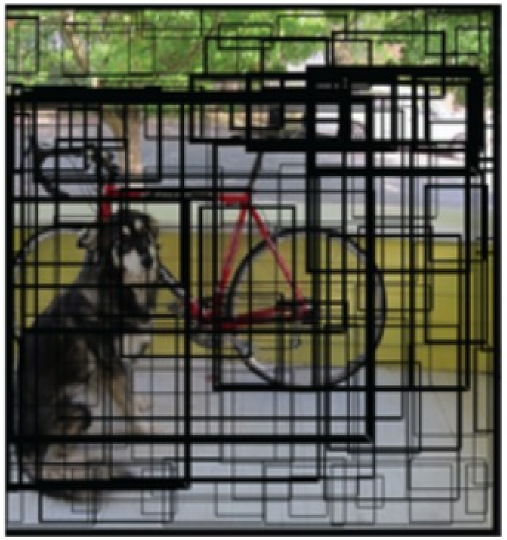
    － Bounding Box Refinement

    － Compare Predicted bounding box and ground truth box

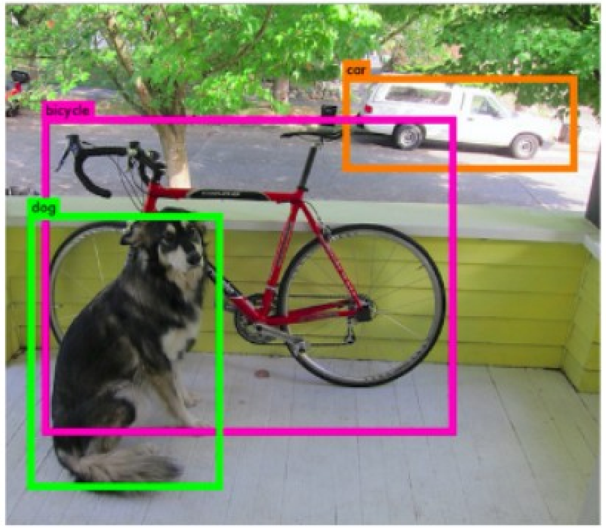
$$L(\{p_i\},\{t_i\}) = \overbrace{\underset{⑦}{\frac{1}{N_{cls}}} \sum_i \underset{⑤}{L_{cls}}(\underset{①}{p_i}, \underset{②}{p_i^*})}^{\text{분류 손실}} + \overbrace{\underset{⑨}{\lambda} \underset{⑧}{\frac{1}{N_{reg}}} \sum_i p_i^* \underset{⑥}{L_{reg}}(\underset{③}{t_i}, \underset{④}{t_i^*})}^{\text{회귀 손실}}.$$

Method

- Faster R – CNN
  - NMS(Non-Maximum Suppression)
    - Removes duplicate boxes
      - Boxes with IoU above the threshold are removed.



Multiple Bounding Boxes



Final Bounding Boxes

Method

- Faster R – CNN
  - Training RPN
    - SGD, back-propagation >> End to end(Learn everything from input to output at once)
  - Sampling
    - Randomly sample 256 anchors
    - Positive(object, Iou>0.7) : Negative(background < 0.3) = 1 : 1

Experiments

- Faster R – CNN
  - mAP(mean Average Precision)：evaluates the **precision and recall of object detection models**
    - **Better mAP : Selective Search < RPN**
    - **Variable data set -> Better mAP**

Pascal Voc 2007：train/test images 5000,5000

train-time region proposals method	# boxes	test-time region proposals method	# proposals	mAP (%)
① SS	2000	SS	2000	58.7
② EB	2000	EB	2000	58.6
③ RPN+ZF, shared	2000	RPN+ZF, shared	300	**59.9**

method	# proposals	data	mAP (%)
SS	2000	07	66.9†
SS	2000	07+12	70.0
RPN+VGG, unshared	300	07	68.5
RPN+VGG, shared	300	07	69.9
RPN+VGG, shared	300	07+12	**73.2**
RPN+VGG, shared	300	COCO+07+12	**78.8**

Experiments

- Faster R – CNN
  - FPS by Model
    - VGG : RPN usage offers higher FPS compared to Selective Search.
    - ZF : Increased FPS compared to the VGG model.

model	system	conv	proposal	region-wise	total	rate
VGG	SS + Fast R-CNN	146	1510	174	1830	0.5 fps
VGG	RPN + Fast R-CNN	141	**10**	47	**198**	**5 fps**
ZF	RPN + Fast R-CNN	31	**3**	25	**59**	**17 fps**

Experiments

- Faster R – CNN
  - Hyper parameter: learning rate, batch size, anchor size ratio
    - mAP that is independent of hyperparameters.
  - mAP based on anchor scales and ratios.

λ	0.1	1	10	100
mAP (%)	67.2	68.9	69.9	69.1

settings	anchor scales	aspect ratios	mAP (%)
1 scale, 1 ratio	128^2	1:1	65.8
	256^2	1:1	66.7
1 scale, 3 ratios	128^2	{2:1, 1:1, 1:2}	68.8
	256^2	{2:1, 1:1, 1:2}	67.9
3 scales, 1 ratio	$\{128^2, 256^2, 512^2\}$	1:1	**69.8**
3 scales, 3 ratios	$\{128^2, 256^2, 512^2\}$	{2:1, 1:1, 1:2}	**69.9**

Experiments

- Faster R – CNN
  - Analysis of Recall to IoU
    - Recall : The proportion of ground-truth objects covered by proposals with IoU.

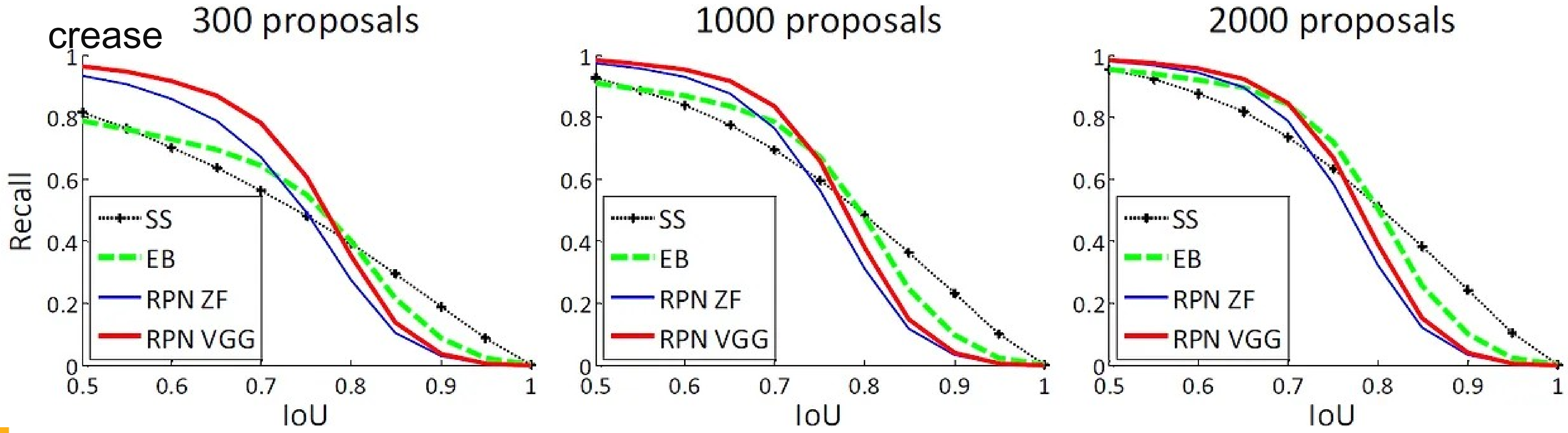
예측 결과 (Predict Result)

실제 정답 (Ground Truth)		Positive	Negative
	Positive	TP (True Positive) 있다고 올바르게 판단	FN (False Negative) 없다고 잘못 판단
	Negative	FP (False Positive) 있다고 잘못 판단	TN (True Negative) 없다고 올바르게 판단

- Precision (정확도)
  - 올바르게 탐지한 물체의 수(TP) / 모델이 탐지한 물체의 개수(TP + FP)
- Recall (재현율)
  - 올바르게 탐지한 물체의 수(TP) / 실제 정답 물체의 수(TP + FN)

Experiments

- Faster R – CNN
  - Analysis of Recall to IoU
    - Evaluate the **quality of proposal boxes**
    - Diffrence **proposal methods** (SS, EB, RPN) and **model types** (ZF, VGG)
    - Relationship Between Proposal and Recall : Proposal 300 → 1000 → 2000 >> Recall in-crease



Conclusion

R CNN
: Selective Search방식 사용하고 CNN 사용을 통해서 높은 정확도를 달성하였습니다

FAST R CNN
: ROI Pooling을 통해서 속도를 개선하였으나, Selective Search방식을 사용하기 때문에 느린 문제

Faster R CNN
1. RPN도입으로 end to end 학습이 가능하게 되어 속도와 정확도 모두 올릴 수 있었습니다.
2. 적은 데이터셋에서 뛰어난 성능을 보입니다.
3. Anchor Box을 이용하기 때문에 다양한 객체 검출이 가능합니다.