

AL5.A.01 - Techniques IA

Deep learning-based object detection

Minh-Tan Pham

BUT3-INFO, IUT Vannes, Université Bretagne-Sud

minh-tan.pham@univ-ubs.fr

Table of contents

1 Object detection: introduction

2 Two-stage detectors

3 One-stage detectors

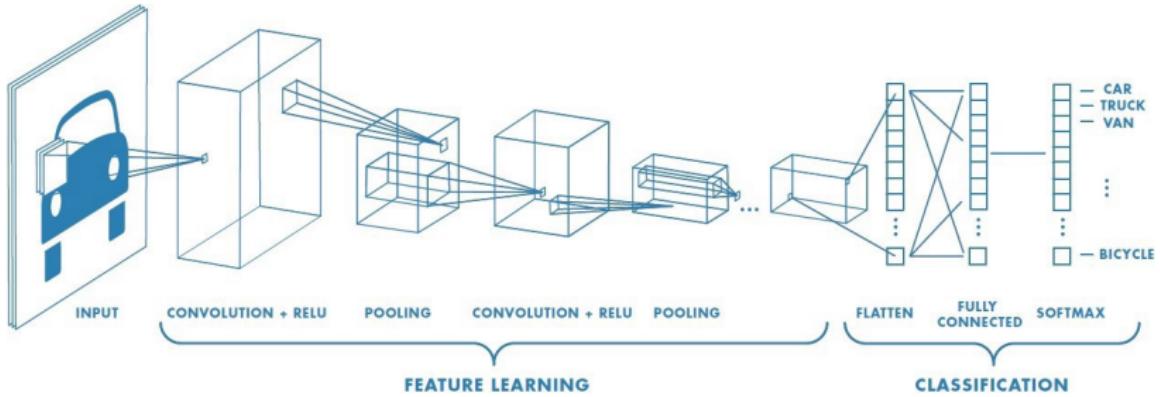
Table of contents

1 Object detection: introduction

2 Two-stage detectors

3 One-stage detectors

Introduction So far: image classification using CNN



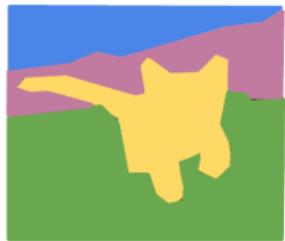
Maybe some remarks? questions?

Introduction

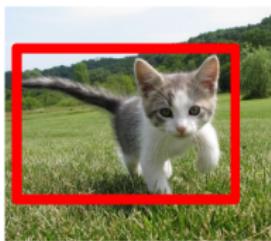
But: that's just the most simple task!

Other computer vision tasks:

Semantic Segmentation



Classification + Localization



Object Detection



Instance Segmentation

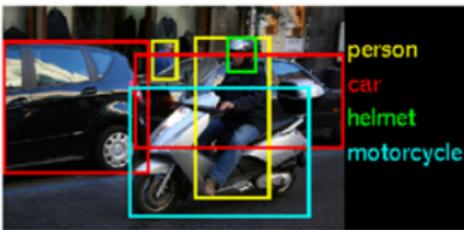
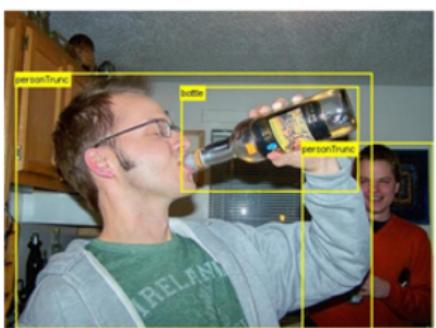


Applications of object detection:

- Computer vision:
 - facial detection and recognition
 - industrial quality check
 - autonomous driving cars
 - people counting and tracking, etc.
- Remote sensing:
 - wildlife animal detection
 - vehicle detection and counting from space
 - ship detection, etc.
- And many others in medical imaging, robotics, astronomy, etc.

Introduction Data annotation

In general: bounding box + class for each object within the image¹



¹Zhengxia Zou et al. Object Detection in 20 Years: A Survey

In computer vision²

Dataset	train		validation		trainval		test	
	images	objects	images	objects	images	objects	images	objects
VOC-2007	2,501	6,301	2,510	6,307	5,011	12,608	4,952	14,976
VOC-2012	5,717	13,609	5,823	13,841	11,540	27,450	10,991	-
ILSVRC-2014	456,567	478,807	20,121	55,502	476,688	534,309	40,152	-
ILSVRC-2017	456,567	478,807	20,121	55,502	476,688	534,309	65,500	-
MS-COCO-2015	82,783	604,907	40,504	291,875	123,287	896,782	81,434	-
MS-COCO-2018	118,287	860,001	5,000	36,781	123,287	896,782	40,670	-
OID-2018	1,743,042	14,610,229	41,620	204,621	1,784,662	14,814,850	125,436	625,282

²Zhengxia Zou et al. Object Detection in 20 Years: A Survey

How to evaluate the predicted bounding boxes ?

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


(a) Intersection over Union (IoU)



(b) $\text{IoU} > 0.5$
(accept)



(c) $\text{IoU} < 0.5$
(reject)

What could be other criteria to evaluate the predicted bounding boxes?

Non Maximum Suppression: If predicted bounding boxes overlap, only consider the most confident.

Bounding Box Regression: Regression used to find bounding box parameters (usually the box's center, width and height)

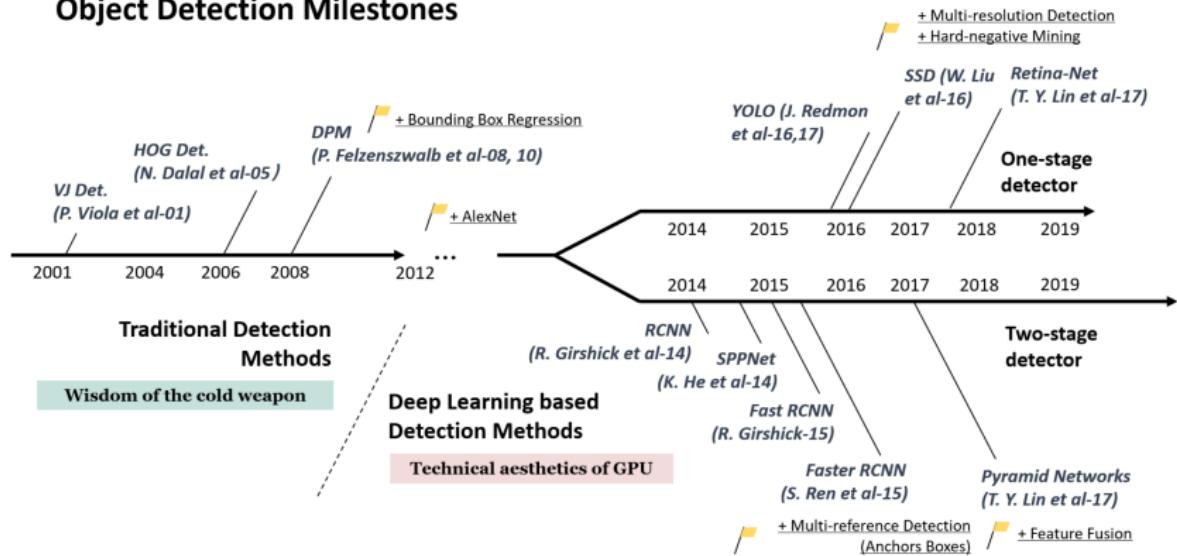
Example Loss Function

$$\sum_{i \text{ in } I} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2]$$

I is the set of all matching bounding boxes (Highest IoU with ground truth)

A road map of object detection in computer vision³

Object Detection Milestones



³Zhengxia Zou et al. Object Detection in 20 Years: A Survey

Before deep learning: traditional detection methods

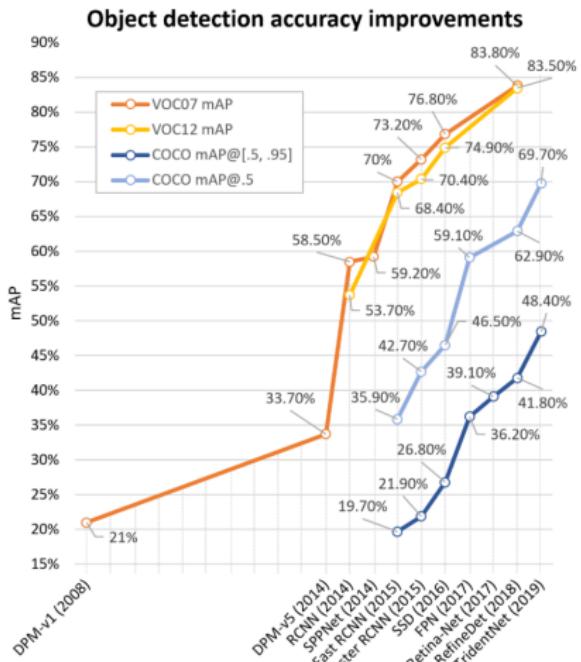
- Hand-crafted feature + classifier
- Features: HOG (Histogram of Gradients), LBP (Local Binary Pattern), textural features, etc.
- Classifier: SVM, Random Forest, etc.

Modern deep-learning object detection³

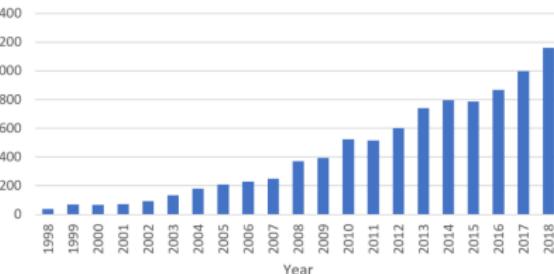
- Two-stage detectors
- One-stage detectors

³<http://zoey4ai.com/2018/05/12/deep-learning-object-detection/>

The evolution of deep-learning based object detection³



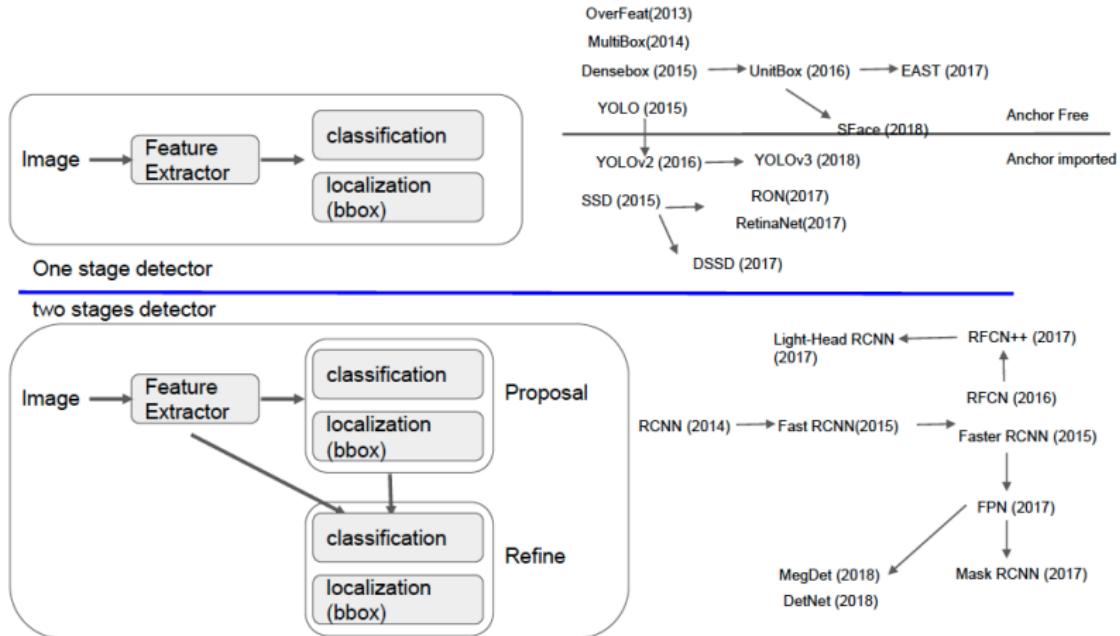
Number of Publications in Object Detection



³Zhengxia Zou et al. Object Detection in 20 Years: A Survey

Introduction State-of-the-art

One-stage vs Two-stage detectors³



³Image source: Gang Yu, An Introduction to Modern Object Detection

Table of contents

1 Object detection: introduction

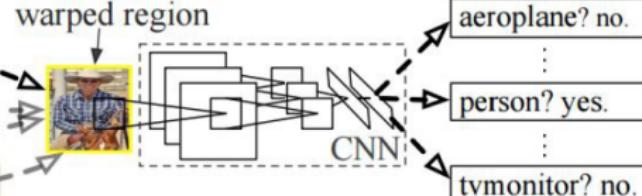
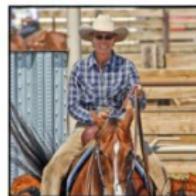
2 Two-stage detectors

3 One-stage detectors

R-CNN: Region proposal + CNN⁴

- Use **Selective Search** to come up with regional proposal
- Then classification + bounding box regression for each proposal
- First object detection method using CNN

R-CNN: *Regions with CNN features*



1. Input image

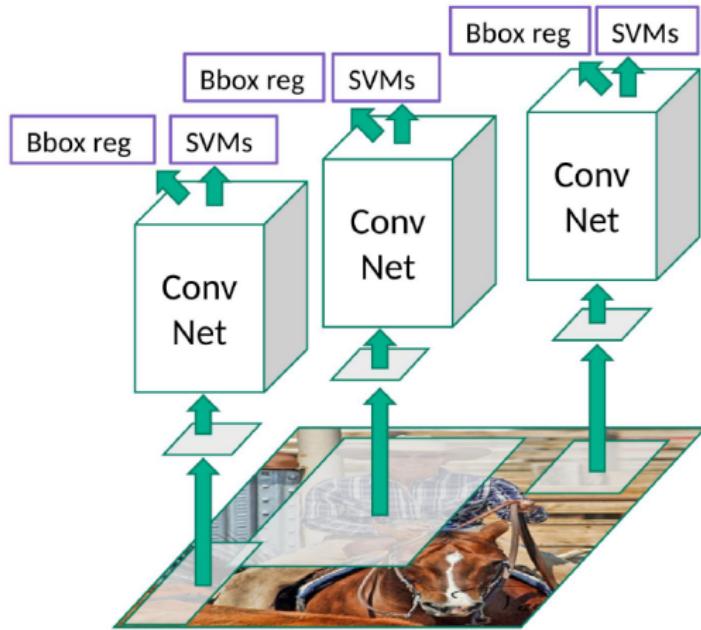
2. Extract region proposals (~2k)

3. Compute CNN features

4. Classify regions

⁴Ross Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.

R-CNN: Region proposal + CNN⁴



⁴Ross Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.

Two-stage detectors R-CNN

Example:⁵

extract feature



Crop & Warp



Convolution
and Pooling

Store all the features
after pool 5 layer and
save to disk

It is about ~ 200G
features

⁵Sihao Liang, Jiajun Lu and Kevin Perkins, Lecture on Object detection.

Two-stage detectors R-CNN

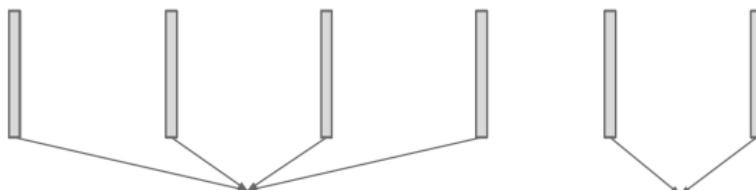
Example:⁵

train SVM for each class

Crop /
Warp
image



Features
from last
step



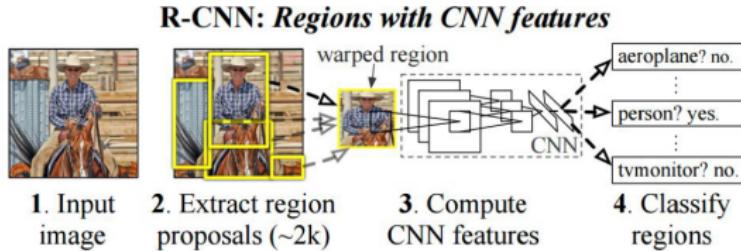
Negative
samples for
Bicycle

Positive
samples for
Bicycle

⁵Sihao Liang, Jiajun Lu and Kevin Perkins, Lecture on Object detection.

R-CNN: Summary⁶

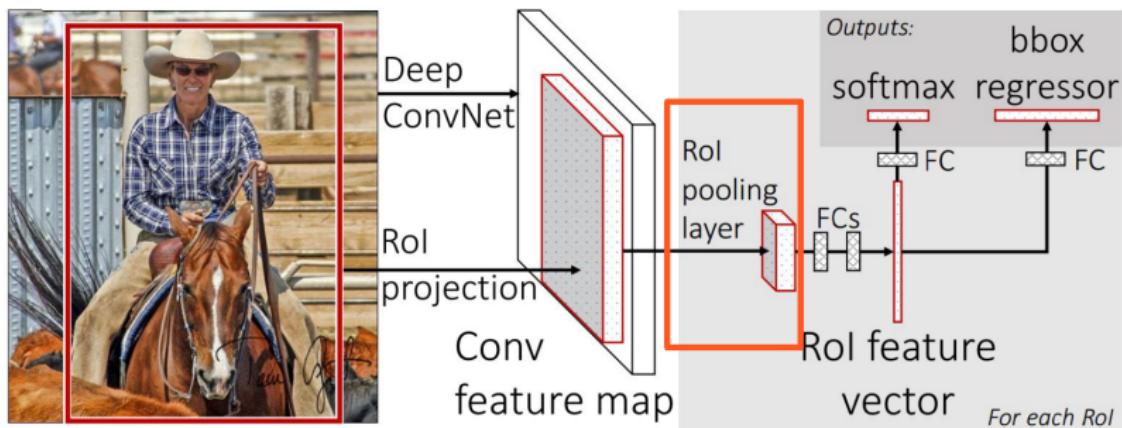
- Simple pipeline!
- Use the CNN as a feature extractor
- Warp cropped regions to make them “square”
- Run an SVM (one per category) on the deep features (stocked on disk ⇒ very heavy!!!)
- Accuracy 😊 but **very slow** 😞
- **Not end-to-end** 😞



⁶Ross Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014.

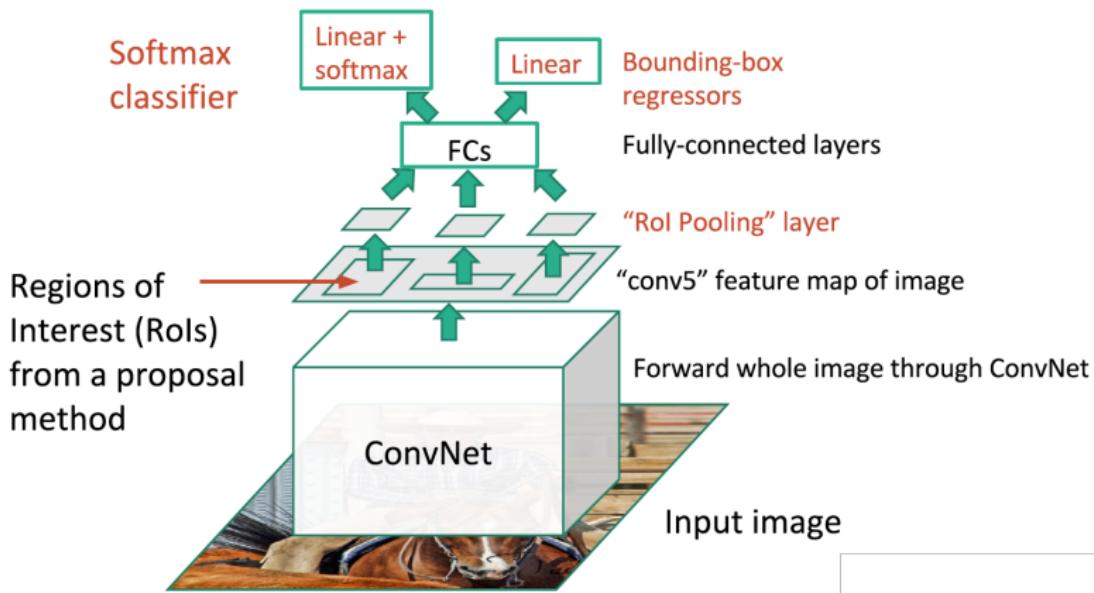
Fast R-CNN: architecture⁷

- Share convolution layers for proposals from the same image
- Introduce the ROI Pooling
- Faster and more accurate than R-CNN



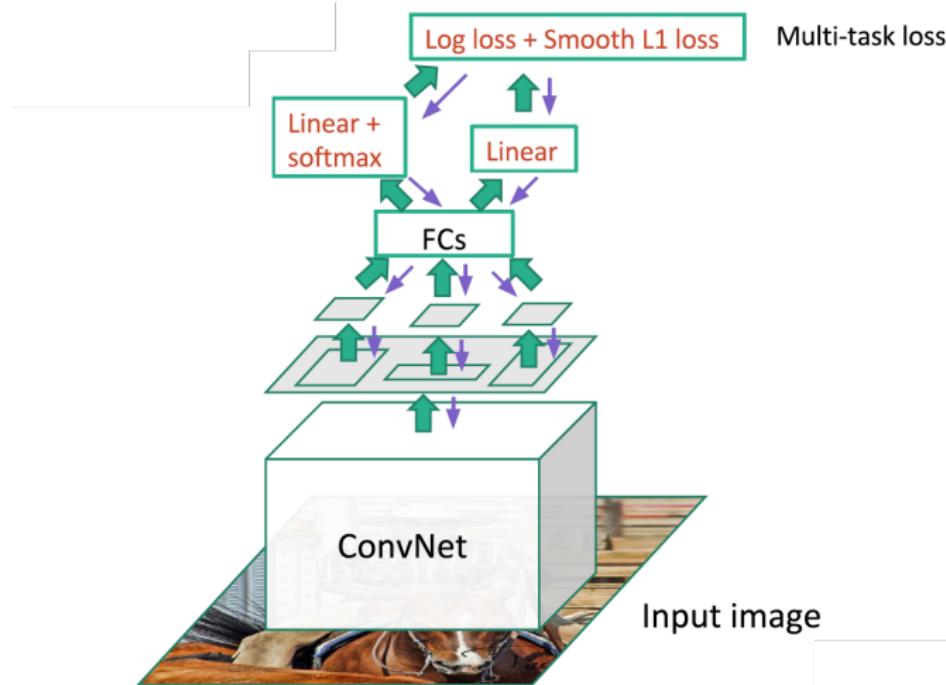
⁷Ross Girshick et al. Fast RCNN, ICCV 2015.

Fast R-CNN: architecture⁷



⁷Ross Girshick et al. Fast RCNN, ICCV 2015.

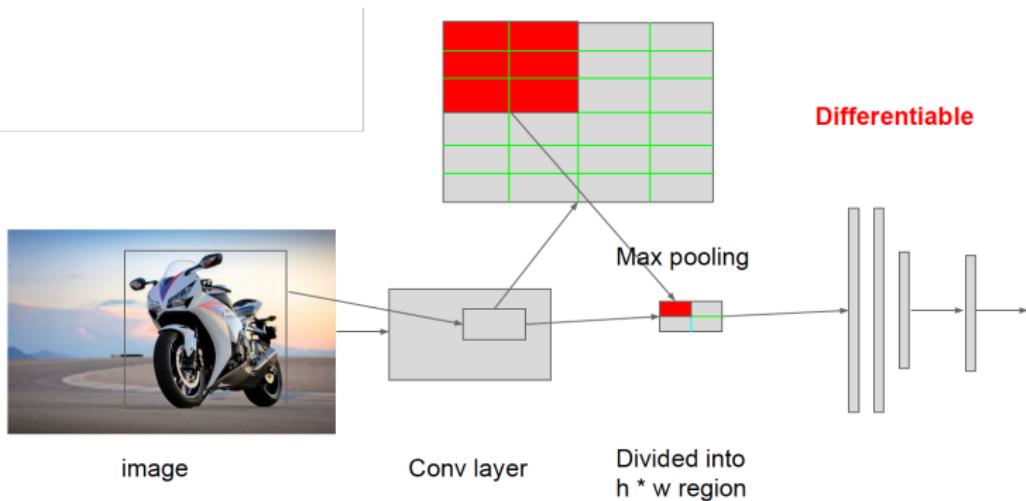
Fast R-CNN: architecture⁷



⁷Ross Girshick et al. Fast RCNN, ICCV 2015.

Explanation: ⁸

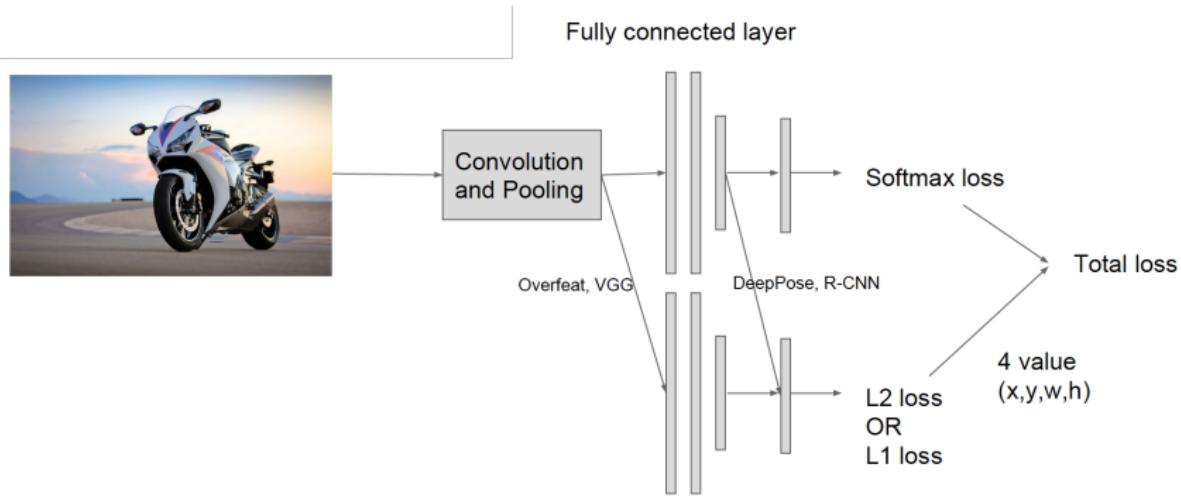
ROI Pooling



⁸Sihao Liang, Jiajun Lu and Kevin Perkins, Lecture on Object detection.

Explanation: ⁸

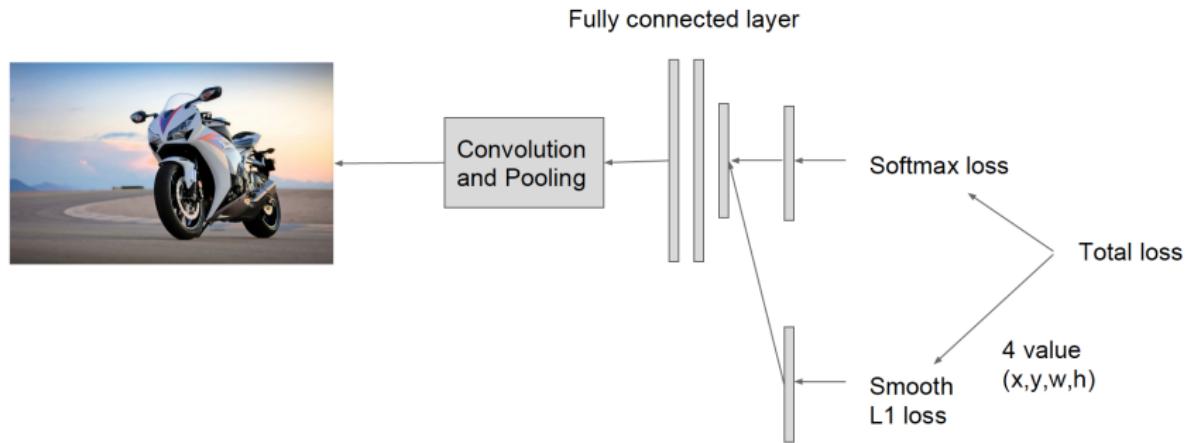
Bbox regressor



⁸Sihao Liang, Jiajun Lu and Kevin Perkins, Lecture on Object detection.

Explanation: ⁸

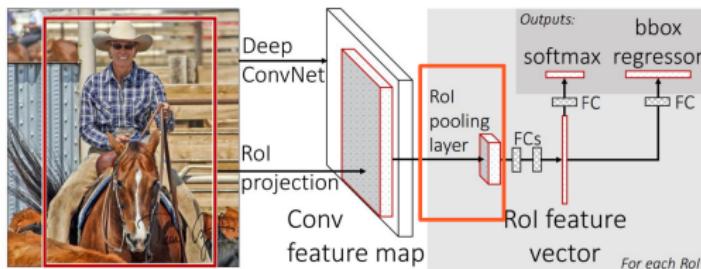
Bbox regressor



⁸Sihao Liang, Jiajun Lu and Kevin Perkins, Lecture on Object detection.

Fast R-CNN: Summary⁹

- Do not compute forward propagate from scratch for each box
- Compute feature maps once and reuse the convolutions for different boxes
- New: ROI pooling
- End-to-end training ☺
- More accuracy and must faster (25x) than R-CNN ☺
- External box proposals needed 😞

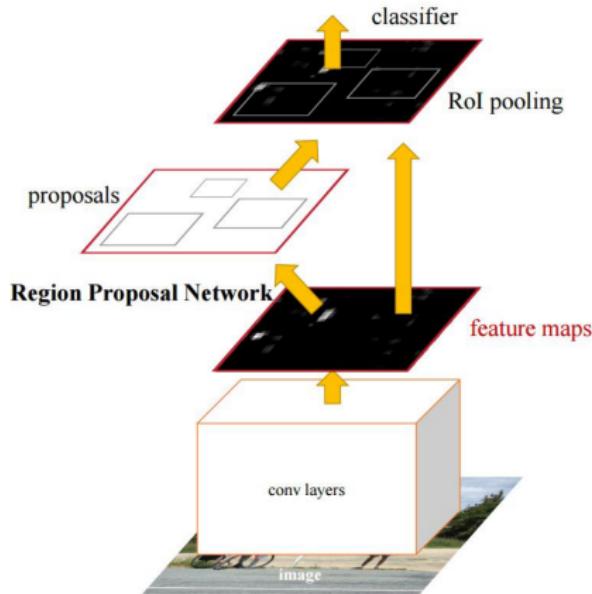


⁹Ross Girshick et al. Fast RCNN, ICCV 2015.

Two-stage detectors Faster R-CNN

Faster R-CNN: architecture¹⁰

- Don't need to have external regional proposals (which is the bottleneck)
- RPN - Regional Proposal Network

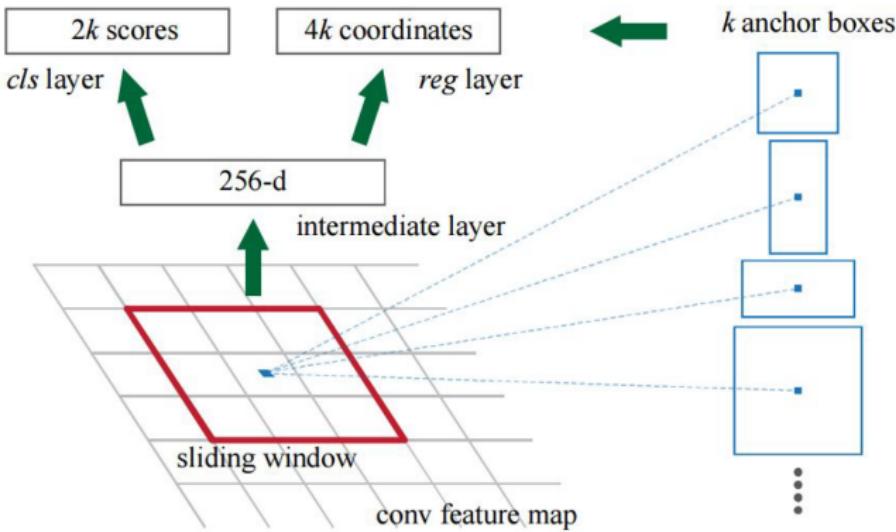


¹⁰ Shaoqing Ren et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NeurIPS'15.

Two-stage detectors Faster R-CNN

Faster R-CNN: architecture¹⁰

- Don't need to have external regional proposals (which is the bottleneck)
- RPN - Regional Proposal Network



¹⁰ Shaoqing Ren et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NeurIPS'15.

Faster R-CNN: architecture¹⁰

- Don't need to have external regional proposals (which is the bottleneck)
- RPN - Regional Proposal Network

	Faster R-CNN	Fast R-CNN	R-CNN
Test time/image With proposal	0.2S	2s	50s
-test speedup	250x	25x	1x
mAP	66.9	66.9	66

(Training on PASCAL VOC 2007 dataset)

¹⁰ Shaoqing Ren et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NeurIPS'15.

Other SOTA two-stage detectors

- R-FCN: Object Detection via Region-based fully convolutional networks (NeurIPS 2016)
- FPN: Feature pyramid networks for object detection (CVPR 2017)
⇒ great performance for detecting objects with a wide variety of scales

Remarks

- Faster R-CNN ⇒ most exploited in the literature
- FPN ⇒ a basic building block of many latest detectors
- **Faster R-CNN + FPN** ⇒ an excellent choice to work on two-stage object detector for your applications !

Table of contents

1 Object detection: introduction

2 Two-stage detectors

3 One-stage detectors

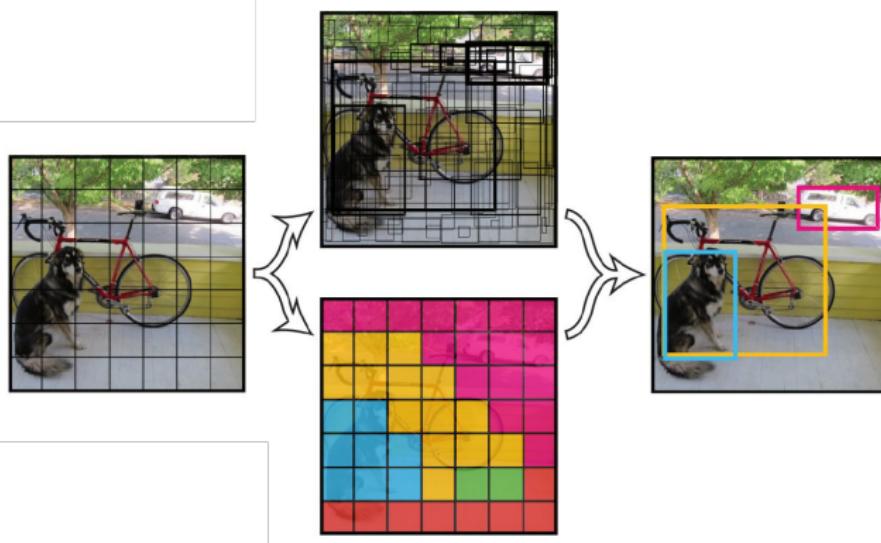
Problems with two-stage detectors

- Complex Pipeline
- Hard to optimize each component
- Main: **Slow** (Cannot run in real time)

One-stage detectors YOLO

YOLO: You Only Look Once¹¹

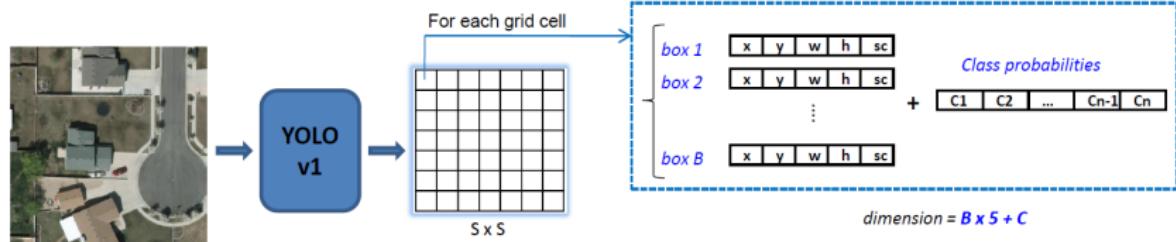
- Consider detection a regression problem
- Use a single CNN
- Runs once on entire image ⇒ Very Fast!



¹¹ Joseph Redmon et al., You only look once: Unified, real-time object detection, CVPR 2016.

One-stage detectors YOLO

YOLO: You Only Look Once¹¹

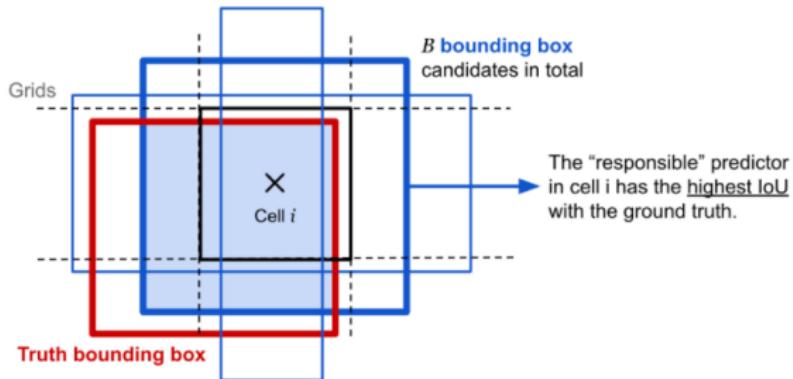
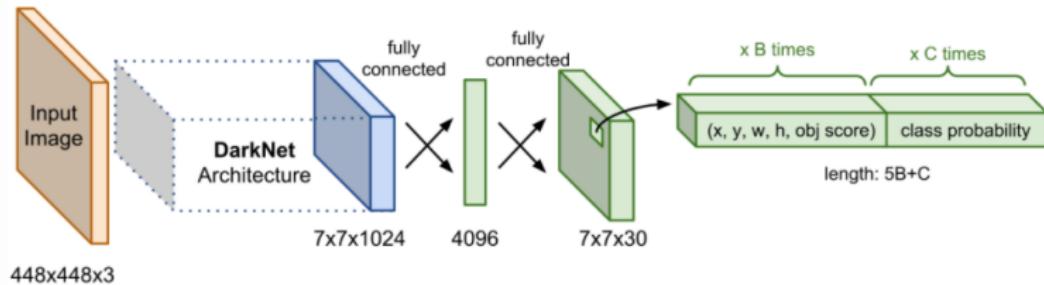


In YOLOv1, the output is a tensor of dimension $(S, S, B \times 5 + C)$ with (S, S) the size of the grid, B the number of predicted boxes for each cell and C the number of classes. By default, $S = 7$, $B = 2$ and $C = 20$ for the PASCAL VOC dataset. For an input image of size 448×448 pixels, the output is a tensor of size $7 \times 7 \times 30$.

¹¹ Joseph Redmon et al., You only look once: Unified, real-time object detection, CVPR 2016.

One-stage detectors YOLO

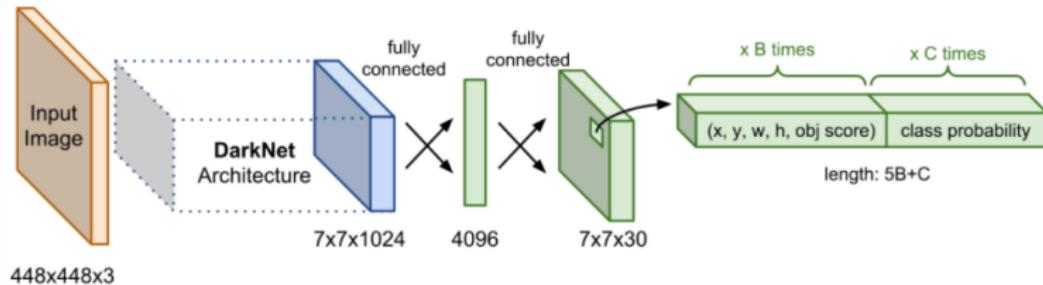
YOLO: You Only Look Once¹¹ Let's take a closer look!



¹¹ Joseph Redmon et al., You only look once: Unified, real-time object detection, CVPR 2016.

One-stage detectors YOLO

YOLO: You Only Look Once¹¹ And the loss function!



$$\mathcal{L}_{\text{loc}} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{\text{obj}} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2]$$

$$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{S^2} \sum_{j=0}^B (1_{ij}^{\text{obj}} + \lambda_{\text{noobj}}(1 - 1_{ij}^{\text{obj}})) (C_{ij} - \hat{C}_{ij})^2 + \sum_{i=0}^{S^2} \sum_{c \in \mathcal{C}} 1_i^{\text{obj}} (p_i(c) - \hat{p}_i(c))^2$$

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{cls}}$$

¹¹ Joseph Redmon et al., You only look once: Unified, real-time object detection, CVPR 2016.

YOLO: Very fast but many limitations

- Struggles with small objects
- Struggles with unusual aspect ratios
- Poor localization

	Yolo	Faster R-CNN (VGG-16)
mAP	63.4	73.2
FPS	45	7

Trained on Pascal VOC 2007 + 2012 dataset

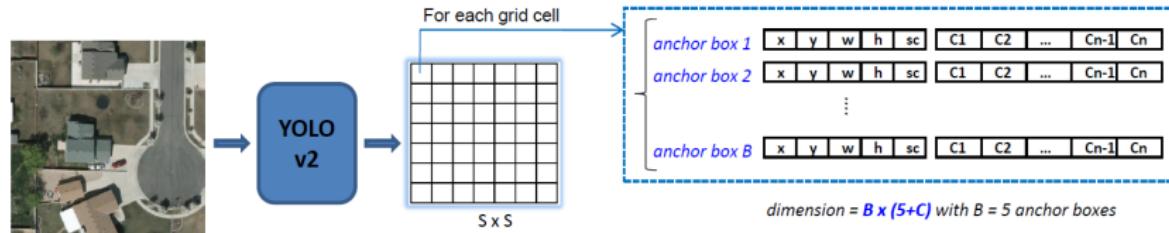
YOLO9000: better, faster, stronger¹²

	YOLO								YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓	✓
anchor boxes?					✓	✓			
new network?						✓	✓	✓	✓
dimension priors?							✓	✓	✓
location prediction?							✓	✓	✓
passthrough?								✓	✓
multi-scale?								✓	✓
hi-res detector?									✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	78.6

¹² Joseph Redmon et al., YOLO9000: better, faster, stronger, CVPR 2017.

One-stage detectors YOLOv2

YOLO9000: better, faster, stronger¹²

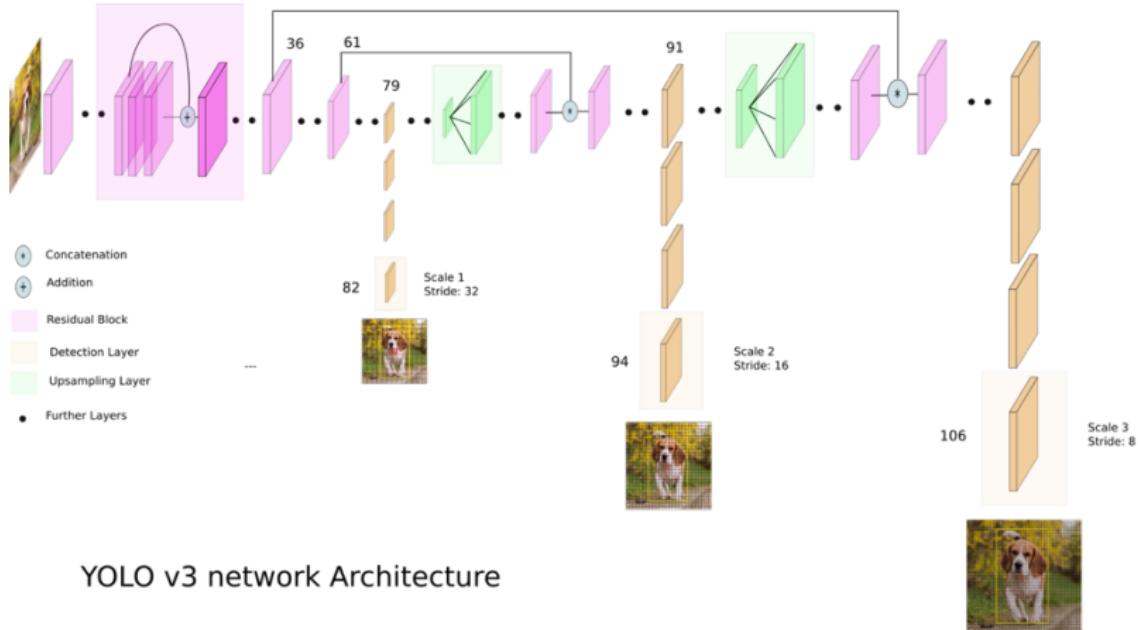


In YOLOv2, the output is a tensor of dimension $(S, S, B \times (5 + C))$. The difference is that the class probabilities are calculated for each anchor box. By default, $S = 13$, $B = 5$ anchor boxes and $C = 20$ for the PASCAL VOC dataset. For an input image of size 416×416 pixels, the output is a tensor of size $13 \times 13 \times 125$.

¹² Joseph Redmon et al., YOLO9000: better, faster, stronger, CVPR 2017.

One-stage detectors YOLOv3

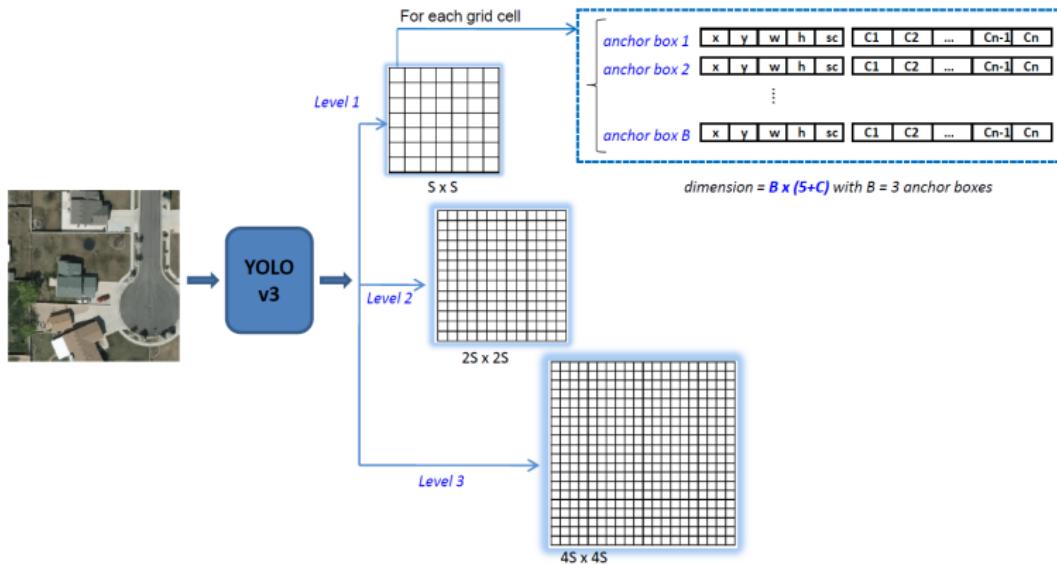
YOLOv3: multi-level detection¹³



¹³ Joseph Redmon et al., Yolov3: An incremental improvement, Arxiv 2018.

One-stage detectors YOLOv3

YOLOv3: multi-level detection¹³



In YOLOv3, the output consists of 3 tensors of dimension $(S, S, B \times (5 + C))$, $(2S, 2S, B \times (5 + C))$ and $(4S, 4S, B \times (5 + C))$ which correspond to the 3 detection levels (scales). By default, $S = 13$, $B = 3$ anchor boxes and $C = 80$ for the COCO dataset. For an input image of size 416×416 pixels, the outputs are three tensors of size $13 \times 13 \times 255$, $26 \times 26 \times 255$ and $52 \times 52 \times 255$.

¹³ Joseph Redmon et al., Yolov3: An incremental improvement, Arxiv 2018.

Other SOTA one-stage detectors

- SSD: Single shot multibox detector (ECCV 2016) and its variants!
- RetinaNet (ICCV 2017): focal loss
- EfficientDet (CVPR 2020)
- YOLOv4: Optimal Speed and Accuracy of Object Detection, Arxiv 2020
- Scaled-YOLOv4: Scaling Cross Stage Partial Network, CVPR 2021
- YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, Arxiv 2022
- YOLOv8, YOLOv9, YOLOv10 Ultralytics
- YOLO-world (Open-vocabulary Object Detection)

Your remarks ???

References I

Books & articles:

1. Zhengxia Zou et al. **Object Detection in 20 Years: A Survey**. Arxiv, 2020.
2. Liu, Li, et al. **Deep learning for generic object detection: A survey**. International journal of computer vision 128.2 (2020): 261-318.

Courses:

1. Lecture 11: Detection and Segmentation, Fei-Fei Li, Justin Johnson & Serena Yeung
<http://cs231n.stanford.edu/slides/2017/>
2. Lecture 07: Object Detection, Sihao Liang, Jiajun Lu & Kevin Perkins
<http://slazebni.cs.illinois.edu/spring17/>
3. Lecture 6: Convnets for object detection and segmentation, Deep Learning at UvA
<https://uvadlc.github.io/lectures/feb2016/>