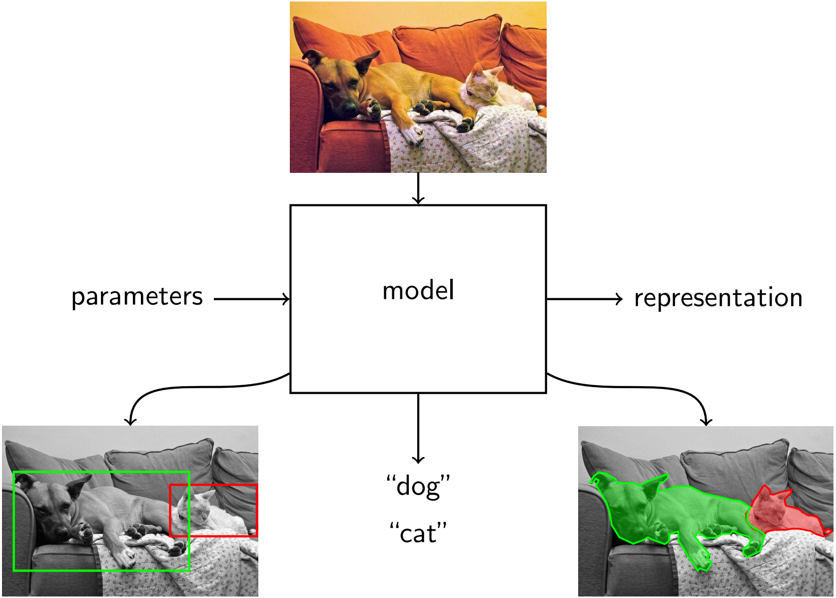
Computer vision - Detection

Ronan Sicre
Credits to Yannis Avrithis https://sif-dlv.github.io/

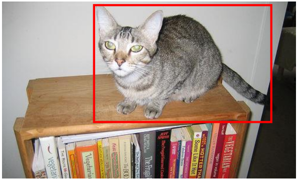
Outline

background
two-stage detection
one-stage detection

data-driven approach

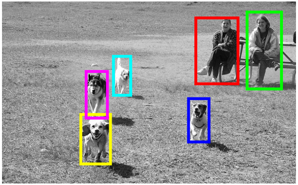


beyond classification

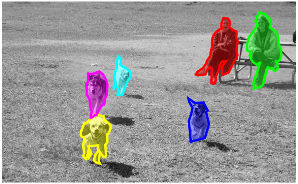


object localization
classify + regress
bounding box (x, y, w, h)

semantic segmentation
pixel-wise classify



object detection
per region: classify + regress
bounding box (x, y, w, h)



instance segmentation
per region: pixel-wise classify

selective search (SS)

[van de Sande et al. 2011]

input image

ground truth

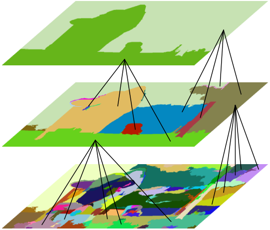
van de Sande, Uijlings, Gevers and Smeulders. ICCV 2011. Segmentation As Selective Search for Object Recognition.

selective search (SS)

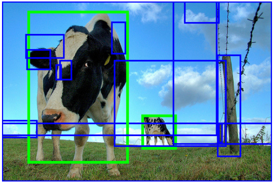
[van de Sande et al. 2011]

input image

ground truth



hierarchical grouping

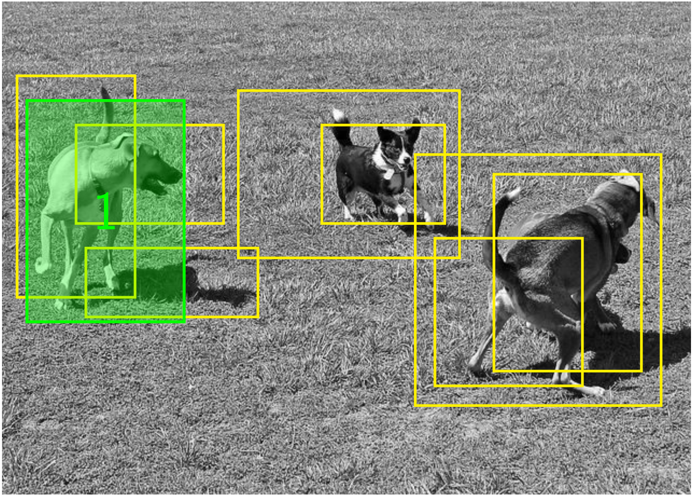


object proposals

van de Sande, Uijlings, Gevers and Smeulders. ICCV 2011. Segmentation As Selective Search for Object Recognition.

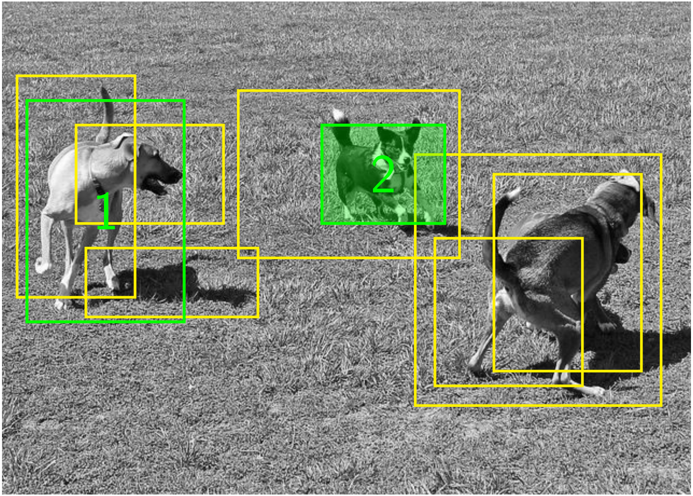
non-maximum suppression (NMS)

non-maximum suppression (NMS)



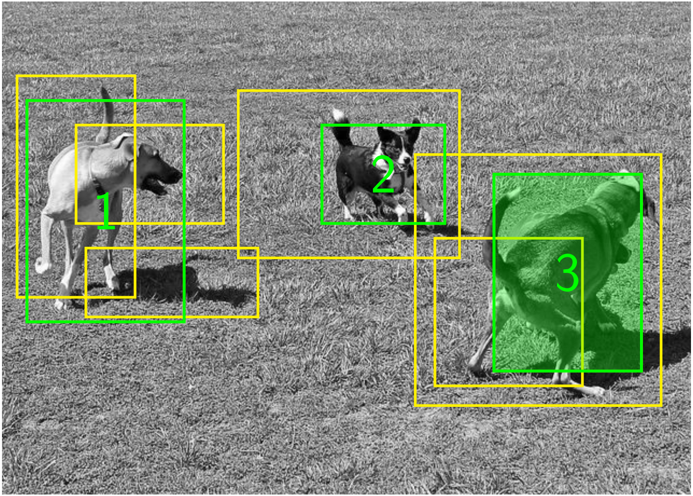
region 1 remains

non-maximum suppression (NMS)



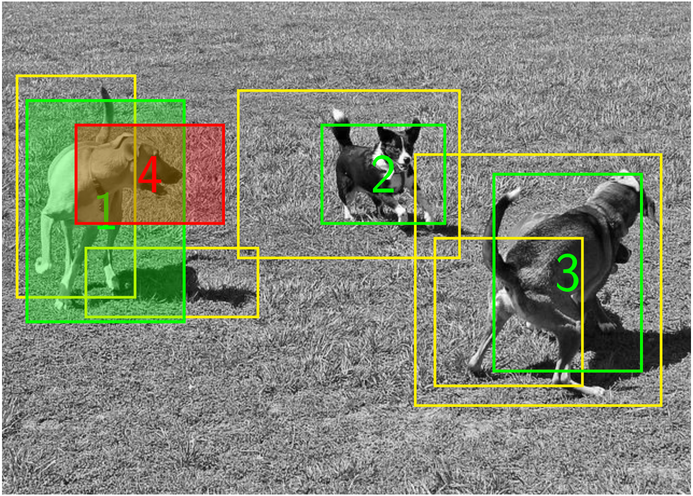
region 2 remains

non-maximum suppression (NMS)



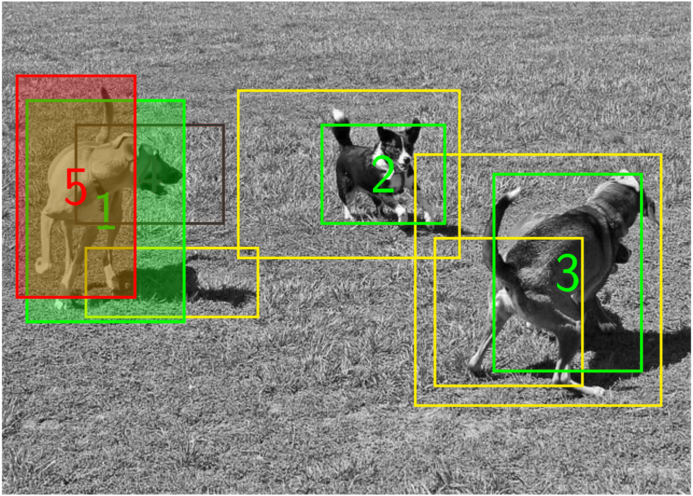
region 3 remains

non-maximum suppression (NMS)



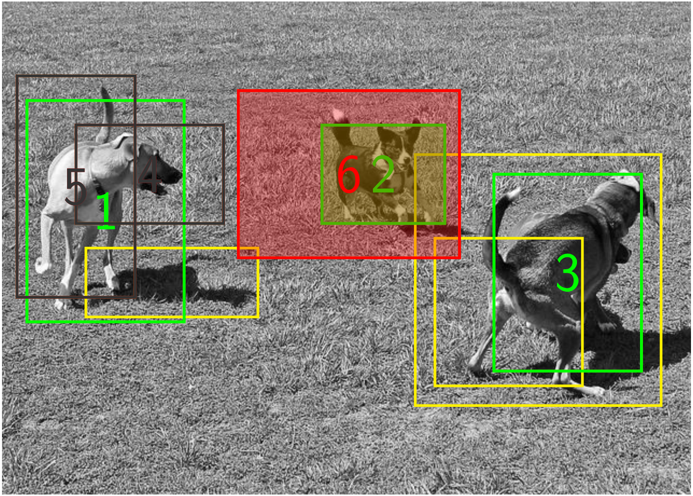
region 4 is rejected because $J(r_4, r_1) = 0.2750 > 0.25$

non-maximum suppression (NMS)



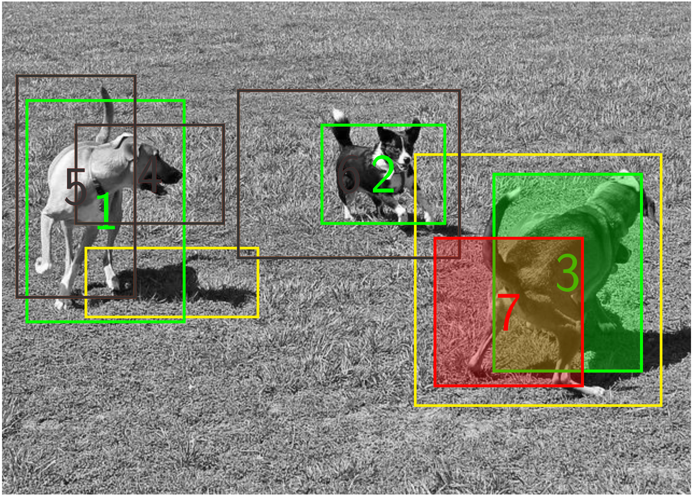
region 5 is rejected because $J(r_5, r_1) = 0.5366 > 0.25$

non-maximum suppression (NMS)



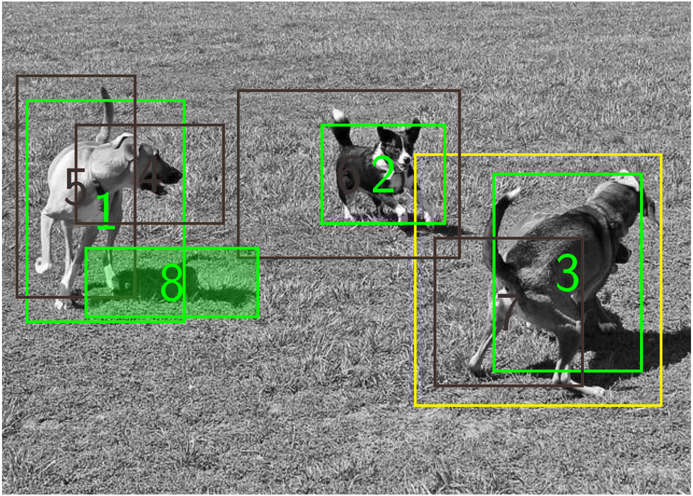
region 6 is rejected because $J(r_6, r_2) = 0.3268 > 0.25$

non-maximum suppression (NMS)



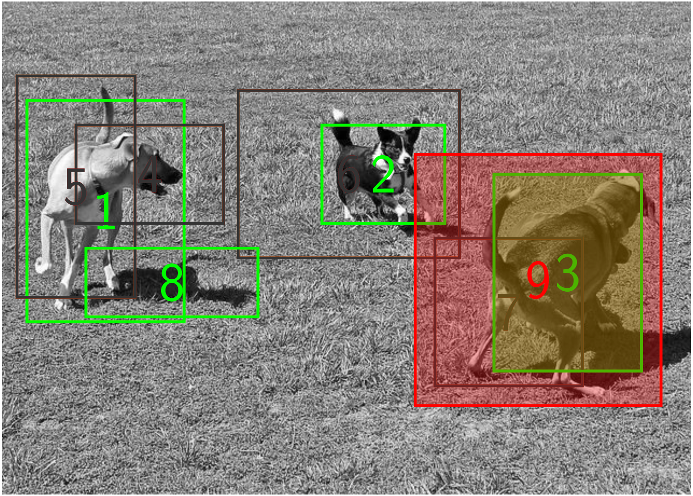
region 7 is rejected because $J(r_7, r_3) = 0.3011 > 0.25$

non-maximum suppression (NMS)



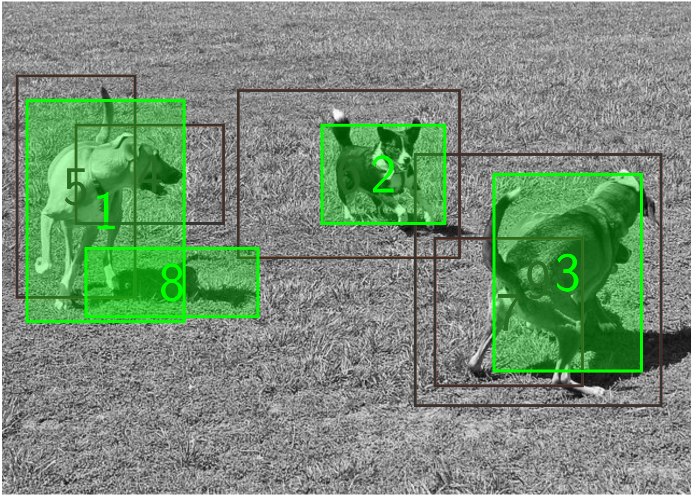
region 8 remains

non-maximum suppression (NMS)



region 9 is rejected because $J(r_9, r_3) = 0.4706 > 0.25$

non-maximum suppression (NMS)



in the end, regions 1, 2, 3, 8 remain

non-maximum suppression on regions

- given regions $r_1, r_2, \ldots$ of each class independently, ranked by decreasing order of confidence score
- for $i = 2, 3, \ldots$, reject region r_i if it has intersection-over-union (IoU) overlap higher then a threshold τ

$$J(r_i, r_j) > \tau$$

with some higher scoring region r_j with $j < i$ that has not been rejected

detection evaluation

[Russakovsky et al. 2015]

- for each image and for each class independently, rank predicted regions by descending order of confidence and assign each region r to the ground truth region $g^* = \arg\max_g J(r, g)$ of maximum overlap if $J(r, g^*) > \tau$ and mark it as true positive, else false
- each ground truth region can be assigned up to one predicted region
- now for each class independently, rank predicted regions of all images by descending order of confidence and compute average precision (AP) according to true/false labels
- the mean average precision (mAP) is the mean over classes

Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg and Fei-Fei. IJCV 2015. Imagenet Large Scale Visual Recognition Challenge.

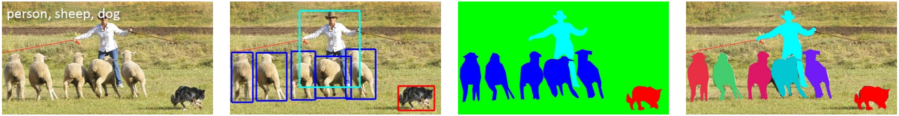
detection evaluation

[Russakovsky et al. 2015]

- for each image and for each class independently, rank predicted regions by descending order of confidence and assign each region r to the ground truth region $g^* = \arg\max_g J(r, g)$ of maximum overlap if $J(r, g^*) > \tau$ and mark it as true positive, else false
- each ground truth region can be assigned up to one predicted region
- now for each class independently, rank predicted regions of all images by descending order of confidence and compute average precision (AP) according to true/false labels
- the mean average precision (mAP) is the mean over classes

Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg and Fei-Fei. IJCV 2015. Imagenet Large Scale Visual Recognition Challenge.

object detection datasets



- **PASCAL** VOC 2007-12: 20 classes; images $5\text{-}11\text{k}$ train/val, $5\text{-}11\text{k}$ test (public for 2007)
- **ImageNet** ILSVRC 2013-14: 200 classes (subset or merged from classification task); images $400\text{-}450\text{k}$ train (partially annotated), 20k val, 40k test
- **COCO** 2014-17: 80 classes; images 80k train, 40k val ($115\text{k}/5\text{k}$ in 2017), 40k test, 120k unlabeled; smaller objects

Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg and Fei-Fei. IJCV 2015. Imagenet Large Scale Visual Recognition Challenge.

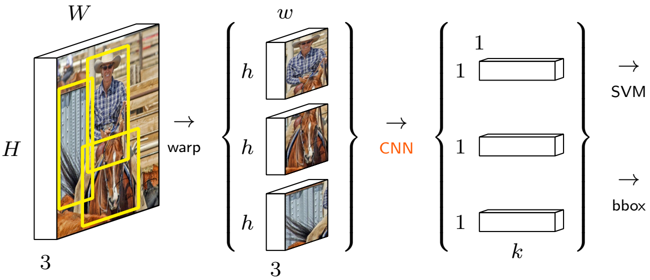
Everingham, Eslami, van Gool, Williams, Winn and Zisserman. IJCV 2015. The PASCAL Visual Object Classes Challenge: a Retrospective.

Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick. ECCV 2014. Microsoft COCO: Common Objects in Context.

two-stage detection

regions with CNN features (R-CNN)

[Girshick et al. 2014]

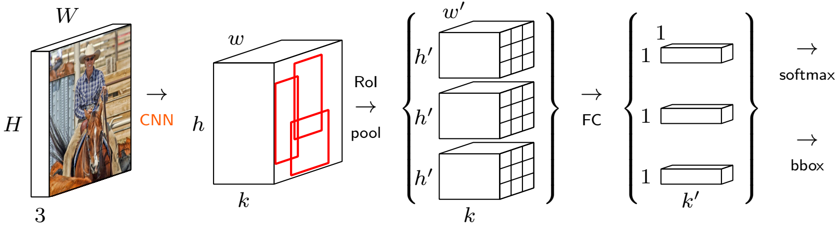


- 3-channel RGB input, fixed width $W = 500$ pixels
- ~ 2000 SS region proposals warped into fixed $w \times h = 227 \times 227$
- each proposal yields a $k = 4096$ dimensional feature by CaffeNet
- each feature is classified into c classes by c one-*vs.* -rest SVMs and localized by bounding box regression

Girshick, Donahue, Darrell and Malik. CVPR 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.

fast R-CNN (FRCN)

[Girshick 2015]



- 3-channel RGB input, arbitrary size
- input yields a single $k = 4096$ dimensional feature map by VGG-16
- ~ 2000 region proposals, projected onto feature maps and RoI-pooled into fixed size $w' \times h' \times k = 7 \times 7 \times k$
- several fully-connected layers follow, for each pooled map
- each pooled map is classified into $c + 1$ classes (c + background) by single softmax and localized by bounding box regression

fast R-CNN (FRCN)

pros

- fast (0.32s/image; $9\times$ training, $213\times$ test speedup *vs.* R-CNN): image forwarded through network only once, only few layers are region-specific
- 2 stages: only region proposals are separate; features, classifier and regressor are trained end-to-end with multi-task loss
- better performance

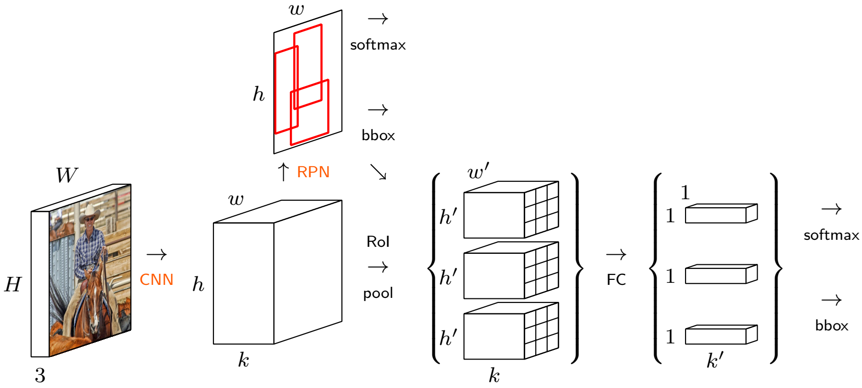
cons

- region proposals are still needed for performance, but are now the bottleneck (~ 2s/image)
- single-scale

Girshick. ICCV 2015. Fast R-CNN.

faster R-CNN

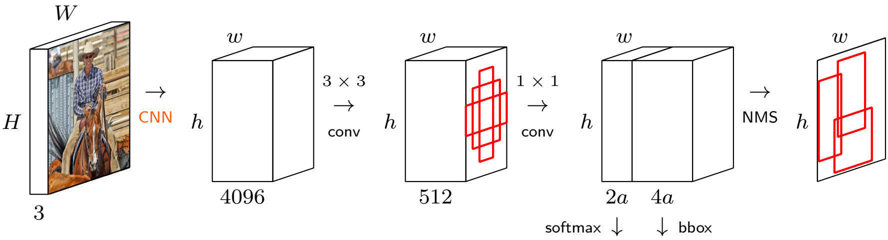
[Ren et al. 2015]



- same input, same VGG-16 feature maps as Fast R-CNN
- proposals detected directly on feature maps by RPN and max-pooled
- same classifier, same bounding box regression, but now also for RPN

Ren, He, Girshick and Sun. NIPS 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

region proposal network (RPN)



- same input, same feature maps, dimension reduced to 512
- $a = 9$ anchors at each position, for 3 scales and 3 aspect ratios
- $2a$ classification (object/non-object) scores and $4a$ bounding box coordinates relative to anchor at each position
- softmax on scores, regression loss on coordinates
- region proposals by non-maxima suppression

Ren, He, Girshick and Sun. NIPS 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

faster R-CNN

pros

- faster (0.2s/image including proposals; $10\times$ test speedup *vs.* fast R-CNN): only few layers are used for RPN and region-specific classification and regression
- trained end-to-end including features, region proposals, classifier and regressor
- more accurate: region proposals are learned, RPN is convolutional

cons

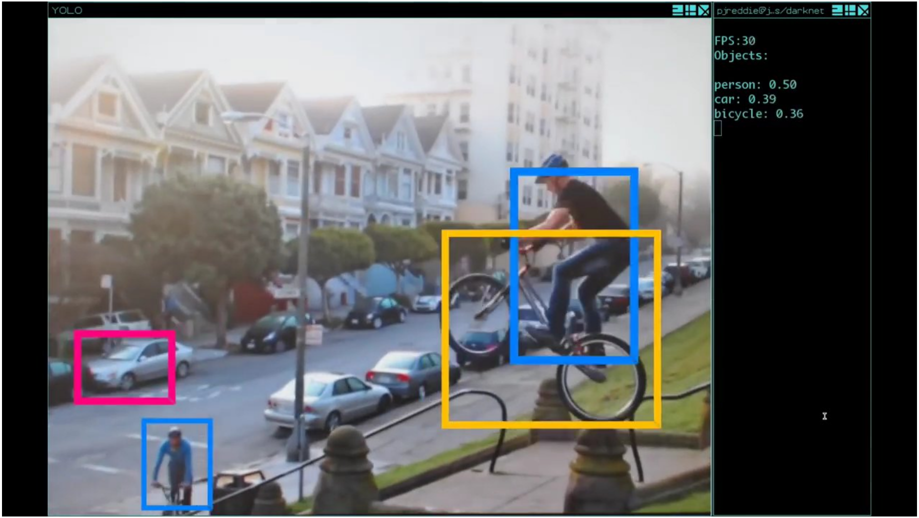
- still, several fully-connected layers needed for region-specific tasks
- still single-scale

Ren, He, Girshick and Sun. NIPS 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.

one-stage detection

“you only look once” (YOLO)

[Redmon et al. 2016]



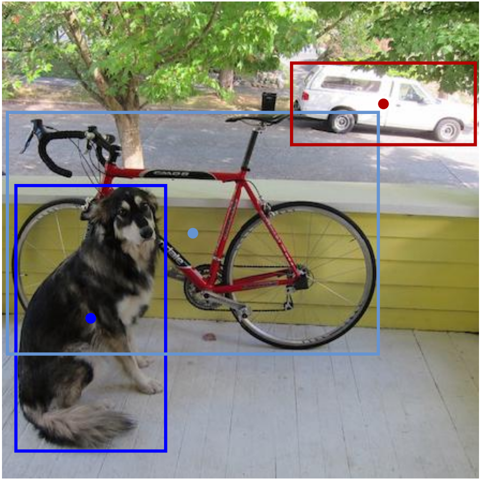
Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)

- input image

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

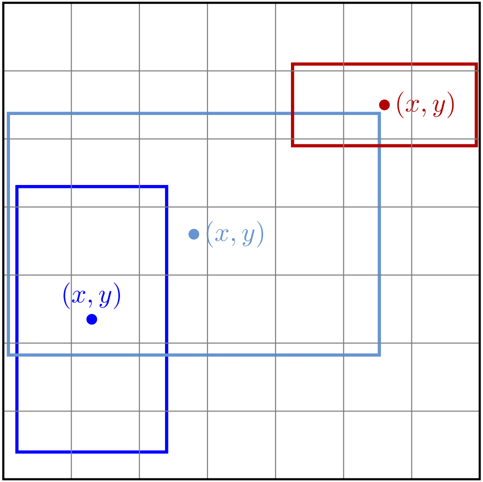
“you only look once” (YOLO)



- groung truth bounding boxes and their centers

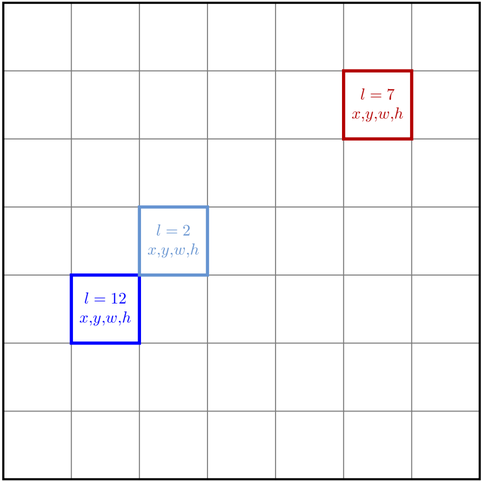
Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)



- image partitioned into 7×7 grid and center coordinates assigned to cells

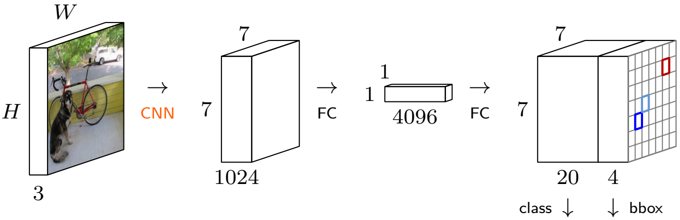
"you only look once" (YOLO)



- network learns to predict up to one object per cell, including class label l, center coordinates x, y and bounding box size w, h

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

“you only look once” (YOLO)

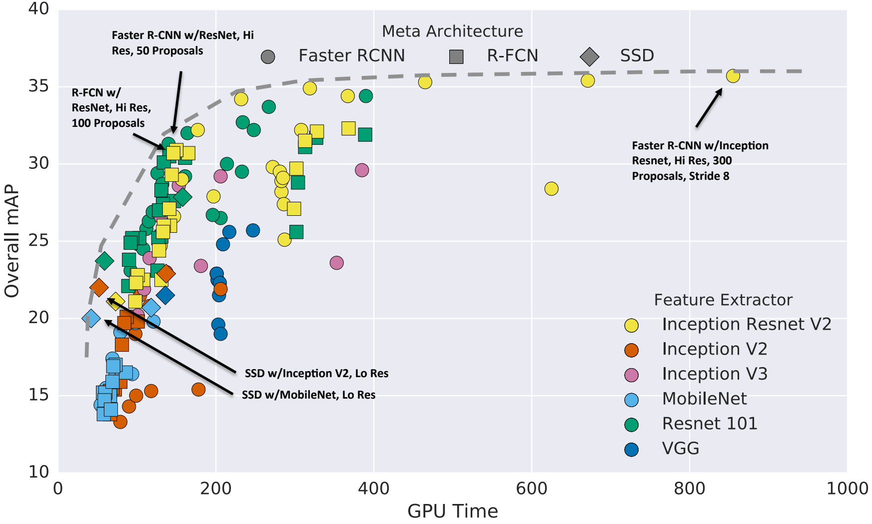


- 3-channel input $W = H = 448$, 24-layer NiN-like network
- fully connected layer, increasing to 4096 features
- $c = 20$ class scores and 4 bounding box coordinates at each position
- in a single stage, network performs regression from the image to a $7 \times 7 \times 24$ tensor encoding detected classes and positions
- regression (ℓ_2) loss on both class scores and coordinates
- “objectness” score makes it look like two-stage

Redmon, Divvala, Girshick and Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

speed-accuracy trade-offs

[Huang et al. 2016]



Huang, Rathod, Sun, Zhu, Korattikara, Fathi, Fischer, Wojna, Song, Guardarrama and Murphy 2016. Speed-Accuracy Trade-Offs for Modern Convolutional Object Detectors.

what is wrong with dense detection?

- in a two-stage detector, the classifier is applied to a sparse set of candidate object locations, which are found by binary classification (object/non-object)
- in a one-stage detector, the classifier is applied to a dense set of locations (*e.g.* a regular grid), which introduces extreme class imbalance between foreground-background
- there is a vast number of easy negatives that can overwhelm the detector
- as an alternative to OHEM, design the loss function such that it does not penalize well-classified examples

one-stage vs. two-stage

- two-stage fights class imbalance; alternatively, use batch sampling, hard negative mining, or a better loss function
- two-stage defines regions at different scales; alternatively, use multiple scales from a feature pyramid
- two-stage pools resamples regions at different aspect ratios, or with deformable parts; this has not been explored with feature pyramids or one-stage detectors yet

Lin, Goyal, Girshick, He and Dollar. ICCV 2017. Focal Loss for Dense Object Detection.