香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

DDA4220 Deep Learning and Applications

Lecture 5 Computer Vision Basic --- Part II

Ruimao Zhang

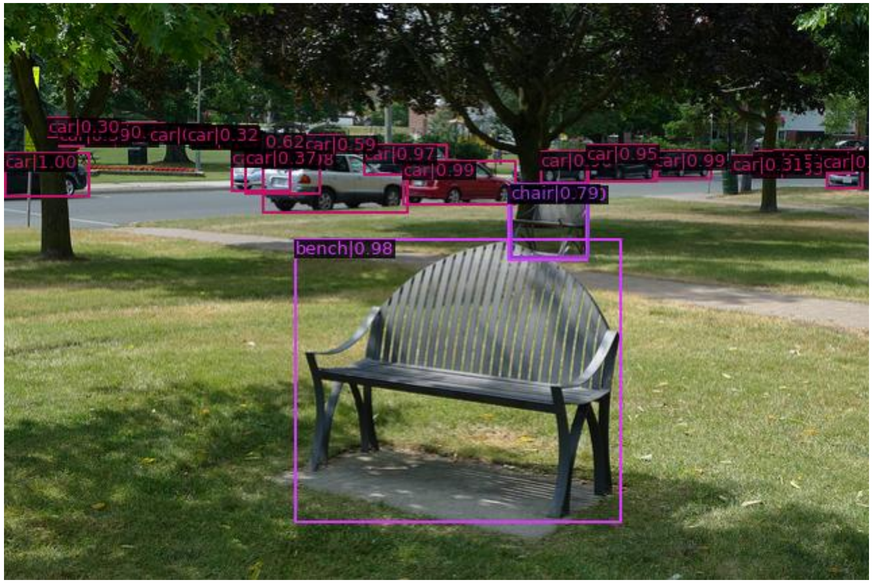
zhangruimao@cuhk.edu.cn

School of Data Science

The Chinese University of Hong Kong (Shenzhen)

Slides partially credited to Prof. Hongsheng Li at CUHK

What is object detection?



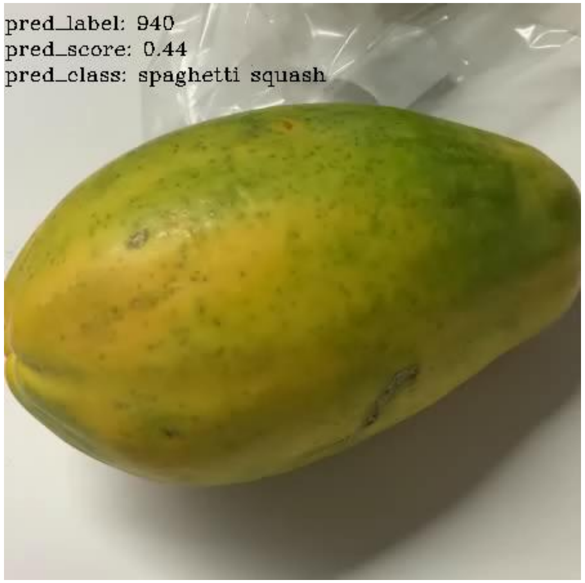
Given an image

Localize all interested objects in a bounding box

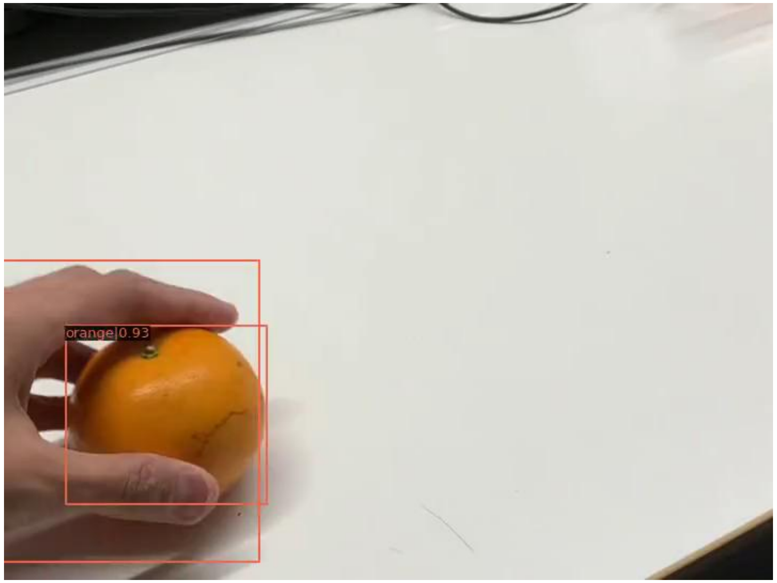
and Predict the class labels of the objects

Image Classification v.s. Object Detection

Image Classification



Object Detection



Differences

Image Classification:

Usually only one object
Usually in the center of the image
Usually occupies the main area

Object Detection:

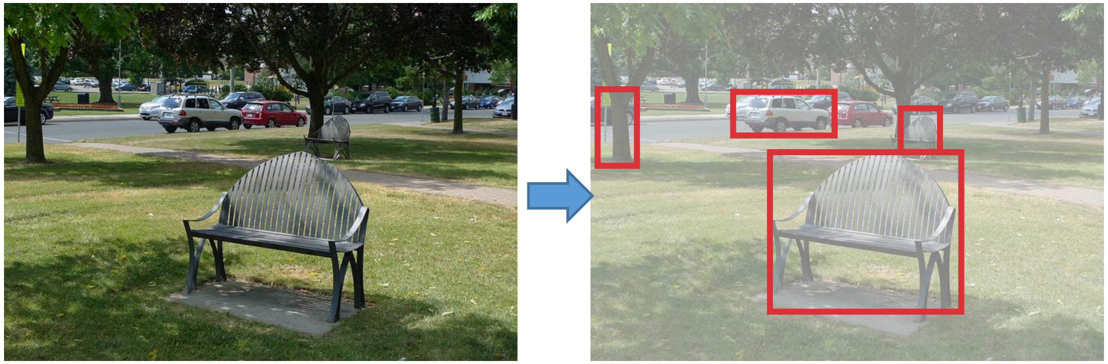
The number of objects is not fixed
Object position is not fixed
Object size is not fixed

Similarities

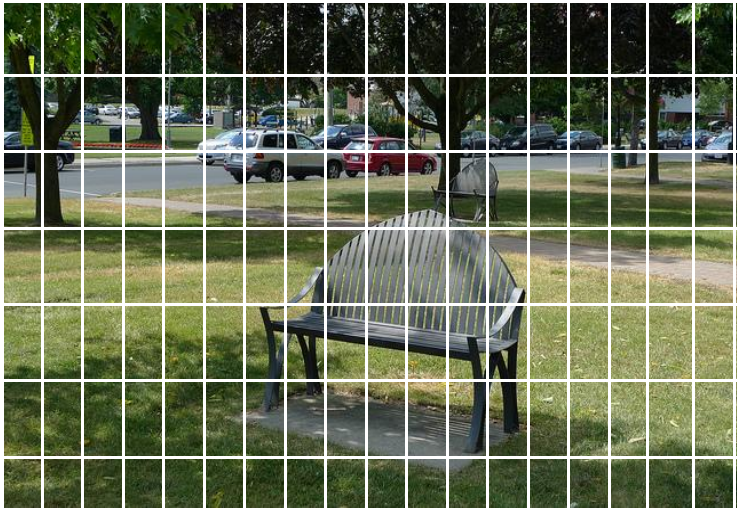
Need algorithm to "understand" the content of the image

� Deep neural network implementation

From Image Classification to Object Detection



Problem: Where do we look in the image for the object?



Idea: Exhaustively search for objects.

Problem: Extremely slow, must process tens of thousands of candidate objects.

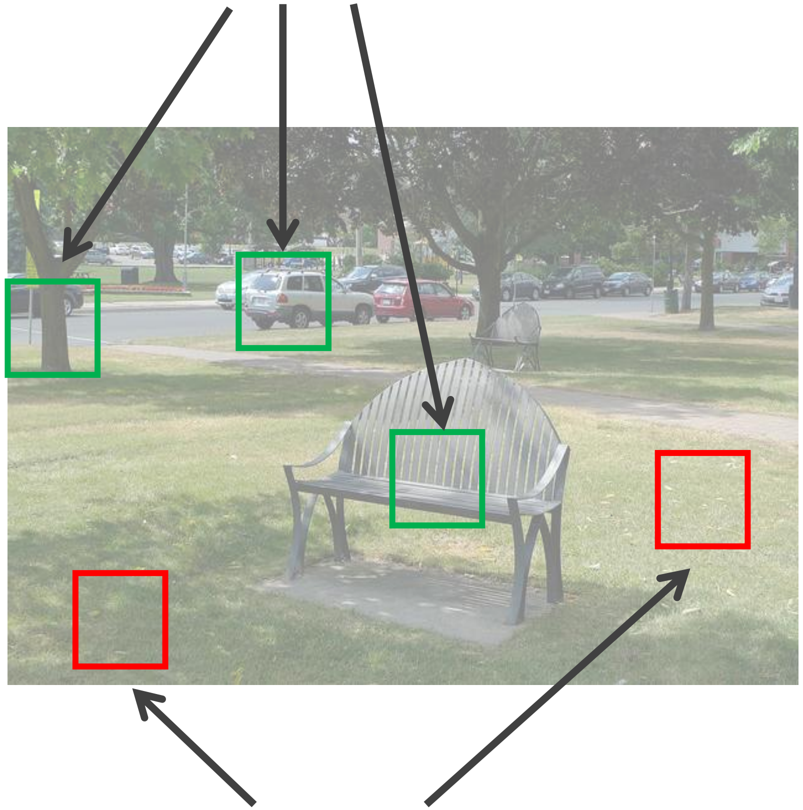
One Solution

Idea: Running a scanning detector is cheaper than running a recognizer, so do that first.

1. Exhaustively search for candidate objects with a generic detector.

2. Run recognition algorithm only on candidate objects.

Problem: What about oddly-shaped objects? Will we need to scan with windows of many different shapes?



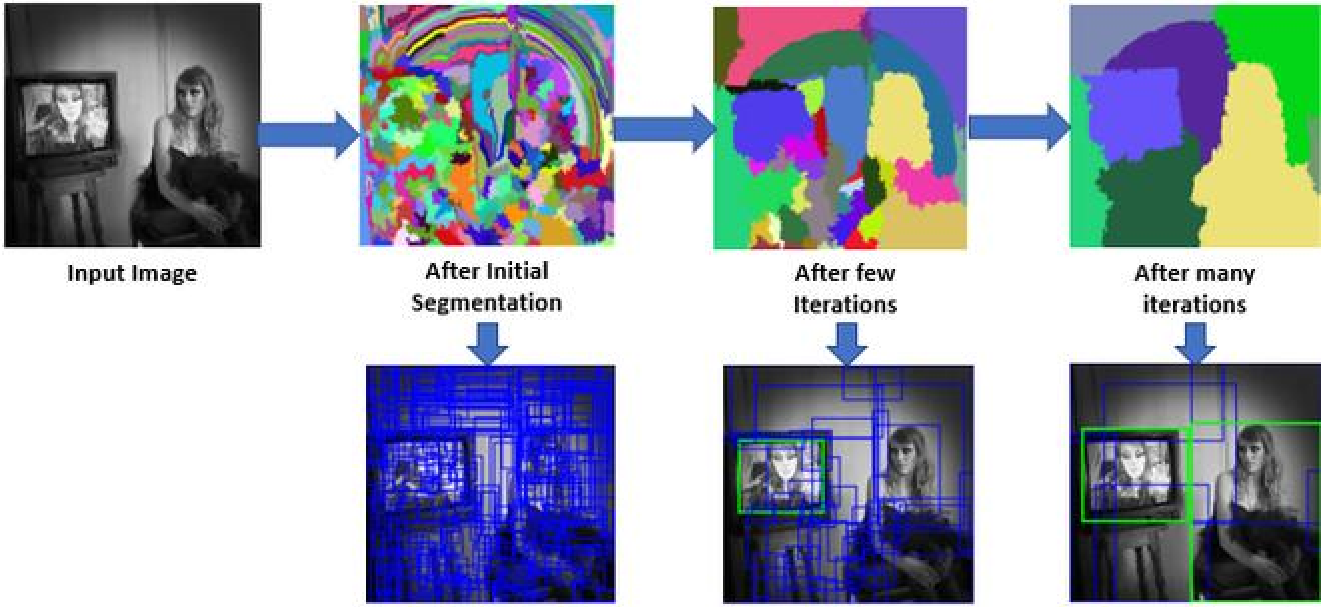
Might be objects

Not objects

Measuring the objectness of image windows, IEEE transactions on Pattern Analysis and Machine Intelligence, 2012.

Selective Search for Object Detection

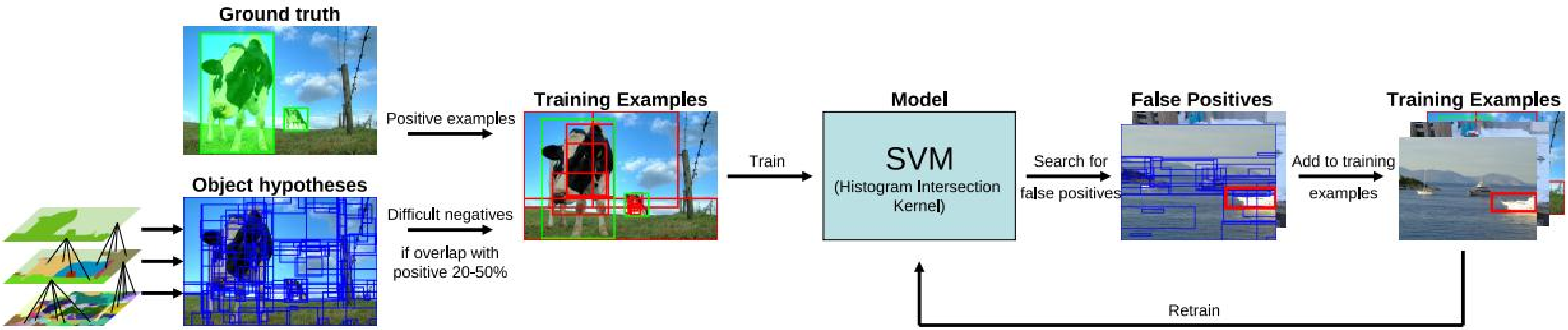
- Selective search is a method for generating a series of object proposals or object hypotheses (candidate bounding boxes that are likely to enclose objects of inerest) from the input image
- It uses hierarchical clustering on pixels to obtain clusters and then obtains object proposals



Selective Search for Object Recognition, International Journal of Computer Vision, 2013

Selective Search for Object Detection

- An image patch is cropped from each object proposal. A multi-class classifier (SVM in the original paper) is then used to classify each each patch into the background or one of the foreground (object of interest) classes



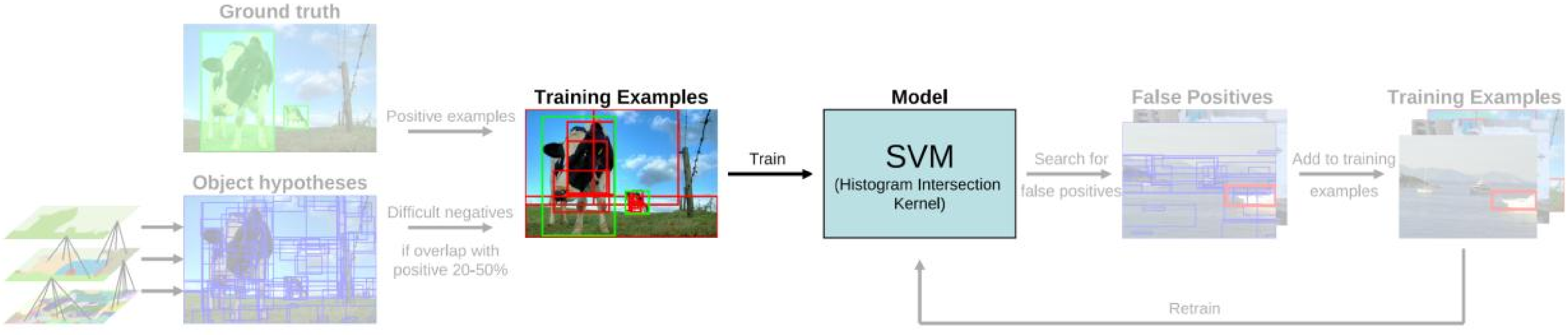
Selective Search for Object Recognition, International Journal of Computer Vision, 2013

Selective Search for Object Detection

Step 1: Train Initial Model

Positive Examples: From ground truth.

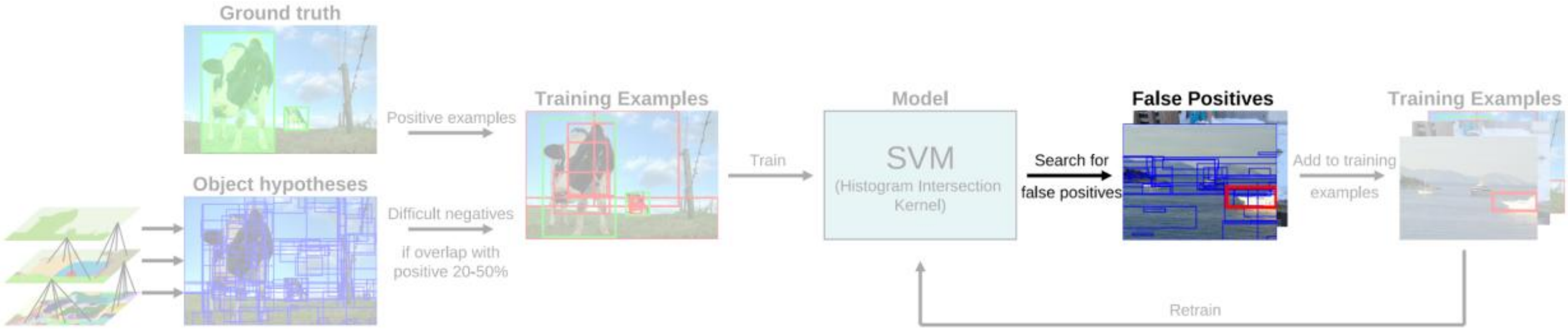
Negative Examples: Sample hypotheses that overlap 20-50% with ground truth.



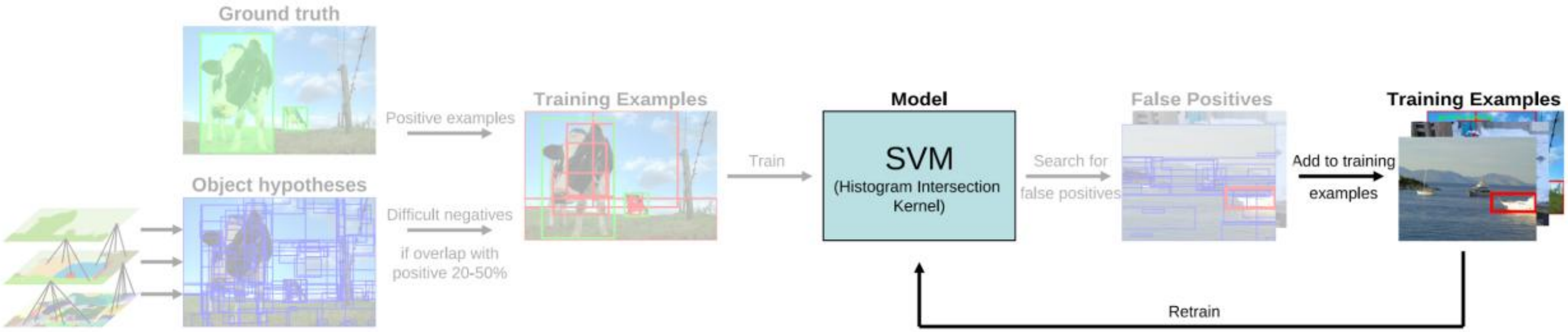
Selective Search for Object Recognition, International Journal of Computer Vision, 2013

Selective Search for Object Detection

Step 2: Search for False Positives Run model on image and collect mistakes.



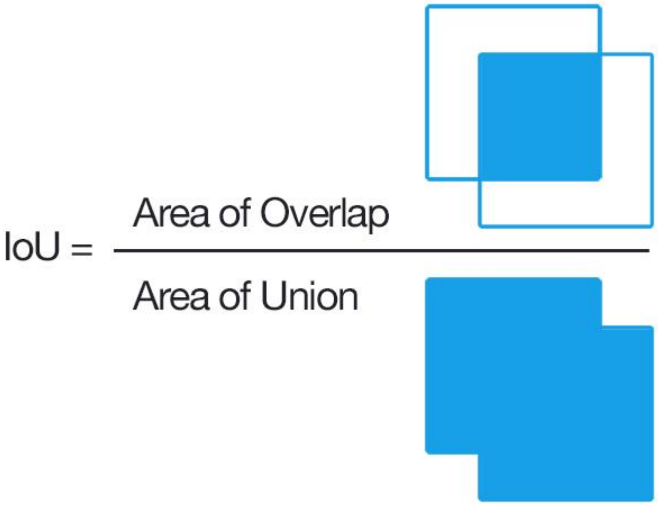
Step 3: Retrain Model Add false positives as new negative examples, retrain.



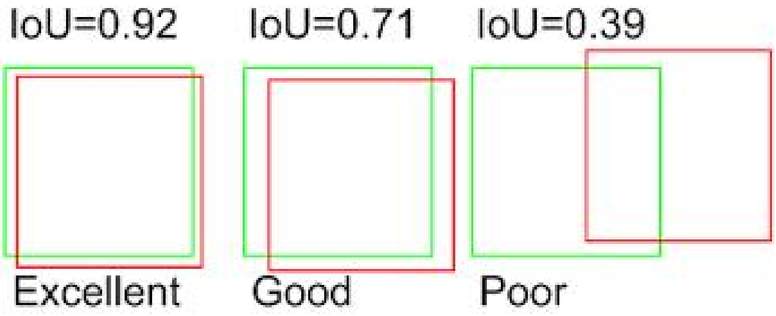
Selective Search for Object Recognition, International Journal of Computer Vision, 2013

Box Overlapping

- During the training, we would like to determine positives and negatives from the object hypotheses with ground-truth bounding boxes
- Intersection-over-Union is used to determine the overlapping between two bounding boxes
- IoU = 1 means 100% overlapping between two boxes, and IoU = 0 means 0% overlapping between two boxes
- The selective search paper uses proposals of IoU∈ [0.2, 0.5] with ground-truths as difficult (hard) negatives

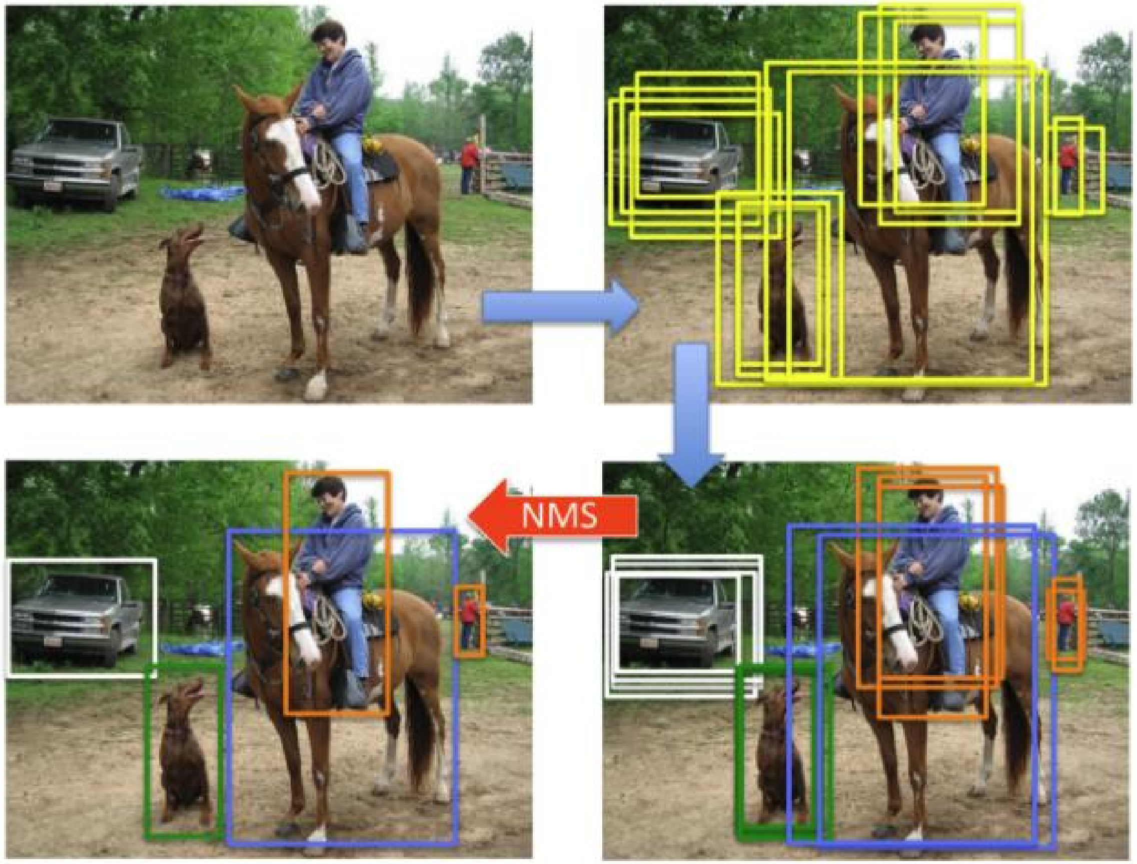


$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Non-maximum Suppression

- During the testing stage, a ground-truth object might be covered by multiple bounding box
- **Non-Maximum Suppression (NMS)** is introduced to let higher-scoring boxes to “kill” lower-scoring boxes

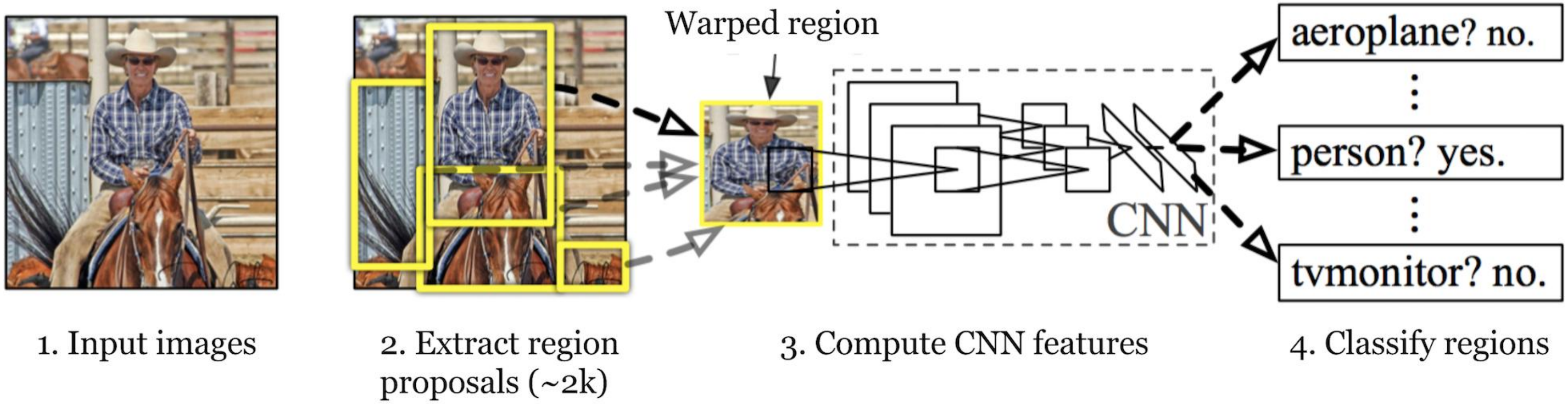


Detection Datasets

- **PASCAL Visual Object** Classes (VOC) 2012 dataset contains 20 object categories. It is split into three subsets: 1,464 images for training, 1,449 images for validation and a private testing set
- **COCO** is a large-scale detection benchmark has 80 object categories with trainval35k split (115k images) for training, minival split (5k images), and test dev split (20k iamges) without any annotation
- **OpenImage** is released by Google. Its latest v6 version has 1.6M boxes of 600 categories on 1.9M images, making it the largest existing dataset with object location annotations. However, currently only some large companies“play” it as it requires a huge amount of computational resources

R-CNN

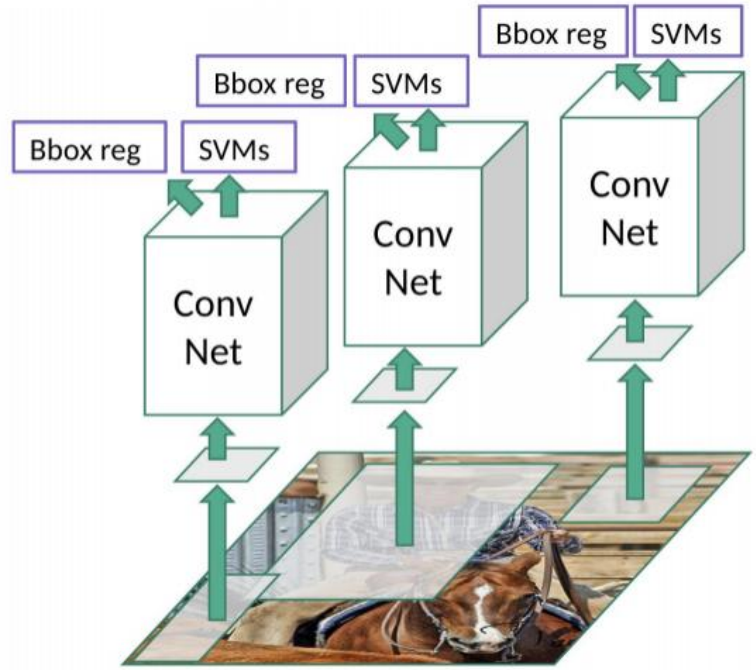
- R-CNN is short for "Region-based Convolutional Neural Networks". It follows the pipeline of the previous conventional method but replaces the SVM classifier with the CNN classifier



- Category-independent region proposals are obtained via selective search
- Region proposals are resized to a fixed size as required by a CNN
- Train the CNN for (K + 1)-class classification (background + K foreground classes). Abandon the last FC (classification) layer (in their original paper)
- Given every image region, one forward propagation through the CNN generates a feature vector, which is then fed into a binary SVM trained for each class independently

Problems with R-CNN

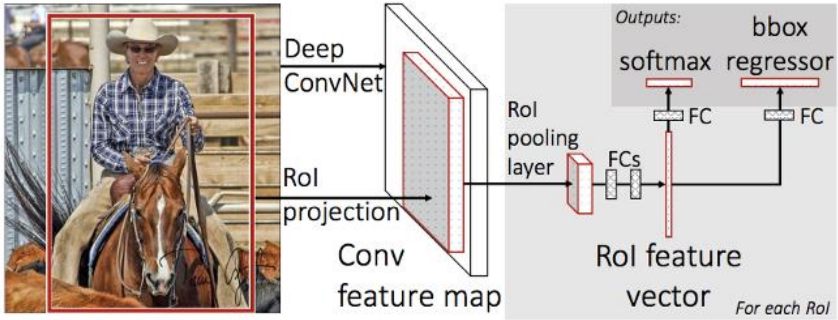
- It still takes a huge amount of time to train the network as you would have to classify 2,000 region proposals per image
- It cannot be implemented in real time as it takes around 47 seconds for each test image
- The selective search algorithm is a fixed algorithm. Therefore, no learning is happening at that stage. This could lead to the generation of bad candidate region proposals

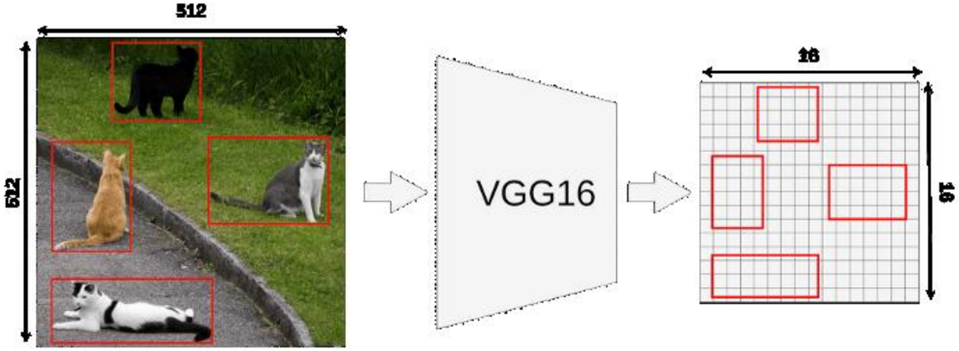


Rich feature hierarchies for accurate object detection and semantic segmentation, In CVPR 2014

Fast R-CNN

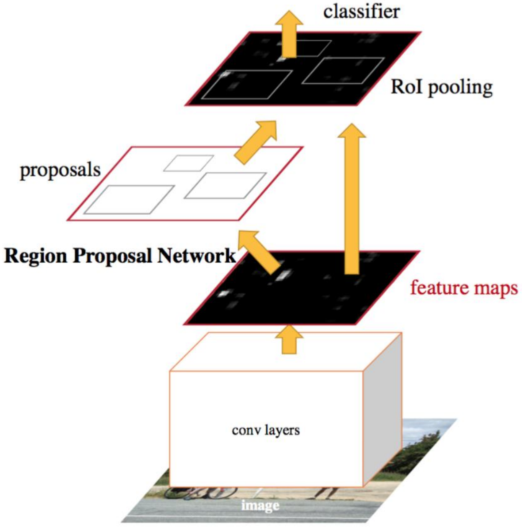
- Fast R-CNN is proposed to solve some issues of the R-CNN
- Instead of cropping and resizing image patches from the input image, the input image is first processed by a CNN to obtain its feature maps, whose spatial size is generally smaller than the original image (e.g., 1/32 of the input image size)
- Given a region proposal, it is resized to match the size of the feature map
- The RoI pooling operation is introduced to crop the feature maps with the resized object proposal, and the cropped feature maps are then further resized to a fixed spatial size (7 × 7 × 512 in the original paper)
- Positive samples: IoU$\geq$ 0.5; Negative samples: IoU$\in$ [0.1, 0.5). Each mini-batch consists of 1:3 positive and negative samples

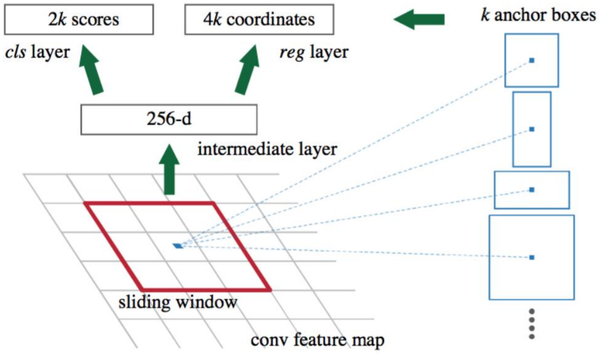




Fast R-CNN

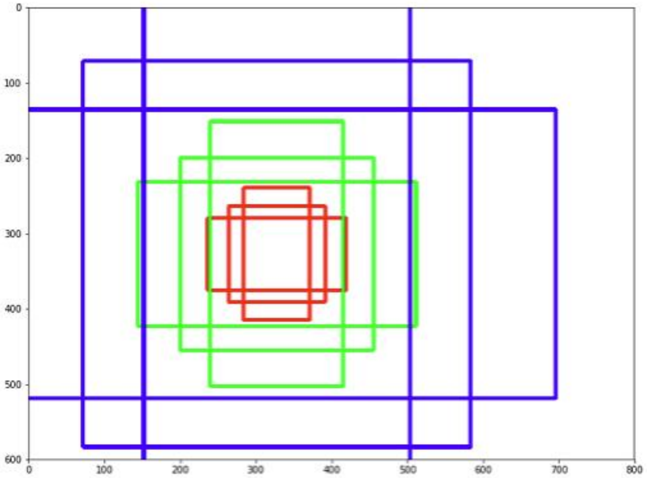
- **Problems with Fast R-CNN:**
  - RoI pooling (we skip details here) quantizes RoI coordinates into image coordinates. The pooled features are not accurate
  - Region proposals are not learned from training data
- Faster R-CNN was proposed to make the region proposals generated from the CNN
- The first part of the CNN is called **Region Proposal Network (RPN)** to generate object proposals. The RPN first generates an n × n feature maps, which are then processed by 3 × 3+1 × 1 conv to generate

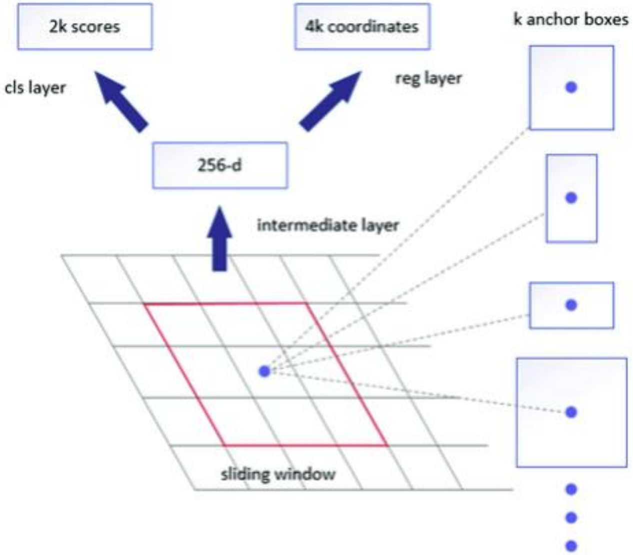




Faster R-CNN

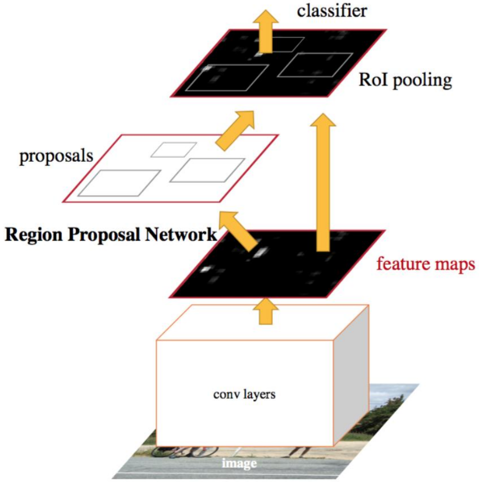
- At each of the n × n spatial location of the topmost feature maps from the RPN, $k = 9$ anchor boxes are used as initial boxes for estimating object proposals, i.e., offsets are predicted based on the initial boxes to predict the object proposals
- 3 scales and 3 aspect ratios (3 × 3 = 9 anchor boxes) are considered at each spatial location
- For each spatial location, 3 × 3 conv + 1 × 1 conv layers are built up on the RPN topmost feature maps, to predict $2k$ proposal (yes/no) confidence scores and $4k$ proposal regression (center, size, width, height)
- After the region proposals are generated, NMS (IoU threshold 0.7) is performed for each class independently to remove duplicate region proposals with the predicted proposal confidence scores

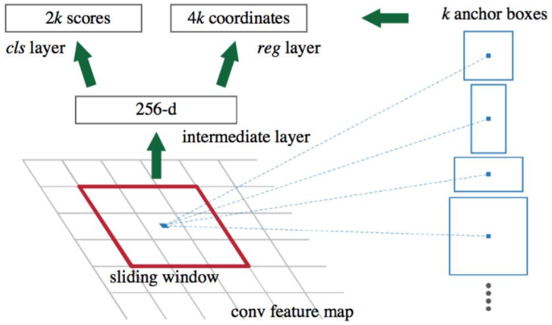




Faster R-CNN

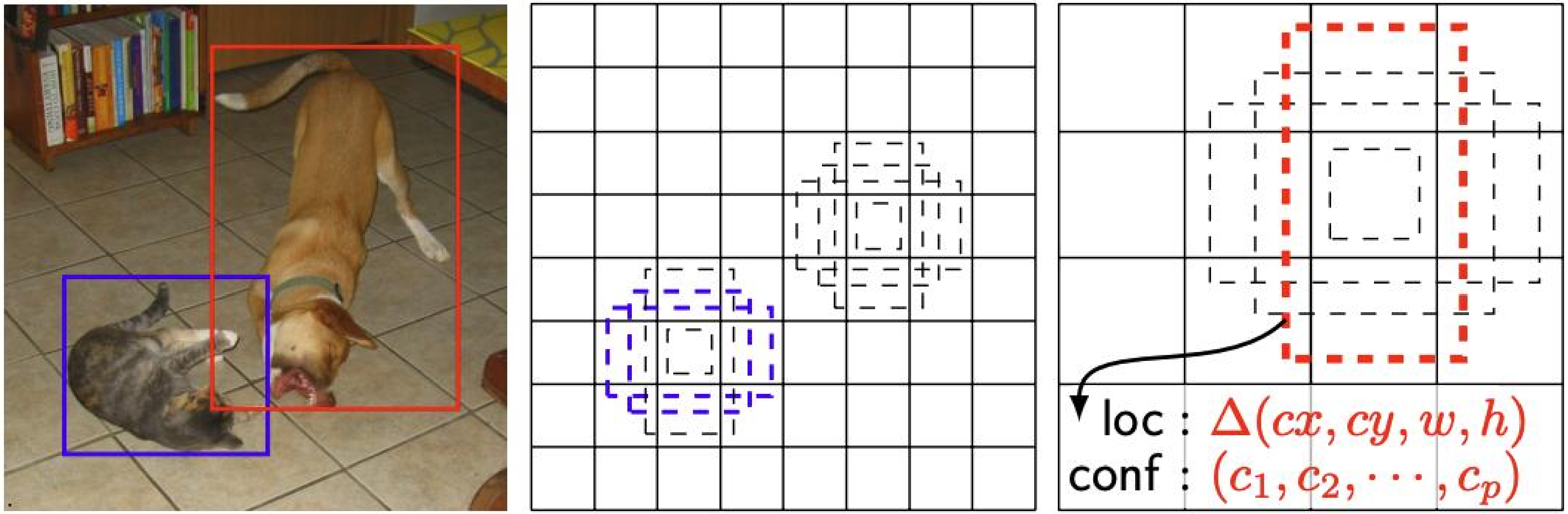
- For region proposal, two types of positive labels: 1) the anchor(s) with the highest IoU overlap with a ground-truth box; 2) an anchor that has an IoU overlap higher than 0.7. Negative labels: an anchor has an IoU lower than 0.3 for any ground-truth box
- As there are generally more negative samples than positive samples in each image. Each image contains 1:3 to 1:1 positive and negative samples. If not enough positive samples, pad the mini-batch with the negative ones
- After the RPN is trained, it can generate region proposals, which RoI pools features from the proposals for training a Fast R-CNN model with a classification head and a regression head
- Faster R-CNN shares parameters for both the RPN and the Fast R-CNN model, and jointly traines both RPN and Fast R-CNN





Single Shot MultiBox Detector (SSD)

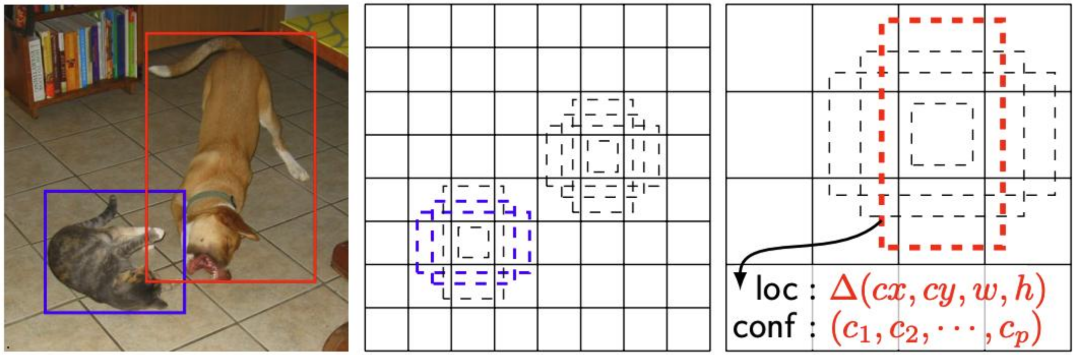
- The previous methods are called two-stage methods, as they generate region proposals first, and use an additional classification-and-regression head for refining the proposals
- Single-shot Detector (SSD) is an one-stage detector. It can be viewed as just using the RPN part of Faster R-CNN to generate object detection boxes


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

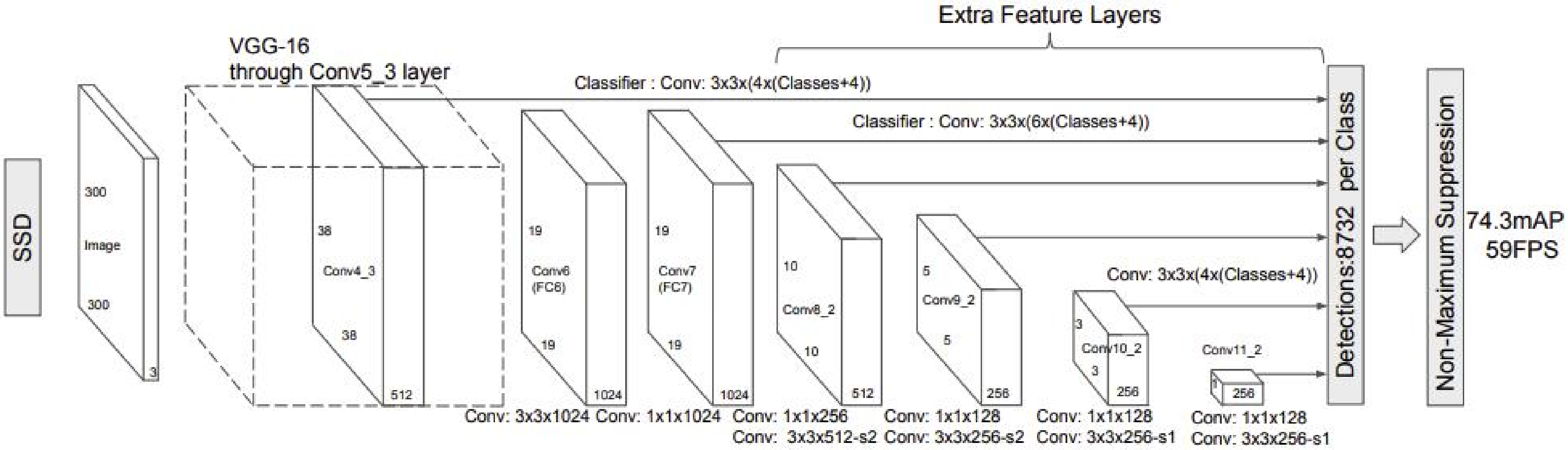
(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Single Shot MultiBox Detector (SSD)

- It doesn’t adopt FPN but still uses feature maps of different scales to generate object detection boxes corresponding to different scales

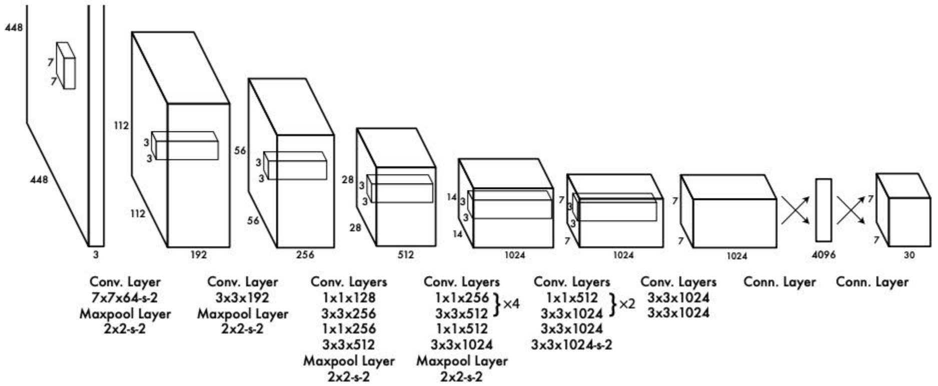

loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map



YOLO (You Only Look Once): A Real-time One-stage Detector

- The input image is converted into 7 × 7 × 1024 feature maps. S × S cells (7 × 7 in the original paper) are used to estimate object detection boxes
- Each cell predicts multiple B (B = 2) bounding boxes and their confidence scores. If a cell covers a portion of an object, it will be potentially considered to be responsible for detecting the object during training
- At training time we only want one bounding box predictor to be responsible for each object. We assign one predictor to be “responsible” for predicting an object based on which prediction has the highest current IOU with the ground truth
- There is a series of YOLOv2, v3, v4 ... that further improve both accuracy and efficiency



Conv. Layer
7x7x64-s-2
Maxpool Layer
2x2-s-2

Conv. Layer
3x3x192
Maxpool Layer
2x2-s-2

Conv. Layers
1x1x128
3x3x256
1x1x256
3x3x512
Maxpool Layer
2x2-s-2

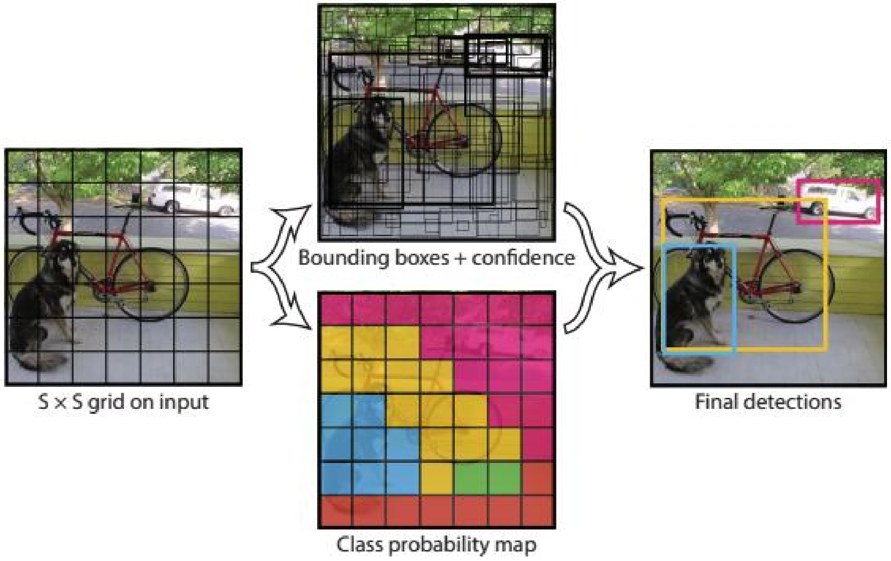
Conv. Layers
$\left.\begin{matrix}1x1x256\\3x3x512\end{matrix}\right\}\times 4$
1x1x512
3x3x1024
Maxpool Layer
2x2-s-2

Conv. Layers
$\left.\begin{matrix}1x1x512\\3x3x1024\end{matrix}\right\}\times 2$
3x3x1024
3x3x1024-s-2

Conv. Layers
3x3x1024
3x3x1024

Conn. Layer

Conn. Layer



S × S grid on input

Bounding boxes + confidence

Class probability map

Final detections

References

- https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html
- Uijlings, J. R., Van De Sande, K. E., Gevers, T. and Smeulders, A. W., 2013. Selective search for object recognition. International journal of computer vision, 104(2), pp.154-171.
- Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.
- Girshick, Ross. “Fast R-CNN”. In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.
- Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards real-time object detection with region proposal networks.” arXiv preprint arXiv:1506.01497 (2015).
- Shrivastava, Abhinav, Abhinav Gupta, and Ross Girshick. “Training region-based object detectors with online hard example mining.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 761-769. 2016.

Summary

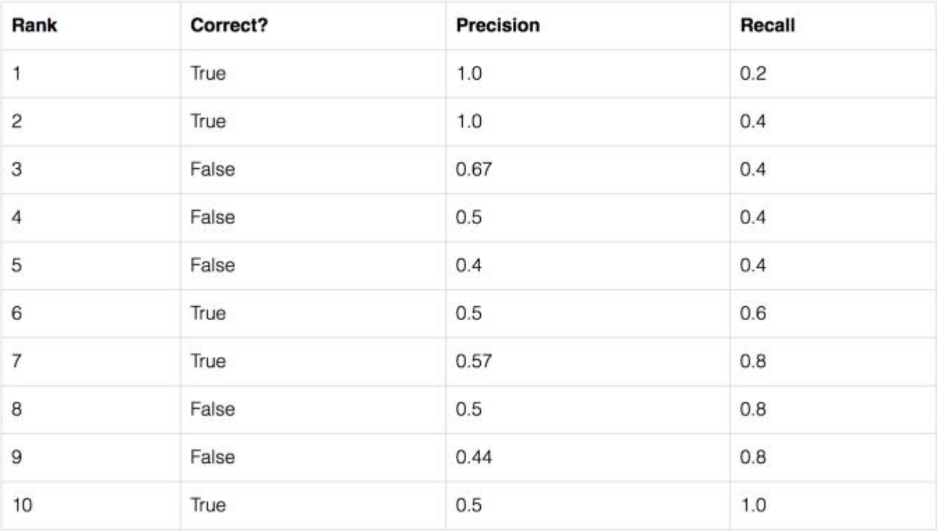
- Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. “SSD: Single shot multibox detector.” In European conference on computer vision, pp. 21-37.
- Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar. “Focal loss for dense object detection.” In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988. 2017.
- Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788. 2016.
- Tian, Zhi, Chunhua Shen, Hao Chen, and Tong He. “FCOS: Fully convolutional one-stage object detection.” In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627-9636. 2019.
- He, Kaiming, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick. “Mark R-CNN” In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969. 2017.

Evaluation Metric

- For a given object detection box, it is considered correct if its IoU with a ground-truth box is over 0.5 (sometimes other thresholds are also used). If the IoU> 0.5, the detection box is considered true positive (TP), otherwise it is considered as false positive (FP)
- Precision and recall are defined as:

$$\text{Preciesion} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

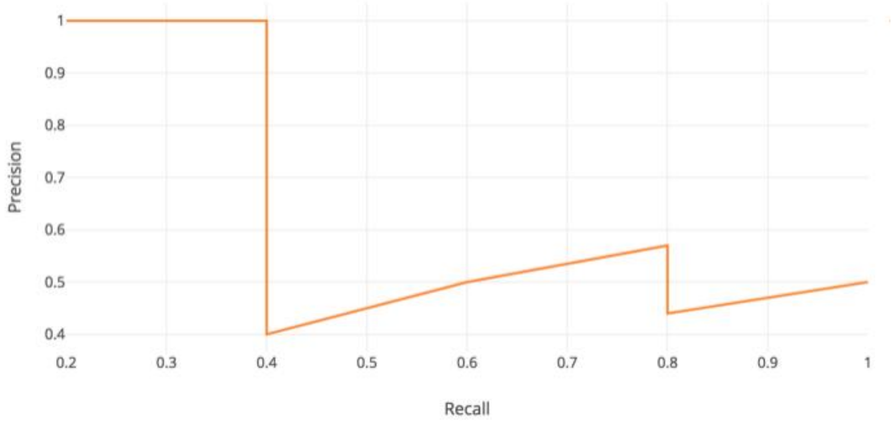
- Assume that there are 5 objects in an image and a model generates 10 boxes
- The boxes are first ranked according to their predicted confidence scores



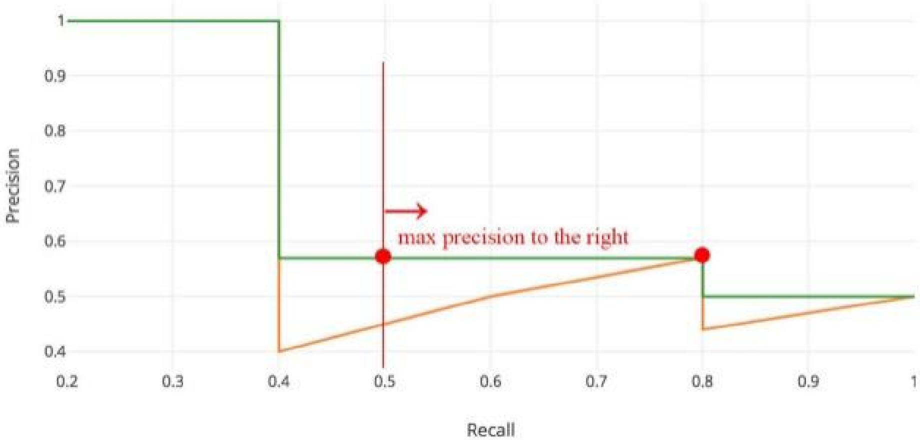
Rank	Correct?	Precision	Recall
1	True	1.0	0.2
2	True	1.0	0.4
3	False	0.67	0.4
4	False	0.5	0.4
5	False	0.4	0.4
6	True	0.5	0.6
7	True	0.57	0.8
8	False	0.5	0.8
9	False	0.44	0.8
10	True	0.5	1.0

Evaluation Metric

- If we count increasing numbers of detection boxes as the final results, the recall increases and the precision gradually decreases



- We smooth out the P-R curve



Bounding Box Regression in R-CNN

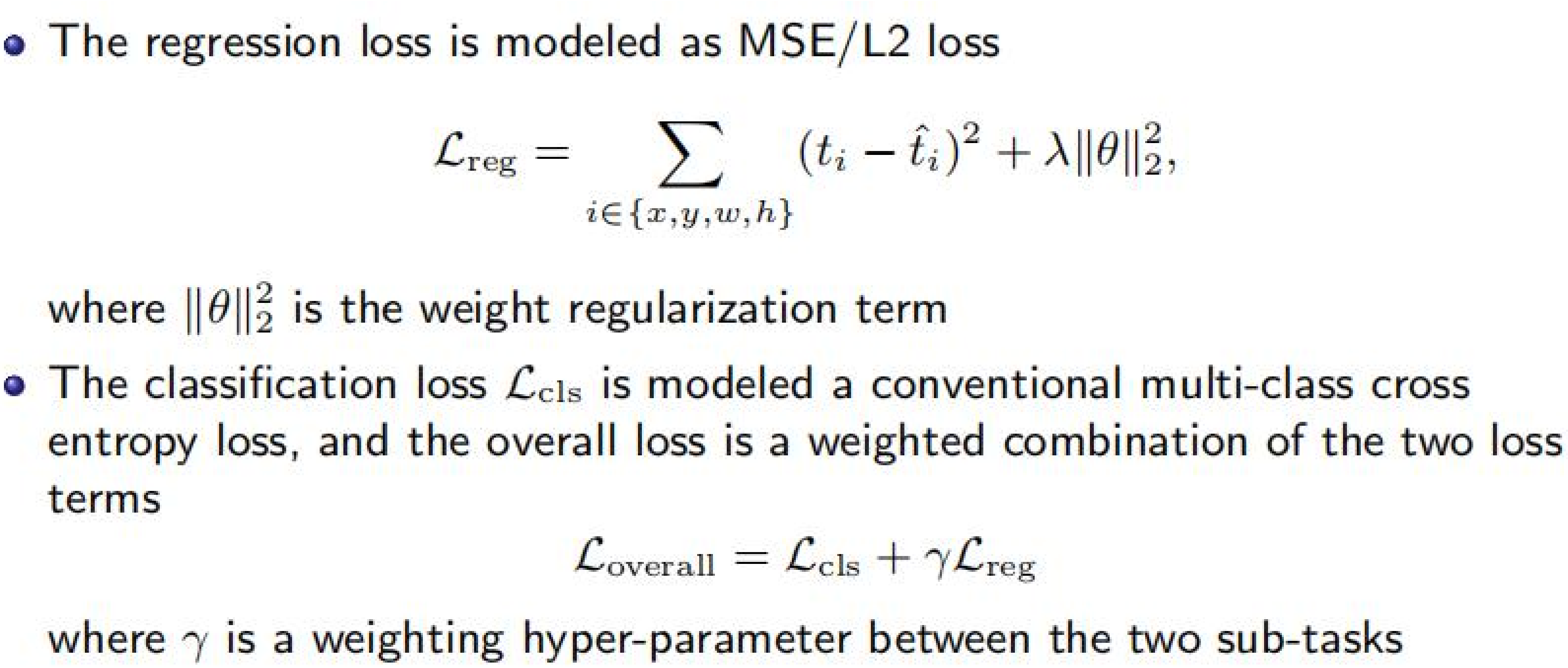
- The positive samples are proposed regions with IoU≥ 0.3, and negative samples are irrelevant others.
- The positions of the candidate boxes generated by selective search might not be accurate
- A regression head is trained to correct the predicted detection box by estimating correction offsets using CNN features
- Given a region proposal's box coordinate $\mathbf{p} = (p_x, p_y, p_w, p_h)$ (center coordinate, width, height) and its corresponding ground truth box coordinates $\mathbf{g} = (g_x, g_y, g_w, g_h)$, the regressor is trained to predict the following targets

$$\begin{aligned} t_x &= (g_x - p_x)/p_w \\ t_y &= (g_y - p_y)/p_h \\ t_w &= \log(g_w/p_w) \\ t_h &= \log(g_h/p_h) \end{aligned}$$

- Given the predicted $\hat{t}_x, \hat{t}_y, \hat{t}_w, \hat{t}_h$, the modified box $(\hat{g}_x, \hat{g}_y, \hat{g}_w, \hat{g}_h)$ becomes

$$\begin{aligned} \hat{g}_x &= p_w\hat{t}_x + p_x \\ \hat{g}_y &= p_h\hat{t}_y + p_y \\ \hat{g}_w &= p_w \exp(\hat{t}_w) \\ \hat{g}_h &= p_h \exp(\hat{t}_h) \end{aligned}$$

The regression loss and the overall loss of R-CNN



- The regression loss is modeled as MSE/L2 loss

$$\mathcal{L}_{\text{reg}} = \sum_{i \in \{x,y,w,h\}} (t_i - \hat{t}_i)^2 + \lambda ||\theta||_2^2,$$

where $||\theta||_2^2$ is the weight regularization term

- The classification loss $\mathcal{L}_{\text{cls}}$ is modeled a conventional multi-class cross entropy loss, and the overall loss is a weighted combination of the two loss terms

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{reg}}$$

where γ is a weighting hyper-parameter between the two sub-tasks

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Thanks for Listening !

Ruimao Zhang

Room 517, Daoyuan Building, The Chinese Univeristy of Hong Kong, Shenzhen
zhangruimao@cuhk.edu.cn
ruimao.zhang@ieee.org