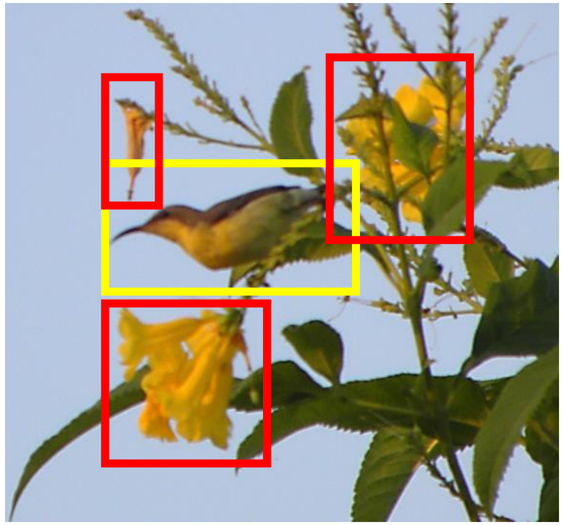
Object Detection

State-of-the-Art-Algorithms

CS8004: Deep Learning and Applications

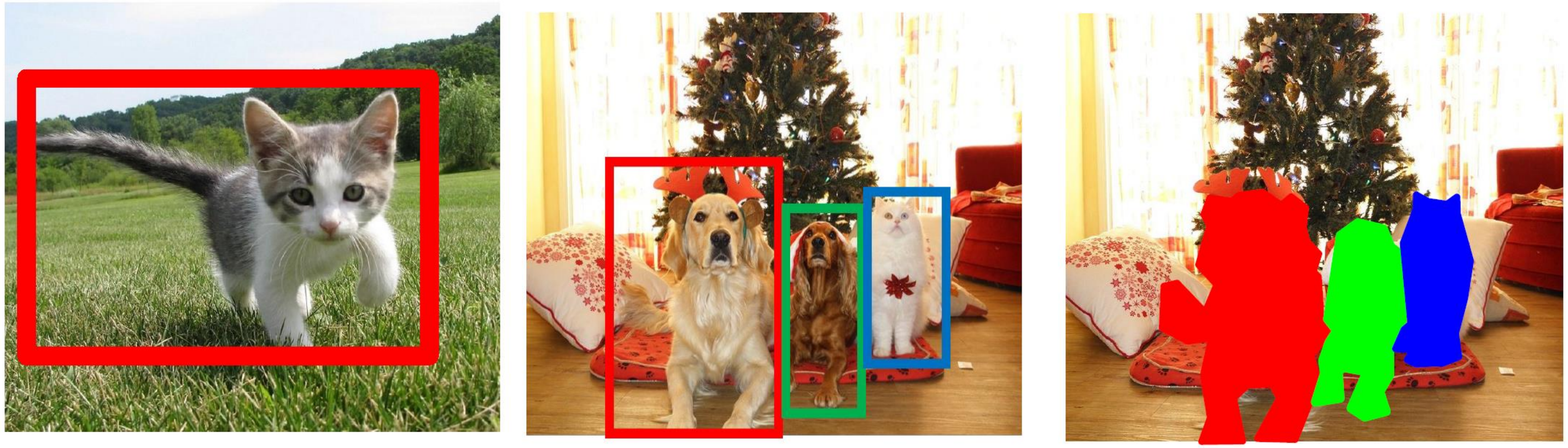


Object Detection

- What all objects are in the scene?
- Can you locate them ?
- How did you locate them ?

Object Classification, Object Detection and Segmentation

Classification + Localization | **Object Detection** | **Instance Segmentation**



CAT | DOG, DOG, CAT | DOG, DOG, CAT

Single Object | Multiple Object

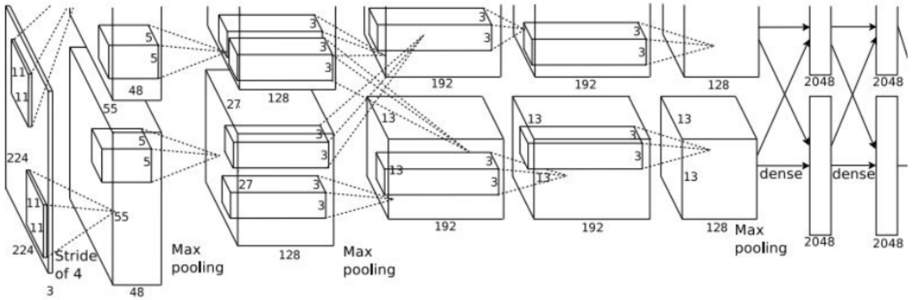
Classification and Localization

Classification with localization

Flower with bounding box

- Suppose there are five categories of objects with their corresponding labels.
  - Flower (1: [1,0,0,0,0])
  - Fruit (2: [0, 1,0,0,0,])
  - Bird (3: [0, 0, 1,0,0])
  - Insect (4: [0,0,0,1,0])
  - Background only (none of the above) (5: [0,0,0,0,1])
- CNN output would be ‘flower’ with bounding box:
  - centre, height and width.

Classification and Localization



Fully connected
4096 to 5

Class Scores

Flower: 0.92
Fruit : 0.023
Bird: 0.02
Insect: 0.07
Background: 0. 32

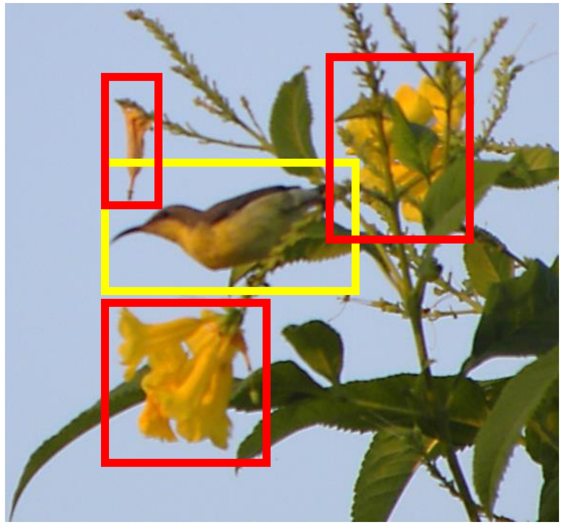
Vector:
4096

Fully Connected:
4096 to 4

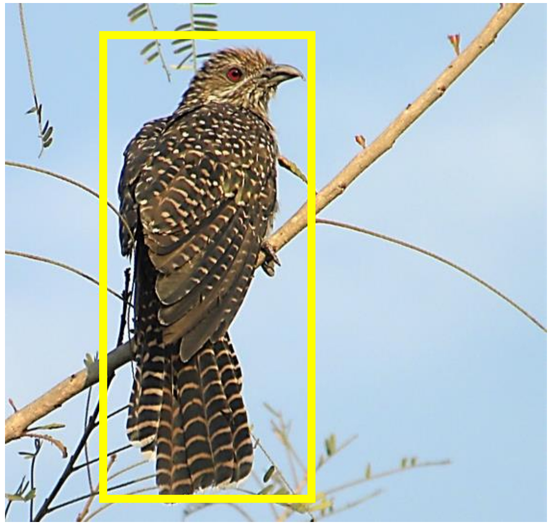
Box Coordinates
(x, y, w, h)

Localisation is a regression problem !

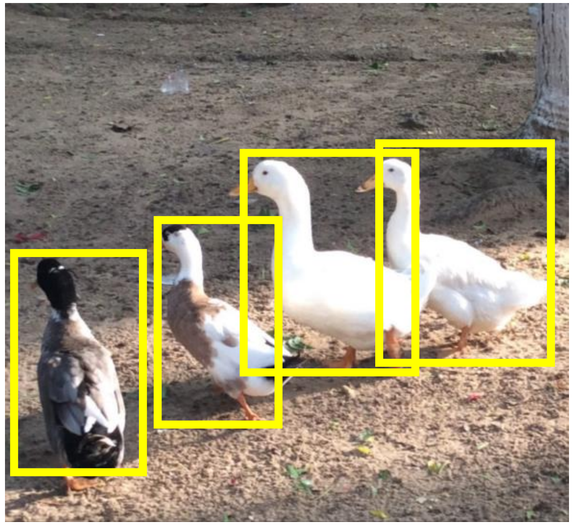
Detection as a Regression Problem



2 classes
4 boxes



1 class
1 box

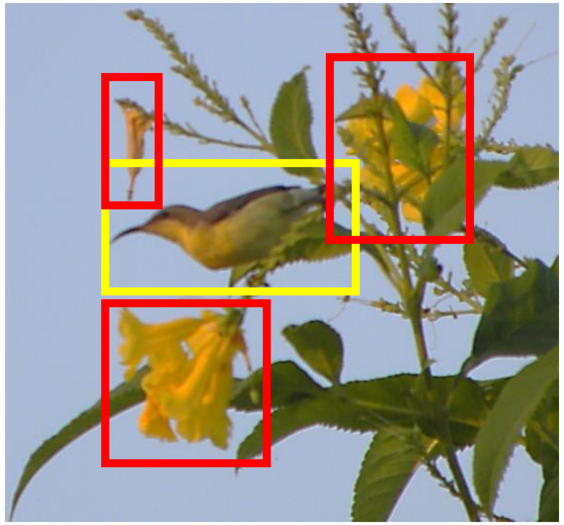


1 class
4 boxes

Each image can give different number of outputs !

Object Detection: Data Labels

- Two classes, four instances.
- How will you label?
- Five classes dataset:



([1,0,0,0,0], bounding box of flower location 1,
[0,0,1,0,0], bounding box of bird,
[1,0,0,0,0], bounding box of flower location 2,
[1,0,0,0,0], bounding box of flower location 3)

Object Detection Methods using CNN

- Two types of methods
  - Two stage methods : Initial feature extraction locally and then classification of each segmented \local region.
  - Single stage methods : both object localisation and their classification by a single pass through CNN.

Region Based CNN (R-CNN): Two Stage Method

- **Region proposal** : Propose category-independent regions of interest by selective search (~ 2000 per image)
- **Classification of regions :** Use CNN for feature extraction and SVM for **classification**

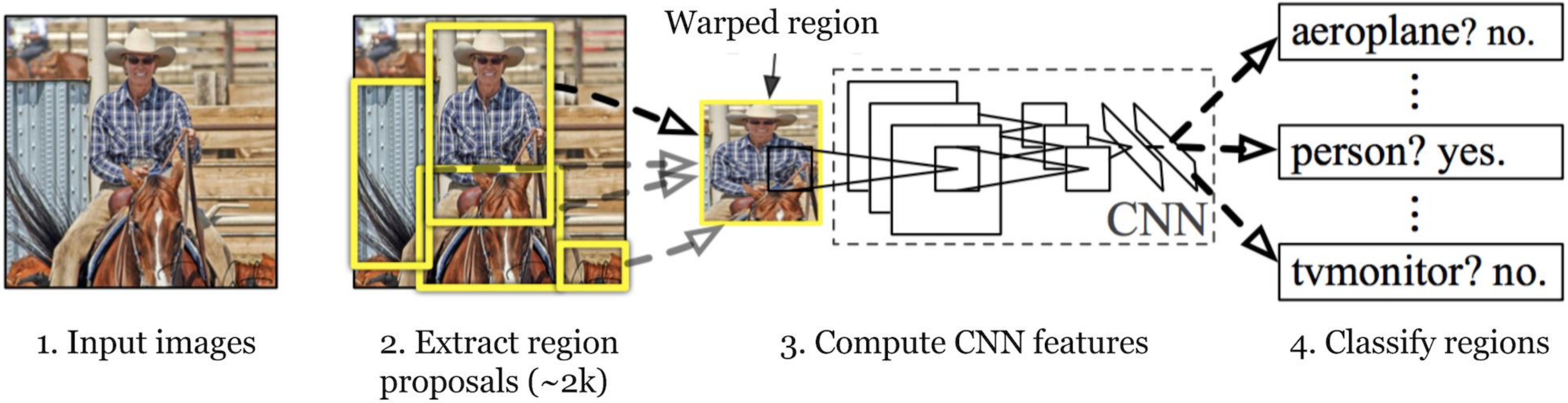


Image source: Girshick et al., 2014

Region Based CNN (R-CNN)…

- Category-independent region proposals:
  - Defines the set of candidate detections for the detector.
- A large convolutional neural network:
  - Extracts a fixed-length feature vector from each region.
- A set of class specific linear SVMs: provides binary classification for each proposal.

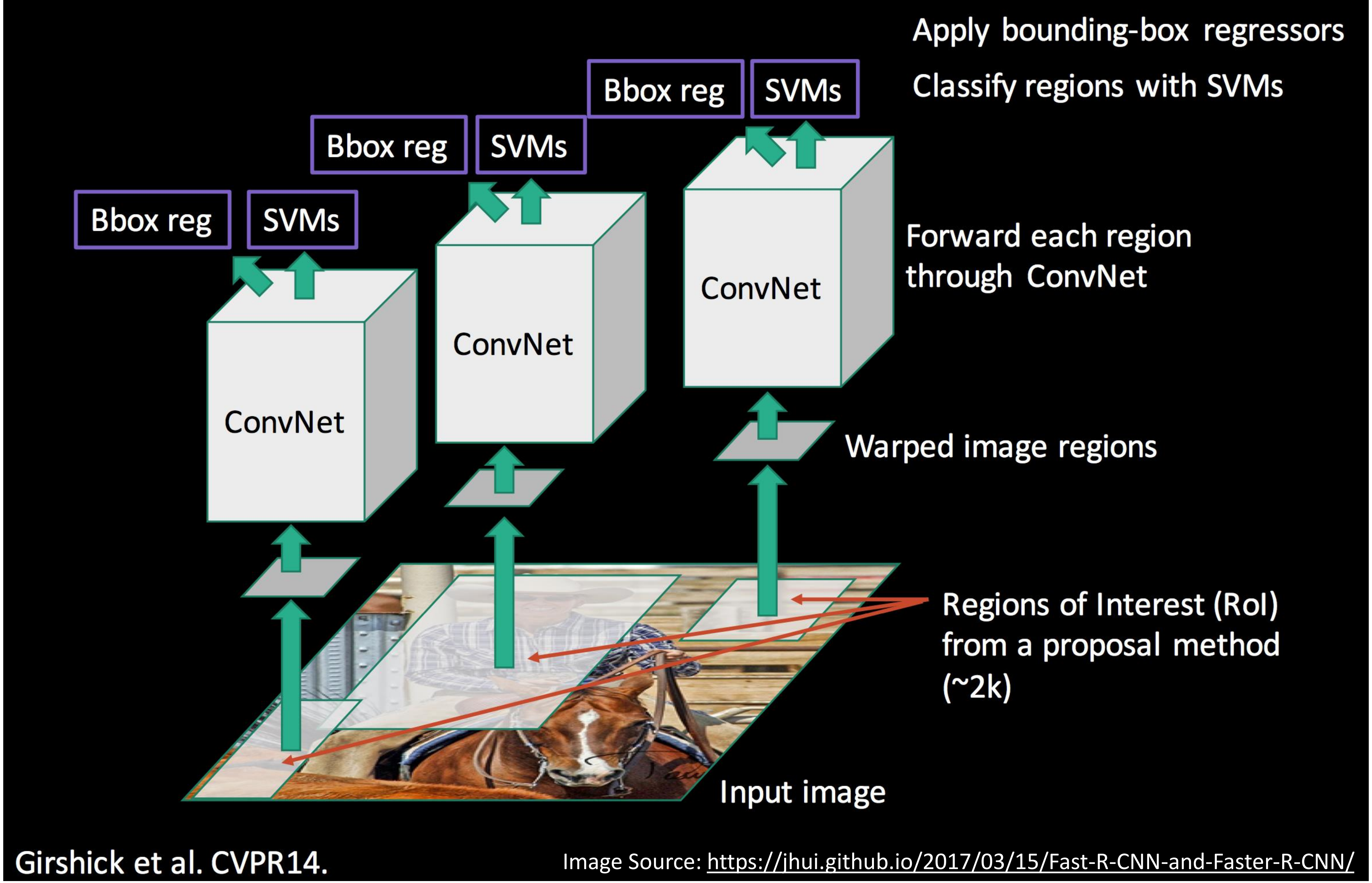
Region Based CNN (R-CNN)…

- Selective search for region proposals
  - Start with thousands of tiny initial regions. (divide the image into a grid and process the grid cells for extracting information)

  - Use a greedy algorithm to grow a region. Similar regions are merged with a similarity measure S between regions a and b defined as:

$$S(a,b) = S_{size}(a,b) + S_{texture}(a,b)$$

Selective search for region proposals



Apply bounding-box regressors
Classify regions with SVMs
Bbox reg
SVMs
Bbox reg
SVMs
Bbox reg
SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image
Girshick et al. CVPR14.
Image Source: https://jhui.github.io/2017/03/15/Fast-R-CNN-and-Faster-R-CNN/

Main Drawback of R-CNN & Fast R-CNN as an Improvement

- Very slow in training and inference.
- Nearly 2,000 region proposals are needed to be processed by a CNN to extract features.
- Therefore R-CNN repeats the CNN feature extraction process approx. 2,000 times.
- Fast R-CNN was introduced by Girshik et al, (2015) to overcome this processing issue.

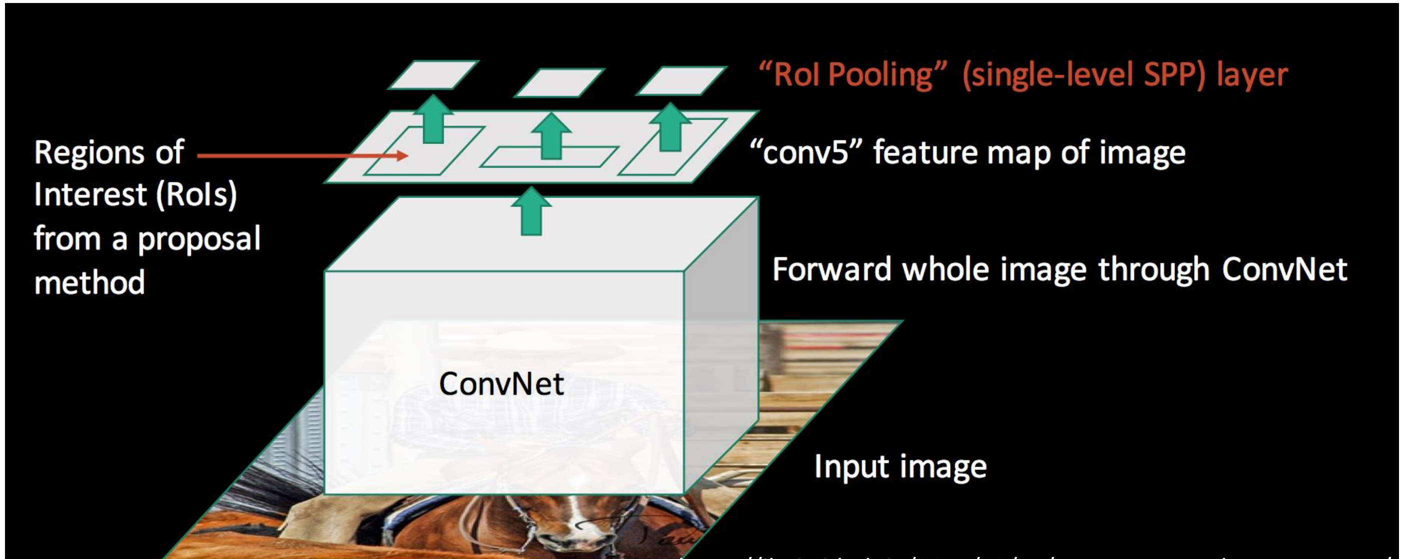


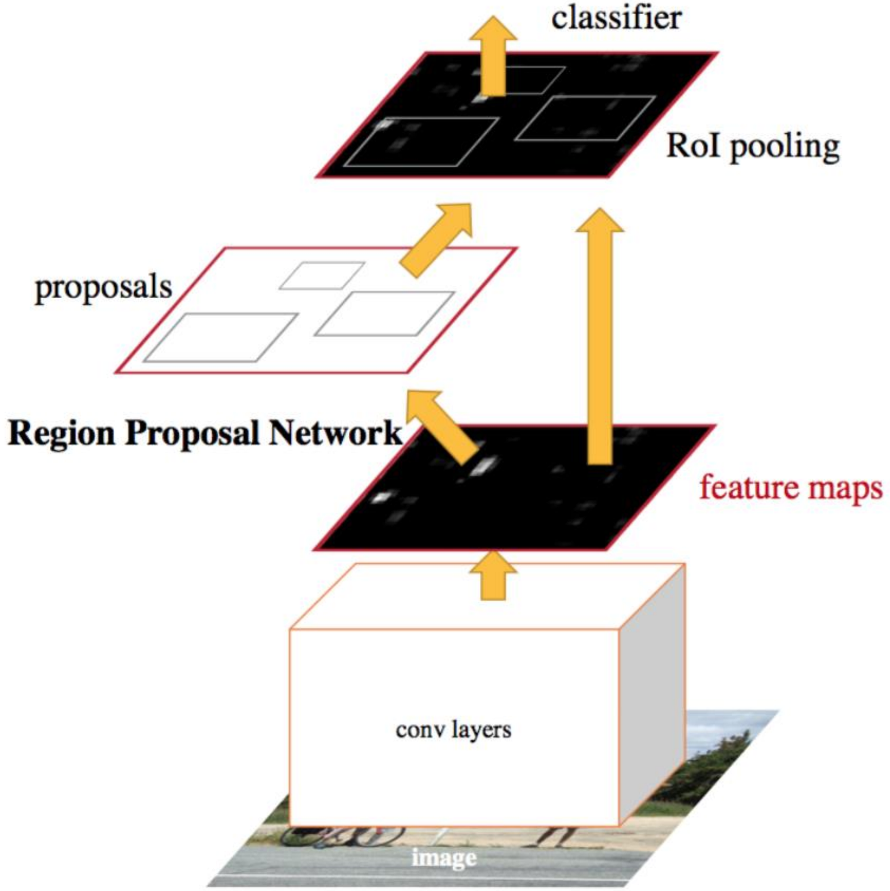
Image Source: https://jhui.github.io/2017/03/15/Fast-R-CNN-and-Faster-R-CNN/

Faster RCNN

- Faster R-CNN does not use a special region proposal method to create region proposals.

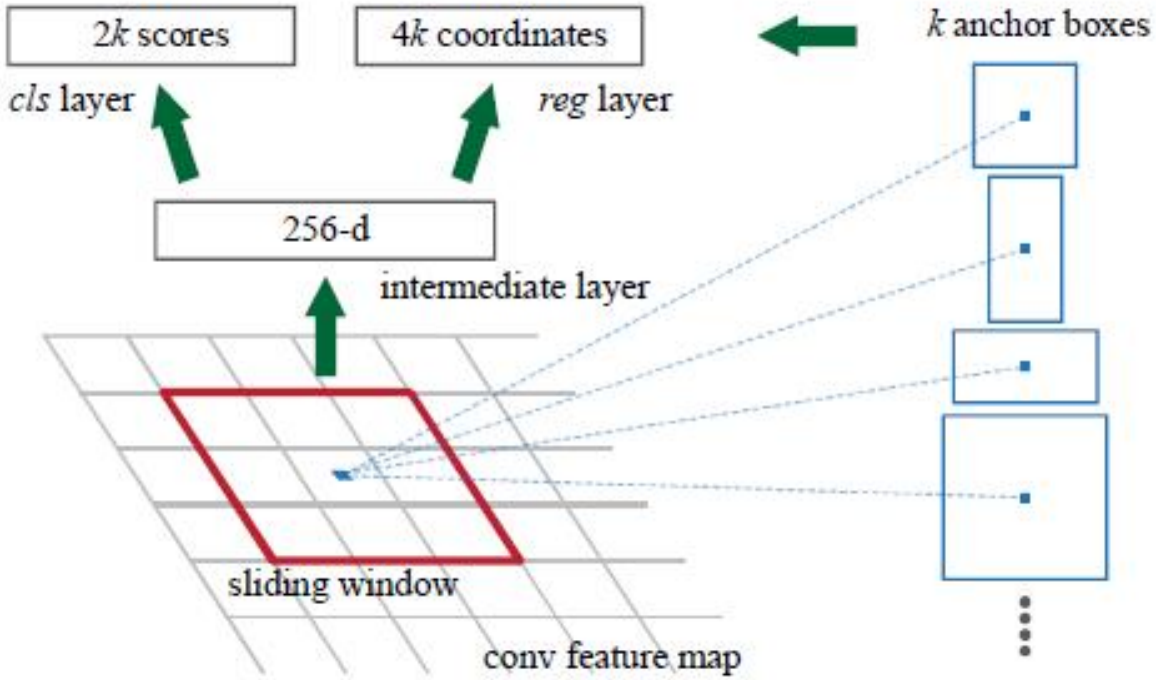
A region proposal network is trained to extract region proposals from the feature maps.

These proposals are then fed into the Region of interest (RoI) pooling layer in the Fast R-CNN type network.



Ren et al, {2015)

Region Proposal Network



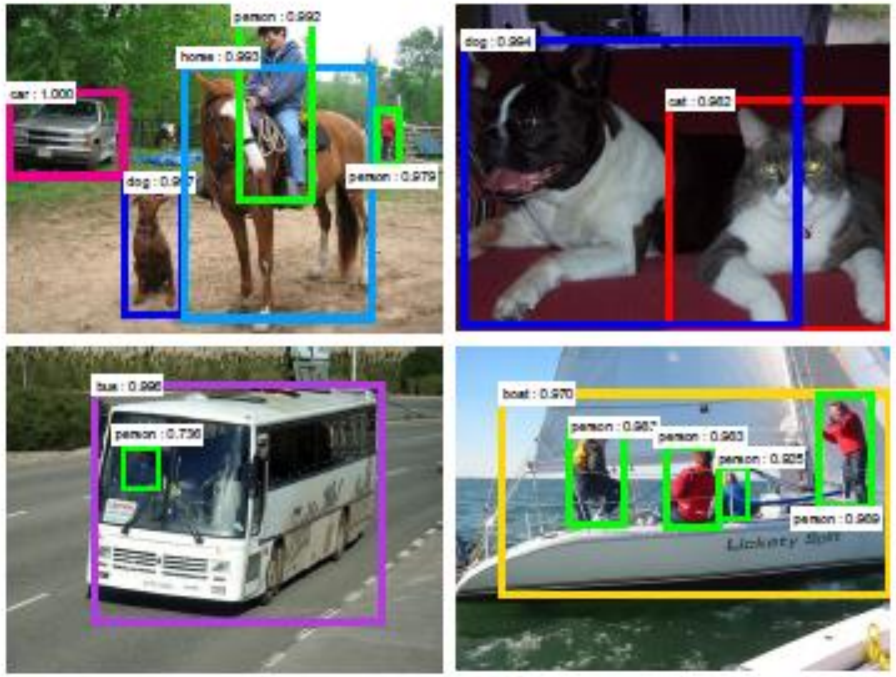


Image source: Ren et al, {2015)

Exhibited a good performance of upto 17 frames per second fps processing and 70% mAP 9 mean average precision.

Yet not suitable for real time applications.

Next

- Single Shot Object Detection

References

- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(1), 142-158. (first appeared in 2014).
- Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
- Andrew Ng Course on Convolutional Neural Networks, Coursera (deeplearning. ai)
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
- Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., & Ling, H. (2019, July). M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9259-9266).