Review: R-CNN (Object Detection)

SH Tsang Follow
Aug 31, 2018 · 4 min read

Region-CNN (R-CNN) [1] is one of the state-of-the-art **CNN-based deep learning object detection approaches**. Based on this, there are **fast R-CNN** and **faster R-CNN** for faster speed object detection as well as **mask R-CNN** for object instance segmentation. On the other hand, there are also other object detection approaches, such as **YOLO** and **SSD**.

To know deep learning object detection approach well, R-CNN is a must read item. And it is a **2014 CVPR paper with about 6000 citations** at the moment I was writing this story. (SH Tsang @ Medium)

To have object detection, we need to know the class of object and also the bounding box size and location.

Conventionally, for each image, there is a **sliding window** to search every position within the image as below. It is a simple solution. However, different objects or even same kind of objects can have **different aspect ratios and sizes** depending on the object size and distance from the camera. And **different image sizes** also affect the effective window size. This process will be **extremely slow** if we use deep learning CNN for image classification at each location.

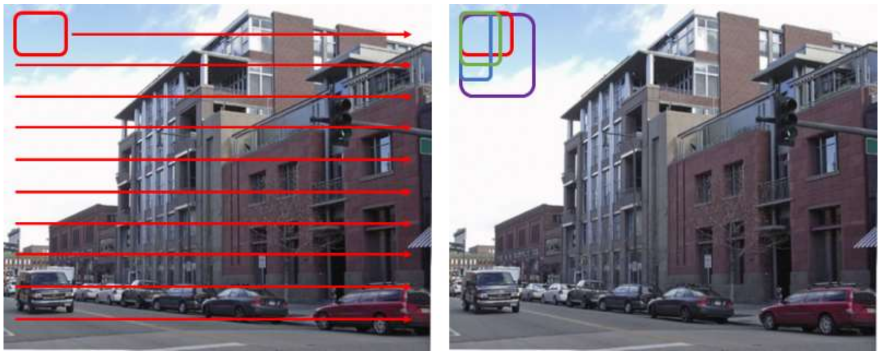
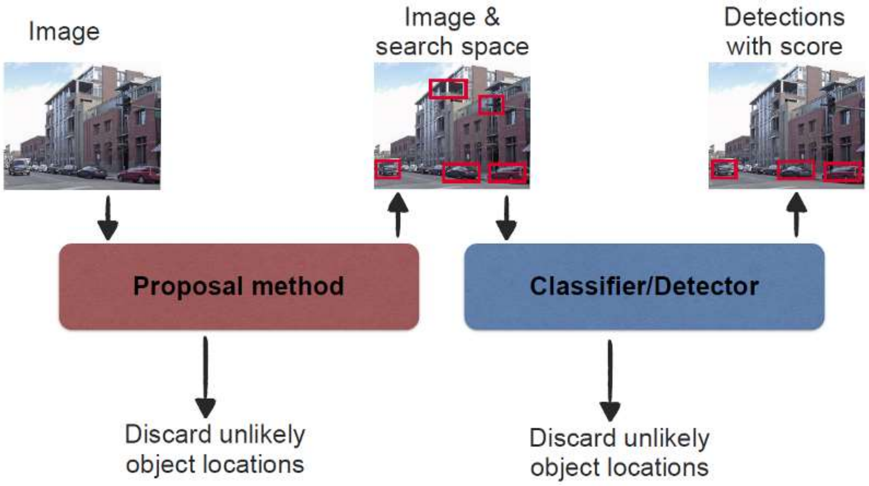


Illustration of Sliding Window (Left) with Different Aspect Ratios and Sizes (Right)

1. First, R-CNN uses selective search by [2] to **generate about 2K region proposals**, i.e. bounding boxes for image classification.
2. Then, for each bounding box, image classification is done through CNN.
3. Finally, each bounding box can be refined using regression.



R-CNN Flowchart

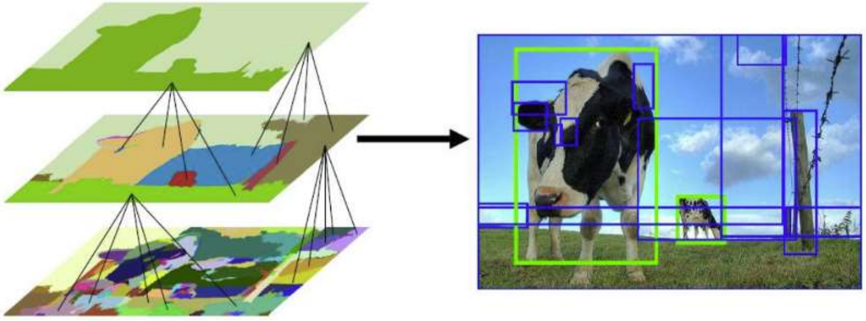
· · ·

What will be covered:

1. Selective Search
2. CNN-based Classification and Scoring
3. Results

· · ·

1. Selective Search



Selective Search

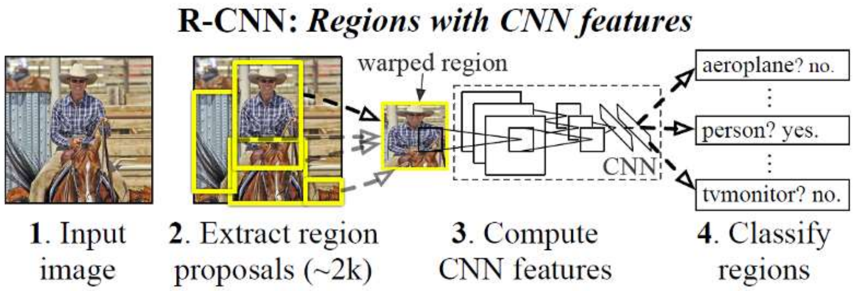
Selective search is proposed by [2].

1. First, color similarities, texture similarities, region size, and region filling are used as **non-object-based segmentation**. Therefore we obtain **many small segmented areas** as shown at the bottom left of the image above.
2. Then, bottom-up approach is used that **small segmented areas are merged together to form larger segmented areas.**

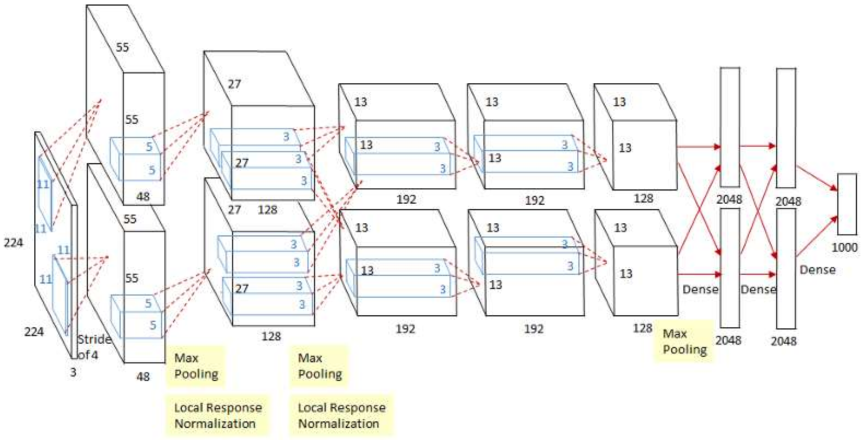
3. Thus, **about 2K region proposals (bounding box candidates) are generated** as shown in the image.

· · ·

2. CNN-based Classification and Scoring



R-CNN Flowchart with More Details



Original AlexNet

AlexNet [3] is used to extract the CNN features.

For each proposal, a 4096-dimensional feature vector is computed by forward propagating a mean-subtracted 227×227 RGB image through five convolutional layers and two fully connected layers.

The input has the fixed size of 227×227 while bounding boxes have various shapes and sizes. So, **all pixels in a tight bounding box are warped to 227×227 size.**

The feature vector is scored by SVM trained for each class.

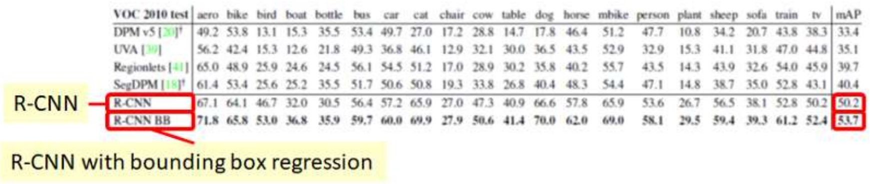
For each class, **High IoU (Intersection over Union) overlapping bounding boxes are rejected** since they are bounding the same object.

The **predicted bounding box can be further fine-tuned** by another bounding box regressor.

· · ·

3. Results

3.1 VOC 2010



VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

R-CNN

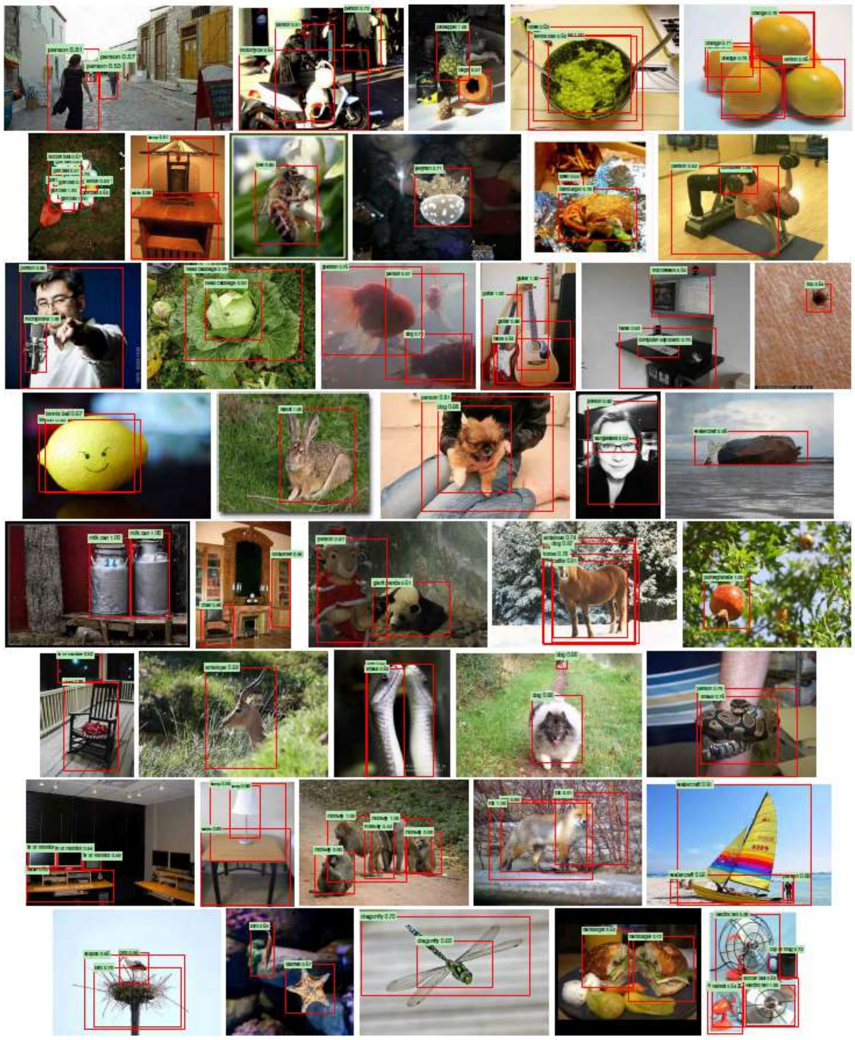
R-CNN with bounding box regression

VOC 2010

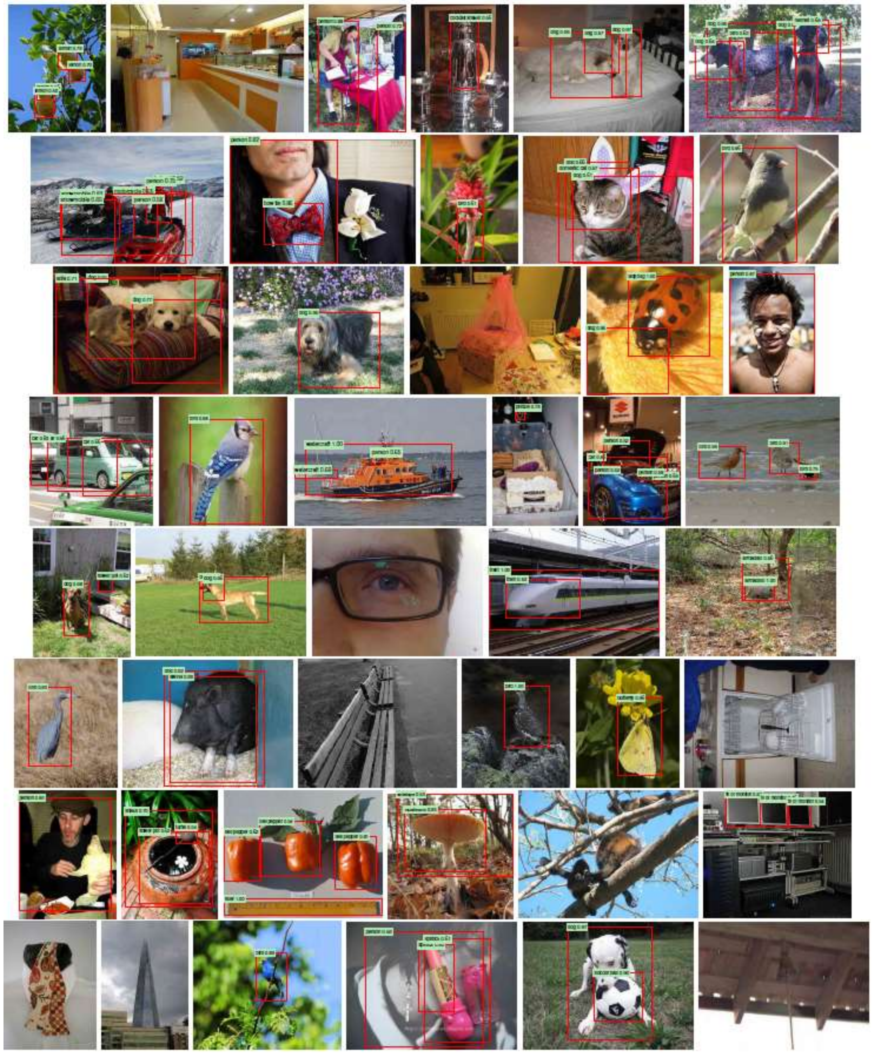
R-CNN and R-CNN BB obtain the highest mAP (mean average prediction).

· · ·

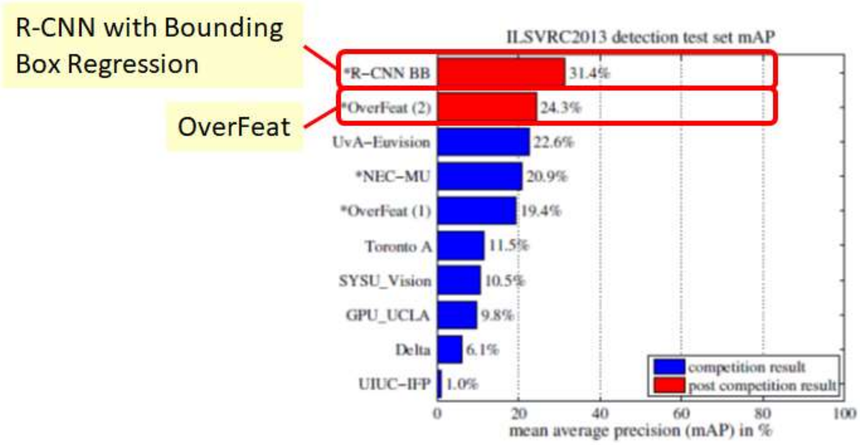
3.2 ILSVRC 2013



Some Amazing ILSVRC 2013 Results



Some ILSVRC 2013 Results with Some Missing Detections

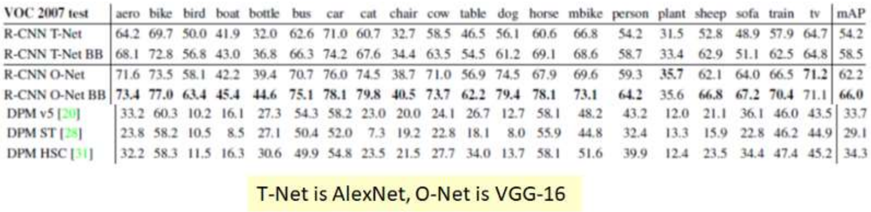


ILSVRC 2013

R-CNN BB even outperforms OverFeat [4], which is the winner of ILSVRC 2013 localization task!

3.3 VOC 2007

Some examples with high activations in VOC 2007



VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN T-Net	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN T-Net BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5
R-CNN O-Net	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	**35.7**	62.1	64.0	66.5	**71.2**	62.2
R-CNN O-Net BB	**73.4**	**77.0**	**63.4**	**45.4**	**44.6**	**75.1**	**78.1**	**79.8**	**40.5**	**73.7**	**62.2**	**79.4**	**78.1**	**73.1**	**64.2**	35.6	**66.8**	**67.2**	**70.4**	71.1	**66.0**
DPM v5 [20]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [28]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [31]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

T-Net is AlexNet, O-Net is VGG-16

VOC 2007

As you may already know, **the CNN used in R-CNN can be changed to any CNNs used in image classification.**

When R-CNN BB uses VGG-16 [5] which is a 16-layer VGGNet, mAP is even increased to 66.0%.

. . .

If interested, please read also my reviews about AlexNet, VGGNet, and OverFeat. (Links at the bottom)

And I will write more reviews for other state-of-the-art deep learning approaches.

. . .

References

1. [2014 CVPR] [R-CNN]
   Rich feature hierarchies for accurate object detection and semantic segmentation

2. [2013 IJCV] [Selective Search]
   Selective Search for Object Recognition

3. [2012 NIPS] [AlexNet]
   ImageNet Classification with Deep Convolutional Neural Networks

4. [2014 ICLR] [OverFeat]
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks

5. [2015 ICLR] [VGGNet]
Very Deep Convolutional Networks for Large-Scale Image Recognition

My Reviews

1. Review: AlexNet, CaffeNet—Winner of ILSVRC 2012 (Image Classification)

2. Review: OverFeat—Winner of ILSVRC 2013 Localization Task (Object Detection)

3. Review: VGGNet—1st Runner-Up of ILSVRC 2014 (Image Classification)