Weakly Supervised Region Proposal Network and Object Detection

Shubham Sharma(18I190002)
Under the guidance of Prof. P Balamurugan

November 23, 2019

Index

1. Object Detection Problem
2. Weakly Supervised Object Detection
3. Before Mid-term Work
4. Dataset used
5. Step by step project performed

Object Detection

- Object detection in images are computer vision problems that deals with detecting instances of semantic objects of a certain class

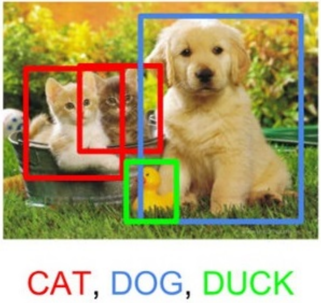

Figure 1: Object detection; Source:google

Weakly Supervised Object Detection(WSOD)

- WSOD is the task of training object detectors with only image tag supervisions
- It is laborious and expensive to collect bounding box annotations
- Image level annotations whether an image belongs to an object class or not are much easier to acquire

Network Architecture

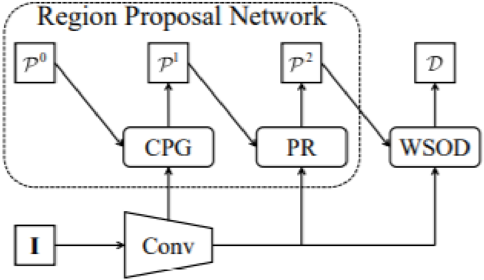

Figure 2: The overall architecture. "**I**": input image: "$\mathcal{P}^0$": The initial proposal by sliding window [1], "$\mathcal{P}^1$: the proposals from first stage of the network, "$\mathcal{P}^2$": the proposals from second stage of the network, "$\mathcal{D}$": the detection results. "Conv": convolution layers, CGP: coarse proposal generation, "PR": proposal refinement, "WSOD": weakly supervised object detection

Coarse Proposal Generation

- $\mathcal{P}^0 = \{(b_n^0, o_n^0)\}_{n=1}^{N^0} \rightarrow$ Exhaustive set of sliding window [1] boxes with various sizes and aspect ratio
- After obtaining the edge like response map, objectness score of $\mathcal{P}^0$ is calculated using Edge Boxes[2]
- Proposals with higher objectness scores are taken for $\mathcal{P}^1 = \{(b_n^1, o_n^1)\}_{n=1}^{N^1}$

Coarse Proposal Generation

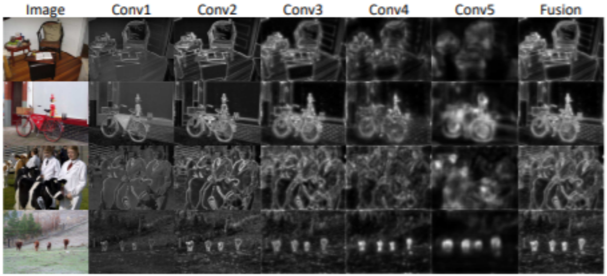

Figure 3: The responses of different convolutional layers from the VGG16[3] network trained on the ImageNet[4] dataset using only image level-levelannotations. Results from left to right are the original image. responses from the first to fifth layer, and the fusion of responses from the second layer to the forth layer

Proposal Refinement

- $\mathcal{P}^1 = \{(b_n^1, o_n^1)\}_{n=1}^{N^1}$ are still noisy as there are high responses on the background regions of the edge-like response map
- The task of PR is to find $f(\mathbf{I}, b_n^1)$ i.e., the probability of b_n^1 covering an object in image $\mathbf{I}$
- We evaluate $\tilde{o}_n^1 = h(o_n^1, f(\mathbf{I}, b_n^1)) = o_n^1.f(\mathbf{I}, b_n^1)$ to reject the proposals with low scores.
- we use faster rcnn[5] to find $f(\mathbf{I}, b_n^1)$

Proposal Refinement

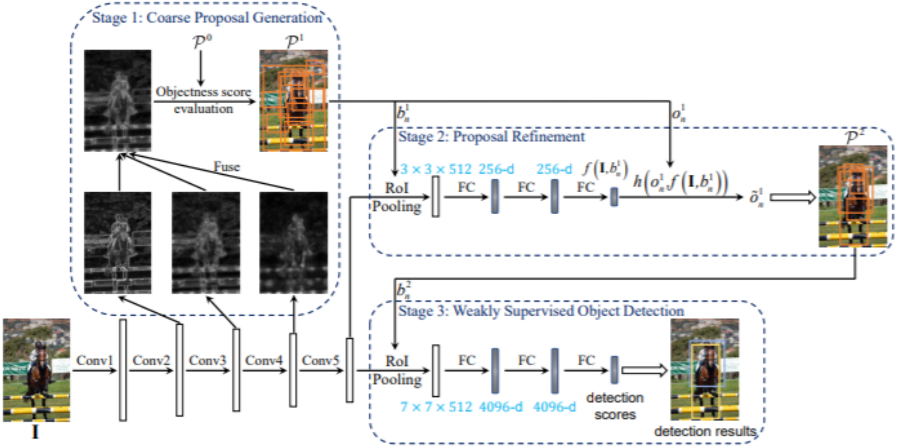

Figure 4: The detailed architecture of the network

Dataset used

- ImageNet and PASCAL VOC detection datasets
- ‘Experiments on Google Open Images dataset V4 is ain process
- This dataset file cover the 600 boxable object classes, and span the 1,743,042 training images and 125,436 testing images sets.

Step done in project

The Project has been done in 6 main steps:

- **Step 1**: Generation of bounding boxes for $\mathcal{P}^0$
- **Step 2**: Designing of the network
- **Step 3**: Preparation of the data-set for the network
- **Step 4**: Generating $\mathcal{P}^1$ from $\mathcal{P}^0$
- **Step 5**: Generating $\mathcal{P}^2$ from $\mathcal{P}^1$
- **Step 6**: Weakly Supervised Object Detection

Generation of bounding boxes for $\mathcal{P}^0$

- Used selectivesearch package for making the bounding boxes

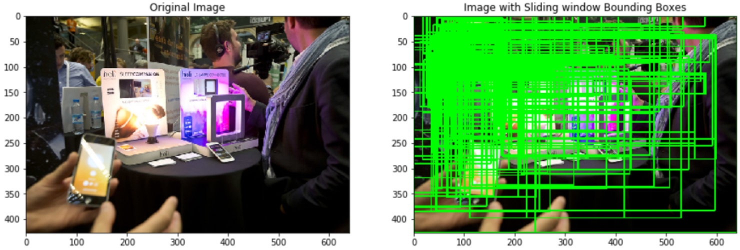

Figure 5: Results of sliding window or $\mathcal{P}^0$

Designing of the network

- A network has been designed in such a way that only this network can be used in every stage
- The network predicts two things:
  - Whether the region is object or not
  - Classifies between different types of object and none of them.
  - The network is also used in generation of proposals $\mathcal{P}^1$

Preparation of the data-set for the network

- We have extracted the regions with objects from all of these images and resized to (256,256,3) with:
  - 224 car images
  - 117 phone images
  - 281 person images
  - 622 object images
  - 269 none of these images
- These data-sets are saved in npy format to be used in the network

Generating $\mathcal{P}^1$ from $\mathcal{P}^0$

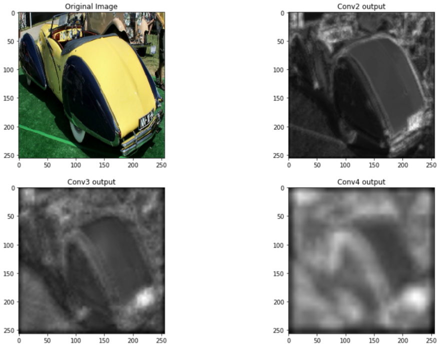

Figure 6: Outputs from conv2, conv3, conv4 of the network

Generating $\mathcal{P}^1$ from $\mathcal{P}^0$

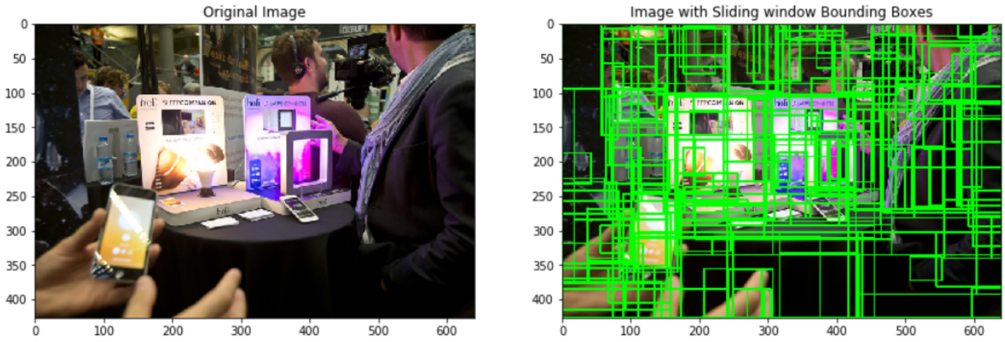

Figure 7: Proposals P^1 of the image

Generating $\mathcal{P}^2$ from $\mathcal{P}^1$

- Used 1^{st} output from the network to generate $\mathcal{P}^2$

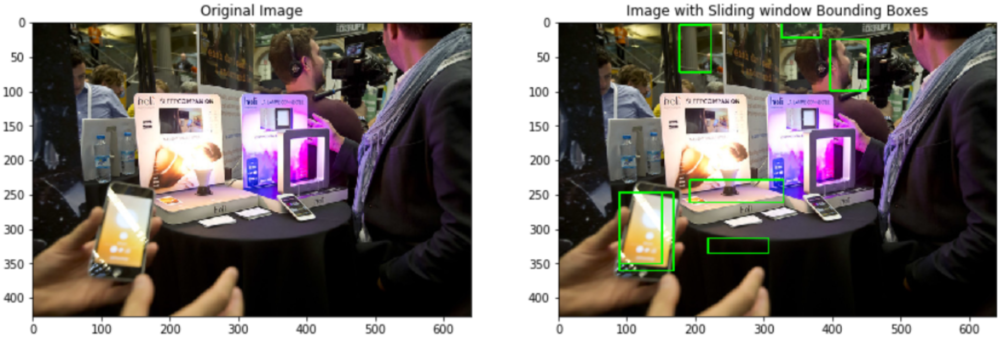

Figure 8: Proposals P^2 of the image

Weakly Supervised Object Detection

- Object detection results from $\mathcal{P}^2$

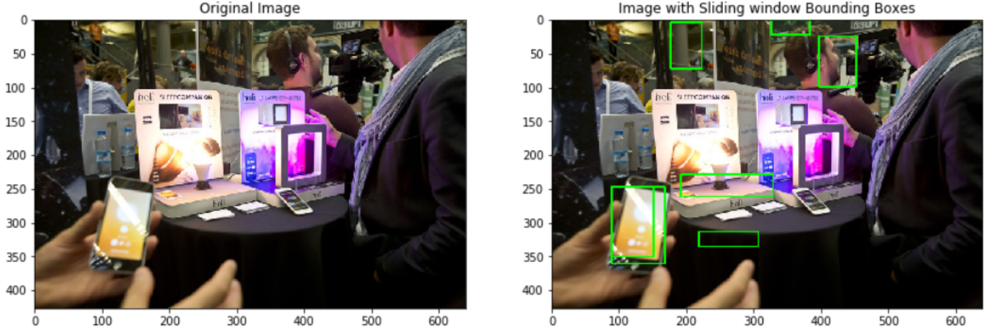

Figure 9: Object Detection results

Challenges and future works

- **Challenges**
  - The network is not very well trained as we are still having some regions that it is falsely detecting
  - Would work better if we would have some pre-trained network available
- **Future works**
  - Can check for a bigger dataset with a pre-trained model like Image-Net
  - can use auto-encoders for feature extraction for making $\mathcal{P}^1$
  - can use Canny edge detection to score the proposals $\mathcal{P}^0$

References

[1] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[2] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.