IE643: Deep Learning : Theory and Practice **Aug-Dec 2020**

Quiz 3

Lecturer: P. Balamurugan *Shashank Kumar (170050031)*

1 Question 1

1.1 Part (a)

Deep convolutional networks like VGGNet, ResNet which are trained on datasets like Pascal VOC and Imagenet cannot be used to perform the object detection tasks. Environmental conditions like mist, pollution, smoke and dust lead to poor quality images. The objects in the images are not clearly visible due to smoke and mist. Such conditions hinder the visibility level in an image and therefore, bounding boxes are either not generated around the less visible objects or the box isn't accurate and the object is not recognizable in the image. Many state-of-the-art object detection algorithms have failed to generate the desired output with such images.

Figure 1: Faster R-CNN fails to detect objects

1.2 Part (b)

The architecture that would help with the task at hand can use a combination of image enhancement neural network followed by classical object detection and recognition networks.

1. **Image enhancement:**
   Deep neural networks for image enhancement have been proposed like [1] and [2]. The network in [2] accepts a foggy image and renders an enhanced version. The datasets to be used are described in part (c). The output image generated by this network is compared with the ground truth images. Squared-error loss function is used in this case. The results can be seen in figure 2.
2. **Object recognition and detection:**
   The enhanced images obtained can be used for object detection tasks with the deep convolutional networks like Faster-RCNN and YOLO. A few other research paper like [3] use saliency map for better estimation of bounding boxes. Saliency map is a topographical representation of the visually alluring locations in an image. The saliency map of the images under foggy conditions have distinct and clear boundaries.

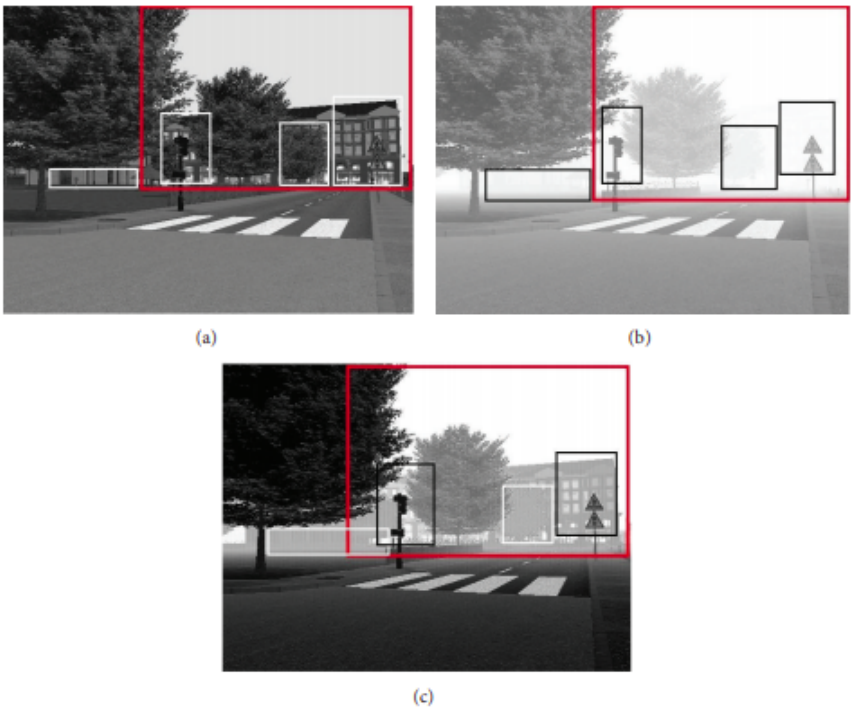

Figure 2: (a) Original scene without fog. (b) Scene visibility degraded under fog. (c) Visibility enhancement by deep neural network.

Proposed neural network architecture:
The figure 3 shows the neural network architecture proposed in [3]. Instead of using Faster-RCNN, YOLO architecture can be used for faster real-time performance. The loss function consists of a penalty term for bad localization of the center of the bounding box, a penalty for wrong height and width estimation of the bounding box and squared loss for the confidence scores

1.3 Part (c)

FRIDA (Foggy Road Image DAtabase) image database comprises of 90 synthetic images of 18 urban road scenes. FRIDA2 comprises 330 synthetic images of 66 diverse road scenes. SHIA (Spectral Hazy Image database for Assessment) is composed of 1540 images with 10 levels of fog each and their corresponding fog-free (ground-truth) images. Pascal VOC dataset can be used after preprocessing the images. A blur can be applied to each of the image while maintaining the same bounding box. The blur intensity is subjective to the image. Image enhacement neural network can use all the three datasets along with the ground truth images. The object detection neural network can be trained on the modified Pascal VOC dataset with blur.

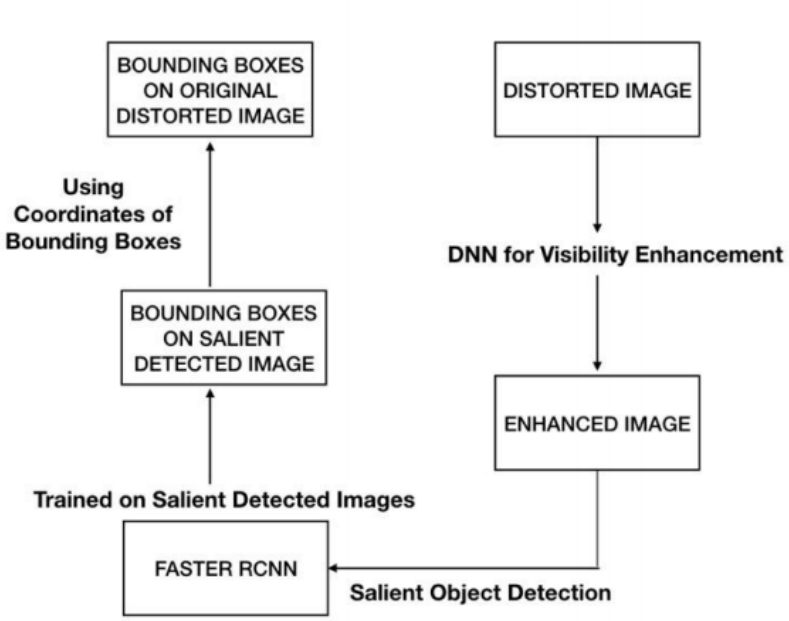

Figure 3: Neural network with image enhancement and object detection

2 Question 2

2.1 Part (a)

YOLO (You Only Look Once) algorithm [4] can be used for the detection of animated characters in the frames. Since we are considering real-time detection i.e. capturing frames from the video, YOLO is the obvious choice owing to the high detection speed. The architecture is depicted in the figure 4.

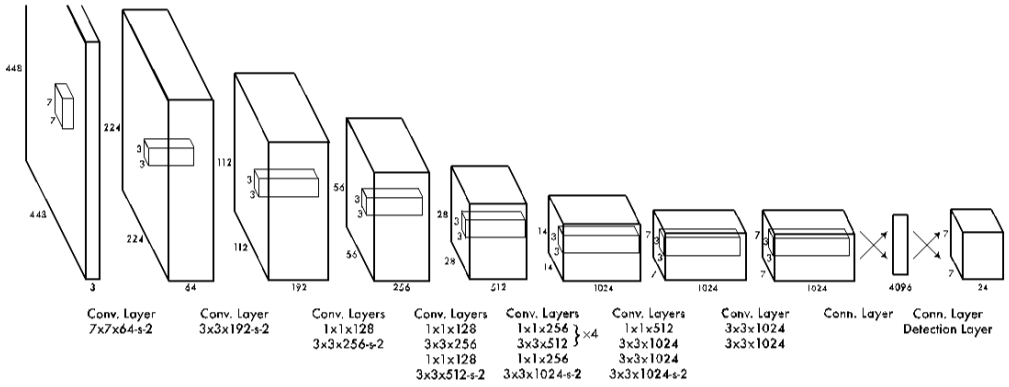

Figure 4: YOLO Architecture

Architecture:
In our case, the output of the algorithm are the probability that there is an animated character in the frame, the coordinates of the center, and the height and width of the bounding box relative to the image size. There is just one class label since we only want to detect the animated objects and not classify them. The algorithm segments the image into cells and each cell is assigned the class probability. A maximum over the class probabilities is then assigned to each cell. The image then looks as shown in figure 5. The next step involves non-max suppression which involves combining overlapping bounding boxes using intersection-over-union method. This step is depicted in figure 6 which results in the final bounding box.

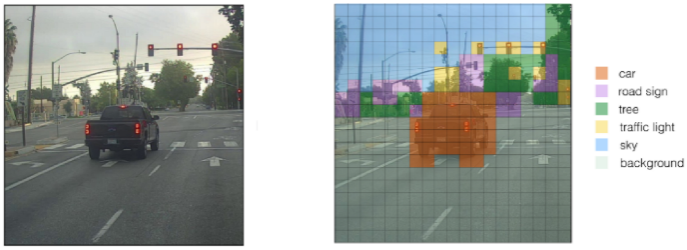

Figure 5: Image divided into cells with class probabilities

Figure 6: Non-max suppression

Loss function:
The loss function used in YOLO algorithm comprises of the sum of the following three components:

1. **a classification loss** over all the cells with the conditional class probabilities given by

$$\sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in C} (p_i(c) - \hat{p}_i(c))^2$$

where
$1_i^{obj} = 1$ if the object is in i-th cell, otherwise 0
$\hat{p}_i(c)$ denotes the conditional probability that a class c appears in the i-th cell

2. **a localization loss** corresponding to the bounding box which is as follows

$$\lambda_{coord}\left(\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[(x_i-\hat{x_i})^2+(y_i-\hat{y_i})^2+(\sqrt{w_i}-\sqrt{\hat{w_i}})^2+(\sqrt{h_i}-\sqrt{\hat{h_i}})^2\right]\right)$$

where
$1_{ij}^{obj}=1$ if the j-th boundary box in the i-th cell is responsible for detecting the object, otherwise 0
λ_{coord} increases the weight for the loss in the boundary box coordinates

3. **a confidence loss** based on whether the object is present or not given as

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[1_{ij}^{obj}(C_i-\hat{C_i})^2+\lambda_{noobj}1_{ij}^{noobj}(C_i-\hat{C_i})^2\right]$$

where
1_{ij}^{noobj} is the complement of 1_{ij}^{obj}
C_i is the box confidence score of box j in cell i
λ_{noobj} scales down the loss when detecting background

Capturing frames:
To capture the frames with animated character(s), record the frames with atleast one bounding box. Script can be written to take video as input and returning all such frames.

2.2 Part (b)

No publically available datasets for the purpose were found.

2.3 Part (c)

Datasets can be generated using overlapping cartoon characters on image datasets like Pascal VOC. The data, however, needs to be in a predefined format as per YOLO. The center of the animated character image along with the height and width need to scaled with the image dimensions that it is being overlapped with. There can be multiple characters in an image. The class of all the animated characters can be set to 1. This augmented dataset can be mixed with the original one to generate final dataset.

2.4 Part (d)

The dataset generated above can be mixed before splitting it into 3 sets (considering the dataset size of approximately 40000) -

- **a training set** to train the neural network, consisting of approximately 80 percent of data
- **a dev set** to test the trained model and tune the model based on the observations, approximately 10 percent of data
- **a test set** to test the finally ready deployable mode, with the rest 10 percent of data

The trained model can be run on the movie and it would produce bounding boxes for the frames with the animated characters.

2.5 Part (e)

The above proposed network already generates the bounding boxes around the animated characters in the frame. No change is required with respect to the datasets, the network architecture and the training procedure.

References

[1] *Deep Photo: Model-Based Photograph Enhancement and Viewing* Johannes Kopf, Boris Neubert, Billy Chen, Michael F. Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, Dani Lischinski

[2] *Visibility Enhancement of Scene Images Degraded by Foggy Weather Conditions with Deep Neural Networks* Farhan Hussain and Jechang Jeong

[3] *Intelligent Detection in Less Visibility by Saliency Techniques and Faster Region-based Convolutional Neural Networks* Sanjay Kumar, Dipti Lohia, Darsh Pratap and Ashutosh Krishna

[4] *You Only Look Once: Unified, Real-Time Object Detection* Joseph Redmon, Santosh Divvala, Ross Girshick and Ali