BTech Project Report (BTP 2)

By

Fazal Ahmad
160070043

Under the guidance of
Prof. Amit Sethi

Department of Electrical Engineering
Indian Institute of Technology Bombay

Object detection

Introduction

Given an image or a video stream, an object detection model can identify which of a known set of objects might be present and provide information about their positions within the image.

An object detection model is trained to detect the presence and location of multiple classes of objects. For example, a model might be trained with images that contain various pieces of fruit, along with a *label* that specifies the class of fruit they represent (e.g. an apple, a banana, or a strawberry), and data specifying where each object appears in the image.

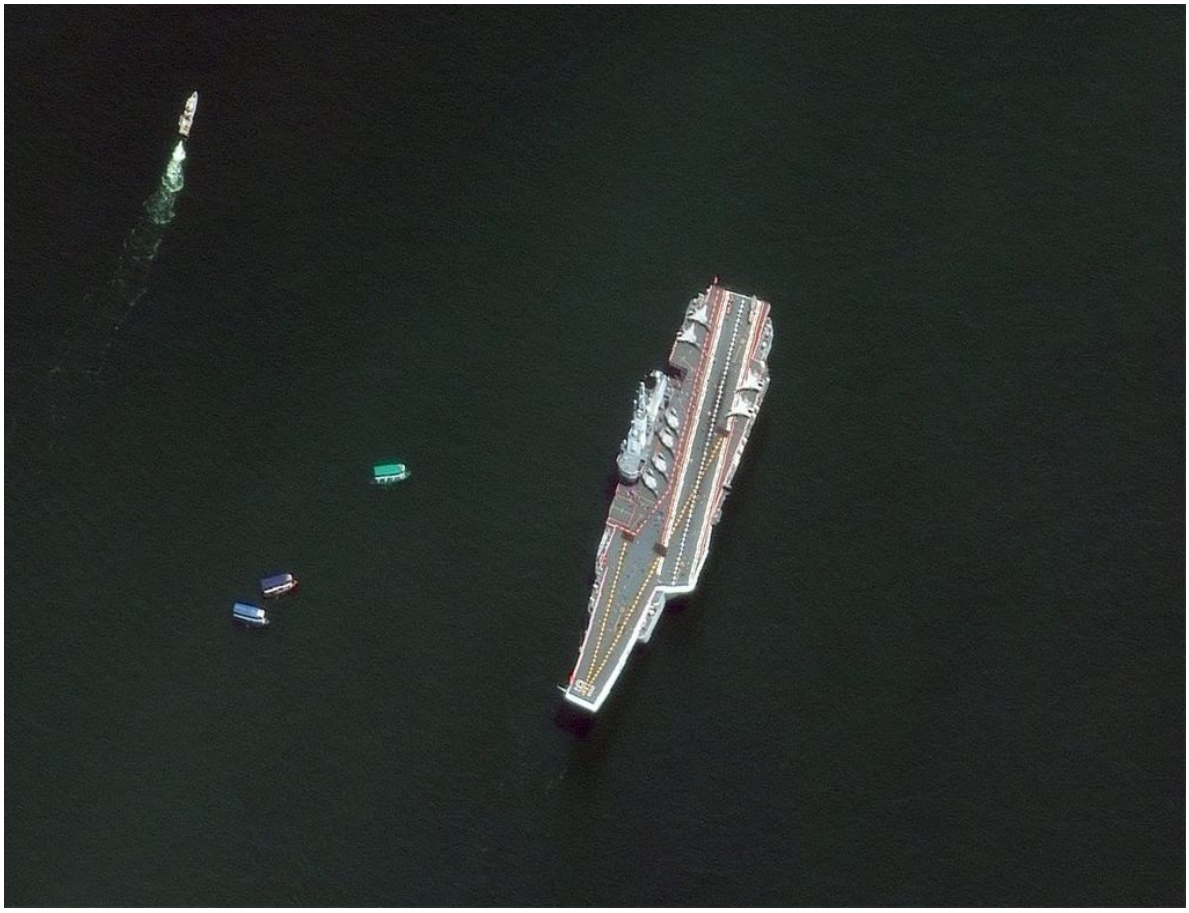
When we subsequently provide an image to the model, it will output a list of the objects it detects, the location of a bounding box that contains each object, and a score that indicates the confidence that detection was correct.

Method

The purpose of this is to explain how to train your own convolutional neural network object detection classifier for multiple objects. Here I use tensorf low object detection API to train a classifier for a single object (Ship). I used the model Fast RCNN object Detection.

Data Collection and labelling

TensorFlow needs hundreds of images of an object to train a good detection classifier. To train a robust classifier, the training images should have random objects in the image along with the desired objects, and should have a variety of backgrounds and lighting conditions. There should be some images where the desired object is partially obscured, overlapped with something else, or only halfway in the picture.
Our Data set contains navy ships but we have few images to train and test the model. Also In 1 or 2 images some merchant ships were also there to get the insight of false cases.



One of the image from the Dataset

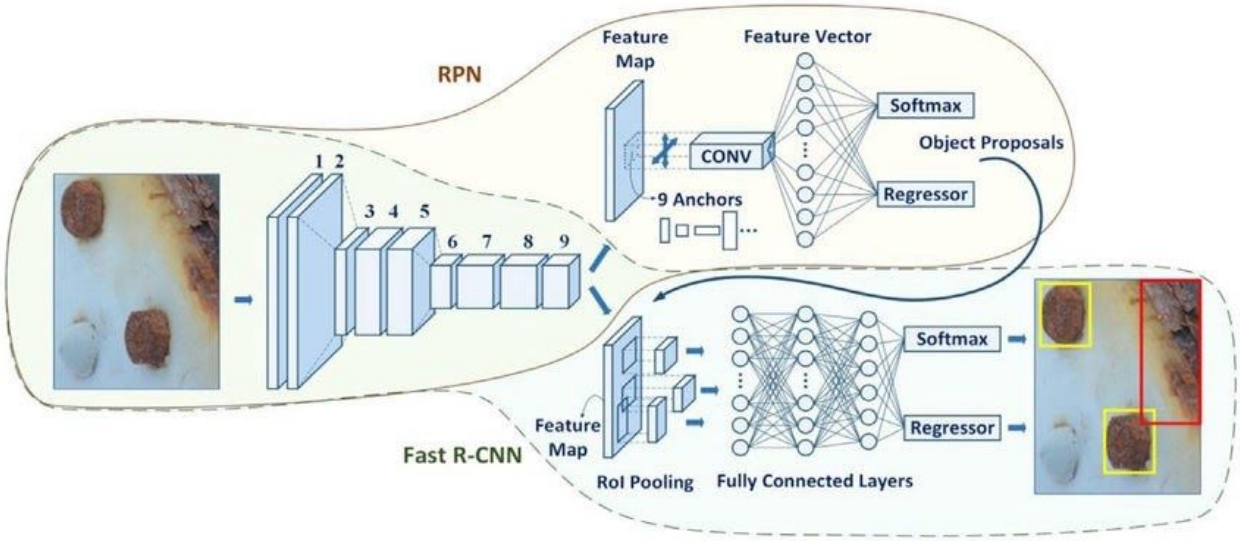
With all the pictures gathered, it's time to label the desired objects in every picture. Labelmg is a great tool for labeling images

Draw a box around each object in each image. Repeat the process for all the images.

Faster RCNN MODEL

Faster RCNN is an object detection architecture presented by Ross Girshick, Shaoqing Ren, Kaiming He and Jian Sun in 2015, and is one of the famous object detection architectures that uses convolution neural networks

Faster RCNN is composed from 3 parts



Faster RCNN is an object detection architecture presented by Ross Girshick et al.

- ## Part 1 : Convolution layers
  In this layer we train filters to extract the appropriate features of the image, for example let's say that we are going to train those filters to extract the appropriate features for a human face, then those filters are going to learn through training shapes and colors that only exist in the human face.

- ## Part 2 :Region Proposal Network (RPN)
  RPN is a small neural network sliding on the last feature map of the convolution layers and predicting whether there is an object or not and also predict the bounding box of those objects.

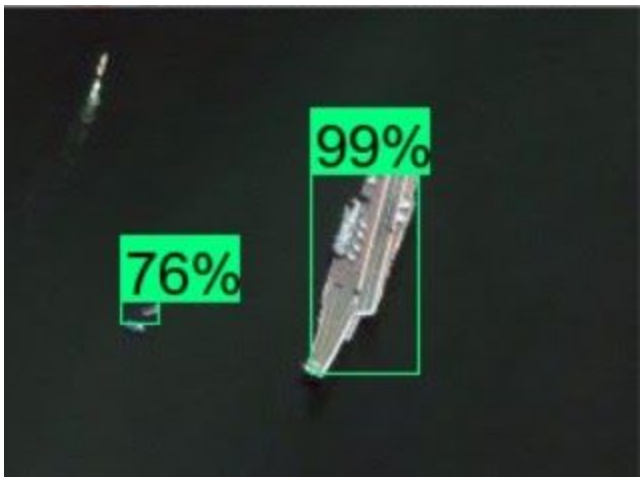
- ## Part 3 :Classes and Bounding Boxes prediction
  Now we use another Fully connected neural network that takes as an inpt the regions proposed by the RPN and predict object class (classification) and Bounding boxes (Regression).

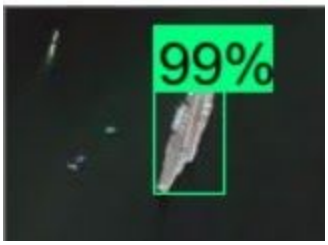
Training

To train this architecture, we use SGD to optimize convolution layer filters, RPN weights and the last fully connected layer weights.

Results

We Tried to reduce the Resolution of the test Image





The model is not detecting the image of size below 650 bytes

Video Super Resolution

Introduction

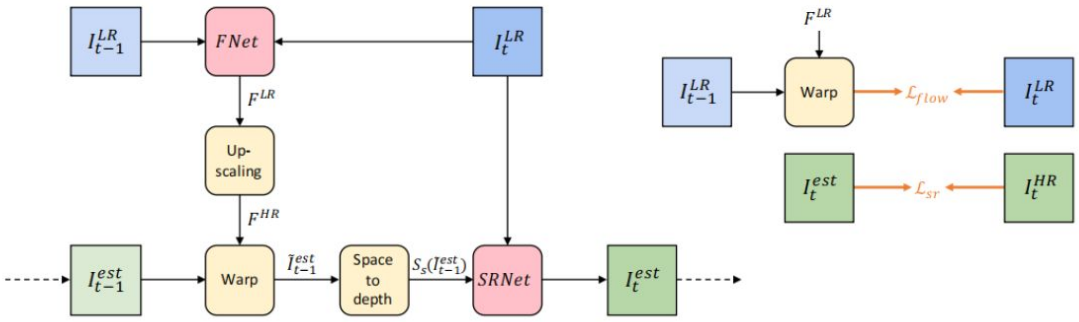
The technique of retrieving high frequency features which are vanished in its low resolved consecutive frames is called Video Super-Resolution. It aims to enhance the important information, due to which it can be applied to a variety of fields like video surveillance, high-definition television, satellite imagery,etc.
The major task of video super resolution can be divided into 2 broad problems - (1) The alignment of consecutive frames to have accurate correspondence, and (2) Combine such matched frames to generate high resolution output.

Single Image and Multi-Image super resolution

In SISR, the algorithm is applied to each frame of a video individually whereas in VSR, to multiple consecutive frames. SISR concentrates on calculating unknown pixel intensities by applying some kernel i.e. directly estimating the mapping from low to high resolution whereas VSR concentrates on exploiting the non-redundant inter-frame features. For a given pixel, SISR only has the information of that pixel and its neighbours' intensities (a 2D input domain) whereas VSR has an additional information in temporal domain (so overall a 3D input domain) However, VSR has the need of motion estimation, image registration and one additional temporal dimension which makes it computationally costly.
Previously, only classical approaches were used like interpolation which includes bilinear, bicubic and other variants . As it is a SISR method, the estimation of new pixels is done using some kernel based mapping from low resolution(LR) to high resolution(HR) frames. As it is a closed form solution, it requires lesser computation i.e. it is extremely fast. However, it lacks generation of high-frequency information, which are lost in down-sampling the given high resolution images . Now, with the advancement in neural networks, the focus is shifted from classical to learning based approaches. And also, due to recurrent based networks, the solution of how to effectively use a sequence of information i.e. temporal information, has made VSR possible.

Frame Recurrent VSR (FRVSR)

To SR any given frame , their model used the current LR frame, the previous LR frame and the estimated previous HR frame.

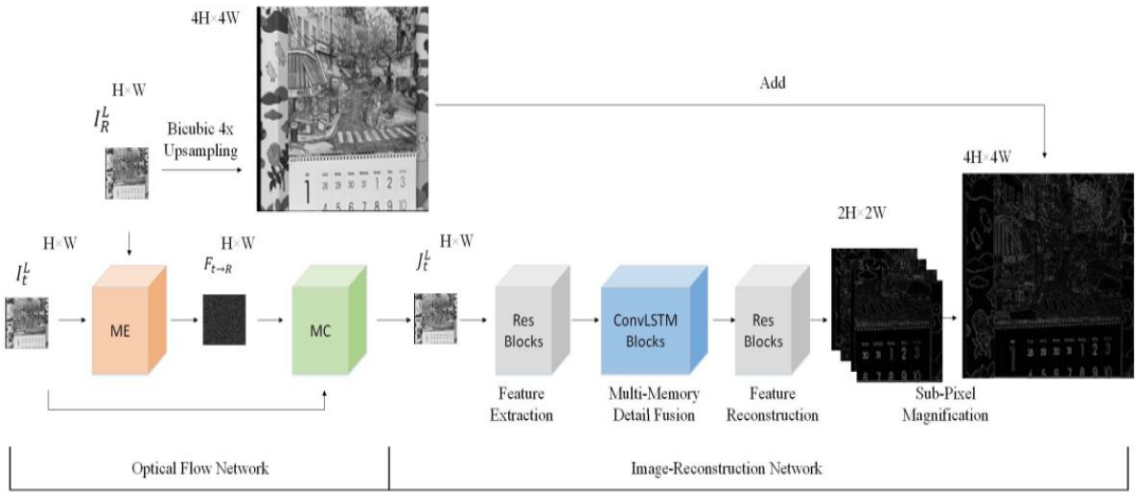


Frame Recurrent VSR frame work by Sajjadi et al.

- For motion estimation, an optical flow estimator FNet was used. FNet consists of various CNN blocks to compute the flow. For motion compensation, a differentiable function using bilinear interpolation was used.
- They concatenate the current frame with the estimated previous frame and pass it through a series of residual CNN blocks to get the HR frame.
- Used a combination of 2 losses - (1) L2 Loss between ground truth HR estimated output (2) L2 Loss between current LR image and previous LR image warped using flownet.
- The disadvantage of this method is the large cost of computation and limited utilization of spatial and temporal information.

Multi Memory CNN for VSR

The paper proposes a multi-memory residual block to extract and store the feature maps which are correlated in the consecutive frames.



- For motion estimation and compensation, the model has used optical flow estimator ’FlowNet’ and VESPCN’s multi scale spatial transformer method for motion compensation.
- For Feature extraction and reconstruction, a series of Densely connected Convolutional layers were used. 6
- LSTMs are good as predicting if the input is a sequence of 1D data and CNN are good if the input is 2D data. So, a combination of both such networks which is ConvLSTM are used to magnify LR images.
- To enlarge the output, the most common way is to use transposed convolution.
- The optical flow module was trained with L1 loss and the image reconstruction module was trained with difference loss.
- Disadvantage of this method is its large computation cost.

Future Work

Object Detection

- Try to Train the model for more images and which includes more false cases
- Augment the Images of training dataset and retrain the model with more images

Video Super Resolution(VSR)

- Effect of using a better motion estimator or motion compensator, should be studied.
- Effect of using different number of frames as an input i.e. by varying the sequence length of the input

References

[1] *Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks"*

[2]https://towardsdatascience.com/faster-rcnn-object-detection-f865e5ed7fc4#:~:text=Faster%20RCNN%20is%20an%20object,SSD%20(%20Single%20Shot%20Detector).

[3]https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/latest/training.html

[4] M. S. M. Sajjadi, R. Vemulapalli, M. Brown, "Frame-Recurrent Video Super-Resolution," in Computer Vision and Pattern Recognition, Jan - 2018.

[5] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu and J. Ma, "Multi-Memory Convolutional Neural Network for Video Super-Resolution," in IEEE Trans. Image Process., May - 2019, pp. 2530 - 2544.

[6] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video superresolution," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4482–4490.