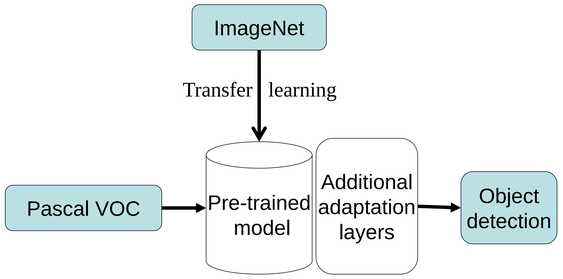
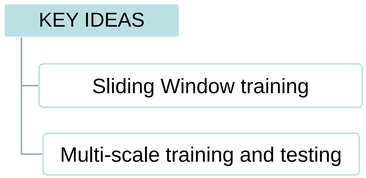
Weakly supervised object detection using CNN

L S Vishnu Sai Rao, Saurabh Kataria

CS676A (Computer Vision) course project

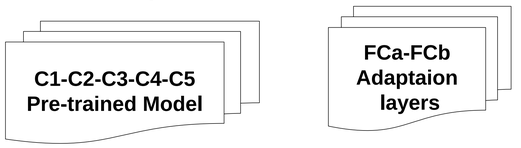
Overview



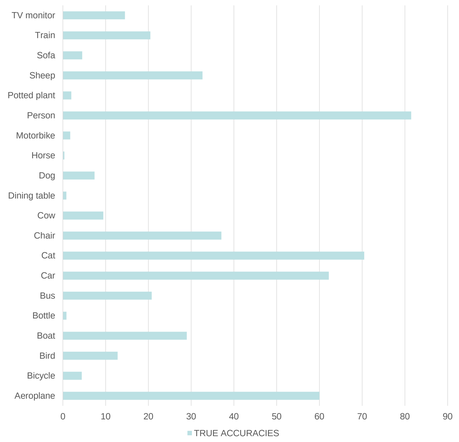


Implementation

- Training has been done on the train dataset of Pascal Voc 2012 and testing on the test dataset of Pascal Voc 2007.
- Memory constraints required us to cut down the pre-trained model network architecture from 7 convolutional layers to 5 convolutional layers.

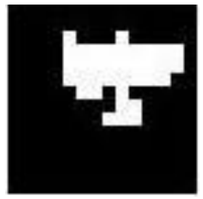


Results



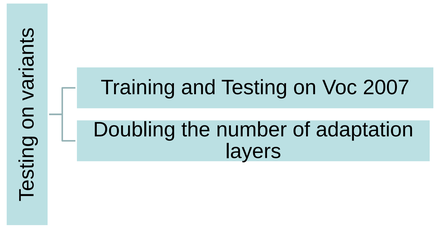
Final Layer Binary Visualization

Input image



Visualization of final layer activation

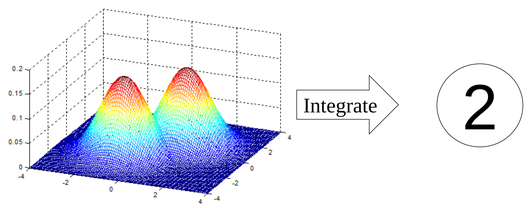
Variants



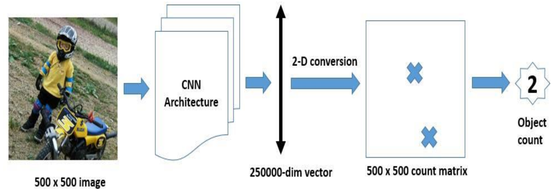
Supervised Object Counting

Idea

- Prepared a labelled dataset for each image such that a Gaussian functions are present on the image wherever the object was present.
- Integrated the labelled data would then give the object count. Currently performed irrespective of the classes
- The two-dimensional labelled data was transformed to a one dimensional vector to incorporate it within the CNN architecture



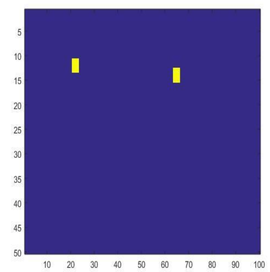
Flow chart for object count predication



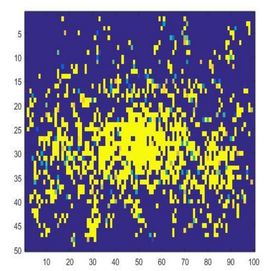
Experiments

(a) Inherent Scaling between the inputs and the targets

- Input images were of size 500 X 500 where as the target was of dimension 1500 X 1 (formed from a 50 X100 2-D).
- This was done because of the memory constraints. This experiment was repeated for various target dimensions



2-D desired target



2-D observed output

(b) Inputs and Targets are of similar dimension – two variants

1. Images from Pascal Voc were resized to 100 X 50. Targets also of dimension 1500 X 1 (from 2D 100 X 50)
2. Images from Pascal Voc were resized to 224 X 224. Targets also of dimension 50176 (from 2D 224 X 224)

(c) MESA based cost function

- Currently a sum of logistic regression based cost function was used.
- This cost function can be replaced with MESA distance.
- Given an image I, the MESA distance DMESA between two functions F1(p) and F2(p) on the pixel grid is defined as the largest absolute difference between sums of F1(p) and F2(p) over all box subarrays in I as follows

$$\mathcal{D}_{MESA}(F_1, F_2) = \max_{B \in \mathbf{B}} \left| \sum_{p \in B} F_1(p) - \sum_{p \in B} F_2(p) \right|$$

Conclusion

- Weakly supervised training of CNN does give free information about the location of objects. The results significantly depends on the number of CNN size (number of layers), epoch for training, training size.
- The performance of CNN greatly reduce when last large layers are dropped.
- Main challenge of the project was to train and test using minimum hardware requirements.
- Upon feedforwarding of a single image, the output of last convolutional layer contains rough information about the count of number of objects.

References

- Oquab, Maxime, et al. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- Oquab, Maxime, et al. "Learning and transferring mid-level image representations using convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
- Lempitsky, Victor, and Andrew Zisserman. "Learning to count objects in images." Advances in Neural Information Processing Systems. 2010.