Seeking Methods to Improve Small Object Detection

Shiming Luo[*]
Department of Electrical and Computer Engineering
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
shl666@eng.ucsd.edu

Yu Zhong
Department of Computer Science
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
yuz871@eng.ucsd.edu

Abstract

The following paper discusses several attempts, including enlargement of receptive field, specific data augmentation, and anchor boxes refinement to increase the confidence level of detecting small objects in one picture, especially human being detection and faces detection. Throughout the entire experiment, reasoning of different methodologies and results are provided in the following sections.

1 Introduction

1.1 Motivation

Given the history of computer vision, machine learning has become one of the most popular approaches to achieve solutions to different vision-related tasks included image classification, object detection, video motion tracking, super resolution, and style transfer. In the field of object detection, multiple neural network based solutions have been widely accepted and applied in many situations with increasing speed and accuracy. Given the need for higher accuracy, experimenters decide to further improve the accuracy on objection detection, especially on small objects detection.

1.2 Overview

1.2.1 History

Apply instant object detection has always been one of the goals of computer vision. Starting from 1970s, researchers had been researched into human vision and discovered a series of connections between neuroscience and vision system. Therefore, theories such as primal sketches had attracted enough academic attention and early computer vision started around 1970s. Later, feature-base methods such as using pose consistency(also called alignment), and other geometry primitives to analyze images and extract features. in 1990s, sliding window approaches was among one of the most popular way to recognize human faces. Meanwhile, appearance-based methods also showed potentials in mimicking biological vision system and can be therefore a key method in machine vision. Instead of early canny edge detection technique, features of an image now include more and more features and thus the variety of feature matching became heated again. In early 2000s, local features were utilized in recognition for object instances and then bag of features model prevailed, while in the same time, new features description such as scale-invariant feature transform (SIFT) and histogram of oriented gradients became some of the most common feature description until the era of machine learning. In the recent decade, application of convolution neural network based studies became the most popular topic in computer vision field and most studies were directly related to techniques or methodologies depended on deep learning. Given the huge success brought by the introduction of neural network, most vision tasks apply the methodology of end-to-end learning with significant amount of training data.

[*] Use footnote —*not* for

Preprint. Work in progress.

Provided the successful object classification network, it was reasonable to naively extract all possible bounding boxes in one image and classify each of them to select the one with highest confidence level. However, such algorithm was not practical based on hardware capabilities. Hence, the bounding problem was abstracted into a regression problem and later became a learnable process. Early attempts such as Overfeat did show promising results in tasks combining classification and localization. As more comprehensive datasets such PASCAL, ILSVRC, MSCOCO were available for researchers, training and testing became more convenient and thus more time could be devoted onto discovery of new models. Region proposals became the most promising method to solve bounding box issue in detection. However, its repetition of computing features slowed down the training process and therefore new algorithm and learning procedure were then invented to greatly improve the speed of region-based convolutional neural network. In year 2016, proposal-free object detection pipeline was invented and named "You Only Look Once." (YOLO) Such procedure include default boxes prediction on given number of cells and results were surprisingly fast. At the same year, the Single Shot Multibox Detector was invented and combined "Look Once" idea from YOLO and anchor boxes from Faster R-CNN and made such algorithm one of the fastest object detection algorithm so far.

1.2.2 Issue

Due to the fact that object detection network often apply artificial neural network, in which convolution and pooling are fundamental operations of neural network, the tensor height and width will shrink continuously during those operations, and therefore expand the receptive field in tensors corresponding to the original dimension in the input image. Consequently, if objects appear small in the input image, the detection network may neglect or yield a marginal confidence level for such object. Therefore, researchers would like to seek a possible method to increase the confidence level in small object detection.

2 Experiment Proposals

Based on the given code base, the SSD algorithm is implemented through addition of layers upon VGG network, one of the most popular classification network invented in year 2014. The code transforms the final fully connected layer into additional convolutional layers with no dropout layer. For bounding box prediction, author extracts different features in between different convolutional layers. Based on the dimension of feature map and precalculated anchor sizes, anchor boxes are projected back to the original picture and bounding boxes and prediction are made. Such methodology greatly enhances the small object detection, which is to some extent compromised in YOLO structure. Based on the aforementioned chain of thoughts, we researchers intend to further enhance the confidence level of small object detection by the following various experiments 1.

2.1 Advancing Feature Maps to Early Layers

After multiple layers of convolution and pooling, the dimensions of the feature maps that are used for anchor box prediction are $38 \times 38, 19 \times 19, 10 \times 10, 5 \times 5, 3 \times 3, 1 \times 1$. In this project, the effects of advancing feature maps to earlier block, for example, using an early feature map in between VGG with a dimension of 150×150 or 75×75 will be discussed later in the report 2.

2.2 Addition of Anchor Boxes

The default anchor boxes set include 6 predetermined anchor sizes and aspect ratios, which was hypothesized and tested extensively by researchers. In addition to advance feature maps to early layers, insertion of another anchor box set may affect the final result as well. The results will be discussed in the later section 3.

2.3 Training Set Variation

To further exploit the structure of VGG-based neural network to realize small object detection, researchers focus on a binary human faces recognition due to the availability of human faces datasets. Therefore, instead of PASCAL VOC 2007 dataset, experimenters apply the WIDER FACE dataset to verify the effects of small object detection.

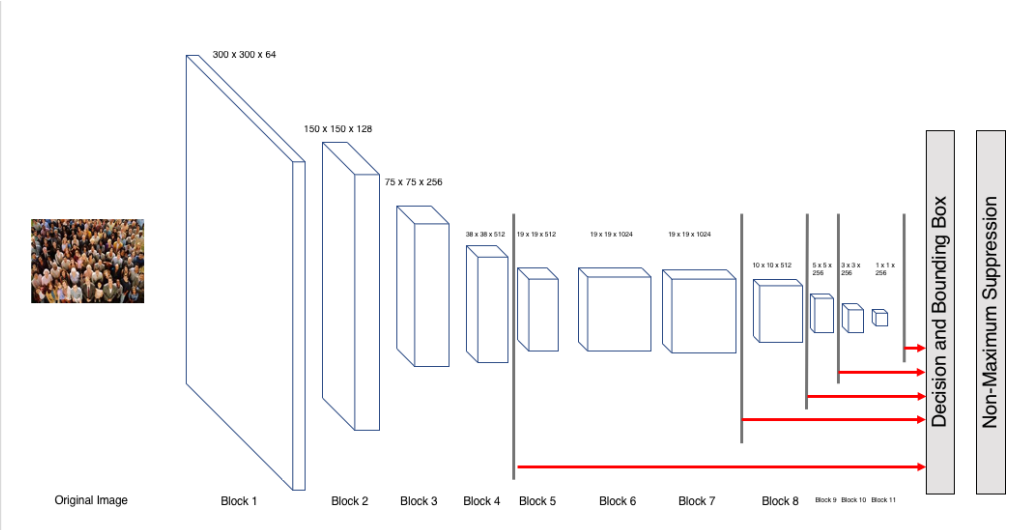

Figure 1: VGG Based SSD Network Structure

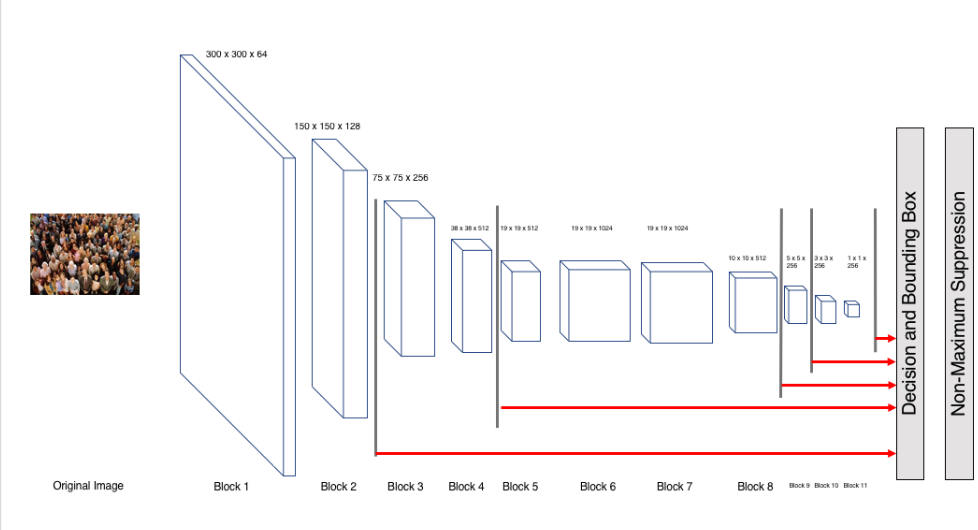

Figure 2: VGG Based SSD Network Structure with Early Feature Prediction

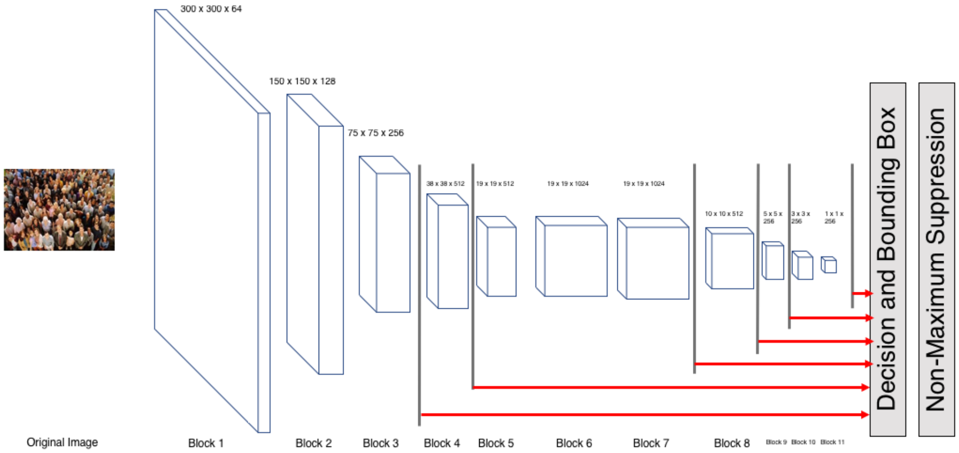

Figure 3: VGG Based SSD Network Structure with 6 Feature Maps Prediction

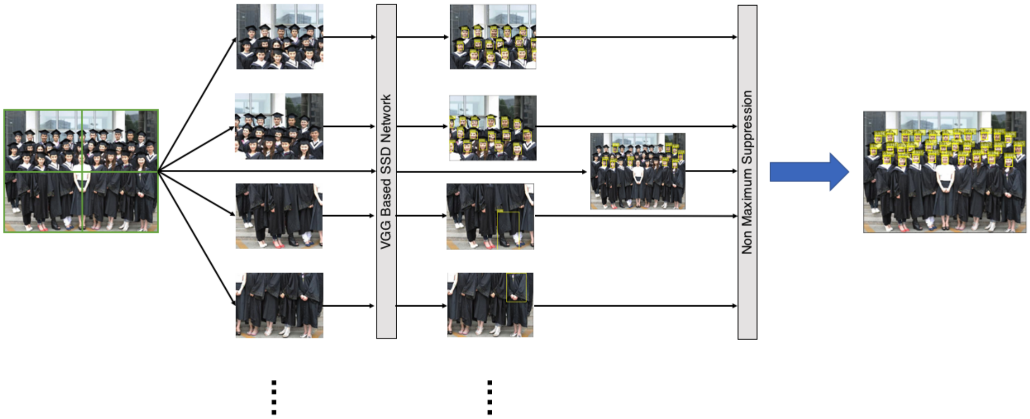

Figure 4: Input Image Processing Pipeline

2.4 Input Image Processing

Given the high speed of object detection by SSD, researchers also experiment on the trade-off between speed and accuracy to improve small faces detection. Cropping the images into sections and send cropped and rescaled images into the network to achieve the final results 4.

3 Results

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use more than eight pages as long as the additional pages contain *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.