A REPORT

ON

OBJECT DETECTION IN AERIAL IMAGES

BY

AYUSH JAIN

2017A7PS00093P

B.E. (HONS.) COMPUTER SCIENCE AND ENGINEERING

GUIDED BY

DR. PRATIK NARANG

AT

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(December, 2019)

ACKNOWLEDGEMENT

I would like to thank BITS PILANI for providing me an opportunity to take up this study oriented project. I want to thank my guide and the in-charge for the project, Dr. Pratik Narang for guiding me with all the project work, encouraging me and providing the necessary aid with all the technologies and explanations of the work I did for the project.

I want to acknowledge the role of my colleagues and friends who helped me during the project and my parents for their immense support throughout the project.

TABLE OF CONTENTS

1. INTRODUCTION

Object detection refers to a problem of computer vision, where given an image, we try to locate a prominent object(cars, pedestrians, etc.) by drawing a bounding box around the object. We then proceed to give a "label" to the object, hence completing the object detection task.

Object detection has been one of the most widely studied topics in computer vision. A lot of advancement has already been made in this field, where we can detect dogs, cats and other objects in an image with accuracy, sometimes even higher than humans. But aerial images becomes a big challenge for machines due to many reasons such as the small size of objects, haziness, low resolution and much more.

The applications of leveraging the expertise out of object detection in aerial images are far-fetched. Drones can be widely deployed in the agricultural field, aerial surveillance systems, and many places which are hazardous for humans. Drones, with high capability of object detection, can provide a significant boost to technology space.

We are trying to build a robust model which could address this issue effectively. We have started with AISKYEYE dataset, which contains images captured from low flying drones with 11 different classes. We will finally be evaluating our model on a freshly curated dataset made with the help of our drones.

2. DATASET

For the initial testing purposes, we are using VisDrone dataset, a large-scale benchmark with carefully annotated ground-truth for various important computer vision tasks. The VisDrone2019 dataset is collected by the AISKYEYE team at Lab of Machine Learning and Data Mining, Tianjin University, China. The benchmark dataset consists of ***288*** video clips formed by ***261,908*** frames and ***10,209*** static images, captured by various drone-mounted cameras, covering a wide range of aspects including location (taken from 14 different cities separated by thousands of kilometers in China), environment (urban and country), objects (pedestrians, vehicles, bicycles, etc.), and density (sparse and crowded scenes). Note that, the dataset was collected using various drone platforms (i.e., drones with different models), in different scenarios, and under various weather and lighting conditions. These frames are manually annotated with more than ***2.6 million*** bounding boxes of targets of common interests, such as pedestrians, cars, bicycles, and tricycles. Some essential attributes including scene visibility, object class and occlusion, are also provided for better data utilization [1]

3. CHALLENGES

1. **Low Spatial Resolution**: The objects which need to be detected are generally quite small in size, and it poses a difficulty to identify them with precision [3].
2. **Variability in Orientation**: In object detection, we choose some anchor boxes, generally a small number, which we believe, could capture all the orientation. This approach works typically fine because the number of objects is quite less and hence they fit in one or the other anchor boxes. But in aerial images, due to so many

purposes, there is high variability in orientation, which becomes a challenge for existing models [3].

3. **Lighting/Shadowing Changes**: Aerial images, being inherently difficult for object detection, becomes more susceptible to lighting and shadowing changes [4].
4. **Occlusion/Hazing**: Aerial images captures the top view of the scenery. Some objects might get occluded due to other surrounding purposes [4].

4. PRELIMINARY EXPERIMENTS:

4.1. RETINANET IMPLEMENTATION

Retinanet is a single shot detector. It uses Focal Loss as a loss function, which is distinctive for its ability to place higher loss for difficult images. It uses Resnet and FPN(Feature Pyramid Network for Region Proposal) for feature extraction and two task-specific subnetworks for classification and bounding box regression[2]

We plan to implement this vanilla retinanet architecture to see the obtained accuracy followed by a detailed analysis of its result, which could motivate further modifications in the network.

There are mainly 3 components in the retinanet architecture:

1. Retinanet backbone: The vanilla retinanet model uses a resnet50 [5] model as a backbone for performing feature extraction on the input images. The input images of very high dimension are convoluted to get a low resolution image, but with high semantic value. The resnet model is first pre trained on imagenet [6]. The resnet50

architecture is a heavy model which is computationally expensive to train. Feature Pyramid Network(FPN) is built on Resnet architecture. It is comprised of two parts: Bottom up parsing and top-down parsing. Bottom up parsing is a five layers network on top of the image for capturing feature which are semantically stronger but resolution wise weaker.

From the topmost layer of the Bottom up parsing network, using nearest upsampling, new layers in the top down parsing are being generated till the second last layer. On every layer, we build the classification and regression subunits. The layers which are on top covers a larger amount of pixel area than the layers below. Hence the bottom layers are more suitable for object detection of small sized images.

Classification Subnet: It us a fully convolutional network comprising of four 3 X 3 conv layers, each having 256 filters. The output is a W X H X KA filters followed by sigmoid layers. K is the number of classes and A are number of anchors.

Regression Subnet: It is a W X H X 4A output network, where A are the number of anchors. It is also made up of 3 X 3 Conv layers similar to classification subunit.

Ground truth labels: In regression, the ground truth labels are the 4 numbers which are offset between anchor box and ground truth. In classification, ground truth is the one hot encoded vector.

IOU and the prediction: Intersection Over Union [7] is one of the most commonly used metric in object detection. For the retinannet model, if the IOU is greater than 0.5, then the prediction is that the anchor box has detected an object. If IOU is less than 0.4, then prediction is that the object is the background, and if the IOU is in between 0.4 and 0.5, there is no prediction as the model is confused.

5. Conclusion and Future Work

1. Developing an understanding of the peculiar annotation design of Aiskyeye dataset: The annotations provided has two additional annotations for truncation and occlusion. We plan to read some research papers using the aiskyeye dataset and analyze how can we exploit it.
2. Adjusting the anchor size and orientation: This is the part which we have to definitely change. Anchors are one of the most important parts of the object detection models, and designing them correctly is important for good performance. We would have to experiment with different anchor sizes. That being said, there are some automated ways also to get to good anchor sizes, some of them being the KNN sampling and neural network-based approach.
3. Using an earlier layer in Feature Extraction Network: In the retinanet architecture, a base resnet50 is used on top of which a feature pyramid layer is attached. Generally, the topmost layer is used for object prediction. But due to the small size of aerial images, the topmost layer might have lost a lot of resolution and hence perform poorly. We can look towards experimenting with a lower layer. One idea I had was to use skip connections from lower layers to the upper layers too. This might help in passing on the high-resolution information to the top. The kind of backbone used seems to have the most dramatic effect on the results on VisDrone. Which layers, how many of them, residual or not, skip connections, multi-scale, etc. are mostly

experimental outcomes. We need to explore using multiple backbones and somehow fuse the information to create pyramids. Also, right now, resnet50, densenet [8], xceptionNet[9], all image classification backbones are used for feature learning.

REFERENCES:

1. Zhu, Pengfei, et al. "VisDrone-VDT2018: The vision meets drone video detection and tracking challenge results." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
2. Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.
3. Sakla, Wesam, Goran Konjevod, and T. Nathan Mundhenk. "Deep multi-modal vehicle detection in aerial ISR imagery." *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017.
4. Sommer, Lars Wilko, Tobias Schuchert, and Jürgen Beyerer. "Fast deep vehicle detection in aerial images." *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017.
5. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
6. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
7. Nowozin, Sebastian. "Optimal decisions from probabilistic models: the intersection-over-union case." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

8. Iandola, Forrest, et al. "Densenet: Implementing efficient convnet descriptor pyramids." *arXiv preprint arXiv:1404.1869* (2014).

9. Tariq, Shahroz, et al. "Detecting both machine and human created fake face images in the wild." *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. ACM, 2018.