DIP Final - Multiple Objects Detection in Aerial Images

Tsung-Lin Tsou
National Taiwan University
Email: r10922081@ntu.edu.tw

Guo-Qing Tan
National Taiwan University
Email: b08902118@ntu.edu.tw

I. Introduction

In this project, our goal is to detect objects in aerial images. Object detection in aerial images is a challenging task as the objects in aerial images are displayed in arbitrary directions and are usually densely packed. Different from detecting objects in COCO dataset [1] or PASCAL VOC dataset [2], aerial images often capture objects from bird's eye view, which are smaller in scale and have lower resolution. Consequently, we first compare the performance of models [3], [4], [5] that are pre-trained on COCO dataset and models [6] that are pre-trained on DOTA dataset [3]. Then, we analyze the pros and cons of all models and discover that different models have an edge over different tasks. Finally, we use some late-fusion techniques [7], [8] to fuse bounding boxes at the decision level to improve the performance. (Note that we do not use Visdrone dataset [9] is because testing image labels are usually not available. Hence, we use other techniques such as late-fusion to improve model performance.)

II. Methods

First of all, we implement four pre-trained models to test on the testing aerial images. Among them, M2Det [3], FCOS [4], and YOLOv7 [5] are pre-trained on COCO dataset [1]. BBAVectors [6] is pre-trained on DOTA dataset [3]. Then, we analyze the pros and cons of all models and discover that different models have an edge over different tasks. Finally, we use some late-fusion techniques [7], [8] to fuse bounding boxes at the decision level to improve the performance. The following is a more detailed description of all methods.

A. M2Det

Despite the fact that feature pyramids are widely exploited by both state-of-the-art one-stage object detectors and two-stage object detectors, they have some limitations as they only simply construct the feature pyramid according to the inherent multi-scale, pyramidal architecture of the backbones which are originally designed for object classification task. Consequently, M2Det proposed a novel method called Multi-Level Feature Pyramid Network (MLFPN) to construct more effective feature pyramids for detecting objects of different scales.

B. FCOS

A majority of state-of-the-art object detectors rely on pre-defined anchor boxes. In contrast, FCOS is anchor box free and proposal free. By eliminating the pre-defined set of anchor boxes, FCOS completely avoids the complicated computation related to anchor boxes such as calculating overlapping during training. FCOS also avoids all hyper-parameters related to anchor boxes, which are often very sensitive to the final detection performance. It is the first detection framework that can achieve improved detection accuracy.

C. YOLOv7

YOLOv7 designs several trainable bag-of-freebies methods that helps real-time object detection to greatly improve the detection accuracy without increasing the inference cost. YOLOv7 also proposes "ex-tend" and "compound scaling" methods for the real-time object detector that can effectively utilize parameters and computation. Specifically, it can effectively reduce about 40 percent of parameters and 50 percent of computation of state-of-the-art real-time object detectors and has faster inference speed and higher detection accuracy.

D. BBAVectors

BBAVectors is an oriented object detection methods testing in aerial images. Current oriented object detection methods mainly rely on two-stage anchor-based detectors. However, the anchor-based detectors typically suffer from a severe imbalance issue between the positive and negative anchor boxes. To address this issue, BBAVectors extend the horizontal keypoint-based object detector to the oriented object detection task. Specifically, BBAVectors first detect the center keypoints of the objects, based on which we then regress the box boundary-aware vectors to capture the oriented bounding boxes.

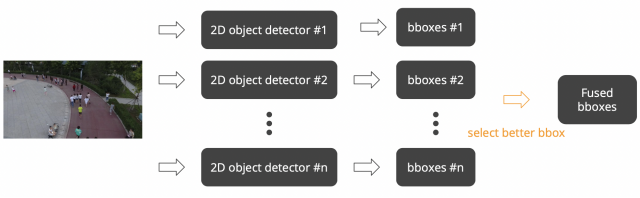

Fig. 1. The framework of late-fusion model ensembling.

E. Late-fusion

Based on the pros and cons analysis of all models, we find out that different models have an edge over different tasks. For instance, BBAVectors is good at predicting ships, vehicles, and objects which are densely packed. On the other hand, YOLOv7 and FCOS are good at predicting person, benches, and the categories which do not frequently appear in aerial image dataset. As a result, we use some late-fusion techniques [7], [8] to fuse bounding boxes at the decision level to improve the performance (Fig. 1 shows the framework of late-fusion model ensembling technique). For the score fusion part, we use Non-Maximum Suppression (NMS) [7] and probabilistic ensembling (ProbEn) [8]. As for bounding box fusion part, we use Non-Maximum Suppression (NMS) [7], averaging, and weighted sum.

III. Results and Discussions

Fig. 2 and Fig. 3 show that BBAVectors is good at predicting ships, vehicles, and objects which are densely packed. Fig. 4 and Fig. 5 show that YOLOv7 is good at predicting person, benches, and the categories which do not frequently appear in aerial image dataset. Fig. 6 and Fig. 7 show that different models have an edge over different tasks. Fig. 8 shows the result after using late-fusion model ensembling techniques Non-Maximum Suppression (NMS). Fig. 9 shows the result after using late-fusion model ensembling techniques probabilistic ensembling (ProbEn). It is proven by our experiment that late-fusion model ensembling techniques aggregate the advantages of each model and improve the model performance.

IV. Division of works

Guo-Qing Tan is responsible for M2Det [3] implementation. Tsung-Lin Tsou is responsible for FCOS [4], YOLOv7 [5], BBAVectors [6] and late-fusion [8] implementation.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[3] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9259–9266.

[4] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[5] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[6] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2150–2159.

[7] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 850–855.

[8] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *European Conference on Computer Vision*. Springer, 2022, pp. 139–158.

[9] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

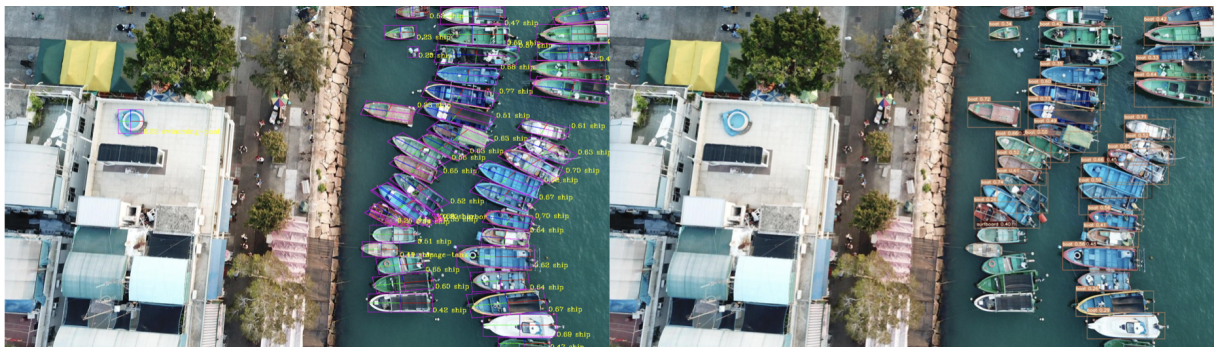

Fig. 2. Qualitative results of BBAVectors (left) and YOLOv7 (right). This figure indicates that BBAVectors is good at predicting ships.

Fig. 3. Qualitative results of BBAVectors (left) and YOLOv7 (right). This figure indicates that BBAVectors is good at predicting vehicles.

Fig. 4. Qualitative results of BBAVectors (left) and YOLOv7 (right). This figure indicates that YOLOv7 is good at predicting person.

Fig. 5. Qualitative results of BBAVectors (left) and YOLOv7 (right). This figure indicates that YOLOv7 is good at predicting person.

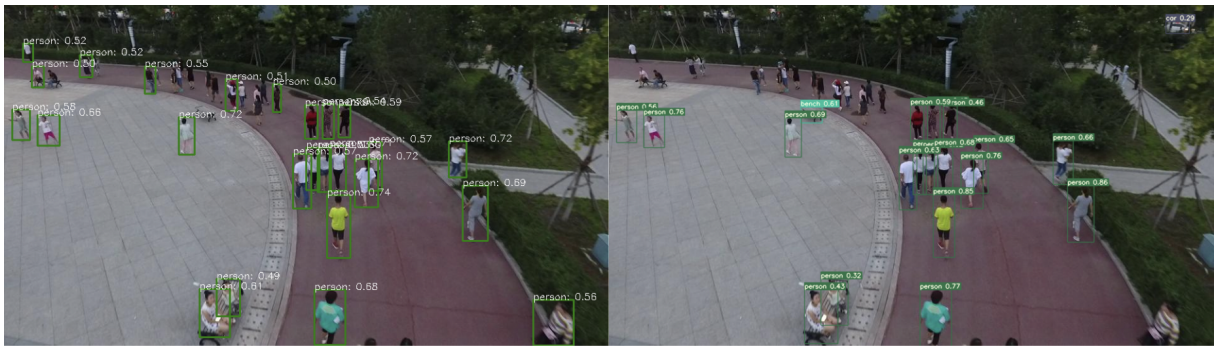

Fig. 6. Qualitative results of FCOS (left) and YOLOv7 (right). This figure indicates that different models have an edge over different tasks.

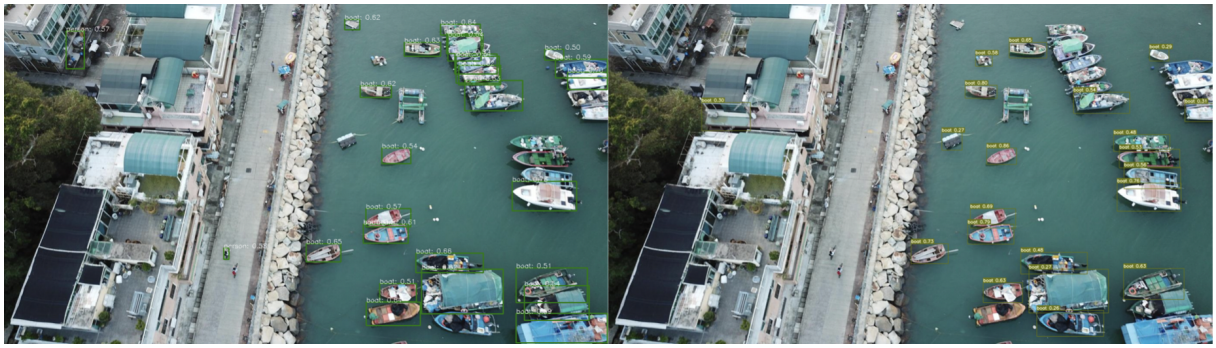

Fig. 7. Qualitative results of FCOS (left) and YOLOv7 (right). This figure indicates that different models have an edge over different tasks.

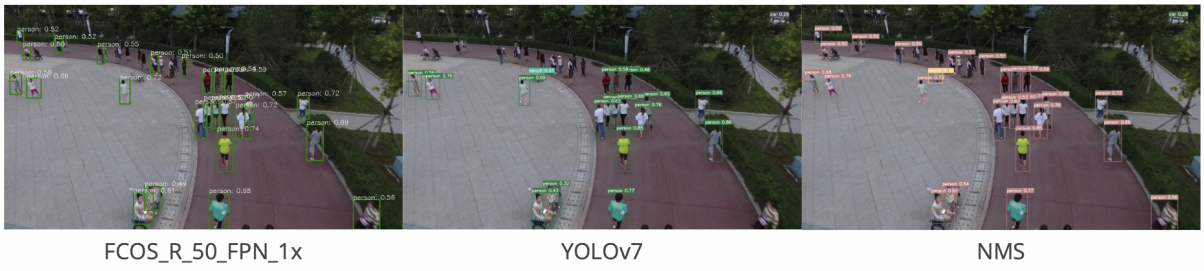

Fig. 8. Qualitative results of FCOS (left), YOLOv7 (middle), and late-fusion model ensembling technique Non-Maximum Suppression (NMS) (right).

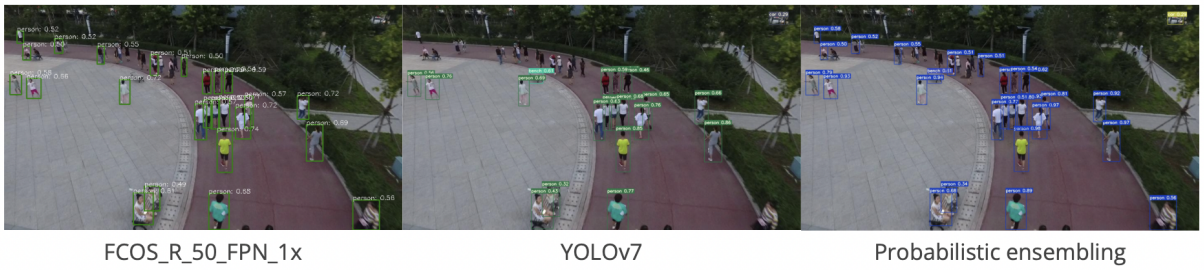

Fig. 9. Qualitative results of FCOS (left), YOLOv7 (middle), and late-fusion model ensembling technique probabilistic ensembling (ProbEn) (right).