Object Detector Combination for Increasing Accuracy and Detecting More Overlapping Objects

Khaoula Drid[1](✉), Mebarka Allaoui[2], and Mohammed Lamine Kherfi[1,3]

[1] Kasdi Merbah University Ouargla, Ouargla, Algeria
khaouladrid22@gmail.com
[2] LAGE Laboratory, Kasdi Merbah University Ouargla, Ouargla, Algeria
[3] LAMIA Laboratory, Université du Québec à Trois-Rivières, Trois-Rivières, Canada

Abstract. Object detection is considered as the cornerstone of many modern applications such as Drone vision and Self-driven cars. Object detectors, mainly those which are based on Convolutional Neural Networks (CNNs) have received great attention from many researchers because they were able to yield remarkable results. However, most of them fail when it comes to detecting overlapping and small objects in images. There are two families of detectors: the first family detects more objects but with imprecise bounding boxes, while those of the second family do the opposite. In this paper, we propose a solution to this problem by combining the two families, in a way similar to classifier combination. Our solution has been validated through the combination of two famous detectors, Faster R-CNN which detects more objects and YOLO which produces accurate bounding boxes. However, it is more general and it can be applied to other detectors. The evaluation of our method has been applied to the PASCAL VOC dataset and it gave promising results.

Keywords: Object detector combination · Detecting overlapping objects · Convolutional Neural Networks (CNNs) · YOLO · Faster R-CNN

1 Introduction

Detecting and identifying the different objects in an image is very useful. Object detection is an active research field in computer vision. It deals with two main issues: object classification and object localization. Humans and many other animals use a process called visual perception to quickly decide which locations of an image should be processed in detail and which can be ignored. This allows us to deal with the huge amount of visual information and to employ the capacities of our visual system efficiently. Concerning computer vision, researchers have to deal with the same problems. Therefore, learning from human's behavior provides a promising way to improve existing algorithms. Object detection has

A. El Moataz et al. (Eds.): ICISP 2020, LNCS 12119, pp. 290–296, 2020.
https://doi.org/10.1007/978-3-030-51935-3_31

received great attention from researchers in several areas and many interesting applications such as Drone vision systems, Self-driven cars, Video surveillance, robot navigation, and many other applications which require object and scene recognition.

Recently, CNN-based detectors achieved good results in object detection. They can broadly be categorized into two categories: two-stage family such as Faster R-CNN [16], and one-stage family such as YOLO [14]. However, those detectors suffer from several limitations. The two-stage family has two main problems: it has an expensive computational cost, and yields bounding boxes that are not precise and which may contain more than one object. This latter problem occurs especially in images with overlapping and small objects. As for the second family, it usually detects fewer objects than those of the first family. In this work, we have developed an algorithmic solution for detecting overlapping and small objects by combining the advantages of the two families. Our solution combines Faster R-CNN detector which belongs to the first family and YOLO which belongs to the second, but can be generalized to other detectors. The paper is organized as follows: In Sect. 2, we present the related work briefly. After that, we describe our algorithm in Sect. 3. In Sect. 4, we give details about the experiments we conducted to validate the proposed solution and compare it against YOLO and Faster R-CNN. We finish the paper with a short conclusion.

2 Related Work

Object detection aims at locating and classifying existing objects in any given image and surrounds them with bounding boxes. Object detection methods can roughly be subdivided into two categories: (i) Two-stage detectors: In the first step, a Region Proposal Network is used to generate regions of interest that have high probability of being an object. In the following step, they perform the final classification and bounding-box regression of objects by taking these regions as input. Examples of this family include R-CNN [7], SPP-net [9], Fast R-CNN [6], Faster R-CNN [16], R-FCN [1], FPN [11] and Mask R-CNN [8]. (ii) One-stage detectors: they process object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates simultaneously. This family includes detectors like MultiBox [2], AttentionNet [18], G-CNN [13], YOLO [14], SSD [12], YOLOv2 [15], DSSD [5] and DSOD [17]. Both families have drawbacks. The two-stage detectors are composed of several correlated stages, and therefore they need to obtain shared convolution parameters between RPN and detection network. As a result, the time spent in handling different components becomes the bottleneck in real-time application. As for one-stage detectors, they have difficulty in dealing with small objects. To address these limitations, we focus on two relatively successful solutions, namely, (YOLO) [14] and Faster region-based convolutional network (Faster R-CNN) [16]. We propose an algorithm for combining them, which aims to increase both the accuracy and the number of the detected objects.

3 Methodology

Images with overlapping and small objects are a real challenge for most object detectors. As we mentioned before, object detectors belong to two families: one-stage and two-stage detectors. Each family deals with overlapping and small objects differently. The one-stage family predicts less bounding boxes compared to the two-stage family. However, the one-stage family's bounding boxes are more accurate than the two-stage family ones. To benefit from the advantages of both families and at the same time limit their shortcoming, we propose to combine them. Our combination guaranteed at least to preserve the accuracy of the bounding boxes and augments the number of predictions. In this work, we focus on two of the most efficient models: YOLO and Faster R-CNN. YOLO belongs to one-stage family detectors whereas Faster R-CNN is a two-stage detector. In Fig. 1, we give an example of an image with overlapping objects, and the objects predicted by YOLO and Faster R-CNN.

Our Algorithm. As mentioned above that YOLO's bounding boxes are more accurate and tight. Moreover, its false prediction is less than the false prediction of Faster R-CNN. The main idea of our algorithm is as follows: for every bounding box predicted by Faster R-CNN, we check to see if YOLO predicts a similar box. If it does, we retain that prediction based on the probability predicted by YOLO. In the opposite case, the objects predicted by Faster R-CNN are retained. Our idea is illustrated in Fig. 2.

In more technical details:

1. We preserve the entire YOLO's bounding boxes,
2. For every Faster R-CNN's bounding box, we compute the intersection over union between it and YOLO's bounding boxes.
   (a) If the result is larger than a certain threshold so that is mean YOLO detects this object and we ignore the Faster R-CNN's bounding box.
   (b) Otherwise, we accept it.

Our solution is summarized in Algorithm 1.

4 Experimental Evaluation

To validate our idea, we compare our combined model with YOLO and Faster R-CNN in terms of accuracy on a benchmark dataset. In this section, we will present the used materials and metrics to get the desired results.

4.1 The Used Dataset

PASCAL VOC is a collection of datasets with 20 classes for object detection. The most common combination for benchmarking is using 16551 samples from both PASCAL VOC 2007 [4] and 2012 [3] for training, and 4952 sample from PASCAL VOC 2007 for the testing.

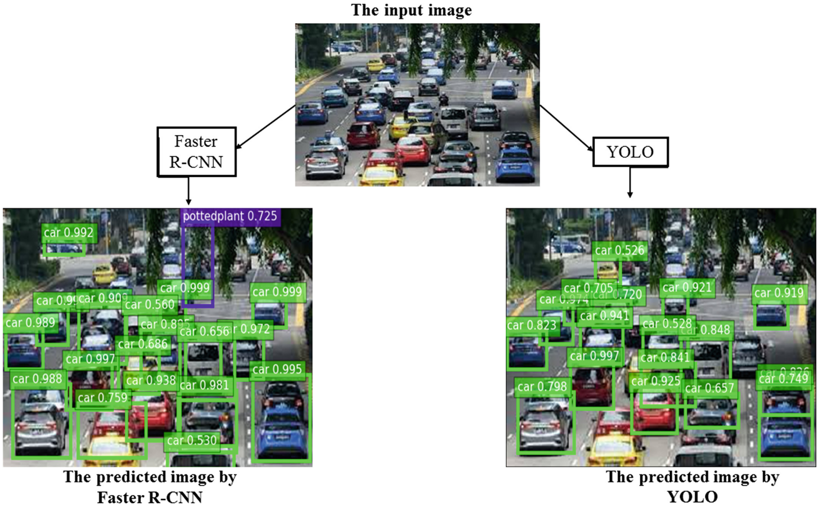
The input image
Faster R-CNN
YOLO
The predicted image by Faster R-CNN
The predicted image by YOLO
pottedplant 0.725

Fig. 1. YOLO and Faster R-CNN predictions.

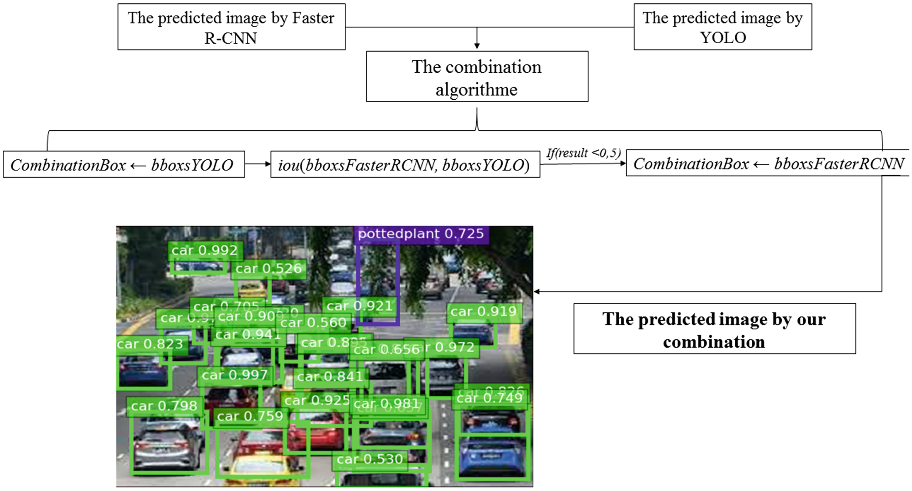
The predicted image by Faster R-CNN
The predicted image by YOLO
The combination algorithme
CombinationBox ← bboxsYOLO
iou(bboxsFasterRCNN, bboxsYOLO)
If(result <0,5)
CombinationBox ← bboxsFasterRCNN
The predicted image by our combination
pottedplant 0.725

Fig. 2. Our combination method [10], the figure illustrates how our method can combine the predicted bounding boxes from both considered object detectors. In our combination, we preserve YOLO's bounding boxes as they are because they are more accurate than those of Faster R-CNN's. Then we compute the intersection over union (IOU) between the YOLO's boxes and Faster R-CNN's boxes. If the IOU result greater than a certain threshold, we ignore this box, elsewhere we accept it, because this means YOLO does not predict a bounding box similar to this one.

Algorithm 1. Combination Algorithm

```
function COMBINATION($bboxsFasterRCNN, bboxsYOLO$)
    $CombinationBox \leftarrow []$
    for $i = 0$ to $len(bboxsYOLO)$ do
        $CombinationBox \leftarrow bboxsYOLO[i]$
    end for
    for $i = 0$ to $len(bboxsFasterRCNN)$ do
        $count \leftarrow 0$
        for $j = 0$ to $len(bboxsYOLO)$ do
            $result \leftarrow iou(bboxsFasterRCNN, bboxsYOLO)$
            if $result >= 0.5$ then
                $count \leftarrow count + 1$
            end if
        end for
        if $count < 1$ then
            $CombinationBox \leftarrow bboxsFasterRCNN[i]$
        end if
    end for
    return $CombinationBox$
end function
```

4.2 Evaluation Metric

We evaluated our methods using average precision (AP) [1] and its mean (mAP), which is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, SSD, etc. Average precision computes the average Precision values for Recall values from 0 to 1. For the precision it measures how accurate are the predictions, i.e. the percentages of how much the predictions are correct, and the recall measures how good is the prediction. The mAP can be computed by calculating the average precision (AP) separately for each class, then the average over the classes.

4.3 Experiments Results

In this work, the used YOLO and Faster R-CNN detectors are pre-trained on 16551 samples from PASCAL VOC 2007/2012. And the experiment conducted between our combination method against YOLO and Faster R-CNN on test samples from PASCAL VOC 2007. The obtained results are presented in Table 1 and Table 2.

4.4 Discussion

The proposed combination focuses on images with high overlapping and small objects. As we see in Table 1, our method has the highest precision for most classes of the dataset except for motorbike and bird. Indeed YOLO, For the motorbike class it has been able to give 86% and for the bird class, it gave

Table 1. Comparison between Faster R-CNN, YOLO and our combination method on PASCAL VOC2007/2012 dataset according to mAP metric. The table represents the achieved score for the three models in the 11 classes from the dataset in term of average precision.

	Areo	bike	Bird	Boat	bottle	Bus	Car	Cat	Chair	Cow	table
Faster R-CNN	0.78	0.75	0.47	0.53	0.40	0.90	0.77	0.88	0.34	0.78	0.58
YOLO	**0.90**	**0.81**	**0.70**	0.63	0.42	**0.95**	**0.87**	0.81	0.53	0.79	0.60
Combination	**0.90**	**0.81**	0.69	**0.65**	**0.54**	**0.95**	**0.87**	**0.90**	**0.59**	**0.88**	**0.65**

Table 2. Comparison between Faster R-CNN, YOLO and our combination method on PASCAL VOC2007/2012 dataset according to mAP metric. The table represents the achieved score for the three models in the 9 rest classes from the dataset in term of average precision, and the mAP column represents the achieved score for each model.

	dog	Horse	mbike	person	Plant	sheep	Sofa	Train	TV	**mAP**
Faster R-CNN	0.75	0.77	0.74	0.75	0.42	0.79	0.61	0.78	0.79	0.68
YOLO	0.79	**0.81**	**0.86**	0.78	0.57	**0.88**	0.65	**0.80**	**0.84**	0.75
Combination	**0.87**	**0.81**	0.85	**0.83**	**0.64**	**0.88**	**0.67**	**0.80**	**0.84**	**0.78**

70%. Our method gave the best results because it combines the advantages of the two models with some improvements since we filter the bounding boxes predicted by both YOLO and Faster R-CNN and only keep bounding boxes with a high probability of existence. As a consequence, the combination accuracy obtained is 78% mAP. As expected, our proposed method succeeds in detecting the maximum instances of the existing objects, and that is a good achievement. However, we should notice that in term of speed, our method speed is less than YOLO. Since we run each model separately and then combine the results. Since YOLO is so fast it does not add any significant computational time compared to Faster R-CNN.

5 Conclusion

In this paper, we proposed a combination framework to deal with the shortcoming of object detectors when we detect overlapping and small objects in an image. We tested our model by combining two of the most famous detectors which are Faster R-CNN and YOLO. Our method succeeded to increase the number of detected objects with accurate bounding boxes. In general, the obtained results were good and promising as seen in the experimental section. Our method was able to improve the performance of existing detectors, which makes it useful for several applications where overlapping objects are omnipresent like in self-driven cars. However, our method still needs to be improved in terms of computation time. In the future, this method could be extended to video object detection.

References

1. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
2. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2147–2154 (2014)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes challenge 2012 (VOC2012) results (2012). http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes challenge 2007 (VOC2007) results (2007)
5. Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
6. Girshick, R.: Fast R-CNN. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
8. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
10. Kherfi, M.L., Drid, K., Zouaoui, O.K.: Object detection and recognition fom images. Master thesis, University Kasdi Merbah Ouargla, Algeria (2019)
11. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
12. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
13. Najibi, M., Rastegari, M., Davis, L.S.: G-CNN: an iterative grid based object detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2369–2377 (2016)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
17. Shen, Z., Liu, Z., Li, J., Jiang, Y.-G., Chen, Y., Xue, X.: DSOD: learning deeply supervised object detectors from scratch. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1919–1927 (2017)
18. Yoo, D., Park, S., Lee, J.-Y., Paek, A.S., Kweon, I.S.: Attentionnet: aggregating weak directions for accurate object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2659–2667 (2015)