Pondering about Task Spatial Misalignment: Classification-Localization Equilibrated Object Detection

Yudong Zhang [1,2] Wei Lu [1,2] Xu Wang [1,2] Pengkun Wang [1,2] Yang Wang [1,2,*]

[1]University of Science and Technology of China (USTC) [2]Suzhou Institute for Advanced Research, USTC

Introduction

Motivation

Object detection is a fundamental task in computer vision, consisting of both classification and localization tasks. Previous works mostly perform classification and localization with shared feature extractor like Convolution Neural Network. However, the tasks of classification and localization exhibit different sensitivities with regard to the same feature, hence the **"task spatial misalignment"** issue. This issue can result in a hedge issue between the performances of localizer and classifier.

Contribution

To address these issues, we first propose a novel **Dynamic Coefficient Loss** to simultaneously consider and balance the performances of classification and localization tasks. To well address anchor label misjudgment issue in irregular-shaped object detection, we define a new **classification-aware IoU metric** to assign anchors intelligently. Finally, we further introduce the localization factor into NMS by proposing a **Classification-Localization balanced NMS**. Extensive experiments on MS COCO and PASCAL VOC demonstrate that our proposals can improve RetinaNet by around 1.5% AP with various backbones.

Literature Review

To date, extensive efforts have been made to address the issues of object detection with deep learning technologies, and existing solutions can be divided into two categories, anchor-based and anchor-free models. Nevertheless, both anchor-based and anchor-free models employ CNN as an feature extractor, then feed extracted features into diverse localizer and classifier to respectively solve localization and classification problems. As a consequence, for an individual detector, its localizer and classifier share a set of identical features that are extracted by a same CNN. On the other hand, previous works have already pointed out that the sensitivities of classifier and localizer to same features are different. For instance, the existence of the phenomenon so called "task spatial misalignment" can be obviously observed in Figure 1. Specifically, from this figure, we discover that localization task is more interested in the marginal details of an object, while classification task focuses more on the specific robust features of an object which are mostly non-existent in marginal areas.

Research Problems

In conclusion, the aforementioned phenomenon impacts the performance of object detection in three ways:

- **Problem 1:** There exists an obvious hedge within the performances of localizer and classifier, i.e., the performance of localizer is excellent while the performance of classifier is rather poor, or vice versa.
- **Problem 2:** For anchor-based methods, the labels of anchors are determined by the maximum overlapping between the anchors and the Ground Truth boxes. Regarding the task of irregular-shaped object detection, the aforementioned hedge within classifier and localizer may directly lead to the misjudgment of the anchors which are with poor localization performance but contain abundant classification information.
- **Problem 3:** Non-Maximum Suppression (NMS) algorithms are usually employed to suppress a portion of bounding boxes after predicting the bounding boxes with an object detector. However, most traditional NMS algorithms determine which part of bounding boxes should be reserved only by taking the single metric of classification score in account, and the phenomenon of "task spatial misalignment" indicates that this kind of operation is irrational.

Study Methodology

.

1. **Dynamic coefficient loss:** The classification gradient should be relatively large in case that the localization error is relatively small. We define a dynamic coefficient, which is the reciprocal of localization error, as follows,

$$\lambda\left(x_i, \widehat{x_i}\right) = \frac{1}{\text{Sigmoid}\left[\left(c_i - \widehat{c_i}\right)^2\right] + \text{Tanh}\left(\left|\frac{w_i}{h_i} - \frac{\widehat{w_i}}{\widehat{h_i}}\right|\right) + \varepsilon} \tag{1}$$

We then normalize the coefficient λ by employing the classification loss function, i.e.,

$$\lambda'\left(x_i, \widehat{x_i}\right) = \lambda\left(x_i, \widehat{x_i}\right) \frac{\sum_{j=1}^{N} L_{cls}(p_j, \widehat{p_j})}{\sum_{j=1}^{N} \lambda\left(x_j, \widehat{x_j}\right) L_{cls}(p_j, \widehat{p_j})} \tag{2}$$

The overall loss function of object detection can be denoted as,

$$\mathcal{L} = \lambda'\left(x_i, \widehat{x_i}\right) \sum_{i=1}^{N} L_{cls}(p_i, \widehat{p_i}) + \sum_{i=1}^{M} L_{cls}(p_i, \widehat{p_i}) + \sum_{i=1}^{N} L_{reg}(x_i, \widehat{x_i}) \tag{3}$$

2. **IoU-classification-aware anchor assignment:** For the i-th anchor, the IoU-classification-aware score can be calculated by,

$$s_i^{iou} = \alpha * \widehat{p_i} + (1 - \alpha) * \frac{area_i^{intersect}}{area_i^{min}} \tag{4}$$

The new protocol for judging the label Γ_i of the i-th anchor can be written as,

$$\Gamma_i = \begin{cases} 1 & in\ case\ IoU_i \geq fg\_threshold \\ & or\ s_i^{iou} \geq score\_threshold \\ 0 & in\ case\ IoU_i \leq bg\_threshold \\ -1 & otherwise \end{cases} \tag{5}$$

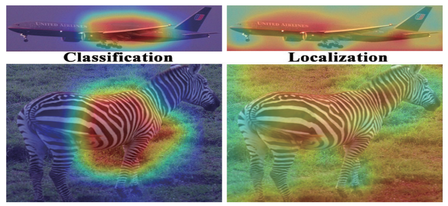
3. **Classification-localization-balanced NMS:** We calculate the weighted average center coordinates of all the boxes by using their classification scores as weights, and take the weighted average center as the approximate center $\widehat{c^0}$ of the ground truth o, i.e.,

$$\widehat{c^o} = \sum_{i=1}^{n} \frac{\exp(f_i^o)}{\sum_{j=1}^{n} \exp(f_j^o)} * c_i^o \tag{6}$$

Regarding a specific bounding box $b_i^o (1 \leq i \leq n)$, the new selection score of NMS algorithm can be written as,

$$s\_nms_i^o = p_i^o + \text{dis}(c_i^o, \widehat{c^o}) \tag{7}$$

Figure 1. **Illustration of the phenomenon called "task spatial misalignment".** The images in the first column are the sensitive locations for classifier, and images in the second column correspond to the sensitive locations for localizer.



Results and Discussion

We use them to modify RetinaNet and compare the modified network with other state-of-the-art detectors by conducting extensive experiments with different backbones. The results are shown in Table 1. In case of using RetinaNet with backbone ResNet-101 and ResNext-101, our methods achieve 1.5% and 1.4% AP improvements respectively, thus **verifying the superiority of our proposed methods**. Worth noting that we use DC Loss and CL-Balanced NMS to modify FCOS with Generalized Focal Loss and achieve an amazing high AP of 48.9%.

Table 1. **performance of alternative detectors with different backbones on MS-COCO test-dev.**

Method	Backbone	AP	AP^{50}	AP^{75}	AP^s	AP^m	AP^l
YOLOv3	DarkNet-53	33.0	57.9	34.4	18.3	35.4	41.9
Faster R-CNN	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Deformable R-FCN	Aligned-Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
RetinaNet	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
IoU-Net	ResNet-101	40.0	59.0	-	-	-	-
Ours	**ResNet-101**	**40.6**	**60.2**	**43.5**	**23.9**	**43.9**	**51.1**
GHM	ResNeXt-101	41.6	62.8	44.2	22.3	45.1	55.3
Faster R-CNN	ResNeXt-101	40.3	62.7	44.0	24.4	43.7	49.8
Mask R-CNN	ResNeXt-101	41.4	63.4	45.2	24.5	44.9	51.8
FCOS	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
CornerNet	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
RetinaNet	ResNeXt-101	40.8	61.1	44.1	24.1	44.2	51.2
Ours	**ResNeXt-101**	**42.2**	**62.5**	**45.1**	**25.5**	**44.9**	**52.3**
ATSS	ResNeXt-101-64x4d-DCN	47.7	66.5	51.9	29.7	50.8	59.4
GFL	ResNeXt-101-32x4d-DCN	48.2	67.4	52.6	29.2	51.7	60.2
Ours	**ResNeXt-101-32x4d-DCN**	**48.9**	**67.8**	**53.2**	**30.0**	**52.2**	**61.7**

To demonstrate the **effectiveness of the three individual components** of our methods, we use these components to modify RetinaNet. The performance of these variations is evaluated with both COCO val2017 and VOC 2007 test. The results are respectively demonstrated in Table 2.

Table 2. **Component analysis on COCO val and PASCAL VOC 2007 test.** The baseline is ResNet-50 RetinaNet.

IAC -score	DC Loss	CL-B NMS	VOC 2007 test mAP	MS COCO 2017 AP	AP^{50}	AP^{75}
			78.2	35.9	56.1	38.6
✓			78.6	36.4	56.5	39.1
✓	✓		79.2	37.1	56.3	39.7
✓	✓	✓	79.7	37.5	56.9	40.0

Conclusion

In this paper, we propose a systematical method to enhance the performance of object detection by energizing existing solutions from three different aspects to neutralize the negative impacts of task spatial misalignment. Specifically, a novel DC Loss is proposed to address the "task spatial misalignment" in object detection. Further, IoU-classification-aware score is devised to involve classification scores during assigning labels of anchors. Finally, CL-Balanced NMS is designed to address the misalignment between classification and localization via adding the localization score of candidate bounding boxes into conventional NMS algorithm.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), and the Academic Leaders Cultivation Program, USTC.