Object Detection with Missing Labels

Wenhao Chai, Han Yang, Chenhao Li, Mu Xie

***Abstract*—Supervised detection methods have gotten a good result with qualified annotation. However, with imperfect bounding box annotation, 30% of missing labels in this project, normal detection method like YOLOv5 doesn't achieve a relatively good result. In our project, we use COCO dataset. And we greatly eliminate the negative influence o n m issing l abels b y u sing a modified l oss f unction a nd d ynamic w eight. O ur m ethod result is close to YOLOv5 under perfect annotation.**

The code is in https://github.com/ChenhaoLi1106/Missing-Label-Detection/tree/main

***Index Terms*—Object Detection, Missing Label, Focal loss**

I. Introduction

Impressive results have been achieved on object detection benchmarks by supervised object detection methods. However, the performance of supervised object detection methods is profoundly affected by the quality of these annotations. For instance, imperfect bounding box annotations or missing annotations of objects in training images can have a drastic impact on its performance. In this project, we try to improve the performance with a missing label train dataset.

We use COCO dataset for this project. The COCO dataset we have contains a small number of train set, with only 1000 images. The train set has 1000 pictures of human with **30% missing bounding box**, some of pictures even miss all bounding box. The validation set has 2693 pictures with complete human labels. We use mAP (mean Average Precision) in different IOU thresholds to evaluate our model.

The baseline we apply is YOLOv5. By analysis the loss function provided by YOLOv5, we find they cannot be used directly in missing label problem. So, in this report, we are going to explore a new Loss function that can help the model better do machine learning problem with missing label.

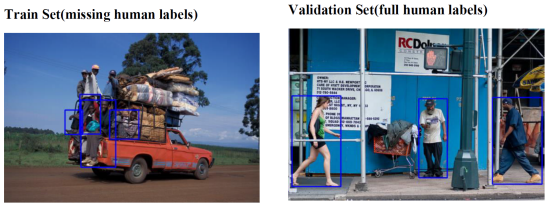

Fig. 1. Illustration of Missing Label Problem

II. Related Work

A. Object Detection

In the generic object detection, proposal based (i.e. two-stage) methods include R-CNN [1], Fast RCNN [2], Faster R-CNN [3], Mask R-CNN [4], etc., are the most popular detectors in many real-world applications, such as surveillance, self-driving, etc. Another branch for general object detection is the one-stage detectors, such as YOLO [5], SSD [6], etc. In this paper, we adopt one-stage detector(YOLOv5) and propose a novel loss design for the missing-label circumstance.

B. Imbalanced or Unlabeled Problem

Re-sampling with imbalanced or unlabeled objects is a common approach for object detection in the recent years. The representative methods among them like GHMC [7], OHEM [8] and Focal Loss [9] have been proposed for sampling RoIs and re-weighting the gradients.

III. Our Method

A. Rethinking Missing-label

Missing label problem actually play a role in two different parts. In the process of computing box loss, we only consider the box's size, which means if there is a lack of ground truth, it just has no enough box to learn, and it makes it slower to train. But for the object loss, the situation is different. The misclassification problem may cause the bad influence on model. In ideal case, it only has seventy percent positive sample to push forward, but thirty percent misclassification samples may add extra fault to the loss function when simply use binary cross entropy.

B. Loss Function

As we choose YOLOv5 as baseline method and want to make a modification on it, one quick solution is to design a new object loss function to fit the task. There is loss function proposed by others, such as the one in the origin code of YOLOv5, it has the form like:

$$L(p_t, g_t) = 1 - \frac{e^{p_t - g_t - 1}}{\alpha + \epsilon} \tag{1}$$

Single stage detectors often depend on a dense coverage of anchors on the image feature maps. While the dense coverage ensures sufficient training samples, it naturally introduces massive class imbalance. Focal Loss is widely favored among single stage detectors as it effectively re-scales the gradients of the anchors. In general, it calibrates and re-scales the gradients for both hard and easy training examples. It has the form below:

$$FL(p_t) = -(1 - p_t)^{\gamma} log(p_t) \tag{2}$$

The term $p_t \in [0, 1]$, represents the predicted confidence score of a anchor. The parameter γ in this formula give the strength

of the equation to balance the easy and hard samples. And consider the bad influence caused by the lack of positive samples. We think it is necessary to add a parameter α to balance the positive and negative samples, which gives:

$$\alpha FL(p_t) = -\alpha(1-p_t)^{\gamma} log(p_t) \tag{3}$$

Actually, these equations can only for discrete ground truth. We propose our loss function with adaptation to the soft label for all the ground truth is actually the IOU between the annotations and the predicted anchors. Assume that the term $g_t \in [0,1]$ representing the ground truth score of a anchor. Our proposed loss function is:

$$L = -\alpha(1-p_t)^{\gamma} g_t log(p_t) - (1-\alpha) p_t^{\gamma} (1-g_t) log(1-p_t) \tag{4}$$

Despite the lack of correct labels, the model can adjust the gradients according to its own well established classifier. Specifically, if the feature map of an anchor region is similar to that of a ground truth object, the classifier naturally assigns a low confidence score pt. If the classifier is trained sufficiently, we can trust the model more with itself instead of the annotations. In this manner, the model would actually learn as if the anchor is positively labeled.

IV. Dynamic Weight

In the baseline training process, the training curve is shown below. Due to the case that the box loss is hard to decrease, and the object loss is quicker, I simply design this linear schedule to allocate the weight. Another thinking is that it's good for the network to learn the object part earlier, which means let them know where is possible position of a person, than focus on what the shape of the bounding box.

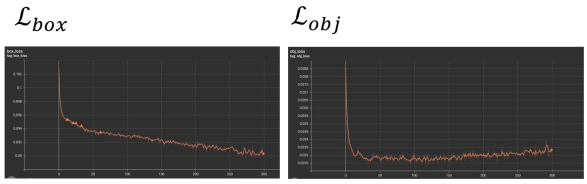

Fig. 2. Baseline Training Curve

The dynamic weight strategy can be represented as:

$$\mathcal{L}_t = (1 - \frac{t}{T})\mathcal{L}_{obj} + \frac{t}{T}\mathcal{L}_{box} \tag{5}$$

The term t is the current epoch and T is the total epoch. The dynamic weight strategy is helpful for loss convergence.

V. Experiments

A. Dataset

We start our experiments with a baseline YOLOv5 on the provided sub COCO dataset only contain human annotation. The COCO dataset we have contains a small number of train set, with only 1000 images. The train set has 1000 pictures of human with 30% missing bounding box, some of pictures even miss all bounding box. The validation set has 2693 pictures with complete human labels. We simply resize all the pictures to the size (640,640).

B. Results

For baseline setting, lr=0.01, batch size=16, epoch=50. The parameters shown below are the parameters in equation (4). DW means that using dynamic weight strategy. The val/train in the parentheses means the baseline is trained on validation set (full annotation) or training set (30% missing label).

case	mAP(0.5)	mAP(0.75)
baseline(val)	0.753	0.560
baseline(train)	0.266	0.107
baseline(train) + DW	0.103	0.000
$\gamma = 2.5, \alpha = 0.25 + DW$	**0.607**	**0.536**
$\gamma = 2, \alpha = 0.25 + DW$	0.583	0.498
$\gamma = 2.5, \alpha = 0.25$	0.621	0.402
$\gamma = 2.5, \alpha = 0.2 + DW$	0.542	0.497

TABLE I
Results

VI. Conclusion

In our project, we use a modified loss function and dynamic weight to improve our model under 30% of missing labels. We choose YOLOv5 as our baseline, and its mAP result under 30% missing labels is only 0.266, 0.107 when the IOU is 0.5 and 0.75 respectively. Our method result is greatly improved compared with baseline. Our result mAP equal to 0.607 and 0.536 when the IOU is 0.5 and 0.75, which is close to YOLOv5 with perfect annotation, 0.753 and 0.560 respectively.

References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Computer Vision and Pattern Recognition, 2014.

[2] Ross Girshick, "Fast r-cnn," in International Conference on Computer Vision (ICCV), 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015.

[4] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick, "Mask r-cnn," in ICCV, 2017.

[5] "You only look once: Unified, real-time object detection," in Redmon, Joseph and Divvala, Santosh and Girshick, Ross and Farhadi, Ali, 2016.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single shot multibox detector," in ECCV, 2016.

[7] Buyu Li, Yu Liu, and Xiaogang Wang, "Gradient harmonized single-stage detector," in AAAI Conference on Artificial Intelligence, 2019.

[8] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining," in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Doll´ar, "Focal loss for dense object detection," in ICCV, 2017.