Reconcile Prediction Consistency for Balanced Object Detection Supplementary Material

Keyang Wang, Lei Zhang(✉)
Learning Intelligence & Vision Essential (LiVE) Group
School of Microelectronics and Communication Engineering, Chongqing University, China
{wangkeyang, leizhang}@cqu.edu.cn

1. The derivation process of effectiveness analysis on regression task from the gradient.

The Harmonic loss for a positive sample x_i is defined as follows in our paper:

$$\mathcal{L}^i_{Har} = (1+\beta_r)CE(p_i, y_i) + (1+\beta_c)\mathcal{L}(d_i, \hat{d}_i) \quad (1)$$

We calculate the partial derivative of Eq.(1) with respect to the predicted regression offset $(d_i - \hat{d}_i)$ as:

$$\frac{\partial \mathcal{L}^i_{Har}}{\partial(d_i - \hat{d}_i)} = \frac{\partial \beta_r}{\partial(d_i - \hat{d}_i)} CE(p_i, y_i) + (1+\beta_c)\frac{\partial \mathcal{L}(d_i, \hat{d}_i)}{\partial(d_i - \hat{d}_i)} \quad (2)$$

Let $t = (d_i - \hat{d}_i)$, we can have

$$\frac{\partial \mathcal{L}^i_{Har}}{\partial t} = \frac{\partial \beta_r}{\partial t} CE(p_i, y_i) + (1+\beta_c)\frac{\partial \mathcal{L}(t)}{\partial t} \quad (3)$$

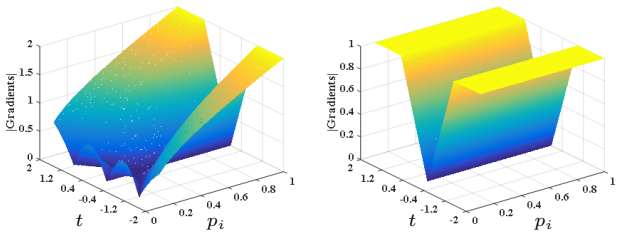
where CE$(\cdot)$ is the cross-entropy loss and y_i is the one-hot label, and there is

$$\begin{aligned} CE(p_i, y_i) &= -y_i \log(p_i) = -\log(p_i), \\ \beta_c &= e^{-CE(p_i, y_i)} = e^{y_i \log(p_i)} = p_i \end{aligned} \quad (4)$$

The definition of β_r and β_c is presented in Eq.(3) of the main paper. By substituting Eq.(4) and $\beta_r = e^{-\mathcal{L}(d_i, \hat{d}_i)} = e^{-\mathcal{L}(t)}$ into Eq.(3), we can have

$$\begin{aligned} \frac{\partial \mathcal{L}^i_{Har}}{\partial t} &= e^{-\mathcal{L}(t)} \log(p_i)\frac{\partial \mathcal{L}(t)}{\partial t} + (1+p_i)\frac{\partial \mathcal{L}(t)}{\partial t} \\ &= (1 + p_i + e^{-\mathcal{L}(t)} \log(p_i))\frac{\partial \mathcal{L}(t)}{\partial t} \end{aligned} \quad (5)$$

From the above Eq.(5), we can observe that the gradient of Harmonic loss with respect to the predicted regression offset $t = (d_i - \hat{d}_i)$ is also supervised by the classification score p_i. That is, the classification score also participates in the optimization of the regression branch. We visualize the absolute value of the gradients of the standard detection loss and the proposed Harmonic loss with respect to $t = (d_i - \hat{d}_i)$, respectively, in Fig. 1.



(a) (b)

Figure 1. Visualization of the absolute value of the gradients of detection losses with respect to the predicted regression offset $t = (d_i - \hat{d}_i)$. (a) is the gradient of our Harmonic loss with respect to t, $|\frac{\partial \mathcal{L}^i_{Har}}{\partial t}|$. (b) is the gradient of the standard detection loss (cross-entropy loss plus smooth L1 loss) with respect to t.

For the standard detection loss, the gradient with respect to t does not change with different classification scores as Fig. 1 (b) shows, which means the optimization of regression branch is absolutely independent of the classification task. But for our Harmonic loss, the gradient is a function simultaneously determined by the two variables p_i and t, as shown in Eq.(5). In other words, for each positive sample, the classification score p_i will supervise the optimization of the regression branch during training phase of the Harmonic loss. Specifically, there is a proportional correlation between p_i and the absolute value of the loss gradient w.r.t. t, as is shown in Fig. 1 (a). This means that the gradient will suppress the regression offsets of samples with low classification scores, which, therefore, guarantees the harmony between classification and regression.

2. Convergence Analysis.

The convergence of the proposed loss is important during training, in the following, we will give the clear convergence analysis of our Harmonic loss. The whole Harmonic loss for

a positive sample x_i is defined as in our paper:

$$\mathcal{L}^i_{Har} = (1+\beta_r)CE(p_i, y_i) + (1+\beta_c)\mathcal{L}(d_i, \hat{d}_i) \quad (6)$$

let us take the classification task as an example. if we let $x = CE(p_i, y_i)$, $y = \mathcal{L}(d_i, \hat{d}_i)$ and $F(x, y) = \mathcal{L}^i_{Har}$, the proposed Harmonic loss can be rewritten as:

$$F(x, y) = (1 + e^{-y})x + (1 + e^{-x})y \quad (7)$$

When we calculate the partial derivative for x, we can get

$$\frac{\partial F(x)}{\partial x} = 1 + e^{-y} - ye^{-x} \quad (8)$$

Then we can find that the $F(x)$ increases monotonically with x when the x satisfies:

$$x > \ln y - \ln(1 + e^{-y}) \quad (9)$$

This is basically always true during training. Therefore, the convergence of our Harmonic loss for the classification task is the same as the convergence of CE loss, which can converge well.

Actually, as we described in the Section 3.1 of the main paper, the partial derivative of Harmonic loss with respect to the predicted score p_i is calculated as:

$$\frac{\partial \mathcal{L}^i_{Har}}{\partial p_i} = \mathcal{L}(d_i, \hat{d}_i) - \frac{(1 + e^{-\mathcal{L}(d_i, \hat{d}_i)})}{p_i} \quad (10)$$

We visualize the gradients of our Harmonic loss with respect to p_i, as shown in Fig. 2. We can find that the gradient value is always negative when $CE(p_i, y_i) > \ln(\mathcal{L}(d_i, \hat{d}_i)) - \ln(1 + e^{-\mathcal{L}(d_i, \hat{d}_i)})$ (This equation is the same as Eq.(9), which is basically always true during training). This means that our Harmonic loss is always monotonic during training. In other words, our Harmonic loss for the classification task can converge well during training. Actually, this is why we define the complete harmonic factors as $(1+\beta_r)$ and $(1+\beta_c)$ instead of β_r and β_c. The constant 1 in harmonic factors can ensure that the basic localization and classification loss always exist during training. This can ensure the Harmonic loss is monotonically decreasing for classification and localization tasks and avoid the optimization contradiction. The convergence of our Harmonic loss for regression is similar to classification task, where no longer go into details.

In order to further verify our above derivation process, we visualize the training loss in our experiments in Fig. 3. We can clearly find that the Harmonic loss converges perfectly during the training process.

3. The derivation of Harmonic IoU loss.

The Harmonic IoU loss is defined as follows:

$$\mathcal{L}^i_{HIoU} = (1 + IoU_i)^{\gamma}(1 - IoU_i) \quad (11)$$

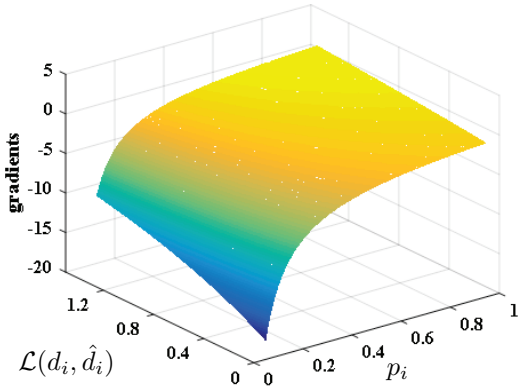


Figure 2. Visualization of the gradient of our Harmonic loss with respect to p_i, $\frac{\partial \mathcal{L}^i_{Har}}{\partial p_i}$.

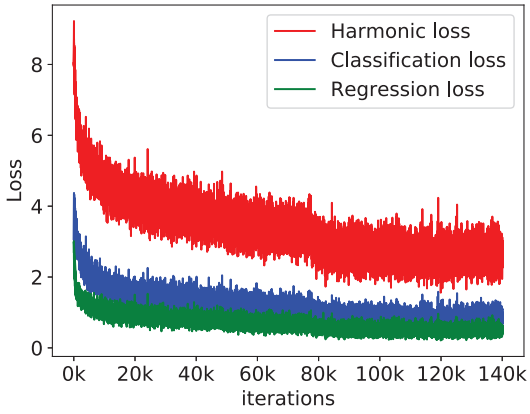


Figure 3. Visualization of the training loss in our experiments.

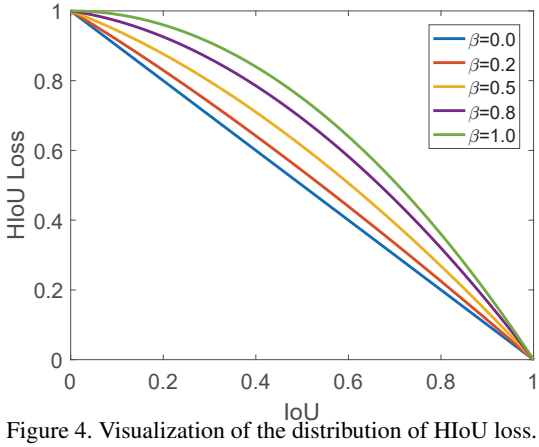


Figure 4. Visualization of the distribution of HIoU loss.

We calculate the partial derivative of Eq.(11) with respect to the IoU_i, we can have

$$\frac{\partial \mathcal{L}^i_{HIoU}}{\partial IoU_i} = (1 + IoU)^{\gamma - 1}((\gamma - 1) - (\gamma + 1)IoU) \quad (12)$$

In order to ensure the HIoU loss is always monotonic, the $\frac{\partial \mathcal{L}^i_{HIoU}}{\partial IoU_i}$ must always satisfy $\frac{\partial \mathcal{L}^i_{HIoU}}{\partial IoU_i} \leq 0$, so we can have:

$$IoU \geq \frac{\gamma - 1}{\gamma + 1} \quad (13)$$

As we all know, the range of IoU is [0,1]. So in order to ensure that the Eq. (13)is always true, the focusing parameter γ in HIoU loss must satisfy $\gamma \leq 1$.

In the Fig. 4, we visualize the distribution of HIoU loss under five different focusing parameters, we can find the HIoU losses show the upwards convex shapes, which means that our HIoU losses increase the weights of examples with high IoUs in regression task. In the end, the contribution of each kind of examples is balanced and the bias of regression can be effectively alleviated.