DR Loss: Improving Object Detection by Distributional Ranking Supplementary

Qi Qian[1] Lei Chen[2] Hao Li[2] Rong Jin[1]
Alibaba Group
[1]Bellevue, WA, 98004, USA
[2]Hangzhou, China
{qi.qian, fanjiang.cl, lihao.lh, jinrong.jr}@alibaba-inc.com

1. Gradient of DR Loss

We have the DR loss as

$$\min_{\theta} \mathcal{L}_{\mathrm{DR}}(\theta) = \sum_{i}^{N} \ell_{\mathrm{logistic}}(\hat{P}_{i,-} - \hat{P}_{i,+} + \gamma)$$

where

$$\ell_{\mathrm{logistic}}(z) = \frac{1}{L}\log(1+\exp(Lz))$$

and

$$\begin{aligned} \hat{P}_{i,-} &= \sum_{j_-}^{n_-} \frac{1}{Z_-}\exp(\frac{p_{i,j_-}}{\lambda_-})p_{i,j_-} = \sum_{j_-}^{n_-} q_{i,j_-}p_{i,j_-} \\ \hat{P}_{i,+} &= \sum_{j_+}^{n_+} \frac{1}{Z_+}\exp(\frac{-p_{i,j_+}}{\lambda_+})p_{i,j_+} = \sum_{j_+}^{n_+} q_{i,j_+}p_{i,j_+} \end{aligned}$$

It looks complicated but its gradient is easy to compute. Here we give the detailed gradient. For p_{i,j_-}, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_{i,j_-}} &= \frac{1}{1+\exp(-Lz)}\frac{\partial z}{\partial p_{i,j_-}} \\ &= \frac{q_{i,j_-}}{1+\exp(-Lz)}(1+\frac{p_{i,j_-}}{\lambda_-} - \frac{1}{\lambda_-}(\sum_{j_-} q_{i,j_-}p_{i,j_-})) \end{aligned}$$

where $z = \hat{P}_- - \hat{P}_+ + \gamma$.

For p_{i,j_+}, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_{i,j_+}} &= \frac{1}{1+\exp(-Lz)}\frac{\partial z}{\partial p_{i,j_+}} \\ &= \frac{q_{i,j_+}}{1+\exp(-Lz)}(-1+\frac{p_{i,j_+}}{\lambda_+} - \frac{1}{\lambda_+}(\sum_{j_+} q_{i,j_+}p_{i,j_+})) \end{aligned}$$

2. Proof of Theorem 1

Proof. First, we give the definition of smoothness

Definition 1. *A function F is called μ-smoothness w.r.t. a norm $\|\cdot\|$ if there is a constant μ such that for any θ and θ', it holds that*

$$F(\theta') \leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta\rangle + \frac{\mu}{2}\|\theta' - \theta\|^2$$

We assume that the loss in Eqn. 9 is μ-smoothness, then we have

$$\begin{aligned} E[\mathcal{L}(\theta_{t+1})] &\leq E[\mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1} - \theta_t\rangle \\ &\quad + \frac{\mu}{2}\|\theta_{t+1} - \theta_t\|_F^2] \\ &= E[\mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), -\frac{\eta}{m}\sum_{s=1}^{m}\nabla\ell_t^s\rangle \\ &\quad + \frac{\mu\eta^2}{2}\|\frac{1}{m}\sum_{s=1}^{m}\nabla\ell_t^s\|_F^2] \end{aligned}$$

According to the definition, we have

$$\forall s, E[\nabla\ell_t^s] = \nabla\mathcal{L}(\theta_t)$$

If we assume that the variance is bounded as

$$\forall s, \|\nabla\ell_t^s - \nabla\mathcal{L}_t\|_F \leq \delta$$

then we have

$$\begin{aligned} E[\mathcal{L}(\theta_{t+1})] &\leq E[\mathcal{L}(\theta_t) - \eta\|\nabla\mathcal{L}_t\|_F^2 \\ &\quad + \frac{\mu\eta^2}{2}\|\frac{1}{m}\sum_{s=1}^{m}\nabla\ell_t^s - \nabla\mathcal{L}_t + \nabla\mathcal{L}_t\|_F^2] \\ &\leq E[\mathcal{L}(\theta_t) - \eta\|\nabla\mathcal{L}_t\|_F^2 + \frac{\mu\eta^2}{2}(\frac{\delta^2}{m} + \|\nabla\mathcal{L}_t\|_F^2) \end{aligned}$$

Therefore, we have

$$(\eta - \frac{\mu\eta^2}{2})\|\nabla\mathcal{L}(\theta_t)\|_F^2 \leq E[\mathcal{L}(\theta_t)] - E[\mathcal{L}(\theta_{t+1})] + \frac{\mu\eta^2\delta^2}{2m}$$

	Focal Loss						DR Loss					
Threshold	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
0.05	36.1	55.0	38.7	19.5	39.5	49.0	37.4	56.0	40.0	20.8	41.2	50.5
0.1	36.1	54.9	38.7	19.4	39.4	49.0	37.4	56.0	40.0	20.8	41.2	50.5
0.2	35.4	53.4	38.2	18.3	38.7	48.6	37.4	56.0	40.0	20.8	41.2	50.5
0.3	33.9	50.2	37.0	16.2	37.1	47.6	37.4	56.0	40.0	20.8	41.2	50.5
0.4	31.6	45.8	35.0	14.1	34.4	45.2	37.3	55.9	40.0	20.7	41.2	50.4
0.5	28.4	39.7	31.7	10.5	30.5	42.1	37.2	55.6	39.8	20.1	41.0	50.3

Table 1. Comparison of different thresholds.

	Focal Loss [2]						DR Loss					
scale	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
400	30.5	47.8	32.7	11.2	33.8	46.1	32.4	49.9	34.5	11.7	34.8	48.0
500	32.5	50.9	34.8	13.9	35.8	46.7	34.5	52.6	36.6	14.7	36.9	48.9
600	34.3	53.2	36.9	16.2	37.4	47.4	36.1	54.6	38.7	17.4	38.5	49.2
700	35.1	54.2	37.7	18.0	39.3	46.4	37.1	55.8	39.7	18.9	39.8	49.2
800	35.7	55.0	38.5	18.9	38.9	46.3	37.6	56.4	40.3	20.1	40.5	48.9

Table 2. Comparison of different input scales. We adopt 1× iterations and ResNet-50 as the backbone in training. Results on the *test-dev* are reported.

By assuming $\eta \leq \frac{1}{\mu}$ and adding t from 1 to T, we have

$$\sum_t \|\nabla\mathcal{L}(\theta_t)\|_F^2 \leq \frac{2\mathcal{L}(\theta_0)}{\eta} + \frac{\mu\eta T\delta^2}{m}$$

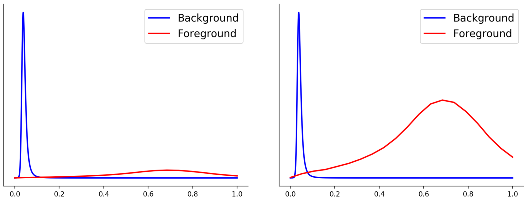
We finish the proof by letting

$$\eta = \frac{\sqrt{2m\mathcal{L}(\theta_0)}}{\delta\sqrt{\mu T}}$$

□

3. Additional Experiments

Effect of DR Loss: We illustrate the empirical PDF of foreground and background from DR loss in Fig. 1. Fig. 1 (a) shows the original density of foreground and background. To make the results more explicit, we decay the density of background by a factor of 10 and demonstrate the result in Fig. 1 (b). It is obvious that DR loss can separate the foreground and background with a large margin in the imbalanced scenario.



(a) Original Density (b) Decayed Density

Figure 1. Illustration of empirical PDF of distributions from DR loss.

Effect of Large Margin: Before non-maximum suppression (NMS), the candidates with low confidence will be filtered to accelerate detection. Since the distribution of foreground from focal loss is close to that of background as illustrated in Fig. 4 of our paper, a small threshold as 0.05 is adopted to eliminate negative examples. The proposed loss function optimizes the distributions with a large margin and can be robust to the selection of the threshold. Table 1 demonstrates the performance with different thresholds. It is obvious that the performance of DR loss keeps almost the same while that of focal loss degrades significantly when increasing the threshold.

Effect of Image Scale: We tune the parameters of DR loss with a single input scale of 800 but the parameters are robust to different input scales. We follow the settings in the ablation study and Table 2 compares the performance on *test-dev* with scales varied in $\{400, 500, 600, 700, 800\}$. We report the results of focal loss from [2]. Evidently, DR loss can consistently improve the performance over focal loss by about 2%. It demonstrates that the proposed loss function is not sensitive to the scale of input images.

Comparison on PASCAL: Finally, we evaluate the proposed DR loss on a different data set: PASCAL VOC2007 [1], which contains $9,963$ images and 20 classes. We adopt the same configurations for RetinaNet as in the ablation study and the same parameters as those on COCO for DR loss and focal loss. We change the initial learning rate to 0.008 and it is decayed at $6,250$ iterations, where the total number of iterations is $8,750$ as suggested by the codebase. Other training settings are the same as the pipeline for

COCO. The detector is trained with the training and validation sets, and Table 3 shows the comparison on the test data. We can observe that with the same parameters on a different task, our method can outperform focal loss with a significant margin. It demonstrates that the proposed loss function can be applicable for different tasks.

Loss	AP	AP_{50}	AP_{75}
Focal	39.5	67.2	40.8
DR	41.2	68.6	42.6

Table 3. Comparison on VOC2007. Results on *test* are reported.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[2] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017.