Supplementary: Prime Sample Attention in Object Detection

Yuhang Cao[1] Kai Chen[1] Chen Change Loy[2] Dahua Lin[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Nanyang Technological University
{yhcao6,chenkaidev}@gmail.com ccloy@ntu.edu.sg dhlin@ie.cuhk.edu.hk

Appendix A: Derivative of CARL

As discussed in Section 4.2, we prove that there is a positive correlation between $\frac{\partial L_{carl}}{\partial p_i}$ and $\mathcal{L}(d_i, \hat{d}_i)$, where

$$\begin{aligned} L_{carl} &= \sum_{i=1}^{n} c_i \mathcal{L}(d_i, \hat{d}_i) \\ c_i &= \frac{v_i}{\frac{1}{n}\sum_{i=1}^{n} v_i} \\ v_i &= ((1-b)p_i + b)^k \end{aligned} \tag{1}$$

By chain rule,

$$\begin{aligned} \frac{\partial L_{carl}}{\partial p_i} &= \frac{\partial \mathcal{L}_{carl}}{\partial c_i}\frac{\partial c_i}{\partial p_i} \\ &= \mathcal{L}(d_i, \hat{d}_i)\frac{\partial c_i}{\partial p_i} \end{aligned} \tag{2}$$

Furthermore,

$$\frac{\partial c_i}{\partial p_i} = \frac{\partial c_i}{\partial v_i}\frac{\partial v_i}{\partial p_i} \tag{3}$$

Denoting $S = \sum_{i=1}^{n} v_i$, we have

$$\frac{\partial c_i}{\partial v_i} = \frac{n}{S}(1 - \frac{v_i}{S}) \tag{4}$$

The batch size is usually large, so $v_i << S$. Thus we have

$$\frac{\partial c_i}{\partial v_i} \approx \frac{n}{S} \tag{5}$$

On the other hand,

$$\frac{\partial v_i}{\partial p_i} = (1-b)k((1-b)p_i + b)^{k-1} \tag{6}$$

We have $0 \leq b < 1$ and $k > 0$, so $\frac{\partial v_i}{\partial p_i} > 0$. Especially when $k = 1$, $\frac{\partial v_i}{\partial p_i} = 1 - b$.

Combining (2)(3)(5)(6),

$$\frac{\partial L_{carl}}{\partial p_i} = \frac{n}{S}\frac{\partial v_i}{\partial p_i}\mathcal{L}(d_i, \hat{d}_i) \tag{7}$$

When $k = 1$, $\frac{\partial L_{carl}}{\partial p_i} = \frac{n(1-b)}{S}\mathcal{L}(d_i, \hat{d}_i)$, indicating that $\frac{\partial L_{carl}}{\partial p_i}$ is proportional to $\mathcal{L}(d_i, \hat{d}_i)$, otherwise $\frac{\partial L_{carl}}{\partial p_i}$ and $\mathcal{L}(d_i, \hat{d}_i)$ are positively correlated.

Appendix B: Implementation details

We use 8 Tesla V100 GPUs in all experiments. For SSD, we train the model for a total of 120 epochs with a minibatch of 64 images (8 images per GPU). The learning rate is initialized as 0.001 and decreased by 0.1 after 80 and 110 epochs. For other methods, we adopt ResNet-50 or ResNeXt-101-32x4d as the backbone. FPN is used by default. The batch size is 16 (2 images per GPU). Models are trained for 12 epochs with an initial learning rate of 0.02, which is decreased by 0.1 after 8 and 11 epochs, respectively. For the random sampling baseline, We sample 512 RoIs from 2000 proposals and the ratio of positive to negative samples is set to 1:3. When OHEM is used, we forward all 2000 proposals and select 512 samples with the highest loss. A variant of OHEM is also explored, where the ratio of positive/negative samples is set to be 1:3 they are mined independently. PISA consists of ISR (ISR-P and ISR-N) and CARL with one exception, *i.e.*, ISR-N is not used in single-stage models because the number of negative samples in single-stage models are much greater than those in two-stage ones, therefore will introduce significant overhead for training time.