Variational Pedestrian Detection: Supplementary Material

Yuang Zhang[1*], Huanyu He[1*], Jianguo Li[2], Yuxi Li[1], John See[3], Weiyao Lin[1†]
[1]Shanghai Jiao Tong University, China, [2]Ant Group, [3]Heriot-Watt University, Malaysia

6. Appendix

6.1. Focal Loss as a Probabilistic Online Hard Example Mining

Problem Setup Consider a set of dense proposals $\mathbf{z}$ with classification score $\mathbf{s}$. Since positives samples for final detection boxes $\mathbf{x}$ can be obtained from the dataset, we are only interested in negative samples in this section. The negative dense proposals will be selected to form a final detection set $\widehat{\mathbf{x}}$ according to the Bernoulli random variable $\mathbf{K}$ with parameter $\mathbf{k}$ where $k_j := P[K_j = 1]$. If and only if $K_j = 1$, the box z_j will be selected and copied to the predicted final detection set $\widehat{\mathbf{x}}$ for training. We denote the copy of the j^{th} dense proposal z_j in the predicted final detection set as $\widehat{x}_j$.

Assume the final detections are independent, i.e., $p(\widehat{\mathbf{x}} = 0) = \prod_j p(\widehat{x}_j = 0)$. Then,

$$p(\widehat{x}_j = 0) = p(\widehat{x}_j = 0, K_j = 0) + p(\widehat{x}_j = 0, K_j = 1). \tag{1}$$

The first term can be rewritten as

$$p(\widehat{x}_j = 0, K_j = 0) = p(\widehat{x}_j = 0|K_j = 0)p(K_j = 0) \tag{2}$$

where $p(K_j = 0) = 1 - p(K_j = 1) = 1 - k_j$ and $p(\widehat{x}_j = 0|K_j = 0) = 1$ from problem setup.

The second term can be rewritten as

$$p(\widehat{x}_j = 0, K_j = 1) = p(\widehat{x}_j = 0|K_j = 1)p(K_j = 1) \tag{3}$$

where $p(\widehat{x}_j = 0|K_j = 1) = p(z_j = 0) = 1 - s_j$ and $p(K_j = 1) = k_j$.

Finally, the likelihood for negative final detection sample is

$$p(\widehat{x}_j = 0) = 1 - k_j + (1 - s_j)k_j = 1 - k_j s_j. \tag{4}$$

Set the negative log likelihood to the stage-of-the-art Focal loss [1]. Then, we have

$$\log p(\widehat{x}_j = 0) = -\text{Focal-loss}(s_j) \tag{5}$$
$$\log(1 - k_j s_j) = s_j^\gamma \log(1 - s_j) \tag{6}$$
$$1 - k_j s_j = (1 - s_j)^{(s_j^\gamma)} \tag{7}$$
$$k_j = \frac{1 - (1 - s_j)^{(s_j^\gamma)}}{s_j}. \tag{8}$$

Then, the relationship between the probability of mining the negative dense proposal z_j and its score s_j is established. The larger the Focal loss coefficient γ, the less likely a well-classified negative sample will be mined. When $\gamma = 0$, all dense proposals will be kept and the Focal loss will deteriorate to binary cross-entropy loss.

6.2. Relaxed Jaccard Index as a Pseudo Detection Likelihood

We demonstrate the advantage of the Auto-Encoding Variational Bayes (AEVB) algorithm in object detection with the relaxed Jaccard Index as a pseudo detection likelihood, which is an alternative to the adopted FreeAnchor [3] likelihood described in Section 3.5.

Jaccard Index is the evaluation metric for the CrowdHuman [2] pedestrian detection competition, which is defined as

$$\text{JI} := \frac{|\text{IoUMatch}(\mathcal{G}, \mathcal{D})|}{|\mathcal{G}| + |\mathcal{D}| - |\text{IoUMatch}(\mathcal{G}, \mathcal{D})|} \tag{9}$$

where $|\mathcal{G}|$ is the number of ground truth boxes, $|\mathcal{D}|$ is the number of predicted boxes, and IoUMatch is the number of truth positives after the optimal match between ground truth and predicted boxes.

The relaxation of Jaccard Index We relax the Jaccard Index to a smoother version for efficient training with the reparametrization trick. Assume we have n ground truth boxes and m dense proposals, the relaxed Jaccard Index is defined as

$$\widetilde{\text{JI}} := \frac{\sum_{i=1}^n \widetilde{M_i}}{n + \sum_{j=1}^m s_j - \sum_{i=1}^n \widetilde{M_i}} \tag{10}$$

*Equally-contributed first authors
†Corresponding author, Email:wylin@sjtu.edu.cn

Table 1: Performance on the CrowdHuman dataset evaluated by Jaccard Index (higher is better).

Method	JI (best)
RetinaNet	0.7031
$\widetilde{\text{JI}}$ + ML	0.6788
$\widetilde{\text{JI}}$ + AEVB	0.6896

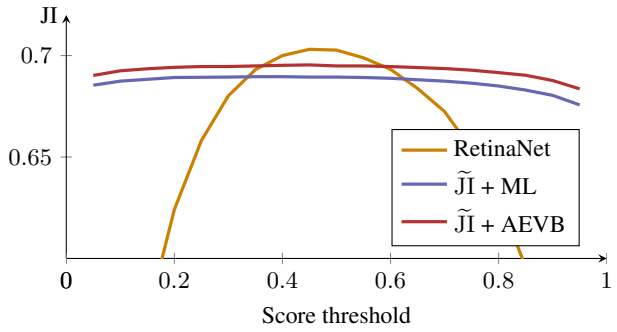

Figure 1: Jaccard Index at different score thresholds.

where the relaxed match $\widetilde{M}_i$ for the i^{th} ground truth box is given by

$$\widetilde{M}_i = 1 - \widetilde{P}[x_i \text{ is missed by all dense proposals } z_j] \quad (11)$$

$$:= 1 - \prod_{j=1}^{m} \left(1 - g(\text{IoU}_{ij}) \cdot s_j\right). \quad (12)$$

IoU_{ij} indicate the IoU between the ground truth box x_i and the dense proposal z_j and g is a transformed sigmoid function acting as a soft IoU threshold. Since a one-to-one match is expected, only a few dense proposal will dominate $\widetilde{M}_i$. Therefore, we only calculate and back-propagate the gradient of the top 8 best-matched dense proposals, which means the m in Equation 12 is replaced with 8 in implementation.

This relaxed Jaccard Index is a smooth measure of similarity between the set of ground truth boxes and detection boxes. It has two major differences from the FreeAnchor likelihood: **First**, the Jaccard Index assumes a binary selection of detection boxes, and the classification score is interpreted as the probability of the existence of an object such that the expected number of objects on one image is the sum of the scores; **Second**, a one-to-one match between dense proposals $\mathbf{z}$ and ground truth boxes $\mathbf{x}$ is expected. As a result, the inference procedure features minimal post-processing steps, i.e., NMS is not required and the detection performance should be insensitive to the score threshold.

Experiment The network architecture and training schedule are kept the same as they are in the main text, and we use the true Jaccard Index (Equation 9) as the evaluation metric and apply the official evaluation code. FreeAnchor + AEVB pre-trained weight is applied. We compare the model trained with the maximum likelihood method and the Auto-Encoding Variational Bayes algorithm with the relaxed Jaccard Index as a detection likelihood, i.e., define $\log \widetilde{p}(\mathbf{x}|\mathbf{z}) := \widetilde{\text{JI}}$ by Equation 10. Performance of RetinaNet [1] trained with the same pre-trained weight and schedule is included for reference. For all the three methods, the Jaccard Index over score thresholds ranging from 0.05 to 0.95 with step size 0.05 is plotted in Figure 1, and the best Jaccard Index is reported in Table 1.

Note that although detectors optimized by the relaxed Jaccard Index are slightly worse than RetinaNet on JI (best), Figure 1 indicates that detectors optimized by the relax JI shows much more stable performance over different score thresholds, which might be an advantage in real-world applications since the score-threshold doesn't need to be tuned carefully. Furthermore, the advantage of minimal post-processing steps makes it an interesting field for future research.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 1, 2

[2] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1

[3] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 147–155, 2019. 1